Dominik Flejter Sławomir Grzonkowski Tomasz Kaczmarek
Marek Kowalkiewicz Tadhg Nagle Jonny Parkes (Eds.)

BIS 2008 Workshops Proceedings

Social Aspects of the Web (SAW 2008)

Advances in Accessing Deep Web (ADW 2008)

E-Learning for Business Needs

11th Conference on Business Information Systems (BIS 2008)
Innsbruck, Austria, 6-7 May 2008
http://bis.kie.ae.poznan.pl/

Editors

Dominik Flejter, Poznań University of Economics, Poland
Sławomir Grzonkowski, DERI, NUI Galway, Ireland
Tomasz Kaczmarek, Poznań University of Economics, Poland
Marek Kowalkiewicz, SAP Research Brisbane, Australia
Tadhg Nagle, DERI, NUI Galway, Ireland
Jonny Parkes, Enterprise Ireland, Ireland

Contact person

Dominik Flejter, D.Flejter@kie.ae.poznan.pl

Published by

Department of Information Systems,
Poznań University of Economics,
http://www.kie.ae.poznan.pl

ISBN

ISBN-10 83-916842-5-3 Dept. of Information Systems, Poznań University of Economics
ISBN-13 978-83-916842-5-2 Dept. of Information Systems, Poznań University of Economics

Copyright

Papers: Copyright © 2008 by individual authors.
Volume: Copyright © 2008 by Dominik Flejter, Sławomir Grzonkowski, Tomasz Kaczmarek, Marek Kowalkiewicz, Tadhg Nagle and Jonny Parkes.
Cover: Copyright © 2008 by Dominik Flejter, Roman Hryniewiecki and Adam Walczak.

Other versions of this book

On-line open-access version of these proceedings (including presentation slides) is also available: Dominik Flejter, Sławomir Grzonkowski, Tomasz Kaczmarek, Marek Kowalkiewicz, Tadhg Nagle, Jonny Parkes (eds.): BIS 2008 Workshop Proceedings, CEUR Workshop Proceedings, ISSN 1613-0073, http://CEUR-WS.org/Vol-333/

Printed in Czech Republic

Contents

Introducing BIS 2008 Workshops on Emerging Web Technologies

Dominik Flejter[1], Sławomir Grzonkowski[2], Tomasz Kaczmarek[1],
Marek Kowalkiewicz[3], Tadhg Nagle[2] and Jonny Parkes[4]

[1] Poznan University of Economics, Poland,
[2] DERI, NUI Galway, Ireland,
[3] SAP Research Brisbane, Australia,
[4] Enterprise Ireland, Ireland

This volume includes papers presented at the workshops held in conjunction with the 11th Business Information Systems Conference, taking place in Innsbruck, Austria on 6-7 May 2008. The conference is a well established knowledge exchange forum, with topics covering development, implementation, application, and improvement of IT systems for business. It has a long tradition of organizing special sessions, tracks, and workshops that focus on new and developing research areas.

This year the conference hosted three workshops: 1st Workshop on Advances in Accessing Deep Web (ADW 2008), Workshop on E-Learning for Business Needs, and 2nd Workshop on Social Aspects of the Web (SAW 2008). The common denominator of these diverse workshops is the research on the application of the emerging technologies (particularly in the Web sphere). This topic is approached from different directions by each of the workshops: SAW concentrates on the influence that technologies exert on the societies, and on emergence of social knowledge and structures in Web-based IT solutions. ADW participants discuss how to use potential that lies in the Deep Web to enable more thorough analyses, broader information integration and stimulate outbreak of new information services. Finally, E-Learning participants ponder on the best ways to utilize new technologies to speed up knowledge acquisition and increase its quality for the e-learning solutions users.

The observation that the Web has recently moved from a simple one-way channel, to a complex social communication space was a direct motivation behind SAW 2008. Today, the distinction between the authors and audience is becoming blurred and new ways to create, share and use knowledge in a social way emerge.

SAW papers investigate a variety of aspects of this change of paradigm, that transforms our interactions with other people, our relationships, ways of gathering information and doing business. Bojārs et al. [1] and Guns [2] focus on bridging semantic technologies research with social aspects. Massa and Souren [3] analyze complex subject of trust measurement in the Web environment. Three papers show how Web 2.0 technologies may be used in knowledge management (Bibikas et al. [4]), customer support (Nguyen et al. [5]) and in development of mobile cultural services (Coppola et al. [6]). Jacquemin et al. [7] study how user conflicts may be handled in Wikipedia. Stocker and Tochtermann [8] analyze usage of weblogs in business scenarios and Ettinger et al. [9] demonstrate possible usages of job boards.

In parallel to advancement of Social Web, a significant growth of complexity of Web information systems can be observed. The growth is giving rise to the Deep Web phenomenon. While the main way of accessing content on contemporary Web is by means of search engines, they do not index significant portion of modern Web

content. In many cases these nonindexable information sources, known as Deep Web, are better structured and of better quality than indexed Web. High value and low availability are thus the basic motivation behind ADW 2008. Its focus is on a wide area of Deep Web research, combining challenges from several active research areas, including information retrieval, information extraction, hypertext, Web engineering, data integration, database technologies, and the Semantic Web.

This is adequately represented by ADW papers that present methods for different stages and approaches of Deep Web content acquisition and usage. Kardkovács and Tikk [10] propose a novel approach to identification of Deep Web sources relevant for specific queries by combining NLP and relational database research. Paper by Wang and Hornung [11] focuses on learning Deep Web sources navigational patterns based on user examples. Finally, Algergawy et al. [12] propose a new approach to schema mapping - a critical task for information integration from the Deep Web.

The same changes addressed by SAW and ADW, are reasons for strong need for life-long learning and increased knowledge availability, especially in business. This area of research is central for E-Learning for Business Needs. Its goal is to bridge the gap between human resource management and emerging technologies to create robust e-learning solutions for the knowledge workers. In addition, it aims at providing guidance for organizations to allow them not only to create new e-learning solutions but also implement these solutions as customers. To tackle these issues Papanikolaou and Mavromoustakos [13] propose a framework for designing e-learning applications incorporating social and collaborative aspects of Web 2.0 technologies. Simon et al. [14] present the Evaluate platform that measures the impact of e-learning on organizations. Finally, Bijlani et al. [15] describe the case study on the Amrita Campus and EDUSAT network and study how the integration of a wide range of technologies (including mobile solutions) can increase the effectiveness of e-learning.

References

1. Bojārs, Passant, Breslin, Decker.: Social Network and Data Portability using Semantic Web Technologies. pp. 5-19.
2. Guns.: Unevenness in network properties on the social Semantic Web. pp. 21-30.
3. Massa and Souren.: Trustlet, Open Research on Trust Metrics. pp. 31-44.
4. Organisational Knowledge Management Systems: The case of OrganiK.: Bibikas, Kourtesis, Paraskakis, Bernardi, Sauermann, Apostolou, Mentzas and Vasconcelos. pp. 45-53.
5. Nguyen, Thompson, Hoile.: Hubbub - An innovative customer support forum. pp. 55-67.
6. Coppola, Lomuscio, Mizzaro, Nazzi and Vassena.: Mobile Social Software for Cultural Heritage: A Reference Model. pp. 69-80.
7. Jacquemin, Lauf, Poudat, Hurault-Plantet and Auray.: Managing conflicts between users in Wikipedia. pp. 81-93.
8. Stocker and Tochtermann.: Investigating Weblogs in Small and Medium Enterprises. pp. 95-107.
9. Ettinger, Wilderom, Van Dick.: Transforming Exchange-based Job Boards into Lasting Career Communities. pp. 109-116.
10. Kardkovács, Tikk.: Determining Relevant Deep Web Sites by Query Context Identification. pp. 119-130.
11. Wang and Hornung.: Deep Web Navigation by Example. pp. 131-140.
12. Algergawy, Schallehn and Saake.: Fuzzy Constraint-based Schema Matching Formulation. pp. 141-152
13. Papanikolaou and Mavromoustakos.: Web2Train: a Design Model for Corporate e-Learning Systems. pp. 155-163.
14. Simon, Seirafi, Realfsen, Strembeck and Neumann.: An Innovative Service for Learning Performance Monitoring in Businesses. pp. 165-176.
15. Bijlani, Manoj and Rangan.: A-VIEW: A Framework for Interactive eLearning in a Virtual World. pp. 177-187.

2nd Workshop on Social Aspects of the Web (SAW 2008)

May 6th, 2008,
Innsbruck, Austria
http://www.integror.net/saw/

Workshop Co-Chairs

Dominik Flejter, Poznan University of Economics, Poland
Tomasz Kaczmarek, Poznan University of Economics, Poland
Marek Kowalkiewicz, SAP Research Brisbane, Australia

Workshop Program Committee

Krisztian Balog, University of Amsterdam, the Netherlands
Simone Braun, FZI Karlsruhe, Germany
John Breslin, DERI, NUI Galway, Ireland
Tanguy Coenen, Vrije Universiteit Brussel, Belgium
Jon Dron, Athabasca University, Canada
Davide Eynard, Politecnico di Milano, Italy
Andrew T. Fiore, University of California, Berkeley, the USA
Sebastian Kruk, DERI, NUI Galway, Ireland
Marcin Paprzycki, Polish Academy of Science, Poland
Katharina Siorpaes, STI, University of Innsbruck, Austria
Marcin Sydow, Polish-Japanese Institute of Information Technology, Poland
Jie Tang, Tshingua University, China
Celine van Damme, Vrije Universiteit Brussel, Belgium
Valentin Zacharias, FZI Karlsruhe, Germany

Social Network and Data Portability using Semantic Web Technologies

Uldis Bojārs[1], Alexandre Passant[2,3], John G. Breslin[1], Stefan Decker[1]

[1] DERI, National University of Ireland, Galway, Ireland
firstname.lastname@deri.org
[2] LaLIC, Université Paris-Sorbonne, Paris, France
alexandre.passant@paris4.sorbonne.fr
[3] Electricité de France R&D, Clamart, France
alexandre.passant@edf.fr

Abstract. Social network and data portability has recently gained a lot of interest as one of the issues for social media sites on the Web. In this paper, we will show how Semantic Web technologies and especially the FOAF and SIOC vocabularies can be used to model user information and user-generated content in a machine-readable way. Thus, we will see how data and network information can be reused among various services and applications, at almost zero-cost for developers of such tools.

Key words: Social Media, Semantic Web, Web 2.0, Data Portability, FOAF, SIOC

1 Introduction

Social media sites, including social networking services, have captured the attention of millions of users as well as billions of dollars in investment and acquisition. To better enable a user's access to multiple sites, portability between social media sites is required in terms of (1) identification, personal profiles and friend networks and (2) user's content expressed on each site, whether it is about blog posts, pictures, bookmarks or any type of data. Such portability would allow users to easily exchange content between services, or merge and share their social network between various websites. This requires representation mechanisms to interconnect both people and objects on the Web in an interoperable, machine-understandable, and extensible way. The Semantic Web, which is *an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation* [3], provides those required representation mechanisms for portability between social media sites: it links people and objects to record and represent the heterogeneous ties that bind each to the other. The FOAF[1] initiative [8] provides a solution to the first requirement (1), while the SIOC[2] project [7] can address the latter (2). By

[1] Friend-of-a-Friend - http://www.foaf-project.org
[2] Semantically-Interlinked Online Communities - http://sioc-project.org

using agreed-upon Semantic Web formats like FOAF and SIOC to describe people, content objects, and their connections, social media sites can interoperate and provide portable data by appealing to some common semantics. Moreover, the combination of OpenID and FOAF can be used as a backbone for unique identification and profile definition on social media sites, which can in turn be linked to a user's created content via SIOC.

In this paper, we will discuss the application of these technologies to enhance current social media sites with semantics and to address issues with portability between such services. We will show how FOAF and SIOC can provide smart solutions for data portability amongst various social media sites, allowing one to reuse their data and friends networks from one service on other services, as well as interlinking content from one site to another. We will present theoretical aspects as well as scenarios and implementations of such solutions.

2 Overview of Social Network Portability

2.1 Data Portability History

"Social network portability" is the term used to describe the ability to reuse one's own profile across various social networking sites. Brad Fitzpatrick[3] spoke from a developer's point of view about forming a *"decentralised social graph"* [9] and discussed some ideas for social network portability and aggregating one's friends across sites. However, it is not just friends that may need to be ported across social networking sites (and across social media sites in general), but identity and content items as well. Soon afterwards, *"A Bill of Rights for Users of the Social Web"* [11] was authored for social websites who wish to guarantee ownership and control over one's own personal information. As part of this bill, the authors asserted that participating sites should provide social network portability, but that they should also guarantee users *"ownership of their own personal information, including the activity stream of content they create"*, and also stated that *"sites supporting these rights shall allow their users to syndicate their own stream of activity outside the site"*. The Social Graph API[4] from Google is another related effort that provides methods to query aggregated social graph information from the Web. It currently uses formats like XFN and FOAF, which we will talk about later. More recently, the temporary removal of prominent blogger Robert Scoble from Facebook[5] relaunched interest in data ownership and portability amongst different social media sites. The DataPortability project[6] was launched in 2007, with members from various organisations including Facebook, Google and Microsoft coming together to discuss portability issues from technical and legal standpoints. The OpenSocial foundation, recently proposed by Google, Yahoo!

[3] Founder of the LiveJournal blogging community
[4] http://code.google.com/apis/socialgraph/
[5] http://scobleizer.com/2008/01/03/ive-been-kicked-off-of-facebook/
[6] http://dataportability.org

and MySpace also aims to provide APIs to let developers write social applications that access data and networks from various social media websites.

However, to enable a person's transition and / or migration across social media sites, there are significant challenges associated with achieving such portability both in terms of the person-to-person networks and the content objects expressed on each site. Social media sites should be able to collect a person's relevant content items and objects of interest and provide some limited data portability (at the very least, for their most highly used or rated items). We will refer to these items as one's social media contributions, or SMCs. Through such portability, the interactions and actions of a person with other users and objects (on systems they are already using) can be used to create new person or content associations when they register for a new social media site. Rather than requiring proprietary APIs to access this data from each service, we think that uniform representation mechanisms are needed to represent and interconnect people and objects on the Web in an interoperable, extensible way.

2.2 The Semantic Web and Data Portability

The Semantic Web provides such representation mechanisms: it links people and objects to record and represent the heterogeneous ties that bind us to each other. By using agreed-upon Semantic Web formats, like RDF with existing or new ontologies, to describe people, content objects, and the connections that link them together, social media sites can interoperate by appealing to common semantics. Developers are already using Semantic Web technologies to augment the ways in which they create, reuse, and link content on social media sites, and some of them already provide exports from social networking sites in such machine-readable formats. In the other direction, social media sites can serve as rich data sources for Semantic Web applications. As Tim Berners-Lee said in the ISWC 2005 podcast, Semantic Web technologies can support online communities even as *"online communities ... support Semantic Web data by being the sources of people voluntarily connecting things together"*[7]. Such semantically-linked data can provide an enhanced view of individual or community activity across social media sites (for example, *"show me all the content that Alice has acted on in the past three months"*). Thus, we do not consider Web 2.0 and Semantic Web as opposing candidates, but rather we believe that they can be combined with each other to provide a Social Web where data can be exchanged and interlinked no matter where it comes from[1].

In the next section, we will describe how the Semantic Web, and especially FOAF, can be used to define one's profile and can act as a unique entry point for personal data across different social media sites. We will also place emphasis on how it can be used to define not only personal information, but also decentralised social networks, and how a user could re-use this information within Semantic Web compliant social media websites. The second part of this paper will overview the data portability aspect, thanks to the SIOC ontology that provides a way to

[7] http://esw.w3.org/topic/IswcPodcast

describe all data entries for a given user wherever they come from. We will see through an example how it helps to move data from one platform to another. Finally, we will conclude with various thoughts regarding links between social media sites, the Semantic Web, social networks and data portability.

3 Social Network Representation with FOAF

3.1 Identity and Networking Management Across Social Media Sites

While many social media sites allow people to define their social networks, only a few of them permit users to export their networks so that they can be reused across other applications. Moreover, when this is the case, users have to rely on some specific APIs, which means writing ad-hoc tools for each data provider. The FOAF project provides a way to represent social network data in a shared and machine-readable way, since it defines an ontology for representing people and the relationships that they share. While some sites already offer FOAF export, such as LiveJournal[8], MyBlogLog[9] and Hi5.com[10], there are many other social media sites that do not directly expose their data in RDF. However, developers have created different tools to achieve this goal. For example, user profile information is available in RDF thanks to exporters for Flickr[11], Facebook[12] or Twitter[13]. In the latter, this complements machine-readable social network descriptions already embedded via microformats in their pages.

Using FOAF, people and relationships can be modeled using these principles:

- each person is represented as a `foaf:Person` instance and may be assigned URI(s), their unique identifiers on the (Semantic) Web;
- each person has various properties, such as a name (`foaf:name`), nickname (`foaf:nickname`) or birthdate (`foaf:birthday`);
- people can be related to each other using the `foaf:knows` property.

For example, the following snippet of code represents one of the author's profiles created from Flickr using FOAF:

```
flickr:33669349@N00 a foaf:Person ;
  foaf:name "Alexandre Passant" ;
  foaf:mbox_sha1sum "528b95cc44060ceea571d7498a9fd2c7e3ca8a4c" .
  foaf:knows flickr:32233977@N00 .
```

Leveraging Semantic Web representations of people and social networks using widely-adopted ontologies such as FOAF allows us to use generic RDF parsers

[8] http://livejournal.com
[9] http://www.mybloglog.com/
[10] http://hi5.com
[11] http://apassant.net/blog/2007/12/18/rdf-export-of-flickr-profiles-with-foaf-and-sioc/
[12] http://www.dcs.shef.ac.uk/ mrowe/foafgenerator.html
[13] http://sioc-project.org/node/262

and SPARQL[10] (a RDF query language which recently became a W3C recommendation) to browse and reuse data. Thus, end users can use the same tools to parse their network wherever it comes from. To that extent, FOAF simplifies the process of writing tools for developers of social-networking frameworks, especially since many open-source tools are available for most platforms[14] [15].

3.2 Merging and Querying Social Networks

As introduced previously, FOAF allows us to describe personal profiles, but it can also be used to represent relationships between people. Since various sites can export a FOAF representation of users and social networks using their own URIs schemes as shown in the previous RDF snippet, there is still a need to merge and consolidate distributed profiles. In order to consolidate URIs, network owners may rely on Semantic Web best practices that suggest the use of the following properties to represent the identity of existing objects [4] (1) `owl:sameAs` is used to identify that two resources are the same in spite of different URIs and (2) `rdfs:seeAlso` is used to let crawlers and Semantic Web browsers such as Tabulator [2] know where to find additional RDF statements about the resource. Many Semantic Web tools also follow Linked Data guidelines[16] and try to dereference instance URIs, thus providing another way to find additional RDF data.

Using these properties in a distributed and open multi social-network context allows people to interlink and unify the various URIs that represent themselves. To do so, people can reference a main FOAF URI which can be described via a hand-crafted or automatically generated FOAF profile which links to other existing profiles (and also to interlink distributed social networks from various platforms), as the following snippet and Fig.1 describes:

```
:me owl:sameAs flickr:33669349@N00 ;
    owl:sameAs twitter:terraces ;
    owl:sameAs facebook:foaf-607513040.rdf#me .
```

Providing such an entry point allows any RDF-compliant tool to browse one's complete social network in a simple way, i.e. retrieving relationships from Flickr, Twitter or Facebook (1) with standard libraries and SPARQL queries and (2) without having to crawl the Web for data since everything can be accessed from one FOAF file. As an example, we provide a simple script that renders a users' complete social network in a user-friendly Flash interface[17]. This tool only requires the main URI of the user, and thanks to the interlinkage properties described before, it retrieves other URIs and related social networks to render it, as shown on Fig.2. This application, which requires only a few lines of Python and SPARQL queries to parse the complete network, clearly shows the benefits of using common semantics to describe networks on social media websites.

[14] http://www.mkbergman.com/?page_id=346
[15] http://www.w3.org/2001/sw/SW-FAQ#tools
[16] http://www.w3.org/DesignIssues/LinkedData.html
[17] http://apassant.net/home/2008/01/foafgear

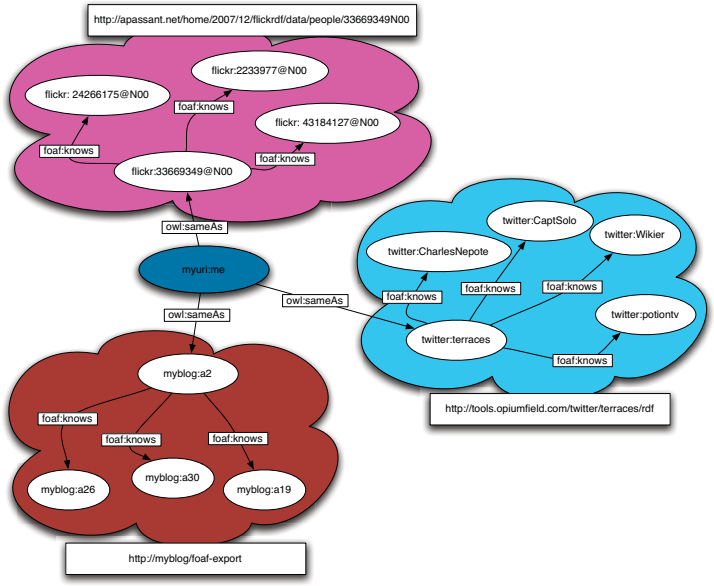


Fig. 1. Interlinking social networks with the Semantic Web

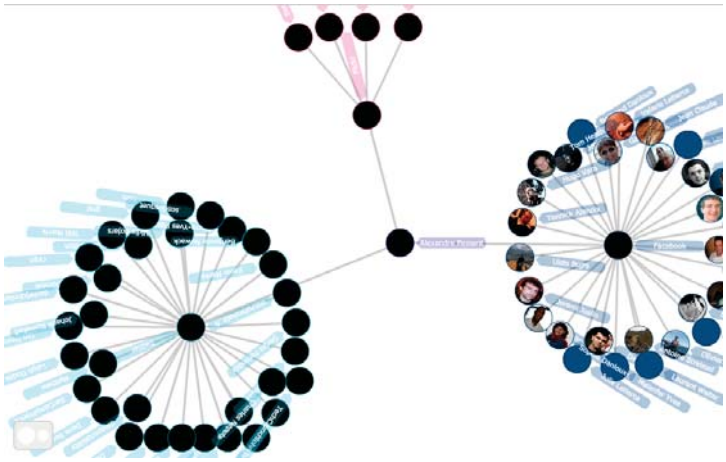


Fig. 2. Browsing a complete social network with a single entry point

Finally, another way to identify uniqueness of someone across various networks is to rely on properties that can uniquely identify him. This is especially useful to merge people among various social networks when they did not explicitly use `owl:sameAs` links. When writing ontologies, OWL offers the ability to define properties as being inverse functional in order to indicate that two RDF descriptions using the same value for this property are "talking" about the same entity. OWL axioms (i.e., `InverseFunctionalProperty`) tell us which

properties can be used in this way, i.e. as indirect identifiers - implementing "reference by description". FOAF uses several properties of this kind, for example, `foaf:mbox_sha1sum` (i.e. scrambled e-mail) and `foaf:openid` (i.e. an OpenID URL). Thus, even if people use different screen names on various websites, as soon as they register with the same email or with the same OpenID URL, they can be uniquely identified across distributed social networks. As an example, the following SPARQL query will retrieve for one user the information about all the people that he or she knows on Flickr and on their weblog (if a contact left some comment there), merging their identities thanks to their `mbox_sha1sum`, whatever their username may be on those platforms.

```
SELECT ?friend ?email
WHERE {
  GRAPH <http://my_flickr_export> {
    :me foaf:knows ?friend .
    ?friend foaf:mbox_sha1sum ?email
  }
  GRAPH <http://my_blog_export> {
    :me foaf:knows ?friend .
    ?friend foaf:mbox_sha1sum ?email
  }
}
```

4 Social Network Portability with FOAF

4.1 Social Networks in Personal Applications

Such semantically-powered social network descriptions can be re-used in existing personal desktop- or web-based tools. For example, Knowee[18] is a web-based application that allows a user to list all of their FOAF URIs, and also includes a microformat parser (which can be used, for example, with Twitter) that features "smushing" capabilities (i.e., identity reasoning), based on user-defined rules as well as on pre-defined ones to uniquely identify people among various social networks descriptions. The network is then browsable using an AJAX-ified user interface. From the desktop point of view, Beatnik[19] provides a semantic address book where you can browse various FOAF profiles and see connections between people. A future development reusing those aspects is the SPARQLPress[20] plug-in for WordPress, that may be used as a personal social network aggregator based on semantic technologies within this popular blogging tool.

[18] http://knowee.org
[19] https://sommer.dev.java.net/source/browse/sommer/trunk/misc/AddressBook/www/
[20] http://wiki.foaf-project.org/SparqlPress

4.2 Reusing Social Networks Across Social Media Site

In order to explain the benefits of such an approach from a portability-between-sites point of view, we will describe the use case of a FOAF-aware social media website.

Bob, a new user, wants to join a (fictional) Networkr service in order to share some pictures and posts with his friends. To do so, he creates an account using his OpenID URL. While the primary advantage with OpenID is that he does not need a new login and password to connect, it allows the site to easily discover his FOAF profile and URI. Indeed, Bob has delegated his OpenID to his own domain name, and added an autodiscovery link in his homepage HTML header to let software agents discover the location of his profile with a single line of code[21]:

```
<head>
<link rel="meta"
  type="application/rdf+xml"
  title="FOAF" href="bob_foaf.rdf" />
</head>
```

The system will then retrieve Bob's profile as well as his URI (thanks to the `foaf:openid` property) and read Bob's social networks to check if any people in one of his existing networks are already registered on Networkr. The service will then ask Bob if he wants to consider all those people as friends on Networkr. Since Bob does not want to grant access to everyone about his activities on this website, he decides to check himself who to add from his existing friends. He then adds photos, and decides to restrict access to only those people he had previously added in Flickr. All of these steps have been efficiently achieved by the website since it just has to query a single profile and the related RDF description of Bob's social graph. Moreover, each time he logs in, Networkr again browses Bob's complete network to retrieve updates and change local access rights if needed. Thus, as soon as Bob adds someone as a friend on Flickr, he will gain access to his new picture gallery. Finally, since Networkr is completely open and consider that the data and social graph belongs to the user, it allows Bob to export his new restricted network, which he can then reuse on other websites. Privacy issues should be considered in those uses cases, to allow more complex access rights definition or restrictions, for example if a user wants certain kind of pictures to be seen only by a subgroup of his Flickr network. Identification and trust may also be a problem to display only relevant information when requesting RDF data from Networkr and the recent RDF authentication discussion[22] could be considered here.

[21] http://wiki.foaf-project.org/Autodiscovery

[22] `http://blogs.sun.com/bblfish/entry/rdfauth_sketch_of_a_buzzword`

5 Data Portability with SIOC

5.1 Describing Social Media Contributions using SIOC

The SIOC initiative was initially established to describe and link discussion posts taking place on online community forums such as blogs, message boards, and mailing lists. As discussions begin to move beyond simple text-based conversations to include audio and video content, SIOC has evolved to describe not only conventional discussion platforms but also new Web-based communication and content-sharing mechanisms [5].

In combination with the FOAF vocabulary for describing people and their friends, and the Simple Knowledge Organization System (SKOS) model for organising thesaurus-like data, SIOC lets developers link user-created content items to other related items, to people (via their associated user accounts), and to topics (using specific "tags" or hierarchical categories). Through its SIOC Type module, SIOC can represent various types of containers (i.e. `Wiki`, `Blog`, `MessageBoard`) and content items (i.e. `WikiArticle`, `BlogPost`, `BoardPost`). Moreover, this is not limited to textual content because SIOC can be also used to represent content such as ImageGallery in the example of the Flickr exporter. Finally, as a good Semantic Web citizen, SIOC reuses and extends existing ontologies such as Dublin Core and FOAF in order to be compatible with RDF data modeled using other existing vocabularies[23].

Various tools, exporters and services have been created to expose SIOC data from existing online communities[24]. These include APIs for PHP, Perl, Java and Ruby, data exporters for systems like WordPress, Drupal, phpBB and BlogEngine.NET, data producers for RFC 4155 mailboxes, SIOC converters for Web 2.0 services like Twitter and Jaiku, and usage in commercial products including Talis Engage and OpenLink Virtuoso.

All of these data sources provide accurate structured descriptions of social media contributions (SMCs) that can be aggregated from different sites (e.g. by person via their user accounts, by co-occurring topics, etc.). Fig.3 shows the process of porting SIOC data from various sources to SIOC import mechanisms for WordPress and future applications. We will now describe the SIOC import plugin for WordPress.

5.2 Importing SIOC Data, with a WordPress Example

The SIOC import plugin[25] for WordPress blog engine is an initial demonstrator for social media portability using SIOC. SIOC import panel in the WordPress administrator user interface (Fig.4) allows a weblog maintainer to import user-created content from another website (described in the form of SIOC data) to their weblog.

[23] SIOC Ontology: Related Ontologies and Vocabularies - http://www.w3.org/Submission/sioc-related/
[24] http://rdfs.org/sioc/applications/
[25] http://wiki.sioc-project.org/w/SIOC_Import_Plugin

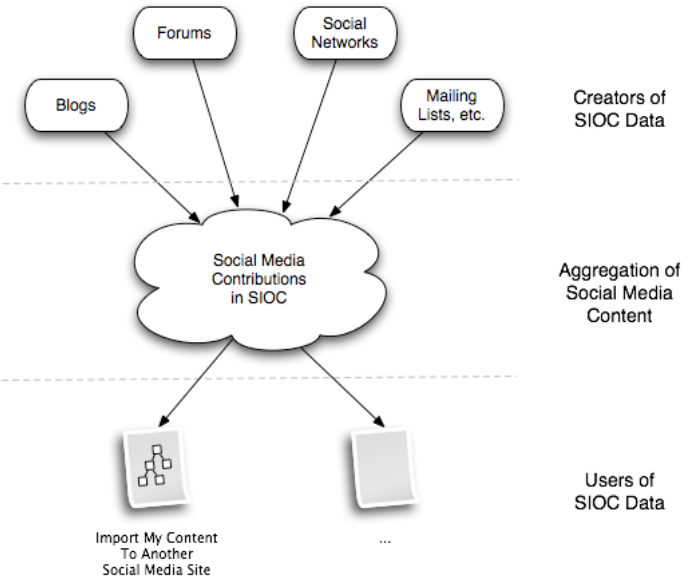


Fig. 3. Porting social media contributions from data providers to import services

Fig. 4. Importing SIOC data in WordPress

Data to be imported can be created from a number of different social media sites using SIOC export tools (as described above) which are at the data creation side of the SIOC "food chain" [6]. Since the only requirement for the importer is to "understand" data modeled with SIOC, it makes no difference wether the data comes from another WordPress blog, a vBulletin message board or your latest updates on Twitter.

For example, a SIOC exporter plugin for a blog engine would create a SIOC RDF representation of every blog post and comment, including information about:

– the content of a post (`sioc:content`)
– the author (`sioc:has_creator`)
– the creation / update date (`dct:created` / `dct:updated`)
– tags and categories (`sioc:topic`)
– all comments on the post (`sioc:has_reply`)
– information about the container blog (`sioc:has_container`)

The use of RDF for data representation enables us to easily extend this data model with new properties when they become necessary, even with needs that are not covered by SIOC itself but that one social media site would require to achieve certain task. We thus benefit from a model which is well-formalised but completely extensible.

The import process implemented by the WordPress SIOC import plugin is the following:

– Parse RDF data (using the open-source ARC[26] RDF parser)
– Find all posts - instance(s) of`sioc:Post` - which exhibit all of the properties required by the target site
– For each post found, it creates a new post using WordPress API calls

The "proof of concept" implementation of SIOC imported worked with a single SIOC file and imports all the posts contained within it. Fig.5 shows an example post imported into WordPress.

Since SIOC is a universal data format and is not specific to any particular site, this pilot implementation already allows us to move content between different blog engines or even between different kinds of social media sites. However, more functionality is needed for data portability and the next iteration of WordPress SIOC importer uses the `sioc:has_reply` property to identify, retrieve (i.e. fetch additional SIOC RDF files describing these comments) and re-create all the comments associated with the blog posts imported. This approach can be further extended by processing all other kinds of objects and information described in source SIOC data.

5.3 Data Portability for a Complete Social Media Site

We will now describe how a SIOC import tool can be extended to port all user-created content from one social media site to another. By starting from a site's main SIOC profile, we retrieve machine-readable information about all the content of this site - starting with the forums hosted therein, and then retrieving the contained posts, comments, and associated users. This extended SIOC import tool retrieves all SIOC data pages (possibly limited by user-defined filters) and to re-create all the data found in this SIOC page on the target social

[26] http://arc.semsol.org

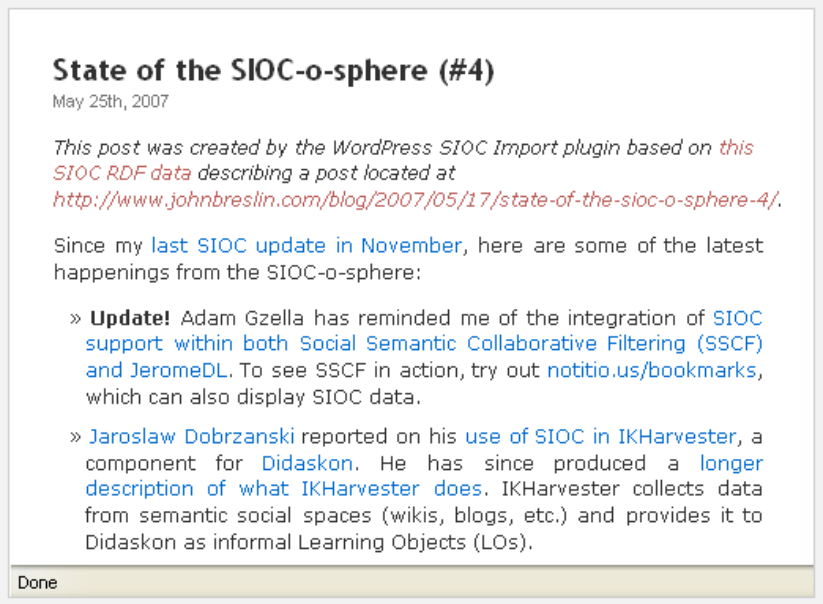


Fig. 5. Imported post in WordPress

media site. Just as can be done with FOAF for social networks, the complete RDF description of a social media site can be found by pointing to a single entry point - a SIOC site profile.

A result of importing the SIOC data for a whole site will be a replica of the original site, including links between objects (e.g. between posts and their comments). Often, a part of the content that a user wants to port is not for public consumption (e.g. if a user is porting some personal information between his accounts). SIOC can be used in this case, but the user will first need to authenticate at the source site and ensure that they have enough privileges to access all the data that need to be migrated.

Another step in social media portability is keeping two sites synchronised (if required): having the same set of users, posts, comments, category hierarchies, etc. In principle, this can be achieved by importing a full SIOC dataset and then monitoring SIOC data feeds for new items added (some SIOC export tools may need to be extended to do this). Implementing this in practice will undoubtedly unfold some interesting challenges, such as real-time synchronisation in two directions, as well as the choice of pushing data or letting services find the data on the Web and update it themselves.

5.4 Perspectives of SIOC Data Portability

An interesting use case for SIOC data portability would be migration between different platforms. For example, this could occur if a person has been using a

mailing list for a particular community, and they then decide that the extended functionality offered to them by a Web-based bulletin board platform is required. Once again, since SIOC can be used to represent various types of containers and items, but using the same format and based on the same high level concepts in the SIOC ontology, moving a mailing-list content to a bulletin board, or a blog post comment to a wiki page can be easily achievable.

While existing web feeds can offer some data portability, the "sliding window" metaphor (providing only last 10-15 items) underlying RSS/Atom contrasts somewhat with the goal of creating a complete archive. SIOC goes further by providing a complete format for representing social media site's data and potentially offering a full archive of site's content. At the same time there remains an overlap in scope here between SIOC and Atom in particular, given that Atom also offers a rich data access and editing protocol.

The discussion-type content items are not the only kind of items that can be ported. The SIOC Types module[27] extends SIOC to be able to describe various Web 2.0 and social media content types. Different types of content items (`Sound`, `MovingImage`, `Event`, `Bookmarks`, etc) can be organised in `sioc:Container`(s) and ported in the same way. While the example in the previous section was based on a WordPress implementation, any social media web site could use the same process to import data from one service to another. For example, using a common data format, a user can post on his blog, others can reply in comments, and then the whole discussion can be moved to a bulletin board or imported as a discussion thread associated with a photo on an image sharing site such as Flickr.

In some use cases, data portability may require selective importing of a specific kind of objects. E.g. a user may decide to port information about all videos (`sioc_t:MovingImage`s) from his website to Youtube, while moving his `sioc_t:Bookmark`s to a bookmark manager such as del.icio.us. What should be kept in mind is that the latter (bookmarks) can be fully expressed in RDF while the former (videos) have a "payload" that will be separate from RDF data and may require specific handling in order to archive it or to transfer from one site to another.

We mainly discussed porting data from one location or site to another, but we can also consider a wider context - personal "life-stream" or activity stream information. Such streams describe all kinds of activities performed by users, including (but not limited to) creation of content items and social network relations. One interesting application using such streams (and of SMC data in SIOC as one kind of activity data) can be "life-stream" archiving where a person retrieves and keeps an archive of all her activities on different social media sites, which can both keep a permanent personal archive and allow different kinds of personal data applications built on top of it.

[27] http://rdfs.org/sioc/types

6 Conclusions and Future Work

In this paper we began by introducing the various challenges regarding social network and data portability and how the Semantic Web can be used to help with achieving this goal. We first introduced the use of FOAF to represent identity and distributed social networks, as well as ways to reuse it across various personal applications or various social media sites. We then demonstrated how SIOC data can be used to represent and port the diverse social media contributions being made by users on various sites. Consequently, both approaches can be merged to retrieve and identify all content objects produced by a single user on various social media sites and services, as well as their friends or social network data, as shown on Fig.6, thus leveraging the transfer of both people and content data between Web 2.0 sites using the Semantic Web.

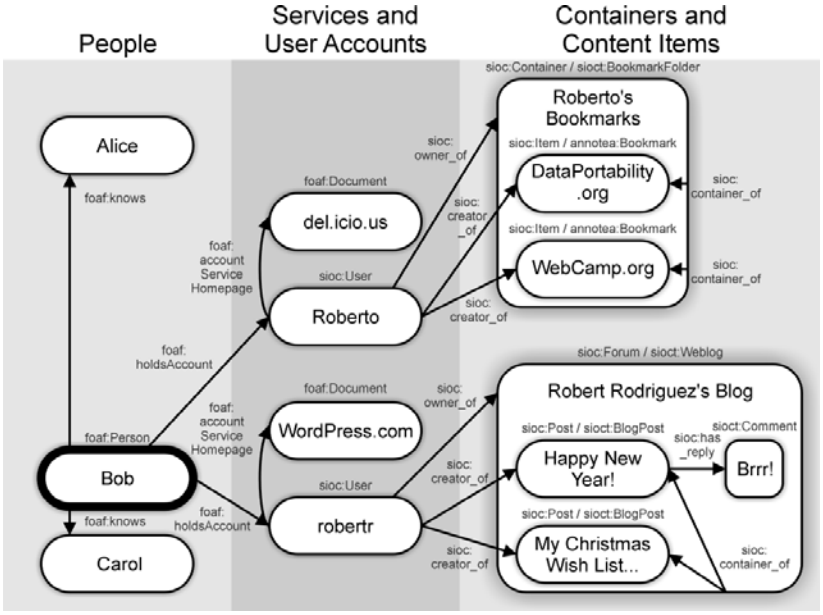


Fig. 6. FOAF, SIOC and Data Portability

For future work, an important issue is who should be allowed to reuse certain data in other sites (as spam blogs are often duplicating other people's content without authorisation for SEO purposes), and what information do people "own" on a social media site and are allowed to port elsewhere. As well as collecting a person's relevant content objects, social media sites may need to verify that a person is allowed to reuse data / metadata from these objects in external systems. This could be achieved by using SIOC and FOAF as representation formats, aggregating content items created by a person (through her user accounts) from various sites, and combining this with some authentication, signature and

trust mechanisms to verify that these items can be reused by the authenticated individual on whatever new sites they choose. When porting content created by others the information about content licenses (e.g. Creative Commons licenses in RDF[28]) will need to be added to SIOC RDF data and taken into account when porting data.

7 Acknowledgments

This material is based upon work supported by Science Foundation Ireland under grant number SFI/02/CE1/I131. Authors would like to thank Dan Brickley, one of the creators of the FOAF vocabulary, for his valuable feedback and suggestions for this paper.

References

1. Anupriya Ankolekar, Markus Krötzsch, Duc Thanh Tran, and Denny Vrandecic. The Two Cultures: Mashing up Web 2.0 and the Semantic Web. *Journal of Web Semantics*, 6(1), FEB 2008.
2. Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.
3. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–44, May 2001.
4. Chris Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web. http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/, 20 July 2007.
5. Uldis Bojārs, John Breslin, Aidan Finn, and Stefan Decker. Using the Semantic Web for Linking and Reusing Data Across Web 2.0 Communities. *The Journal of Web Semantics, Special Issue on the Semantic Web and Web 2.0 (Forthcoming)*, 2008.
6. Uldis Bojārs, John G. Breslin, Vassilios Peristeras, Giovanni Tummarello, and Stefan Decker. Interlinking the Social Web with Semantics. *IEEE Intelligent Systems*, 23(3), May/June 2008.
7. J.G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. In *Proceedings of the Second European Semantic Web Conference, ESWC 2005, May 29–June 1, 2005*, Heraklion, Crete, Greece, 2005.
8. Dan Brickley and Libby Miller. FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.
9. Brad Fitzpatrick and David Recordon. Thoughts on the social graph. http://bradfitz.com/social-graph-problem/, 08 2007.
10. Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. W3C Recommentation, W3C, January 2008.
11. Joseph Smarr, Marc Canter, Robert Scoble, and Michael Arrington. A bill of rights for users of the social web. http://opensocialweb.org/2007/09/05/bill-of-rights/, 4 September 2007.

[28] http://creativecommons.org/ns

Unevenness in network properties on the social Semantic Web

Raf Guns

University of Antwerp, Informatie- en Bibliotheekwetenschap,
CST, Venusstraat 35, 2000 Antwerpen
raf.guns@ua.ac.be

Abstract. This paper studies unevenness in network properties on the social Semantic Web. First, we propose a two-step methodology for processing and analyzing social network data from the Semantic Web, based on the SPARQL query language. After a brief introduction to the notion of unevenness, the methodology is applied to examine unevenness in network properties of real-world data. Comparing Lorenz curves for different centrality measures, it is shown how examinations of unevenness can provide crucial hints regarding the topology of (social) Semantic Web data.

Key words: social network analysis, Semantic Web, SPARQL, unevenness

1 Introduction

The *social Semantic Web* is a broad, non-technical term, referring to data on the Semantic Web (encoded in RDF) that contain social information. The most prevalent ontology on the social Semantic Web is the FOAF (Friend Of A Friend) vocabulary [8]. Yet, FOAF is not alone; in this paper, for instance, we will use a socio-cultural ontology (section 4).

The Semantic Web [5] in general is conceived as a large-scale distributed information system. While some constituents are still in development and its current uptake is relatively modest, the Semantic Web graph already shows the traits of a complex system. As such, it is characterized by [3, 15]:

Skewed degree distribution: The probability $P(k)$ that a node has degree k (is connected to k other nodes) is not randomly distributed. Instead, it follows a power law $P(k) \approx Ak^{-\gamma}$. Moreover, complex systems typically exhibit power law distributions in more than one way. With regard to the Semantic Web, previous research has shown that a diversity of relations — such as the relation between websites and their number of Semantic Web documents or the relation between an ontology and its number of uses — follows a power law [13].

Small world properties: Made famous by Stanley Milgram's [20] letter experiment, the small world notion refers to the fact that the average shortest path

length in a graph is very short (comparable to that of a random graph). More recently, several models have been proposed to account for the small-world effect [21, 27].

High clustering: The neighbours of a given node are likely also neighbours of each other.

Similar traits have been discovered for a variety of social and biological networks [10]. However, these properties also raise several questions. In this paper, we will address two of them. Both questions will be discussed and demonstrated on a real-world socio-cultural data set.

First, how can data on the social Semantic Web be used for Social Network Analysis (SNA)? Significant research in this area has already been performed by, among others, Li Ding and colleagues [12] and Peter Mika [19]. Much work has concentrated on acquiring and aggregating data (often FOAF data), – especially merging information about unique persons turns out to be far from trivial. In the present paper, we concentrate on the development of a methodology for using one single RDF graph as the 'master', which can be used as the basis for several kinds of SNA. Ideally, we want to keep as much information as possible and extract a multitude of potentially interesting relations. This particular aspect has received less attention so far.

Second, it is very rarely examined *how* skewed a distribution is. How can this notion be measured? Quantification of unevenness is crucial for a thorough understanding of a power law distribution; moreover, it can be used for comparison purposes between distributions and between networks.

2 Two-step methodology

Semantic Web data can be stored in many different ways: as a (set of) document(s) in one of the many RDF syntaxes [4]; in a 'classic' relational database; or in a *triplestore*, a dedicated RDF database. For the remainder of this paper, we assume the use of a triplestore (see [17] for an overview of triplestores), using Jena[1] as an example. Triplestores can be queried with a query language like SPARQL [23].

Partly due to its distributed nature, Semantic Web data may appear quite dazzling: many different kinds of data, drawn from several ontologies, between which a multitude of relations exist. How can one make heads or tails out of them?

Assuming the existence of a set of fairly clearly defined questions to be answered, we propose a two-step methodology, which critically depends on SPARQL (or a query language with similar capabilities). In short, the two steps are:

[1] Internally, Jena uses a relational database, but the interface is similar to other triplestores, see `http://jena.sourceforge.net/DB/creating-db-models.html`.

1. Construct an *extraction query* in SPARQL and apply it to the RDF graph. This yields a secondary graph, specifically oriented towards the question(s).
2. Convert the secondary graph to a format intended for SNA.

We will now discuss both steps in greater detail.

2.1 Constructing an extraction query

SPARQL queries are usually `SELECT` queries, which return a table of results. For the extraction query, we employ `CONSTRUCT` queries, which return a new RDF graph. A similar architecture can also be found in the MESUR project [7, 24].

First, we compare the original graph in the triplestore and the questions to be answered. Some questions simply involve the *extraction* of parts of the RDF graph (ignoring the rest). A typical example would be the extraction of all *foaf:knows* relations from a FOAF triplestore. This can actually be done without SPARQL, but for the sake of illustration we give a possible extraction query:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

CONSTRUCT { ?p1 foaf:knows ?p2 }
WHERE {
  ?p1  a           foaf:Person ;
       foaf:knows  ?p2 .
  ?p2  a           foaf:Person .
}
```

Other questions are trickier, in that they require knowledge on how relations in the model interact, — these involve *extraction and combination* of parts of the model. Let's use the IngentaConnect MetaStore project [22], a large-scale database of academic articles, as an example. Fig. 1 shows how article citations are expressed in MetaStore. The citation relation *between authors* can then be queried as follows.

```
BASE <http://metastore.ingentaconnect.com>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX prism: <http://prismstandard.org/namespaces/1.2/basic>
PREFIX ex: <http://example.com/ns/>

CONSTRUCT { ?author1 ex:cites ?author2 }
WHERE {
  ?art1  a                </ns/structure/Article> ;
         foaf:maker       ?author1 ;
         prism:references ?art2 .
  ?art2  a           </ns/structure/Article> ;
         foaf:maker  ?author2 .
}
```

Some remarks are in order. 1) This example query is rather crude and would have to be expanded to handle multiple authorship. 2) Some queries are easier to perform with one or more intermediate extraction queries. 3) Although extraction queries are obviously not as powerful as a dedicated program or full-fledged reasoner, they are often sufficient and much faster to implement.[2]

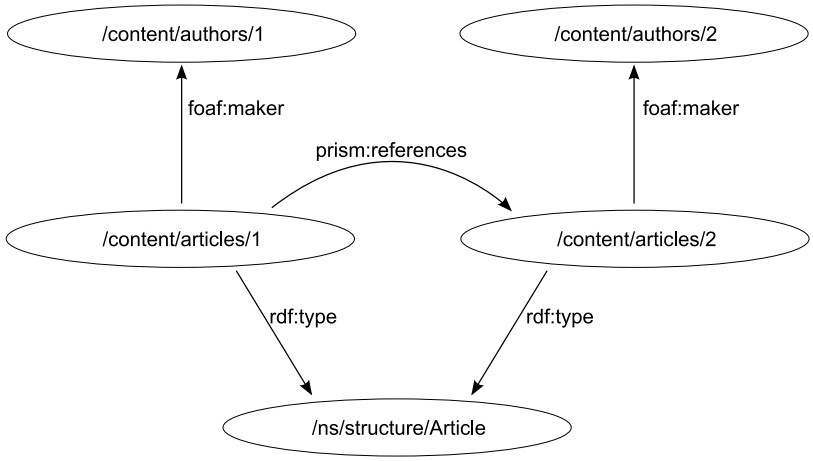


Fig. 1. Citation relation in the IngentaConnect MetaStore [22] with base URI `http://metastore.ingentaconnect.com`

2.2 From secondary graph to SNA format

Once a secondary graph has been obtained, it can be studied. There exist several projects for visualizing and exploring RDF and FOAF data, such as FOAF Explorer,[3] RDF-Gravity[4] and Visual Browser.[5] These tools, however, generally do not provide SNA measures like centrality and clustering. Moreover, they generally do not scale to very large graphs.

Thus, while not strictly necessary, this step ensures compatibility with other SNA efforts and permits techniques that are difficult to perform on plain RDF graphs. We handle these conversions by integrating with pyNetConv, a Python library that can convert to Pajek, NetworkX, CytoScape, GML, ...

[2] Some triplestores, like Jena, also allow custom SPARQL functions.
[3] `http://xml.mfd-consult.dk/foaf/explorer/`
[4] `http://semweb.salzburgresearch.at/apps/rdf-gravity/`
[5] `http://nlp.fi.muni.cz/projekty/visualbrowser/`

3 Unevenness

The distribution of degrees on the Semantic Web is — like many other relations — highly uneven: a small number of nodes has a huge amount of links, while the vast majority has very few. How can this unevenness be quantified?

Unevenness or inequality has been studied extensively in econometrics and informetrics. Since not all existing measures satisfy all necessary requirements [1, 14], we will limit the present discussion to two methods, using the following array as an example: X = (1, 3, 4, 7, 10, 15). These numbers could e.g. express the distribution of wealth or the distribution of degrees for a set of nodes. Clearly, there is some unevenness, but how much exactly?

The *Lorenz curve* [18] is a graphical representation of unevenness. First, we determine the relative amounts:

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{1}$$

resulting in $(\frac{1}{40}, \frac{3}{40}, \frac{1}{10}, \frac{7}{40}, \frac{1}{4}, \frac{3}{8})$. The horizontal axis of the Lorenz curve has the points i/N $(i = 1, 2, \ldots, N)$. The vertical axis of the Lorenz curve has their cumulative fraction: $a_1 + a_2 + \ldots + a_i$. We thus construct the Lorenz curve (Fig. 2). The diagonal line represents the case of perfect evenness. The further the curve is removed from the diagonal, the greater the unevenness. Note that we have ranked our numbers in increasing order, resulting in a convex Lorenz curve. The concave Lorenz curve results from ranking in decreasing order and is completely equivalent. Complete unevenness — one person has everything, and the rest nothing — would be represented as a curve following the bottom and the right side of the plot.

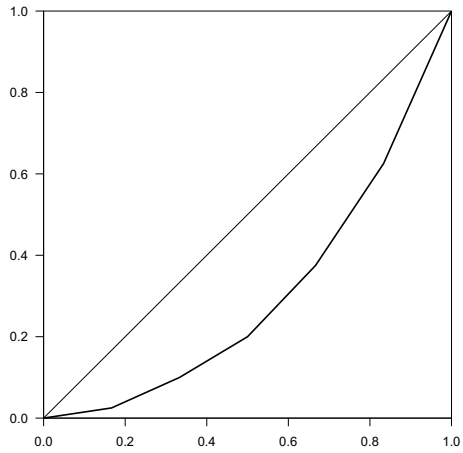


Fig. 2. Convex Lorenz curve of the array (1, 3, 4, 7, 10, 15)

Suppose we want to express this unevenness in a number. A good measure is the *Gini evenness index* G' [25], originally devised to characterize the distribution of wealth and poverty [16],

$$G'(X) = \frac{2}{\mu N^2}\left(\sum_{j=1}^{N}(N+1-j)x_j\right) - \frac{1}{N} \tag{2}$$

with x_j ranked in increasing order and μ the mean of the set x_j. $G' = 2 \times$ the area under the convex Lorenz curve.

Lorenz curves determine a *partial order*: if one convex Lorenz curve is completely below another, then the former expresses less evenness than the latter. It should be stressed that Lorenz curves may 'overlap' or cross each other. In these cases, no order can be determined [25].

4 Example: Agrippa

For this example, we use data from the *Agrippa* database, the catalogue and database of the Archive and Museum of Flemish Cultural Life (AMVC Letterenhuis, Antwerp). Agrippa contains a wealth of information about both the archived materials and the socio-cultural actors that have created them. The RDF version uses existing ontologies like FOAF and Dublin Core, where applicable. The graph is stored in a Jena triplestore and made available via the SPARQL protocol [11] using Joseki.[6] Through this protocol, SPARQL queries can be submitted to a centralized server.

Many secondary graphs can be derived. The following, for instance, constructs a bipartite graph of persons and their affiliations to organizations.

```
PREFIX agrippa: <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?person agrippa:affiliatedWith ?org }
WHERE {
  ?aff agrippa:hasAffiliator ?org .
  ?aff agrippa:hasAffiliatee ?person .
}
```

Agrippa also contains information about 237,062 letters. We construct a simple graph that links author(s) and recipient(s) of each letter:

```
PREFIX agrippa: <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?sender <urn:agrext#writesLetterTo> ?recipient }
WHERE {
  ?context agrippa:hasLetterWriter ?sender .
  ?context agrippa:hasRecipient ?recipient .
}
```

[6] `http://joseki.sourceforge.net`

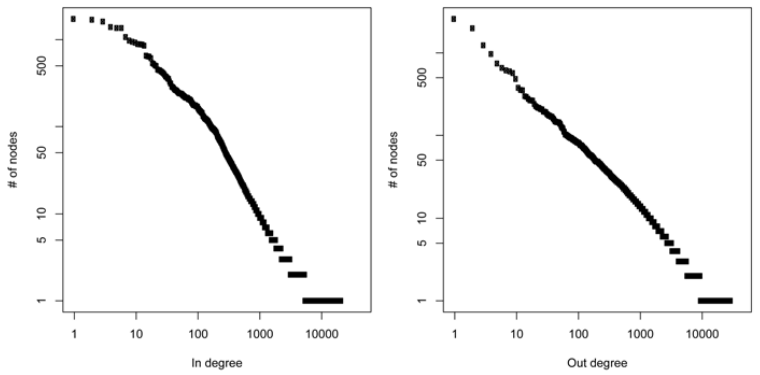


Fig. 3. Zipf distribution for in-degree and out-degree

We will take this author-recipient graph ($N = 40,914$) as an example. Each node is connected by 5.08 links on average, but the actual in- and out-degree follow a Zipf distribution (Fig. 3). Apart from degree centrality (DC), we also consider the following two centrality measures [26]:

Betweenness centrality (BTC): characterizes the importance of a given node for establishing short pathways between other nodes.

Closeness centrality (CC): characterizes how fast other nodes can be reached from a given node.

Comparing the Lorenz curves of the three centrality measures reveals a remarkably diversified picture, shown in Fig. 4. BTC is clearly more uneven than the other two. In spite of the initial appearance, no order can be determined between DC and CC, since the curves overlap slightly at the bottom (recall that the Lorenz curve imposes only a partial order). The Gini evenness indices are: $G'(\mathrm{BTC}) = 0.02 < G'(\mathrm{DC}) = 0.25 < G'(\mathrm{CC}) = 0.98$.

As a tentative explanation, we suggest that these differences may be due to the small-world effect [21, 27]. Even marginal nodes are relatively close to all others, accounting for minimal differences in closeness. Indeed, the length of the diameter — the longest shortest path — is only 11 and the average shortest path length only 3.85! The graph is not fully connected, but the main component ($N = 40,303$) accounts for the vast majority of nodes. The core of the main component is the Largest Strongly Connected Component or LSCC ($N = 9,723$), a component in which any node can be reached (obeying the direction of the links).[7] The LSCC itself has a nucleus of *hubs* [10], nodes with extremely high DC, through which almost all other shortest paths pass. This

[7] As a whole, the graph fits the bow-tie model [6, 9], previously devised for link structure on the World Wide Web.

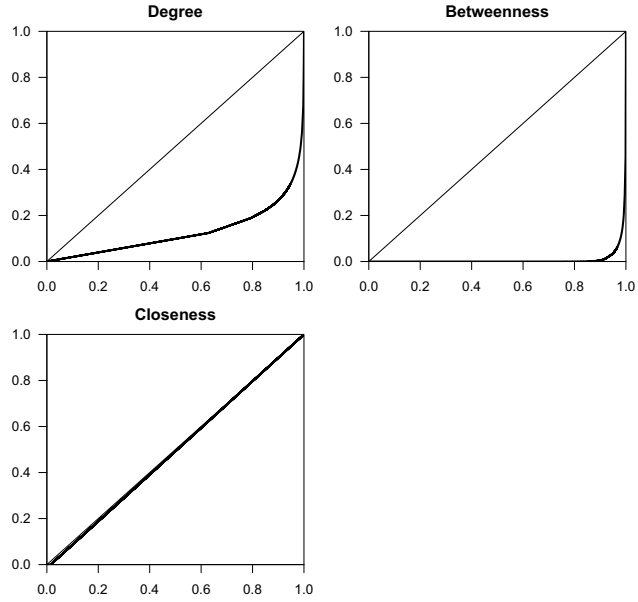


Fig. 4. Lorenz curves for degree, betweenness and closeness centrality

increases closeness for the network as a whole and brings about a very uneven BTC distribution.

5 Conclusions

We have shown how SPARQL can be used in processing social Semantic Web data in a simple two-step methodology, converting the primary graph to a better suited secondary graph. While SPARQL is obviously less powerful than a 'real' reasoning engine or a dedicated program, it is often sufficient and may well prove simpler and faster to implement. RDF tools are generally not geared towards SNA, although Flink [19] incorporates some basic SNA statistics. Generally, conversion to other formats is recommendable but, luckily, straightforward.

The Lorenz curve and the Gini evenness index G' are two excellent methods for studying unevenness. Taking Agrippa as a concrete example, it can be seen that unevenness measures may confirm or enforce hypotheses regarding the network topology. In the example discussed, the massive difference between BTC and CC distribution confirms the small-world hypothesis and reveals the topology of the graph with a small nucleus, through which most other paths must pass.

Most of these results, such as the establishment of the small-world effect, could have been achieved without studying the unevenness of network properties. Consequently, the current paper should be regarded as a first step: it illustrates

how unevenness measures can be used to achieve similar results as existing, well-established methods. In future research, we hope to expand upon these results by studying a greater variety of network properties and (social) networks, including different classes of small-world networks [2].

Acknowledgements: I thank prof. Richard Philips for providing access to the Agrippa dataset and the anonymous reviewers for useful comments on an earlier version.

References

1. Allison, P.D.: Measures of inequality. *American Sociological Review*, 43(6), 865–880 (1978)
2. Amaral, L. A., Scala, A., Barthelemy, M., Stanley, H. E.: Classes of small-world networks. *PNAS*, 97(21), 11149–11152 (2000)
3. Bachlechner, D., Strang, T.: Is the Semantic Web a small world? In: *Proceedings, Second International Conference on Internet Technologies and Applications (ITA 07)*, `http://elib.dlr.de/47899/` (2007)
4. Becket, D.: New syntaxes for RDF. In: *WWW 2004*, `http://www.dajobe.org/2003/11/new-syntaxes-rdf/paper.html` (2004)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American Magazine*, 284(5), 34–43 (2001)
6. Björneborn, L.: *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, RSLIS, Copenhague (2004)
7. Bollen, J., Rodriguez, M. A., Van de Sompel, H., Balakireva, L. L., Hagberg, A.: The largest scholarly semantic network... ever. In: *Proceedings of the 16th International Conference on the World Wide Web*, ACM Press, 1247–1248 (2007)
8. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.91. Namespace Document 2 November 2007 – OpenID Edition, `http://xmlns.com/foaf/spec/` (2007)
9. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the Web. *Computer Networks*, 33(1–6), 309–320 (2000)
10. Christensen, C., Albert, R.: Using graph concepts to understand the organization of complex systems. `http://arxiv.org/abs/q-bio/0609036` (2006)
11. Clark, K. G., Feigenbaum, L., Torres, E.: SPARQL Protocol for RDF. W3C Recommendation 15 January 2008, `http://www.w3.org/TR/rdf-sparql-protocol/` (2008)
12. Ding, L., Finin, T., Joshi, A.: Analyzing social networks on the Semantic Web. *IEEE Intelligent Systems*, 1(9) (2005)
13. Ding, L., Finin, T.: Characterizing the Semantic Web on the web. In: *Proceedings of the International Semantic Web Conference*, Springer (2006)
14. Egghe, L., Rousseau, R.: Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, 42(7), 479–489 (1991)
15. Gil, R., García, R.: Measuring the Semantic Web. In: *Advances in Metadata Research, Proceedings of MTSR '05*, Rinton Press (2006)
16. Gini, C.: Il diverso accrescimento delle classi sociali e la concentrazione della richezza. *Giornale degli Economisti*, 11(37) (1909)
17. Lee, R.: Scalability report on triple store applications. `http://simile.mit.edu/reports/stores/` (2004)

18. Lorenz, M. O.: Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70), 209–219 (1905)
19. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2), p. 211–223 (2005)
20. Milgram, S.: The small world problem. *Psychology Today*, 2(1), 60–67 (1967)
21. Newman, M. E. J.: Models of the small world. *Journal of Statistical Physics*, 101(3), 819–841 (2000)
22. Portwin, K., Parvatikar, P.: Building and managing a massive triple store: An experience report. *XTech 2006: "Building Web 2.0"*, `http://2006.xtech.org/schedule/paper/18/` (2006)
23. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation 15 January 2008, `http://www.w3.org/TR/rdf-sparql-query/` (2008)
24. Rodriguez, M. A., Bollen, J., Van de Sompel, H.: A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In: *JCDL '07. Proceedings of the 2007 Conference on Digital Libraries*, ACM Press, 278–287 (2007)
25. Rousseau, R.: Lorenz curves determine partial orders for comparing network structures. *Critical Events in Evolving Networks (CREEN) Workshop, Brussels* (2007)
26. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences 8, Cambridge University Press (1994)
27. Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442 (1998)

Trustlet, Open Research on Trust Metrics

Paolo Massa and Kasper Souren

FBK/rst
Via Sommarive, 14
Povo (TN) - Italy
{massa,souren}@fbk.eu

Abstract. A trust metric is a technique for predicting how much a user of a social network might trust another user. This is especially beneficial in situations where most users are unknown to each other such as online communities. We think the recent tumultuous evolution of social networking demands for a collective research effort. With this in mind we created Trustlet.org, a platform consisting of a wiki for open research on trust metrics. The goal of Trustlet is to collect and distribute trust network datasets and trust metrics code as free software, in order to facilitate the comparison of different trust metrics algorithms and a more coherent progress in this field. At present we made available some social network datasets and code for some trust metrics. In this paper we also report a first empirical evaluation of different trust metrics on the Advogato social network dataset.

Key words: Trust Metrics, Social network analysis, Wiki, Advogato, Free software, Data acquisition, Science Commons

1 Introduction

In our current society it is more and more common to interact with strangers, people who are totally unknown to us. This happens for example when receiving an email asking for collaboration or advise from an unknown person, when we rely on reviews written by unknown people on sites such as Amazon.com, and also when browsing random profiles on social networking sites such as Facebook.com or Linkedin.com. Even more surprising is the fact a huge amount of commercial exchanges happen now between strangers, facilitated by platforms such as Ebay.com. In all systems in which it is possible to interact with unknown people, it is important to have tools able to suggest which other users can be trustworthy enough for engaging with. Trust metrics and reputation systems [1] have precisely this goal and become even more important, for instance, in systems where people are connected in the physical world such as carpooling systems or hospitality exchange networks (i.e. couchsurfing.com), in which users accept to have strangers into their car or their house.

A commonly cited definition of trust was proposed by Diego Gambetta: "Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before [we] can

monitor each action (or independently of his capacity of ever be able to monitor it) and in a context in which it affects [our] own action" [3]. In all the previous example it is possible to consider the social relationship users can express as a trust statement, an explicit statement stating "I trust this person in this context" (for example as a pleasant guest in a house or as a reliable seller of items) [2].

While research about trust issues spanned disciplines as diverse as economics, psychology, sociology, anthropology and political science for centuries, it is only recently that the widespread availability of modern communication technologies facilitated empirical research on large social networks, since it is now possible to collect real world datasets and analyze them [2]. As a consequence, recently computer scientists and physicists started contributing to this research field as well [4, 5].

Moreover we all start relying more and more on these social networks, for friendship, buying, working, ... As this field become more and more crucial, in the past few years many trust metrics have been proposed but there is a lack of comparisons and analysis of different trust metrics in the same conditions. As Sierra and Sabater put it in their complete "Review on Computational Trust and Reputation Models" [6]: "Finally, analyzing the models presented in this article we found that there is a complete absence of test-beds and frameworks to evaluate and compare the models under a set of representative and common conditions. This situation is quite confusing, specially for the possible users of these trust and reputation models. It is thus urgent to define a set of test-beds that allow the research community to establish comparisons in a similar way to what happens in other areas (e.g. machine learning)" (emphasis added). Our goal is to fill this void and for this reason we set up Trustlet [7], a collaborative wiki in which we hope to aggregate researchers interested in trust and reputation and build together a lively test-bed and community for trust metrics evaluation. A project with similar goals is the Agent Reputation and Trust (ART) Testbed [8]. However ART is more focused on evaluating different strategies for interactions in societies in which there is competition and the goal is to perform more successfully than other players, in a specific context. Our take with Trustlet is about evaluating performances of trust metrics in their ability to predict how much a user could trust another user, in every context. For this reason, we want also to support off-line evaluation of different trust metrics on social network datasets. The two testbeds are hence complementary.

In this paper we describe Trustlet, the reason behind its creation and its goals, we report the datasets we have collected and released and the trust metrics we have implemented and we present a first empirical evaluation of different trust metrics on the Advogato dataset.

2 Trust metrics

Trust metrics are a way to measure trust one entity could place in another. After a transaction user Alice on Ebay can explicitly express her subjective

level of trust in user Bob. We model this as a trust statement from Alice to Bob. Trust statements can be weighted, for example on Advogato [9] a user can certify another user as Master, Journeyer, Apprentice or Observer, based on the perceived level of involvement in the free software community. Trust statements are directed and not necessary symmetric: it's possible a user reciprocates with a different trust statement or simply not at all. By aggregating the trust statements expressed by all the members of the community it is possible to build the entire trust network (see for example Figure 1). A trust network is hence a directed, weighted graph. In fact trust can be considered as one of the possible social relationships between humans, and trust networks a subclass of social networks [4, 5].

Trust metrics are tools for predicting the trust a user could have in another user, by analyzing the trust network and assuming that trust can somehow be propagated. One of the assumptions is that people are more likely to trust a friend of a friend than a random stranger [12, 10, 11, 9].

Trust metrics can either be local or global [10, 12]. A global trust metric is a trust metric where predicted trust values for nodes are not personalized. On the other hand, with local trust metrics, the trust values a user sees for other users depend on her position in the network. In fact, a local trust metric predicts trust scores that are personalized from the point of view of every single user. For example a local trust metric might predict "Alice should trust Carol as 0.9" and "Bob should trust Carol as 0.1", or more formally trust(A,C)=0.9 and trust(B,C)=0.1. Instead for global trust metrics, trust(A,B)=reputation(B) for every user A. This global value is sometimes called reputation [2]. Currently most trust metrics used in web communities are global, mainly because they are simpler to understand for the users and faster to run on central servers since they have to be executed just once for the entire community. For example Ebay and Pagerank [13] are global. However we think that soon users will start asking for systems that take into account their own peculiar points of view and hence local trust metrics, possibly to be run in a decentralized fashion on their own devices.

While research on trust metrics is quite recent, there have been some proposals for trust metrics. We briefly review some of them for later mention in the evaluation presented in Section 4, although our goal is not to provide a complete review of trust metrics here.

Ebay web site shows the average of the feedbacks received by a certain user in her profile page. This can be considered as a simple global trust metric, which predicts, as trust of A in B, the average of all the trust statements received by B [12].

In more advanced trust metrics trust can be extended beyond direct connections. The original Advogato trust metric [9] is global, and uses network flow to let trust flow from a "seed" of 4 users, who are declared trustworthy a priori, towards the rest of the network. The network flow is first calculated on the network of trust statements whose value is Master (highest value) to find who classifies as Master. Then the Journeyer edges are added to this network and the

network flow is calculated again to find users who classify as Journeyer. Finally the users with Apprentice status are found by calculating the flow on all but the Observer edges. The untrusted Observer status is given if no trust flow reached a node. By replacing the 4 seed users for an individual user A, Advogato can also be used as a local trust metrics predicting trust from the point of view of A.

The problem of ranking of web pages in the results of a search engine query can be regarded under a trust perspective. A link from page A to page B can be seen as a trust statement from A to B. This is the intuition behind the algorithm Pagerank [13] powering the search engine Google. Trust is propagated with a mechanism resembling a random walk over the trust network.

Moletrust [12] is a local trust metric. Users are ordered based on their distance from the source user, and only trust edges that go from distance n to distance n+1 are regarded. The trust value of users at distance n only depend on the already calculated trust values at distance n-1. The scores that are lower than a specific threshold value are discarded, and the trust score is the average of the incoming trust statements weighted over the trust scores of the nodes at distance n-1. It is possible to control the locality by setting the trust propagation horizon, i.e. the maximum distance to which trust can be propagated.

Golbeck proposed a metric, TidalTrust [11], that is similar to Moletrust. It also works in a breadth first search fashion, but the maximum depth depends on the length of the first path found from the source to the destination. Another local trust metric is Ziegler's AppleSeed [10], based on spreading activation models, a concept from cognitive psychology.

3 Datasets and trust metrics evaluation

Research on trust metrics started a long time ago, but is somehow still in its infancy. The first trust metric could be even ascribed to the philosopher John Locke who in 1680 wrote: "Probability then being to supply the defect of our knowledge, the grounds of it are these two following: First, the conformity of anything with our own knowledge, observation and experience. Secondly, The testimony of others, vouching their observation and experience. In the testimony of others is to be considered: (1) The number. (2) The integrity. (3) The skill of the witnesses. (4) The design of the author, where it is a testimony out of a book cited. (5) The consistency of the parts and circumstances of the relation. (6) Contrary testimonies" [14]. This quotation can give an idea of how many different models for representing and exploiting trust have been suggested over the centuries. However of course John Locke in 1680 didn't have the technical means for empirically evaluating his "trust metric". Even collecting the required data about social relationships and opinions was very hard in old times. The first contributions in analysis real social networks can be tracked down to the foundational work of Jacob Moreno [15] (see Figure 1) and since then many sociologists, economists and anthropologists have researched on social networks and trust. But the advent of the information age has made it possible to collect,

represent, analyze and even build networks way beyond that what is possible with pen and paper. Computer scientists and physicists have become interested in social networks, now that both huge amounts of data have become available and computing power has advanced considerably [4, 5].

At Trustlet.org we have started a wiki to collect information about research on trust and trust metrics. We hope to attract a community of people with interest in trust metrics. We have chosen to use the Creative Commons Attribution license so that work can easily (and legally) be reused elsewhere. Our effort shares the vision of the Science Commons project [1] which tries to remove unnecessary legal and technical barriers to scientific collaboration and innovation and to foster open access to data. We have also started a repository of the software we create for our analysis, written in Python and available as Free Software under the GNU General Public License [2].

We believe the lack of generally available datasets is inhibiting scientific work. It's harder to test a hypothesis if it has been tested on a dataset that is not easily available. The other alternative is testing the hypothesis on synthesized datasets, which are hardly representative of real-world situations. Prior to the proliferation of digital networks data had to be acquired by running face-to-face surveys, which could take years to collect data of a mere couple of hundreds of nodes. The proliferation and popularity of on-line social networks has facilitated acquiring data, and the implementation of standards like XFN and common APIs like OpenSocial opens up new possibilities for research [2]. A more widespread availability and controlled release of datasets would surely benefit research and this is one of the goal behind the creation of Trustlet.

Trust network datasets are are directed, weighted graphs. Nodes are entities such as users, peers, servers, robots, etc. Directed edges are trust relationships, expressing the subjective level of trust an entity expresses in another entity [2].

We think it is important that research on trust metrics follows an empirical approach and it should be based on actual real-world data. Our goal with Trustlet is to collect as many datasets as possible in one single place and release them in standard formats under a reasonable license allowing redistribution and, at least, usage in a research context. At present, as part of our effort with Trustlet, we collected and released datasets derived from Advogato, people.squeakfoundation.org, Robots.net and Epinions.com[3].

We describe in detail the Advogato dataset since our experiments (section 4) are run on it. Advogato.org is an online community site dedicated to free software development, launched in November 1999. It was created by Raph Levien, who also used Advogato as a research testbed for testing his own attack-resistant trust metric, the Advogato trust metric [9]. On Advogato users can certify each other as several levels: Observer, Apprentice, Journeyer or Master. The Advogato trust metric uses this information in order to assign a global certification level to every user. The goal is to be attack-resistant, i.e. to reduce the impact of

[1] Science Commons http://sciencecommons.org

[2] GNU General Public License http://www.gnu.org/licenses/gpl.html

[3] See http://www.trustlet.org/wiki/Trust_network_datasets

attackers [9]. Precise rules for giving out trust statements are specified on the Advogato site. Masters are supposed to be principal authors of an "important" free software project, excellent programmers who work full time on free software, Journeyers contribute significantly, but not necessarily full-time, Apprentices contribute in some way, but are still acquiring the skills needed to make more significant contributions. Observers are users without trust certification, and this is also the default. It is also the level a user certifies another user at to remove a previously expressed trust certification.

For the purpose of this paper we consider these certifications as trust statements. T(A,B) denotes the certification expressed by user A about user B and we map the textual labels Observer, Apprentice, Journeyer and Master in the range [0,1], respectively in the values 0.4, 0.6, 0.8 and 1.0. This choice is arbitrary and considers all the certifications are positive judgments, except for "observer" which is used for expressing less-than-sufficient levels. For example, we model the fact raph certified federico as Journeyer as T(raph, federico)=0.8.

The Advogato social network has a peculiarly interesting characteristic: it is almost the only example of a real-world, directed, weighted, large social network. However, besides the leading work of Levien reported in his unfinished PhD thesis [9], we are just aware of another paper using the Advogato dataset which is focused on providing a trust mechanism for mobile devices [16].

There are other web communities using the same software powering Advogato.org and they have the same trust levels and certifications system: robots.net, persone.softwarelibero.org, people.squeakfoundation.org, kaitiaki.org.nz. We collected daily snapshots of all these datasets and made them available on Trustlet but we haven't used them for our analysis in this paper, mainly because they are much smaller than the Advogato dataset. Details about the characteristics of the Advogato trust network dataset are presented in Section 4.

The other set of datasets we released is derived from Epinions.com, a website where users can leave reviews about products and maintain a list of users they trust and distrust based on the reviews they wrote [12].

Both released datasets and datasets we are considering for collection are available on Trustlet. Besides aiming at releasing datasets in a coherent format, we also released the Python code we wrote for the main trust metrics presented in section 3 and some baseline trust metrics, under a free software license so that code can be reused and inspected.

4 Initial research outcomes

In the previous sections we highlighted the reasons for creating Trustlet and the way we hope it can develop into a collaborative environment for the research of trust metrics. As a first example of what we hope Trustlet will be able to bring to research on trust metrics, we report our first investigation and empirical findings.

We chose to start studying the Advogato social network because of its almost unique characteristic. Trust statements (certifications) are weighted and this makes it a very useful dataset for researching trust metrics: most networks just

exhibit a binary relationship (either trust is present or not) and the evaluation on trust metrics performances is less insightful.

The Advogato dataset we analyzed is a directed, weighted graph with 7294 nodes and 52981 trust relations. There are 17489 Master judgments, 21977 for Journeyer, 8817 for Apprentice and 4698 for Observers. The dataset is comprised of 1 large connected component, comprising 70.5% of the nodes, the second largest component contains 7 nodes. The mean in- and out-degree (number of incoming and outgoing edges per user) is 7.26. The mean shortest path length is 3.75. The average cluster coefficient [4] is 0.116. The percentage of trust statements which are reciprocated (when there is a trust statement from A to B, there is also a trust statement from B to A) is 33%.

While a large part of research on social networks focuses on exploring the intrinsic characteristics of the network [4, 5], on Trustlet we are interested in covering an area that received much less attention, analysis of trust metrics. We have compared several trust metrics through leave-one-out, a common technique in machine learning. The process is as follows: one trust edge (e.g. from node A to node B) is taken out of the graph and then the trust metric is used to predict the trust value A should place in B, i.e. the value on the missing edge. We repeat this for all edges to obtain a prediction graph, in which some edges can contain an undefined trust value (where the trust metric could not predict the value). The real and the predicted values are then compared in several ways: the coverage, which is a measure of the edges that were predictable, the fraction of correctly predicted edges, the mean absolute error (MAE) and the root mean squared error (RMSE). Surely there are other ways of evaluating trust metrics: for example it can be argued that an important task for trust metrics is to suggest to a user which other still unknown users are more trustworthy, for example suggesting a user worth following on a social bookmarking site such as del.icio.us or on a music community such as Last.fm (for example because she is trusted by all the users the active user trusts). In this case the evaluation could just concentrate on the top 10 trustworthy users. But in this first work we considered only leave-one-out.

4.1 Evaluation of trust metrics on all trust edges

Table 1 reports our evaluation results of different trust metrics on the Advogato dataset. It is a computation of different evaluation measures on every edge present in the social network. The reported measures are fraction of wrong predictions, Mean Absolute Error, Root Mean Squared Error and coverage. We now describe the compared trust metrics. As already mentioned, we released the code and we plan to implement more trust metrics and release them and run the evaluations. We also applied a threshold function in case of trust metrics that can return values in a continuous interval, such as Moletrust and PageRank, so that for example a predicted trust of 0.746 becomes 0.8 (Apprentice).

The compared trust metrics are some trivial ones used as baselines such as Random, which predicts simply a random trust score in the range [0.4, 1] thresholded in the normal way, or the metrics starting with “Always” which

Table 1. Evaluation of trust metrics on all trust edges

	Fraction wrong predictions	MAE	RMSE	Coverage
Random	0.737	0.223	0.284	1.00
AlwaysMaster	0.670	0.203	0.274	1.00
AlwaysJourneyer	0.585	0.135	0.185	1.00
AlwaysApprentice	0.834	0.233	0.270	1.00
AlwaysObserver	0.911	0.397	0.438	1.00
Ebay	0.350	0.086	0.156	0.98
OutA	0.486	0.106	0.158	0.98
OutB	0.543	0.139	0.205	0.92
Moletrust2	0.366	0.090	0.160	0.80
Moletrust3	0.376	0.091	0.161	0.93
Moletrust4	0.377	0.092	0.161	0.95
PageRank	0.501	0.124	0.191	1.00
AdvogatoLocal	0.550	0.186	0.273	1.00
AdvogatoGlobal	0.595	0.199	0.280	1.00

always return the corresponding value as predicted trust score. Other simple trust metrics are OutA which, in predicting the trust user A could have in user B, simply does the average of the trust statements outgoing from user A, and OutB which averages over the trust statements outgoing from user B.

The other trust metrics were already explained in Section 2, here we just report on how we thresholded and how we run them. Ebay refers to the trust metric that, in predicting the trust user A could have in user B, simply does the average of the trust statements incoming in user B, i.e. the average of what all the users think about user B. MoletrustX refers to Moletrust applied with a trust propagation horizon of value X. The values returned by PageRank as predicted trust follow a powerlaw distribution, there are few large PageRank scores and many tiny ones. So we decided to rescaled the results simply by sorting them and linearly mapping them in the range [0.4, 1], after this we thresholded the predicted trust scores. Our implementation of Advogato is based on Pymmetry1. AdvogatoGlobal refers to the Advogato trust metric run considering as seeds the original founders of Advogato community, namely the users "raph", "federico", "miguel" and "alan". This is the version that is running on the Advogato web site for inferring global certifications for all the users. This version is global because it predicts a trust level for user B which it is the same for every user.

AdvogatoLocal refers to the local version of Advogato trust metric. For example, when predicting the trust user A should place in user B, the trust flow starts from the single seed "user A". This version is local because it produces personalized trust predictions which depends on the current source user and can be different for different users. AdvogatoLocal was run on a subset (8%) of all the edges since the current implementation is very slow. Due to the leave-one-

out technique the network will be different for every evaluation and it has to be restarted from scratch for every single trust edge prediction.

The results of the evaluation are reported in Table 1. We start by commenting the column "fraction of wrong predictions". Our baseline is the trust metric named "Random" which produces an incorrect predicted trust score 74% of the times. The best one is Ebay with an error as small as 35% followed by Moletrust2 (36.57%), Moletrust3 (37.60%) and Moletrust4 (37.71%). Increasing the trust propagation horizon in Moletrust allows to increase the coverage but also increases the error. The reason is that users who are near-by in the trust network (distance 2) are better predictors than users further away in the social network (for example, users at distance 4).

Note that Moletrust is a local trust metric that only uses information located "near" the source node so it can be run on small devices such as mobiles which only need to fetch information from the (few) trust users and possibly the users trusted by them. This behaviour is tunable through setting the trust propagation horizon to specific values. On the other hand, Ebay, being a global trust metric, must aggregate the entire trust network, which can be costly both in term of bandwidth, memory and computation power. The AlwaysX metrics depend on the distributions of certifications and are mainly informative of the data distribution.

The fraction of wrong predictions of Advogato (both local and global) is high compared to Ebay and Moletrust. Advogato was not designed for predicting an accurate trust value, but to increase attack-resistance while accepting as many valid accounts as possible. A side effect is that it limits the amount of granted global certifications and assigns a lot of Observer certificates. In the case of AdvogatoGlobal, 45% of the predicted global certifications are marked as Observer which obviously has an impact on the leave-one-out evaluation. Different trust metrics might have different goals, that require different evaluation techniques. Note that the local version of Advogato is more accurate than the global version. The last metric shown in Table 1 is PageRank [13]: the fraction of correct predictions is not too high but again the real intention of PageRank is to rank web pages and not to predict the correct value of assigned trust.

An alternative evaluation measure is the Mean Absolute Error (MAE). The MAE is computed by averaging the difference in absolute value between the real and the predicted trust statement on an edge. There is no need to threshold values because MAE computes a meaningful value for continuous values. The MAE computed for a certain thresholded trust metric is generally smaller than the MAE computed for the same trust metric when its trust score predictions are not thresholded. But in order to compare metrics that return real values and others that return already thresholded values, we consider the MAE only for thresholded trust metrics. The second column of Table 1 reports the MAE for the evaluated thresholded trust metrics. The baseline is given by the Random trust metric which incurs in a MAE of 0.2230. These results are the worst besides the trivial trust metrics that always predict the most unfrequent certification values. Predicting always Journeyer (0.8) incurs in a small MAE because this value is

frequent and central in the distribution. Ebay is the trust metric with the best performance, with a MAE of 0.0855. And it is again followed by Moletrust that in a similar way is more accurate with smaller trust propagation horizons.

A variant of MAE is Root Mean Squared Error (RMSE). RMSE is the root mean of the average of the squared differences. This evaluation measure tends to emphasize large errors, which favor trust metrics that remain within a small band of error and don't have many outlying predictions that might undermine the confidence of the user in the system. For example, it penalizes a prediction as Journeyer when the real trust score should have been Master, or vice versa.

The baseline Random has an RMSE of 0.2839. With this evaluation measure too, Ebay is the best metric with an RMSE of 0.1563 and all the other performances exhibit a pattern similar to the one exposed for the other evaluation measures. However there is one unexpected result: the trivial trust metric OutA is the second best, close to Ebay. Remind that, when asked a prediction for the trust user A should place in user B, OutA simply returns the average of the trust statements going out of A, i.e. the average of how user A judged other users. This trust metric is just a trivial one that was used for comparison purposes. The good performance of OutA in this case is related to the distribution of the data in this particular social setting. The Observer certification has special semantics: it is the default value attributed to a user unless the Advogato trust metric gives a user a higher global certification. So there is little point in certifying other users as Observer. In fact, the FAQ specifies that Observer is "the level to which you would certify someone to remove an existing trust certification". Observer certifications are only when a user changes its mind about another user and wants to downgrade her previously expressed certification as much as possible. This is also our reason for mapping it to 0.4, a less than sufficient level. As a consequence of the special semantics of observer certifications, they are infrequently used. In fact only 638 users used the Observer certification at least once while, for instance, 2938 users used the Master certification at least once. Trust metrics like Ebay and Moletrust work doing averages of the trust edges present in the network (from a global point of view for Ebay and only considering the ones expressed by trusted users for Moletrust) and, since the number of Observer edges is very small compared with the number of Master, Journeyer and Apprentice edges, these predicted average tend to be close to higher values of trust. This means that when predicting an Observer edge (0.4) they lead to a large error. This large error is weighted a lot by the RMSE formula. On the other hand, using the average of the outgoing trust edges (like OutA does) happens to be a successful technique for not incurring in large errors when predicting observer edges. The reason is that a user who used Observer edges tended to use it many times so the average of its outgoing edge certifications is a value that is closer to 0.4 and hence it incurs in lower errors on these critical edges and, as a consequence, in smaller RMSE. This effect can also be clearly seen when different trust metrics are restricted to predict only Observer edges and evaluated only on them. In this case (not shown in Tables), OutA gets the correct value for trust (Observer) 42% of times, while for instance, Ebay only 2.7% of times and

Moletrust2 4%. The fact OutA exhibits a so small RMSE supports the intuition that evaluating which conditions a certain trust metric is more suited for than another one is not a trivial task. Generally knowledge about the domain and the patterns of social interaction is useful, if not required, for a proper selection of a trust metric for a specific application and context.

The last column of Table 1 reports the coverage of the different trust metrics on the Advogato dataset. Sometimes a trust metric might not be able to generate a prediction and the coverage refers to the number of edges that are predictable. The experiment shows that the coverage is always very high. Since local trust metrics use less information (only trust statements of trusted users) their coverage is smaller than the coverage of global trust metrics. Anyway, differently from other social networks [12], it is very high. The Advogato trust network is very dense, so there are many different paths from a user to another user. Even very local trust metrics such as Moletrust2, that only use information from users at distance 2 from the source user, are able to cover and predict almost all the edges.

4.2 Evaluation of trust metrics on controversial users

As a second step in the analysis we concentrated on controversial users [12]. Controversial users are users which are judged in very diverse way by the members of a community. In the context of Advogato, they can be users who received many certifications as Master and many as Apprentice or Observer: the community does not have a single way of perceiving them. The intuition here is that a global average can be very effective when all the users of the community agree that "raph" is a Master, but there can be situations in which something more tailored and user specific is needed. With this in mind we define controversial users as Advogato users with at least 10 incoming edges and standard deviation in received certifications greater than 0.2. Table 2 shows the results of the evaluation of the different trust metrics when they are restricted to predicting the edges going into controversial users. In this way we reduce the number of predicted edges from 52981 to 2030, which is still a significant number of edges to evaluate trust metrics on.

In order to understand better the nature of trust edges under prediction in this second experiment, it is useful to note that, of edges going into controversial users, 1093 are of type Master, 403 of type Journeyer, 115 of type Apprentice and 419 of type Observer. The variance in the values of trust certificates is of course due to the fact that these users are controversial and it is also the reason for which predicting these edges should be more difficult.

We start by commenting the evaluation measures on AlwaysMaster (second row of Table 2) because it presents some peculiarities. Always Master predicts the correct trust value 53.84% (100% 46.16%) of times and, according to the evaluation measure "fraction of correctly predicted trust statements", seems a good trust metric, actually the best one. However the same trust metric, AlwaysMaster, is one of the less precise when RMSE is considered. A similar pattern can be observerd for AdvogatoGlobal. In fact, since in general there

Table 2. Evaluation of trust metrics on trust edges going into controversial users

	Fraction wrong predictions	MAE	RMSE	Coverage
Random	0.799	0.266	0.325	1.00
AlwaysMaster	0.462	0.186	0.302	1.00
AlwaysJourneyer	0.801	0.202	0.238	1.00
AlwaysApprentice	0.943	0.296	0.320	1.00
AlwaysObserver	0.794	0.414	0.477	1.00
Ebay	0.778	0.197	0.240	0.98
OutA	0.614	0.147	0.199	0.98
OutB	0.724	0.215	0.280	0.92
Moletrust2	0.743	0.195	0.243	0.80
Moletrust3	0.746	0.194	0.241	0.93
Moletrust4	0.746	0.195	0.242	0.95
PageRank	0.564	0.186	0.275	1.00
AdvogatoLocal	0.518	0.215	0.324	1.00
AdvogatoGlobal	0.508	0.216	0.326	1.00

is at least one flow of trust with Master certificates going to these controversial users, AdvogatoGlobal tends to predict almost always Master as trust value and since almost half of the edges going into controversial users are of type Master, AdvogatoGlobal often predicts the correct one.

This means that the same trust metric might seem accurate or inaccurate depending on the evaluation measure. This fact once more highlights how evaluating trust metrics on real world datasets is a complicated task and a comparison of same metrics on many different datasets according to different evaluation methods would be highly beneficial for understanding the situation in which one trust metric is more appropriate and useful than another. We already previously explained why OutA is able to have a so small RMSE, the smallest one on controversial users: based on how Observer certifications are used in the system, OutA is the only metric that is able to avoid large errors when predicting the Observer edges.

Arriving at a comparison between a global trust metric such as Ebay and a local trust metric such as Moletrust, we were expecting the latter to be more accurate than the first on controversial users. While on the Epinions dataset, this is what was observed [12], the same is not true here. The reason is partly that in Epinions, the trust values were binary (either trust or distrust) and it was easier to discriminate. Another reason seems to be that on Advogato the user base is not divided in cliques of users such that users of one clique trust each other and distrust users of other cliques. In fact Advogato users are somehow similar and feel part of one single large community. It is future work to analyze if on a social network with a much higher polarization of opinions (such as for

example essembly.com, a political site) the performances of local trust metrics are significantly better than global ones.

5 Conclusions

In this paper we have presented Trustlet [7], an open environment for research on trust metrics. We have claimed that the rapid development of social networking asks for a shared effort in collecting datasets and distributing code of algorithms so that comparisons of different research proposals is easier.

As an initial investigation we have reported our comparison of different trust metrics on the Advogato dataset. The results are partly contradictory and this suggests there is need to run systematically evaluations of different algorithms against the same datasets. As future works we are looking into extending our analysis to more datasets also from different social scenarios, for example the networks of relationships (coediting, talk) among Wikipedia users.

Our goal is to make Trustlet an environment which facilitates this collaborative effort. We believe research on these topics is very needed in a time in which our relationships are starting to move more and more into the "virtual" world and our society and life is affected significantly from the predictions and suggestions produced by many different algorithms.

References

1. Francis Fukuyama. Trust: the Social Virtues and the Creation of Prosperity, 1995. Free Press Paperbacks.
2. Paolo Massa. A survey of trust use and modeling in cu rrent real systems, 2006. Chapter in "Trust in E-Services: Technologies, Practices and Challenges", Idea Group, Inc
3. Diego Gambetta, Can We Trust Trust? In "Making and Breaking Cooperative Relations". 2000
4. M. E. J. Newman, The structure and function of complex networks, SIAM Review 45, 167-256 (2003)
5. A.-L. Barabsi Linked: The New Science of Networks (Perseus, Cambridge, MA, 2002)
6. Sabater, J., and Sierra, C., Review on Computational Trust and Reputation Models. Artificial Intelligence Review (2005)
7. Trustlet, collaborative wiki for trust research. http://www.trustlet.org
8. Fullam, K., T. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. Rosenschein, L. Vercouter, and M. Voss. "A Specification of the Agent Reputation and Trust (ART) Testbed: Experimentation and Competition for Trust in Agent Societies" The Fourth International Joint Conference on Autonomous Agents and Multiagent Systems Utrecht, July 2005
9. Raph Levien, Attack Resistant Trust Metrics. Ongoing PhD thesis. http://www.levien.com/thesis/compact.pdf
10. Cai-Nicolas Ziegler. Towards Decentralized Recommender Systems. PhD thesis, Albert-Ludwigs-Universitaet Freiburg, Freiburg i.Br., Germany, 2005

11. Jennifer Golbeck. Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland, 2005.
12. Paolo Massa, Paolo Avesani, Trust Metrics on Controversial User: Balancing Between Tyranny of the Majority and Echo Chambers. International Journal on Semantic Web and Information Systems 3 (1): 39. (2006)
13. D. Austin, How Google Finds Your Needle in the Web's Haystack, 2006, retrieved on 2008-02-02. http://www.ams.org/featurecolumn/archive/pagerank.html
14. John Locke. An Essay concerning Human Understanding. Harvester Press, Sussex, 1680
15. Jacob Moreno, Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama. Beacon House, Inc. Beacon New York. 1953
16. D. Quercia, S. Hailes, and L. Capra. Lightweight Distributed Trust Propagation. In Proceedings of the 7th IEEE International Conference on Data Mining, 2007

Organisational Knowledge Management Systems in the Era of Enterprise 2.0: The case of OrganiK[1]

Dimitris Bibikas[1], Dimitrios Kourtesis[1], Iraklis Paraskakis[1], Ansgar Bernardi[2], Leo Sauermann[2], Dimitris Apostolou[3], Gregoris Mentzas[4] and Ana Cristina Vasconcelos[5]

[1] South East European Research Centre (SEERC), Mitropoleos 17, 54624 Thessaloniki, Greece
[2] German Research Center for Artificial Intelligence - DFKI GmbH, Knowledge Management Department, P.O.Box 2080, 67608 Kaiserslautern, Germany
[3] Department of Informatics, University of Piraeus, Karaoli and Dimitriou 80, Piraeus, GR-18534, Greece
[4] Institute of Communication and Computer Systems, National Technical University of Athens, 9, Iroon Polytechniou Str., 15780 Zografou, Athens, Greece
[5] Department of Information Studies, The University of Sheffield Regent Court, 211, Portobello Str., S14DP, Sheffield, UK
{dbibikas, dkourtesis, iparaskakis}@seerc.org, {ansgar.bernardi, leo.sauermann}@dfki.de, dapost@unipi.gr, gmentzas@mail.ntua.gr, a.c.vasconcelos@sheffield.ac.uk

Abstract. The increasing need of small knowledge-intensive companies for loosely-coupled collaboration and ad-hoc knowledge sharing has led to a strong requirement for an alternative approach to developing knowledge management systems. This paper proposes a framework for managing organisational knowledge that builds on a socio-technical perspective that considers people and technology as two highly interconnected components. We introduce a knowledge management system architecture that merges enterprise social software characteristics from the realm of Enterprise 2.0, and information processing techniques from the domain of Semantic Web technologies, in order to deliver a KM approach that could assist in reducing the socio-technical gap.

Keywords: knowledge management, socio-technical approach, enterprise social software, semantic web technologies, system architecture.

1 Introduction

Small knowledge-intensive companies are constrained by resource scarcity and cannot compete with large companies in terms of tangible resources, such as capital,

[1] Research project OrganiK (An organic knowledge management system for small European knowledge-intensive companies) is funded by the European Commission's 7th Framework Programme for Research and Technology Development under Grant Agreement 222225 (Research for the benefit of SMEs).

labour, equipment or physical commodities. However, an intangible asset such as knowledge is an invaluable resource that can be utilised by small firms. Knowledge, if properly harnessed, will enable Small-Medium Enterprises (SMEs) to stand out in the competition and outperform their rivals, thus maintaining a competitive edge [1]. Despite this pressing need, it is widely accepted that small companies – even the most knowledge-intensive ones – are characterised by a lack of uptake of knowledge management initiatives, while at the same time many of their large counterparts are effectively practicing knowledge management [2].

1.1 Motivation

The majority of today's enterprise knowledge management tools, techniques and methodologies have been developed with large firms in mind [5], and thus adhere to requirements that are inevitably in conflict with the peculiarities of small knowledge-intensive companies [3]. Current Knowledge Management (KM) systems are not only expensive to purchase, but also necessitate the commitment of significant resources to their deployment, maintenance, and daily operation. The amount of effort required for performing activities core to KM systems, such as designing taxonomies, classifying information, and monitoring functionality [2] is disproportionate to the resource capacity of most SMEs. Moreover, typical KM systems place emphasis on predetermined workflows and rigid "information-push" approaches [4] that reflect the philosophy behind working practices in large enterprises. In contrast, SMEs rely mostly on informal person-to-person communications and people-centric operations [3] that take place in largely ad-hoc and non-standardised ways [2]. By and large, SMEs have a set of distinctive needs that call for the deployment of a new breed of digital environments for generating, sharing, and refining organisational knowledge.

The management of knowledge in idiosyncratic environments such as those of small knowledge-intensive firms can significantly benefit from key characteristics of enterprise social software, like lightweight deployment, flexibility and simplicity of use, emergent and self-organising knowledge structures, and collaboration-oriented philosophy. Nevertheless, in the absence of a knowledge representation scheme to assist in the interpretation of the accumulated information, the evolution of content in a bottom-up fashion may hinder the effectiveness of managing this information and eventually prevent knowledge workers from transforming it into knowledge. To that end, the enhancement of enterprise social software with intelligent information processing capabilities through the use of semantic technologies appears as a rather promising direction. Such a blend would result in considerable improvements to the usability and effectiveness of enterprise social software, and would enable an SME-focused KM system to demonstrate the immediate and profound evidence of benefits needed for knowledge workers to accept it and use it in their every-day activities [2].

The underpinning motivation in this paper is that by leveraging enterprise social software applications with semantic information processing and contextual awareness, we can achieve significant benefits in managing content and knowledge, while allowing for informal, people-centred and ad hoc every-day procedures to be employed.

1.2 Contribution

The aim of this paper is to propose an alternative approach to developing organisational knowledge management systems for small knowledge-intensive companies. In contrast to typical approaches, where knowledge management systems require specific processual use, we suggest that focus should be shifted to delivering solutions that can organically adapt to their every-day work practices and problem solving activities without imposing them from outside or above [6]. This approach to enterprise knowledge management aims at the creation of an environment where encouragement of active social interaction between individuals and teams, empowerment of participation, and self-motivated engagement can promote innovation and assist in attaining sustainable competitive advantage. This perspective suggests a combination of the up to date largely disconnected social and technical organisational system views.

2 Socio-technical Knowledge Management Perspectives

Knowledge management literature has often focused on disjoint approaches of people-centred and technology-centred strategies [7]. These fragmenting perceptions are based upon a focus of discussion and debate on the distinction between explicit and tacit knowledge utilisation: easily codified and documented knowledge should be managed through technology-oriented approaches, whereas knowledge that resides on people's thoughts and beliefs requires people-oriented actions [8]. Nevertheless, it is proposed that overly stressing the importance of either technological or social components of knowledge management can sometimes be misleading and conducive to less effective organisational initiatives, since these two approaches may, in some contexts, be of equal usefulness [9].

This paper adopts the view, following Lytras and Pouloudi [10], of "knowledge management as a socio-technical phenomenon where the basic social constructs such as person, team and organisation require support from Information and Communication Technology (ICT) applications". A socio-technical approach to leveraging organisational knowledge considers people and technology as two highly interconnected components of a single system and is applied to the study of the relationships and interactivities between the social and technical structures of an organisation [11]. Undoubtedly, the tension between the social and technical sub-system can be difficult to harmonise, thus leading to what has become known as the socio-technical gap [6], as illustrated in Figure 1. In particular, it appears that social requirements are often neglected in the process of designing organisational knowledge management solutions.

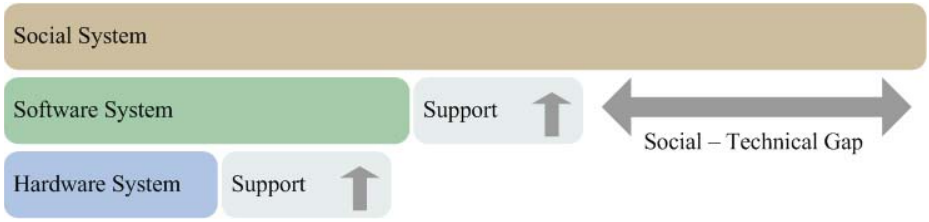


Fig. 1. Socio-technical gap: software and hardware systems provide support for the technical subsystem, while the social subsystem remains virtually unsupported (adapted from Patrick and Dotsika [6]).

We propose an organic perspective to organisational knowledge management system development [6], [12], in which the characteristics of the resulting technical sub-system emerge from a continuous negotiation procedure among the social actors of the organisation and adaptation through user involvement and engagement. This approach attempts to create an iterative dialogic relationship between the social and technical sub-systems that can promote the creation of a collaborative environment for creating, sharing and distilling information in organisational settings.

3 An OrganiK Approach to Knowledge Management: Towards a Socio-technical fit

The vision of the proposed approach is to enable knowledge workers in small knowledge-intensive companies to effectively collaborate and utilise organisational knowledge with the support of an organic knowledge management framework. As stressed above, this approach is founded on a socio-technical perspective, and identifies the effectiveness of interactions among people and technology as a major challenge. As illustrated in Figure 2, the major components of the proposed knowledge management framework are the following:

- A *people-centred* knowledge management conceptualisation focusing on social processes, *ad-hoc* work practices and organisational structures (i.e. individual, team, business units). Situated *innovation management* processes, cultivation of *communities of practice* and project *adaptation procedures* comprise fundamental components of this socially-focused processual approach.
- A *technology-centred* knowledge management conceptualisation focusing on the integration of *enterprise social software* applications (wikis, blogs, collaborative bookmarking tools and search engines) with *semantic technologies* (ontology-based annotation, semantic text analysis, logic-based reasoning).

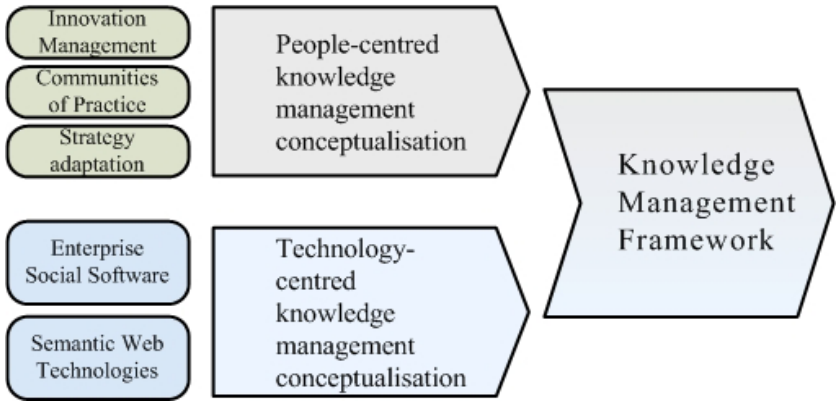


Fig. 2. The proposed OrganiK knowledge management framework.

3.1 Proposed Architecture

The proposed architecture that this paper puts forward for reducing the socio-technical gap among work practices in small knowledge-intensive firms and present-day knowledge management systems, combines key elements from the domains of Enterprise 2.0 and Semantic Web technologies. Regarding the Enterprise 2.0 domain, the system architecture employs the SLATES framework [12]:

- *Search*, to provide mechanisms for discovering information.
- *Links*, to provide guidance to knowledge workers in order to discover the needed knowledge and ensure emergent structure to online content.
- *Authoring*, to enable knowledge workers to share their opinions with a broad audience.
- *Tags*, to present an alternative navigational experience exploiting unhierarchical categorisation of intranet content.
- *Extensions*, to exploit collaborative intelligence and recommend to knowledge workers contextually relevant content.
- *Signals*, to automatically alert knowledge workers for fresh available and relevant content.

The aim is to provide knowledge workers with a collaborative workspace that comprises a set of integrated Web 2.0 applications (a wiki, a blog, a bookmarking system and a search/recommendation engine), augmented with natural language processing and semantic information integration capabilities that enable the combined use of folksonomies and ad-hoc tagging with thesauri and shared ontologies. The use of semantic technologies in the envisaged solution comprises the following key functions:

- *Semantic knowledge representation*: representing knowledge in a formal, machine understandable manner.

- *Semantic resource annotation*: annotating knowledge artefacts and other resources by reference to concepts defined in an ontological model.
- *Semantic inference*: performing automated logic-based reasoning to infer new, implicit knowledge based on what has been already asserted in an explicit manner.
- *Semantic search and discovery*: using ontological terms to describe a search query and rely on logic-based reasoning to derive the matching results.

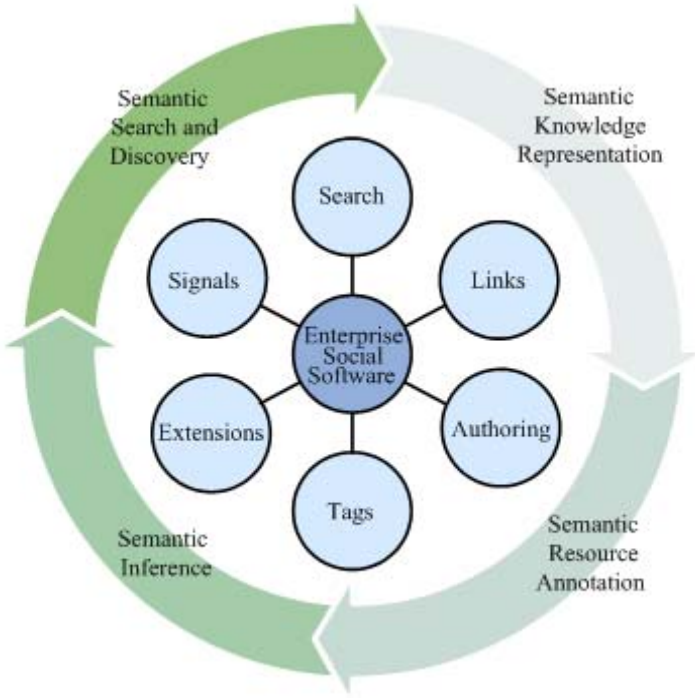


Fig. 3. Integrating components of the SLATES framework with machine processable semantics.

Each of the aforementioned social software functions corresponds to one or more of the components of the SLATES framework, and, as illustrated in Table 1, corresponds to a specific component in our proposed architecture.

Table 1. Association among components in SLATES and our proposed architecture.

SLATES Framework	Proposed Architecture
Search	Semantic Search
Links	Collaborative Bookmarking
Authoring	Wiki and Blog spaces
Tags	Collaborative Bookmarking, Wiki and Blog spaces
Extensions	Recommender System
Signals	Really Simple Syndication (RSS)

A conceptualisation of the proposed architecture is illustrated in Figure 4. As seen in the *Client Interface Layer*, the collaborative workspace that is offered to knowledge workers comprises a wiki, a blog, a social bookmarking tool and a search interface. Each of the client interfaces corresponds to a server-side component in the next layer of the architecture; the *Component Interface Layer*. The server-side building blocks that comprise the *Business Logic Layer* are a recommender system, a semantic text analyser, a collaborative filtering engine and a full-text indexer. The *Metadata Layer* refers to repositories used for the persistence of syntactic and semantic metadata supporting the functionality of all server-side components, while the *Datasources and Back-Office Integration Layer* refers to business information systems and any form of resource container that an enterprise may depend on for its daily operations.

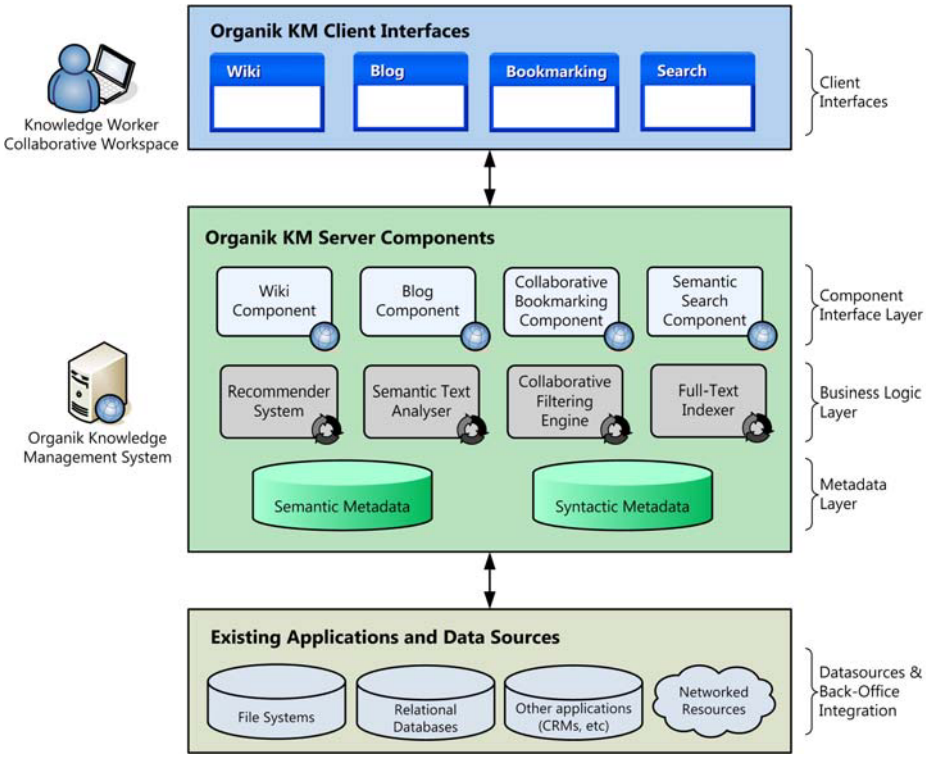


Fig. 4. Proposed conceptual architecture for semantically-enriched enterprise social software.

The functionality of the core components in the proposed architecture is envisaged as follows:

- The *Wiki Component* is a web-based authoring tool allowing knowledge workers to collaboratively create, edit, and share knowledge artefacts such as documents, diagrams, etc.

- The *Blog Component* provides a simple content management tool enabling knowledge workers to build and maintain open project monitoring diaries, complete with links to relevant resources and user commentary.
- The *Social Bookmarking Component* enables knowledge workers to organise and annotate resources relevant to their activities (intranet documents, web resources, wiki entries, blog posts, etc) and share them with their co-workers.
- The *Semantic Search Component* supports browsing, searching, retrieving and displaying knowledge resources leveraging semantic annotation indexing and logic-based inferencing.
- The *Recommender System* focuses on the suggestion of tags and classifications for content added to the system (e.g. wiki entries, bookmarked documents, blog comments, etc), and the suggestion of information items relevant to the search query or feed subscription of a user.
- The *Semantic Text Analyser* employs linguistic and statistical processing functions on the textual content of knowledge artefacts added to the system, in order to perform named entity recognition and term classification. The objective is to identify concepts of interest and establish relationships among resources that can be subsequently used by the Recommender System for suggesting tags and classifications with respect to a taxonomy/ontology.
- The *Collaborative Filtering Engine* enables individual knowledge workers to benefit from the collective experience built within groups of peers. An analysis of subjective views that are explicitly or implicitly expressed by other knowledge workers can assist in the selection and recommendation of resources, as well as influence the ranking of search results.
- The *Full Text Indexer* is an indispensable component of the architecture's Business Logic Layer and complements the content retrieval techniques proposed above.

To summarise, the enhancement of enterprise social software tools with machine-processable semantics and their respective processing techniques is expected to yield significant benefits with respect to efficiency of information management, and contribute towards improving the overall user experience of knowledge workers.

4. Concluding Remarks

This paper theoretically investigates an approach to developing organisational knowledge management systems for small knowledge-intensive companies. In contrast to other approaches employed in present-day knowledge management systems, we suggest that a specific processual use should not be imposed onto knowledge workers, but rather, the provided KM solutions should be able to organically adapt to their every-day work practices and problem solving activities. Despite the fact that the Organik research project is still at a rather initial stage, we envisage a system that is utilised and organically incorporated into every-day *ad hoc* and knowledge-intensive SME work practices. Our objective is to realise a KM system with increased social acceptance and a positive impact on reducing the socio-

technical gap. In particular, we propose an OrganiK knowledge management framework that adopts a socio-technical perspective to leveraging organisational knowledge, and considers people and technology as two highly interconnected components. We adopt the intersection of social software and semantic technologies as the technological baseline towards realising this vision, and present a high-level conceptual architecture of the envisaged solution.

References

1. Wong, K.Y., Aspinwall, E.: An empirical study of the important factors for knowledge management adoption in the SME sector. Journal of Knowledge Management 9(3), 64--82 (2005)
2. Nunes, M.B., Annansingh, F., Eaglestone, B.: Knowledge management issues in knowledge-intensive SMEs. Journal of Documentation 62(1), 101--119 (2006)
3. Desouza, K.C., Awazu, Y.: Knowledge management at SMEs: five peculiarities. Journal of Knowledge Management 10(1), 32--43 (2006)
4. Malhotra, Y.: Integrating knowledge management technologies in organizational business processes: getting real time enterprises to deliver real business performance. Journal of Knowledge Management 9(1), 7--28 (2005)
5. Maguire, S., Koh, S.C.L., Magrys, A.: The adoption of e-business and knowledge management in SMEs. Benchmarking: An International Journal 14(1), 37--58(2007)
6. Patrick, K., Dotsika, F.: Knowledge sharing: developing from within. The Learning Organization 14 (5), 395--406 (2007)
7. Mentzas, G., Apostolou, D., Young, R., Abecker A.: Knowledge networking: a holistic solution for leveraging corporate knowledge. Journal of Knowledge Management 5(1), 94-106 (2001)
8. Bhatt, G., Gupta, J.N.D., Kitchens, F.: An exploratory study of groupware use in the knowledge management process. The Journal of Enterprise Information Management 18(1), 28--46 (2005)
9. Bhatt, G.D.: Knowledge management in organizations: examining the interaction between technologies, techniques, and people. Journal of Knowledge Management 5(1), 68--75 (2001)
10. Lytras, M.D., Pouloudi, A.: Towards the development of a novel taxonomy of knowledge management systems from a learning perspective: an integrated approach to learning and knowledge infrastructures. Journal of Knowledge Management 10(6), 64--80 (2006)
11. Cartelli, A.: ICT and knowledge construction: Towards new features for the socio-technical approach. The Learning Organization 14(5), 436--449 (2007)
12. McAfee, A.P.: Enterprise 2.0: The Dawn of Emergent Collaboration. MIT Sloan Management Review 47(3), 21--28 (2006)

Hubbub - An innovative customer support forum

Duong Nguyen, Simon Thompson, and Cefn Hoile

British Telecommunications Plc
Adastral Park, Martlesham Heath
Ipswich, IP5 3RE, United Kingdom
{duong.nguyen,simon.2.thompson,cefn.hoile}@bt.com

Abstract. Internet user forums have been proven to be effective not just as a community meeting place but also as a supporting tool for various business products. Traditional forums are designed with "browse and read" journey in which users have to select the right sub forum to get into and select topics to read from within. However, we have identified a new trend in forum design toward community question answering systems with an "ask questions first" user journey, a topic-less organisation, search based information retrieval and social network inspired alerting. Here, we report on the implementation and trial of such a forum, Hubbub, that epitomizes the aforementioned trend. It is designed to eliminate key issues found in current forum technologies and has been fielded as a support channel for a BT Softphone product, resulting in a significant reduction in support costs. We then report on the performance of this forum in practice and speculate on the reasons that forum design is taking this direction. Finally, we conclude the paper with some thought about future direction of forums in the Web 2.0 era.

1 Introduction

Modern converged communication products pose considerable support challenges. For example, a VOIP service is dependent on network connectivity, proprietary client software, a VOIP server, PC hardware, a microphone and drivers, a modem and an operating system. Any failure of any part of the solution will leave the product inoperable and the user with a problem.

This is also a problem for the service provider, because users are not able to identify the source of failure for their service. Even articulating the issue in the expected manner is often beyond a non technical user. In addition converged services are often supplied at a very low cost, while they may be profitable there is often insufficient funding from these non traditional services for a large scale traditional support infrastructure. However, in order to make the wider economics of converged service provision work it is critical that these services are provided as part of a complex converged services ecosystem (i.e. the Internet or voice services), which implies all the big company brand risk and customer loyalty dynamic which makes traditional support groups necessary.

Internet forums offer a potential way for companies to handle this support challenge [2]. Community derived support (customers helping each other) offers

zero cost support, one-to-many support (a support agent helping a customer on the forum and many other customers then accessing and reusing the support information) offers a much more efficient support model than single point interactions. Companies working at the top tier of the internet services ecosystem are exposed to a number of risks by internet forum activity.

1. Internet forums can be a vehicle for disruptive customers to damage brand and distort the perception of a product via anecdote and (sometimes) malice. When surveying 25 different forums (selected from the top results of Google searches over forum provision software) we found that on average each post is viewed 119 times (see figure 1, note that the graph is plotted on a log scale). It seems from this observation that the minority of users are posters and any views on the forum may or may not be representative of the experience of the majority. Companies may honestly stand by the view that their customers have a right to publicly state their issues with a product or service, but at the same time it is legitimate to try and ensure that these views are given an appropriate context and appropriate weight [3]. The usage characteristics of internet forums lend themselves to polemic and propaganda and are therefore unsuitable as vehicles for customer advocacy development.

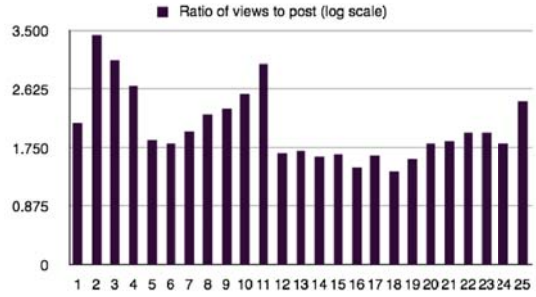


Fig. 1. *Ratio of views to posts on 25 different internet forums. On average each post is read 119 times.*

2. By extension, the use case for internet forum use is overwhelmingly that of customers seeking to discover and read a post that contains the information they want. The rise of internet search engines has demonstrated that users far prefer using search as a mechanism for information retrieval via a web-browser to the experience of information retrieval via web-surfing or browsing.

3. Multiple threads and topics can result in a fragmentation of activity in a forum which militates against the development of a community of expertise that is the commercial objective. Topic structures are initially imposed on forums arbitrarily by their owners, in best-practice cases these structures are frequently revisited in response to the use that is actually being made of the forum by its users. Unfortunately, as soon as a topic structure is created it changes users perception of the forum, and changes their pattern of use.

These three issues provided the non-technical motivations for the implementation of the Hubbub forum [7] which is described in this paper. In addition the development of lightweight Web 2.0 engineering techniques in recent years was a key inspiration to the team. In particular blogging aggregation and tagging engines such as del.icio.us provided models of implementation that changed our view of what could be done with an internet forum.

In the rest of this paper we describe the features of Hubbub which provide enhanced support to the information retrieval use case and reduce the supporting cost for BT business. We provide a qualitative and quantitative evaluation of the forum in use and conclude by pointing to future developments that we hope to make of the technology.

2 Hubbub

In this section we present our community forum solution: Hubbub (see figure 2). It offers a simple means for users to interact with each other, and exploits those interactions to provide functionality, rather than attempting to provide fully automated knowledge management. The knowledge in the system should primarily come from the human contributions, and the system should aim to help users in navigating and coordinating these contributions.

The Hubbub user journey begins with an invitation to the customer to submit their query in their own language and with their own title. This is then parsed and the keywords from the query are extracted and matched to other posts in the forum. When a post is made to the forum the keywords in it are added to the customers profile and an interest in other posts containing that keyword set is registered with the system. Each discussion in Hubbub are indexed via keywords, each of which is a word that has been selected by at least two users of the system.

The posts retrieved via the keyword indexing system are then dynamically compiled into a list for the user creating a "virtual topic" which contains all the relevant posts for the customer to browse. This is in contrast to the static topic structure that is presented by most internet forums. Because the virtual topic is built dynamically for the user it is less likely to be irrelevant. In addition because posts are retrieved with a search function posts which focus on negative issues are unlikely to be retrieved, not because Hubbub filters or removes them, but instead because they are less likely to contain keywords that match the users current post unless the user is looking for negative comments.

If the customer does not find an answer amongst the posts in their virtual topic they can modify their query by interacting with Hubbubs tagging system.

Users are offered tags which are relevant to the post, but which have a relevance below the threshold used for retrieval (see figure 3). If the posts retrieved are inadequate the user is able to select these less relevant tags and will be rewarded with a modified search result. Alternatively the user is also able to add keywords of their own which may not appear in the text of their query. This fea-

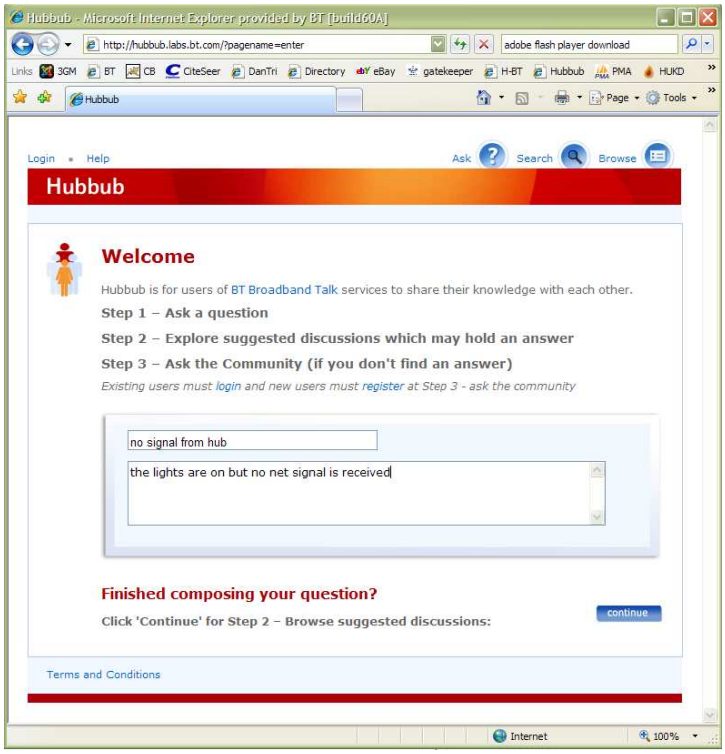


Fig. 2. *Hubbub landing page.*

ture enables tags to be introduced that are separate from the users text enabling the system to link groups of posts in its knowledge base.

If the tag modification process fails to generate an appropriate result they can post their query onto the forum. At this point they must either register or login in our current implementations, this is in order to limit the opportunity for spammers and forum abuse in general and to control the quality of the discussions on the forum.

A key feature of Hubbub is the email alert facility. As mentioned above, author will be automatically notified of any changes related to his/her discussion. Hubbub however provides a mean for end user to subscribe to other thread(s), particular keyword(s) or user(s) so that he or she will not miss any other potential source of information that could provide a solution to his/her problem. Of course, all users have total control over this email alert feature and can decide how the information will be sent to them.

Apart from being able to see relevant discussions, users can also browse the list of discussions on site and view the tag cloud composed of most frequently

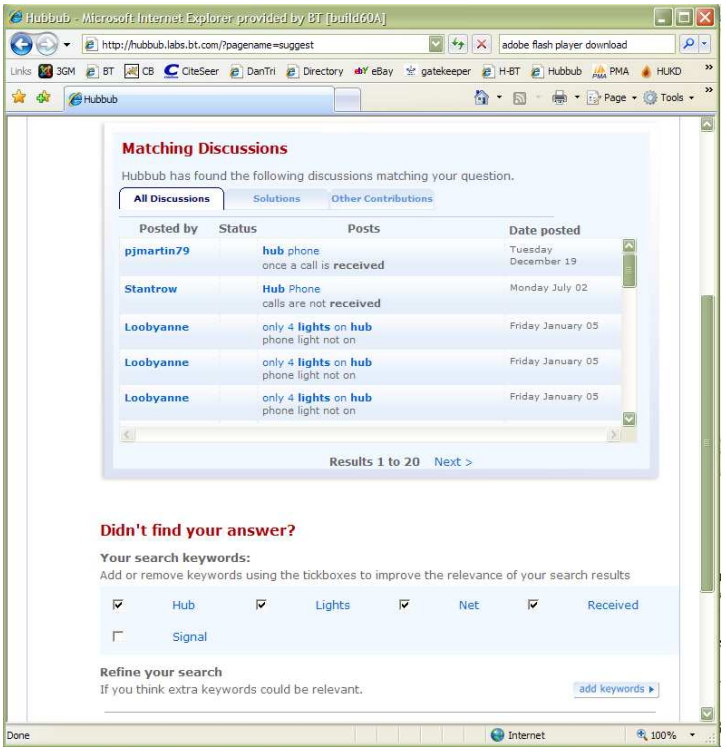


Fig. 3. *Suggested discussions and keywords.*

used keywords[1] (see figure 4). This feature provides a similar facility to that of traditional forum solutions. However, due to the category-less design nature of Hubbub, users will only see a limited subset of all the discussions on site namely latest, unsolved, most read posts. Viewing a keyword allows users to see the list of posts containing that keyword. This list can be either solved or all discussions contain that keyword.

As Hubbub is a community forum, its contents are provided by the members of the community and they do have some features to control this content themselves. For example, if a discussion is considered offensive to an user (i.e. containing an offensive word), he/she can report this discussion as offensive and the system will immediately ban this post from displaying to other users. It can only be displayed again after a customer agent has reviewed and decided that the content is appropriate. Importantly, once the customer support agent or other user with the appropriate administrative rights has asserted that a post is not offensive users will no longer be offered the option of banning it because they find it offensive. This means that it is not possible to use the community

[1] We have not identified a particularly successful mechanism for producing an informative keyword cloud on our data. Nonetheless, we have retained it as a useful navigational feature but its development is beyond the scope of this paper.

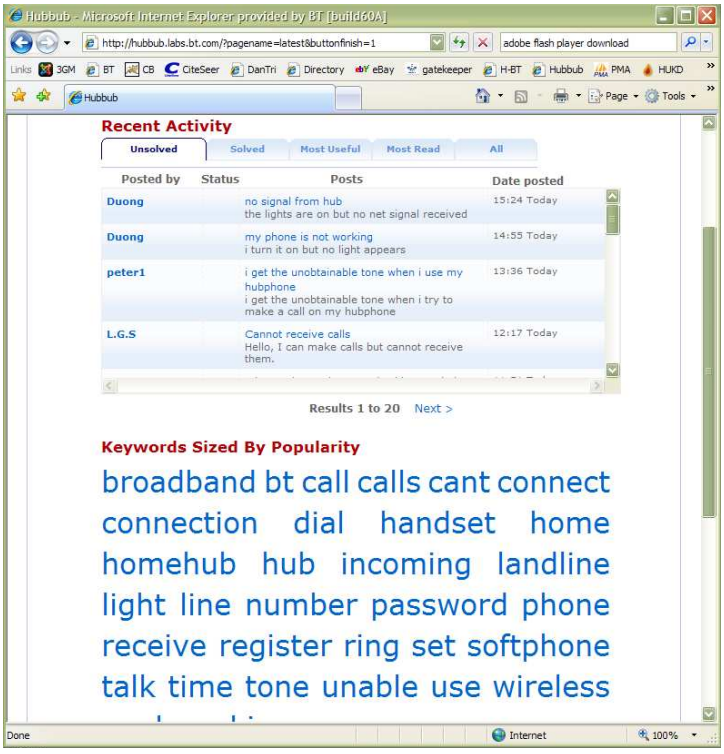


Fig. 4. *Browsing Hubbub.*

moderation feature to remove all the posts from the forum permanently as an act of vandalism.

Administrative users are able to edit the list of keywords to provide automatic deletion of posts that contain certain offensive terms. If any user tries to post a query that contain a banned word, both the query and his/her account will be automatically banned.

Another feature of Hubbub is that it provides customer support agents with the log of activities, including the list of discussions that have not been answered, list of posts/users banned from the site, etc. The agents can then decide to provide answers to such discussions or review the banned posts/users to ensure their validity.

2.1 Technical Design

Figure 5 shows the main components of Hubbub. Hubbub has been built using standard open source solutions including Apache/PHP as the web server and the scripting language, MySQL as the database design and Smarty as the template for the UI. The use of open source software was critical for the Hubbub project because it enabled implementations of the system to be created and run at very

low cost reducing the barrier of investment that potential internal customers were required to commit to.

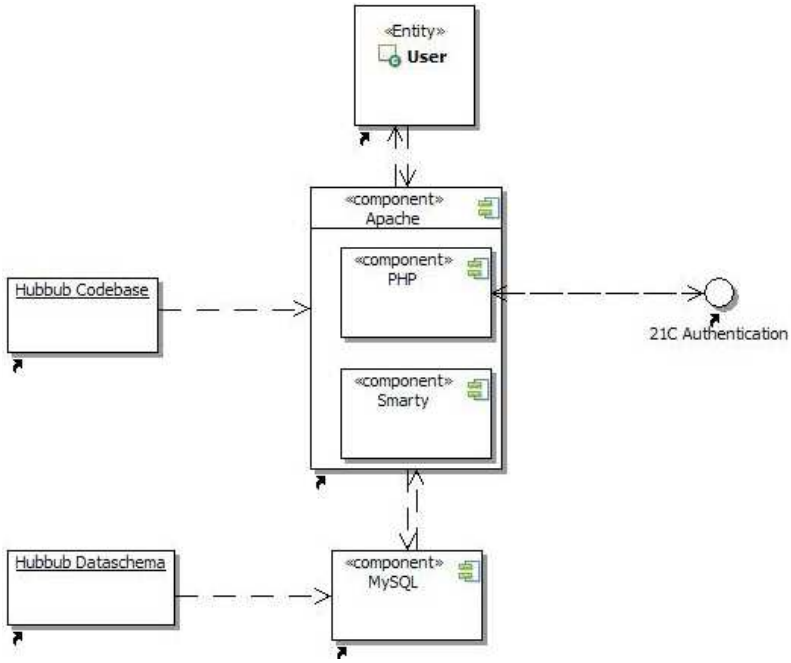


Fig. 5. *Hubbub components.*

A discussion in Hubbub consists of an opening query and is followed by a number of reply posts if they exists. Here a post will be stored with a title and a body. Keywords are automatically extracted from both the title and the body with the rule that any word that is not a stop word or common word will be considered as a keyword. The more times a keyword is ticked by an user, the higher its importance will be.

In terms of database design, there are three nodes in Hubbub which are user, keyword and post. There are number of edges that link these nodes together such as contain_edge links post with keywords (i.e. a post contains few keywords), author_edge links post with user (i.e. authorship). The nodes are indexed for faster data retrieval.

When a query is posted, the keywords are then searched (using simple text matching technique) in the edge tables to find posts sorted by the number of relevant keywords. If the user alters the list of selected keywords the search is restarted. This process is crucial to the functioning of Hubbub since it is the bridge between user queries and potential related solutions. The optimality of this search method cannot be claimed but it provides an adequate, scalable and easily engineered solution.

3 Results

Initially Hubbub has been set up as a trial for 3 months to used as a supporting channel for BT Softphone product on July 2006. After its successful trial, it has been used as a main supporting channel since October 2006. At the time of writing it has more than 5500 registered users and has served more than 4.1 million page requests of which nearly 1.5 million requests come from legitimate users (the other 2.6 million requests are from search bots - which is common for any web site that is live on the internet).

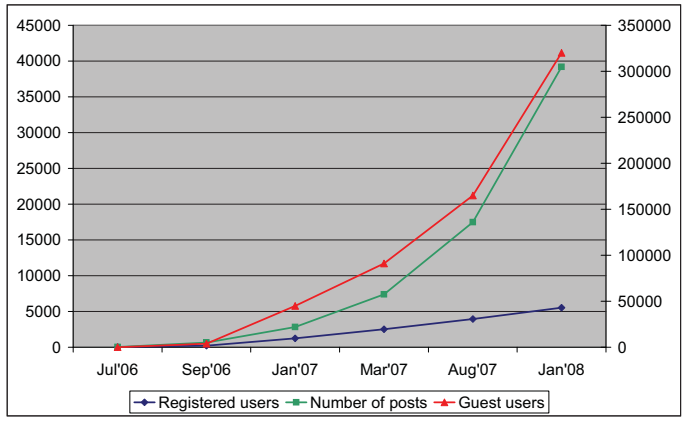


Fig. 6. *Hubbub performance. 2nd axis is for guest users only.*

Figure 6 visualize the statistics of Hubbub since launch. As can be seen, the number of posts and registered users are doubled every 5 months since launch time. In overall, the community uptake of Hubbub is rising steadily.

Table 1 shows the latest average visitor information at this point of writing (Jan 2008). With more than 5000 visitors per day, Hubbub is a significant engine for deflection from BT's contact centers.

5083	visitors per day (excluding bots)
13662	page views per day
190	visitors per day referred from BT.com
510	downloads of robots.txt per day.

Table 1. *Latest Hubbub statistics.*

It is not possible to determine why a particular user decides not to call a contact center, but if we were to measure the value of an interaction via Hubbub as equivalent to a contact center interaction on the basis of published contact

center costs we can quantify something of the business value of this approach. A survey of large organisations in the USA found that the fully loaded cost of a contact center call is between $2.70 and $5.60 [1] on this basis amount of saving per week by Hubbub is approximately £69,000. or approximately £300,000 per month. This is a significant reduction of customer support cost for a relatively low maintenance cost since Hubbub only requires a minimal monitoring (the main server has not been restarted since Oct 2007). Compared to the cost of having a real person behind the phone to give the answer, it is a substantial saving. Furthermore, by going online, the customers will be able to control the amount of information that they can see and have more flexibilities in searching for the right answers. With Hubbub, they are allowed to alter their queries and keywords to search for different posts that might be relevant to them. They might not be able to have such flexibility when dealing with real person behind the phone. In our opinion, it might be beneficial for the supporting agents to actually make use of Hubbub as well but we are not able to try out this idea yet.

Based on the results we gathered during the period of 18 months, it can be seen that Hubbub is quite a success. Users has made 39,000 queries in total and only 12,000 have been registered as a post. Furthermore, of these 12,000 posts, 6000 have at least one answer from other community members and 950 have been marked as solved. Thus, the optimistic reading of the success of Hubbub can be projected at 71.7%[2].

4 A Taxonomy of Internet Forums

There are a number of successful forum solutions available at the time of writing. In this section we will describe some decision making factors that should be considered before selecting the appropriate solution and then outline the main types of forum that are in the market currently.

4.1 Discussion Forums

A number of companies and opensource projects offer complete forum solutions for adoption. Some widely used examples are Jive Forums, mvnforum and Lithium/Rightnow. These solutions can either be installed on a local infrastructure or, in many cases rented on a software as a service basis. The all provide a way for users to read publicly posted messages and to post messages of their own for other peoples consumption, but beyond that basic functionality it is hard to identify their differentiating characteristics.

Wikipedia [5] offers a comparison of 84 forums on the basis of the following features:

[2] Here we assume that if an user enters his or her problem and found a solution based on Hubbub suggestions, that query will not be registered as a forum post. Thus, the success percentage can be measured by (number of unregistered queries + number of solved posts)/total number of queries.

– Flat or Threaded: does the forum support threads or not
– User-selectable themes: is it possible for users to customize the look and feel of the forum
– Calendar functionality: can events or other calendar related information be integrated onto the forum
– Image attachment: can users put images on the forum
– Unread message tracking: can unread messages be detected
– WYSIWYG editing : is there an advanced editor for message formatting on the forum

While the list provided is extensive, these features are not really useful for understanding the usefulness or otherwise of the forums described as channels for customer support. For example, none of the requirements gathering exercises that we undertook in BT identified Calendar, or WYSIWYG functionality as needed. In fact it seems that there is almost complete uniformity in the core features of these systems, which is interesting given the issues that we have identified in this paper.

4.2 Community Question Answering

Yahoo Answers [12] has become one of the most successful sites on the internet by providing a place where users can ask question and receive answers from a community of experts.

In spirit and execution Yahoo Answers is superficially similar to Hubbub, but differs in that Hubbub is specifically designed to serve information rapidly to the user and deflect further enquires, whereas the Yahoo design seems intended to draw users into further interactions. For example, there is no step to present users with similar questions when they post an enquiry on Yahoo answers, and no interactive search mechanism (see Section 3 of this paper for information on how Hubbub does this).

Microsoft have adopted a similar approach with the Microsoft Office question answering system [10] where user can ask a brief question and then be presented with relevant discussion topics. If their question cannot be answered by these topics, they have the option of posting it in the community for other members to reply. It is also similar to Hubbub but lacks the ability for the users to fine tune their query. Here the users are not able to interact with the search process and if the presented solutions are not relevant, users will have to ask another question. On the other hand, Hubbub allows the user to fine tuning their search by adding or removing keywords which in turn will directly affect the list of suggested posts that users can read (see section 2).

Another example of this question answering system is GetSatisfaction [6]. It is a pay-per-support commercial platform that various companies can register to host their support and provide their agents to work on site. GetSatisfaction provides such companies with the forum infrastructure so that end users can search for solutions to their problems on site and get answered by real people if

no solution could be found. At this time of writing, there are about 200 companies with around 40000 posts in total. However, their system has the same disadvantage which is similar to the Microsoft Online where users are not able to interact with the search process.

The Start system from MIT [9] is another question answering system that takes an alternative tack. Whereas Yahoo Answers and Hubbub rely on a human community of experts Start uses advanced AI to attempt to automatically answer user questions from a formal knowledge base. At this time, while interesting, the performance of this kind of system in open domains such as customer support has not been proven.

These examples have clearly demonstrated the new trend in supporting forums in which community question answering systems play a pivotal role. Getting away from the traditional browse and search journey provides user with much greater flexibility and efficiency and allows them to access the right information in much lesser time than previously needed.

4.3 Product Support Forums

Ksamba [8] and Fixya [4] are two well known examples of websites that offer support for products and services from human experts.

They differ somewhat from the Yahoo model in that the objective of the site is to steer the customer into a one-to-one interaction with one of the experts on the site, enabling the expert to provide consultancy in return for remuneration.

This kind of model presupposes that the customer for a product or service does not feel that their relationship with the service provider places the provider under an obligation to support them for free. If there is no chargeable interaction then the site cannot fund itself, and there is no efficiency gain from publishing the expertise elicited from the expert by the customer in the interaction on the site because it is not in the interest of the expert to allow this to happen.

Table 2 summarizes the list of supporting forums that we have discussed in this section.

Features	Lithium Rightnow	Hubbub	Y! Answer	Microsoft GetSatisfaction	MIT - Start	Fixya Kasamba
Query-specific search	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Category-based browsing	*Yes*	*No*	*No*	*No*	*No*	*No*
Direct human contact	*No*	*No*	*No*	*No*	*No*	*Yes*
Email notification						
- posts	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*No*
- keyword/users	*No*	*Yes*	*No*	*No*	*No*	*No*
Solution marking	*No*	*Yes*	*No*	*No*	*No*	*No*
Customize search process	*No*	*Yes*	*No*	*No*	*No*	*Yes*
Administrative functions	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*

Table 2. *Comparison between different forum solutions.*

5 Conclusion and Future Work

Hubbub has been successfully deployed as an innovative customer service solution which combines insights into the fundamental structure of customer forums with web 2.0 technology for tagging and social networking. It has been proven to be popular with customers and BT has plans to use it as a support channel for several new products and services.

Strategically Hubbub will need to be integrated with the BT information architecture under the group wide rule of one for solutions. We are engaging with BT's suppliers and systems architects to attempt to ensure that this happens as efficiently as possible. In parallel we are exploring how to make our solution available for other companies to adopt.

The social and interaction aspects of Hubbub also require further analysis. The motivation of users to participate in Hubbub is complicated by the business objective of the system which is not simply to encourage discussion or community problem solving but also to be an efficient call deflection knowledge base. Therefore the map of corporate ownership and community territory differs for Hubbub when compared to other forums.

A number of research projects are currently underway to develop advanced technology to improve Hubbub's functionality. These include investigations into using the knowledge base as a recommender system for customers and visualization techniques for the management of the forum. We hope to be able to report on these projects in subsequent publications.

In the not too far future, we envisage user forums to have more human like processing power such as the ability to integrate a natural language parser to have a better understanding of user queries, to apply learning techniques in improving both the accuracy and the efficiency of search results for the users. In addition, we also expect internet forum to become a standardized component of an open web application in which end users can fully customize in order to find the right information for themselves with minimal effort.

References

1. C. Barry. *Managing your Cost Per Call.* MultichannelMerchant.com, http://multichannelmerchant.com/opsandfulfillment/contact_center_advisor/cost_percall/, January, 2007.
2. Z. Cui, G. Ducatel, M. Thint, B. Assadian, and B. Azvine. Towards automated customer self-help. *BT Technology Journal*, 24(1):96–106, 2006.
3. J. Dyche. *The CRM Handbook: A Business Guide to Customer Relationship Management.* Addison Wesley, 2001.
4. Fixya - Technical Support, User Guides and Repair Service. http://www.fixya.com.
5. Forum software comparison in Wikipedia. http://en.wikipedia.org/wiki/comparison_of_internet_forum_software.
6. GetSatisfaction - People Powered Customer Service. http://getsatisfaction.com/.
7. Hubbub Community Forum. http://hubbub.labs.bt.com.
8. Kasamba - Live expert advice. http://www.kasamba.com.

9. B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. J. McFarland, and B. Temelkuran. Omnibase: Uniform access to heterogeneous data for question answering. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, June 2002.
10. Microsoft Office Only Community. http://www.microsoft.com/office/community/en-us/wizard.mspx.
11. G. Urban. *Don't just relate - Advocate*. Wharton School Publishing, Upper Saddle River, New Jersey 07458, 2005.
12. Yahoo answers. http://www.yahoo.com/answers.

Mobile Social Software for Cultural Heritage: A Reference Model

Paolo Coppola, Raffaella Lomuscio, Stefano Mizzaro,
Elena Nazzi, and Luca Vassena

Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 — 33100 Udine — Italy
{coppola,lomuscio,mizzaro,vassena,nazzi}@dimi.uniud.it

Abstract. Web 2.0, also known as the Social Web, marks a new philosophy where users are both the main actors and the content producers: users write blogs and comments, they tag, link, and upload photos, pictures, videos, and podcasts. As a step further, Mobile 2.0 adapts Web 2.0 technology to mobile users. We intend to study how Web 2.0 and Mobile 2.0 together can be applied to the cultural heritage sector.
Recently, a number of cultural institutions and museums are introducing in their projects some Web 2.0 applications, but the main knowledge source remains a small group of a few experts. Our approach is different: we plan to let all the users, the crowd, to be the main contents provider. We aim to the crowdsourcing, the long tail power, as fuel of cultural heritage system.
In this paper, we propose a reference model for cultural heritage system that lets users create, share, and use cultural contents including mobile context-aware features.

Key words: Web 2.0, Mobile 2.0, mashup, social, culture, collaboration, crowd, museum, cultural heritage, user-centered

1 Introduction

In this paper we intend to study how Web 2.0 [12] and Mobile 2.0 together can be applied to the cultural heritage sector. With Web 2.0 and social software we represent all web-based services with "an architecture of participation", that is, an architecture featuring a high interaction level among users and allowing users to generate, share, and take care of the content[1]. Mobile 2.0 is the evolution of mobile technology that allows "capturing the content at the point of inspiration", that is, in the exact moment in which the inspiration and the opportunity exists to do it.

Nowadays, Cultural Heritage Organizations (museums, archaeological sites, historical towns, even libraries, etc.) are trying to understand the evolution

[1] http://museumtwo.blogspot.com

of the web and mobile devices, and to exploit the potentialities offered by the new digital instruments. However, these organizations often neglect the social aspects, which are considered by many the true revolution related to these new technologies, and they tend to stick to their traditional role of being the sole owners of knowledge about their collections [8]. Indeed, in this research area, old and new conferences, e.g. Museum and the web[2], International Cultural Heritage Informatics Meeting[3], concentrate on the possible application of Web 2.0 concept and technology to museums, libraries and other cultural heritage institutions.

Our approach is complementary: we want to understand if a fully Web 2.0/Mobile 2.0 approach is viable for the cultural heritage field. We intend to exploit these technologies to let the crowd to be the main contents provider: people are not just passive users, but they are encouraged to create, share, and discuss cultural contents. Web 2.0, the Social Web, and Mobile Web 2.0 provide a lot of useful tools:

- *Wikis* are websites that allow users to create, edit, and link web pages easily, e.g. Wikipedia[4].
- *Blogs* are websites where entries of different types of content are usually displayed in reverse chronological order, e.g. Blogger[5] and MoBlog:UK for mobile devices[6].
- *Tagging (Folksonomy) and social bookmarking* let users use keywords to attach to a digital object to describe it, e.g. del.icio.us which launched the "social bookmarking" phenomenon[7], Mobilicio.us[8] is a "mashup" of del.icio.us or Ma.gnolia[9] online bookmarking services with Google Mobile[10].
- *Multimedia sharing* are services that let the easy storage and sharing of multimedia content, e.g., Flickr for photo[11], YouTube for video[12], Odeo for podcast[13], Twitter[14] and Jaiku[15] for mobile.
- *Virtual worlds* websites that create a virtual parallel world, e.g. Second Life[16].

According to Web 2.0 concepts of remixability and aggregation, the development and adoption of:

[2] http://www.archimuse.com/conferences/mw.html
[3] http://www.archimuse.com/index.html
[4] http://en.wikipedia.org/
[5] http://www.blogger.com/home
[6] http://moblog.co.uk/index.php
[7] http://del.icio.us/
[8] http://mobilicio.us/
[9] http://ma.gnolia.com/
[10] http://www.google.com/mobile/
[11] http://www.flickr.com/
[12] http://youtube.com/
[13] http://odeo.com
[14] http://twitter.com/
[15] http://jaiku.com/
[16] http://www.secondlife.com

– OpenApi[17];
– OpenSocial Api[18];
– DataPortability philosophy[19];

enable websites to interact with each other by using SOAP, Javascript and any other web technology. This approach allows to interconnect websites in a more fluid user-friendly manner not only for programmers but for users too. By reusing and remixing these tools, static content authorities could evolve to dynamic platforms for content generation and sharing.

In this paper, we first survey related experiments aimed at exploiting both Web 2.0 and Mobile 2.0 solutions. We then highlight their limitations and we propose an abstract and general reference model for cultural heritage systems, that lets users create, share, and use cultural contents including mobile context-aware features. Our model is the starting point of an ongoing project, and it will eventually lead to a particular implementation named m-Dvara 2.0. m-Dvara 2.0 represents an evolution of E-Dvara, a previous platform for cultural and scientific contents in digital format[20]. The "m" and "2.0" in m-Dvara 2.0 highlight the mobile and social nature of our platform.

2 Mobile Social Software for Cultural Heritage: Related Work

2.1 Current solutions

Most museums, cultural sites, libraries, and other educational and cultural websites are not involved in Web 2.0 (r)evolution: they are the sole provider of contents, whereas users are only consumers. However, some cultural heritage organization and some educational institutions have introduced Web 2.0 services in their sites:

– Tagging (Folksonomy)
  – Steve[21] is a collaborative research project exploring the potential for user-generated descriptions of the subjects of works of art to improve access to museum collections and encourage engagement with cultural content. A group of US art museums are taking a similar folksonomic approach to their online collections.
  – Trant [14] has explored the potential of social tagging by comparing terms assigned by trained cataloguers and untrained cataloguers to existing museum documentation at The Metropolitan Museum of Art in New York[22].

[17] http://en.wikipedia.org/wiki/Open_API
[18] http://code.google.com/apis/opensocial/
[19] http://dataportability.org/
[20] http://edvara.uniud.it/india
[21] http://www.steve.museum/
[22] http://metmuseum.org

Preliminary results show the potential of social tagging and folksonomy to open museum collections to new and more personal meanings. Untrained cataloguers identified content elements not described in formal museum documentation. Tags assigned by users might help to bridge the semantic gap between the professional discourse of the curator and the popular language of the museum visitor.

- Virtual Worlds
  - Louvre Museum[23], one of the first museums on the web, offers no real Web 2.0 services [6], although it is present on Second Life.
  - Public Library of Charlotte and Mecklenburg County[24] has a teen outreach program that includes a presence in Teen Second Life[25].
- Community Multimedia Sharing
  - Tate museum offers the website youngtate section[26] to young people to create new learning communities, opportunities for input and activity based on personal choice, and innovative forms of interaction with art and artists [3].
  - Brooklyn Museum[27] site has a Community section with blogs, podcasts, forums and a Flickr-based photos sharing service [6].
  - Brooklyn College Library uses MySpace to allow participants to post personal profiles containing their favourite books, movies, photos, and videos[28].

Many projects have been developed to study how to integrate mobile devices in museum visits. Besides common mobile guides, projects for museum co-visits with mobile device [9] involve individual and then collaborative user activities enabling communication, sharing, and collaboration among visitors in their museum experience.

Recent work on cultural institution trying to meet Web 2.0 challenges (e.g., [10]) helps us in a classification based on topics and types of services offered to the virtual and real-world visitor. A list of topics of interest for cultural institution projects are:

- Art cataloguing and description: social services using folksonomies as more efficient way of cataloguing and description of an artwork;
- Collection access: collaborative social services in order to offer access to large collections of cultural content;
- Education: social services for collaborative creation of multimedia content for students, even educational-games, communities;
- Exhibition: resources to enrich user experience of exhibitions;
- History: social services as archival multi-medial or textual resources;
- Promotion and marketing: social services and syndication techniques to promote cultural activities, events or new available contents;

[23] http://www.louvre.fr
[24] http://plcmc.org/
[25] http://plcmc.org/Teens/secondLife.asp
[26] http://www.tate.org.uk/youngtate/
[27] http://www.brooklynmuseum.org/community/
[28] http://www.myspace.com/brooklyncollegelibrary

– Recommendation: social services and monitoring systems to provide recommendation based on user behaviour in order to suggest contents, activities, paths or tours;
– Reference service: social service to improve professionals communities;
– Youth outreach: dedicated social services to gain the close interest of young visitors.

2.2 Limits of current solutions

From these examples it is clear that Web 2.0 technologies are transforming the methods of production and perusal of cultural and educational contents, and also that the heritage sector is evolving towards user generated content. However, all these "Museum 2.0" examples also share the common approach of merely giving to the users the tools to record their personal experience, while a few expert members still are the main content providers. This is different from a full 2.0 approach, in which the users are given the real opportunity of creating contents in a way that makes themselves essential. Another issue is the fragmentation of services offered by current social software projects for cultural heritage: various approaches have been implemented, but none has been able to identify and offer an organic set of mobile and social services to support and stimulate the virtual and real-world visitor experience through collaboration and participation.

3 A Reference Model for Cultural Heritage System

In the previous analysis, we evaluated services provided by different cultural heritage systems. In this section we describe our approach, whose ultimate aim is to let users to be not only visitors of an exposition but the main content creators through a framework of collaboration and participation based on Web 2.0 and Mobile 2.0 technologies. We propose a reference model according to the work presented in [7], for mobile social software for learning. In particular the approach to the model described has been transposed to the cultural heritage field.

Our research questions are: can the crowd become an effective and reliable contents producer? Can the crowd actively participate at the cultural heritage preservation and dissemination process? How users can be motivated to participate? Can we achieve these goals by means of appropriate Web 2.0 and Mobile 2.0 tools already existing?

3.1 Requirements

The reference model we propose describes how existing tools can be used in order to create a Web and Mobile 2.0 system for cultural heritage. In few words, our idea is to combine data from more than one existing source into a single integrated tool. Thus, we suggest a mashup model for cultural heritage system that implies:

– reuse of Web 2.0 technologies;
– reuse of Mobile 2.0 technologies;
– mix of web and mobile services;
– minimum implementation, through mashup of Web 2.0 and Mobile 2.0 services available online.

Also, our model conforms to the main tenets of the Web 2.0 philosophy: it is user centered, it is based on social software, it aims at anywhere and anytime access by means of mobile devices, and it allows and fosters knowledge sharing. We want to ask the user the minimum effort possible so that she can interact with our service platform in an easy and comfortable way. As it has been done partially by Brooklyn Museum and TateYoung, we want to integrate all possible Web 2.0 systems that the user usually uses for her online community activities (MySpace, Flickr, Blogger, etc). By doing so, we will provide an all-in-one familiar set of services for users. Our reference model is an empty box, with many mashup services, where contents have to be inserted by both expert and non-expert users. There is not a central authority that publishes and controls all contents, but the crowd is the real controller of contents. From the development point of view, although the integration of existing services reduces implementation efforts, we cannot ignore its complexity, due to the functional heterogeneity of the same services. This heterogeneity affects also information visualization and user interaction aspects, but this matter is out of the aim of this first work, and it will be dealt with in future steps of our project.

3.2 Functionalities

System functionalities can be classified according to users location and technology being used:

– technology (a user can use a mobile device, desktop, notebook, etc.),
– location (a user can be on-site or off-site).

These user features are used to provide the appropriate services, e.g., a tourist visiting an exhibition will not need a video guide, or will not watch detailed photos on a mobile device, but probably she would like to listen to a location-aware audio guide. In particular, our model describes the way in which both on-site tourists visiting artworks and off-site users interact with the system (figure 1). In both cases, users can add content (posts, comments, etc.), upload new photos, videos, audios. We propose to mashup all these collected information in order to give to user a view of her work (*user view*). Moreover, we can obtain a more complete view about an artwork joining all users views. To accomplish this aim we can aggregate, filter, evaluate, and rate all available contents about an artwork. In this way our system can create an *artwork view*. Conceiving a social tool for cultural heritage in which we could use all available information about registered users, we can capitalize also on the power of the long tail, i.e., on those users that know (or use) only few system functionalities. We can keep track of all events generated by users, (i.e., visited objects, that can be real or

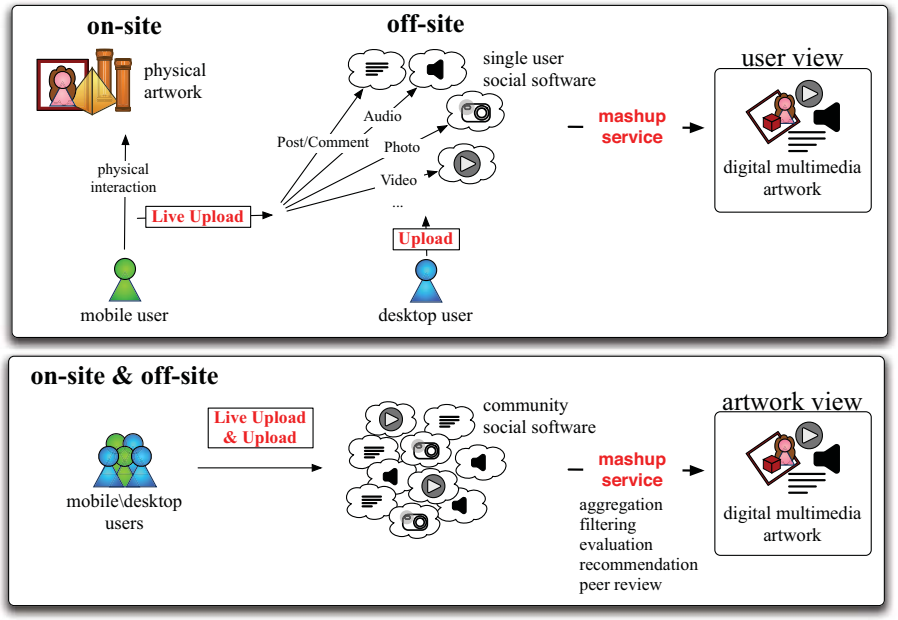


Fig. 1. On-site and off-site user interaction model: mashup creation of user view and artwork view

digital, time spent near each artwork, etc.) and we can create for each of them a *user events cloud* (a kind of user cultural history), that can help us to enjoy new features or improve already existing services (e.g., rank of content to be shown in a social tour or by social guides); see, for example, what we call *custom tour* in Section 3.2.

We can distinguish main system functionality according to Shneiderman's approach to relating human activities and relationships: Activities and Relationships Table (ART) [13]. Table columns represent four activities: collect (information), relate (communicate), create (innovate), and donate (disseminate). The four rows represent relationships, each one describing an increasingly large group (self, family and friends, colleagues and neighbors, citizens and markets) that we generalize to: self, neighbors (including family, friends, and colleagues), and the whole Web (table 1).

We now describe the main system features by means of three possible scenarios.

Scenario 1. On-site user with a mobile device In this scenario we imagine a tourist visiting a museum, an artwork exhibition, an archaeological excavation, etc.

Live Upload The tourist can capture content at the point of inspiration and upload it in real-time on system. Content can be of different kinds: photos,

	Collect	Relate	Create	Donate
Self	Bookmark M-Bookmark Feed Reader		Note Live Upload Upload	
Neighbors	M-Teach Teach	Comment 3D interactive environment	Blog MoBlog Wiki Live Upload Upload M-Meach	
Whole Web	Social Tour	3D interactive environment	Blog MoBlog Wiki Live Upload Upload Live Tagging Social Guides	Recommendation Rating User Events Cloud

Table 1. ART: Activities and Relationships Table of a user in a cultural heritage system

videos, audios, text about an artwork (comments or posts), drawings, etc. She can update her personal page or public page. Twitter, Jaiku technology, and/or YouTube Mobile[29] can be used to upload video. Live Upload differs from simple Upload: the first one take place in real-time, for example while the user is visiting a museum, in contrast with the second one that is related to non real-time experiences.

Live tagging The tourist can tag, using her own mobile device, the artwork she is looking at.

Evaluation and rating Collaboration and participation features involve evaluation mechanisms and for this reason we propose the adoption of social evaluation. Following [11], all contents can be judged by users (e.g., according to accuracy, comprehensibility, etc.). The score assigned to a content item will depend on the combination of the score given by a user and the user's actual score. In addition, every content provider has a dynamic reliability score that depends on the scores of contents she produced. In this way, the crowd is the reviewer of its own contents. Moreover, a tourist can rate every artwork. This rating, combined with the user profile, contributes to improve the artwork profile. In this way the system can suggest to tourists the artworks closer to their preferences.

Social tour The system can help tourists by suggesting a tour. The tourist can request to the system an ideal tour according to her preferences, and/or

[29] `http://youtube.com/mobile`

tourist can select on her mobile device a tour criterion. There are three main kinds of tours: custom, dynamic and contextual tour. For custom tour we mean that system can detect user information keeping track of her actions (e.g. visited places or artworks, commented posts) or it can evaluate user's profile to set her preferences, then system process these information in order to create the user's ideal tour. A dynamic tour does not relate to user's personal information, but it depends on all users actions, thus user can decide to visit the most viewed, most commented, or most voted artworks. In other words, she can visit all the artworks that the crowd (community) advises to see. Finally, in a contextual tour, user can decide to visit only artworks about a specific topic or artworks belonging to the same artist, and so on. In addition, a tourist can change the tour criterion or she can add or remove artworks to visit from the suggested list at any time. To detect user location we intend to integrate Google Mobile with MoBe location features [5, 4].

Social guides A cultural heritage system could be a guide. A tourist can record an artwork description as a guide and listen an audio description from her mobile device about the item she is examining. She can also access a wiki in order to read or use a screen reader to know what she needs. All different descriptions about a certain object are rated according to the crowd opinion (social evaluation). We can use, again, Twitter or Jaiku.

Travel diary The system can keep track of artworks, monuments and places the user has seen, in order to maintain a personal travel diary.

Questions and answers A tourist can post a question, or answer to question posted by other users in the community.

M-Note The tourist can note down on her mobile device whatever she needs to retain about the object she is observing. To this aim we can exploit Google Notebook.

M-Bookmark To bookmark from mobile devices. For this we can integrate Mobilicio.us.

M-Teach Students can use their own mobile devices for educational lab activities.

Scenario 2. Off-site user with a desktop or notebook device User accessing to cultural heritage system from his own desktop or notebook device.

Wiki per topic User can create, add, modify, delete contents about a topic or an object to the open wiki in a collaborative way like, e.g., in Wikipedia.

Wiki per author Each article can be written by a single author and other users can edit it only with permission from the author, like, e.g., in Knol. There are also multiple articles for the same topic, each written by a different author. Readers may rate or comment on the articles. Wiki per author lets users know who wrote what, so they can make better use of Web content.

3D collaborative environment User can visit a 3D museum or a 3D exhibition, interact with other users or a guide in the museum, as in the real word. Moreover we can merge the 3D museum (e.g. Second Life) with wiki, chat, photo, and comments of users. In this way user can visit 3D environment

and talk with other visitors, but she can also update a wiki, write comments, upload photos, videos, etc.

Blog User can write a post about an artwork on her own blog, or on a blog dedicated to a specific topic. Also, she can comment in other blogs.

Bookmark User can bookmark other users Web-pages or artwork dedicated Web-pages.

Personal profile and social network User can manage his social network, defining white and black lists. He can select his "friends" in order to create a personal sub-community. He can also suggest other user he is interested in, in order to be notified of their new posts. Similarly a user can suggest posts or themes he is interested in to be notified of their evolution.

Scenario 3. Off-site user with a mobile device User accessing to cultural heritage system from his own mobile device.

MoBlog User can upload photo, video, text, audio on the blog section. We can exploit MoBlog.

Live Upload Like tourist on-site, also user online with mobile device can live upload content on system.

4 Conclusions and Future Work

In this paper we have presented how various museum evolution projects aim at providing Web 2.0 services for improving user's experience. However, these projects lack of users participation as the central content creators, since the main content creators remain a few institutional experts. We have then described the starting point of an ongoing project, namely a reference model for a more integrated approach. The goal of our project is to produce a service that allows the crowd of users to control (manage) the knowledge flow through collaboration and participation. The service will be developed as an aggregator of Web 2.0 and Mobile 2.0 services for institutions of humanistic field. Users participation and motivation are essential and this leads to the question: "why the user should use our system?". Our system could be an important added value service to the user, but we are going to verify our believes only at the final stage of the project, with appropriate user testings. At this point, we trust in the popularity of the Web 2.0 services we rely on. The project is rather ambitious, and we will be facing many problems. For example, the reuse and remixing of other services involve the direct dependence on their existence (what would happen if some service stops its functioning?). Redundancy and robustness will be key factors. Also, copyright issues are a complex field, dependent on single nation legislation, and should be taken into account when working with cultural heritage contents.

Focusing on the evaluation system, there are several aspects that need future investigations. In particular there should be just one general user's score, or it is better to have a score for each field of contribution? Several scores help to give a more detailed evaluation of a user, that could be an architecture expert

but have no experience in science. On the other side, giving a score for a specific field requires the use of a kind of classification, like, for example, a folksonomy.

A user's score can also change because of an interaction between the community and the user's contributions. Possible parameters are: "how many times a content is read", "how many times a content is updated by the community", "how much time passes between a creation of a content and an update", etc. For example, if a content is frequently read, but never changed, it could be a good content; on the contrary if a content is never read, it is probably not good. Having no control on users' contributions, we hypothesize that this evaluation system could be a way to automatically manage the contents quality.

Acknowledgements

The authors acknowledge the financial support of the Italian Ministry of Education, University and Research (MIUR) within the FIRB project number RBIN04M8S8.

References

1. A. Alain and M. Foggett, (2007). Towards Community Contribution: Empowering community voices on-line. In J. Trant and D. Bearman (eds). Museums and the Web 2007: Proceedings. Toronto: Archives & Museum Informatics, `http://www.archimuse.com/mw2007/papers/alain/alain.html`
2. P. Anderson, (2007). What is Web 2.0? Ideas, technologies and implications for education, JISC Technology and Standards Watch, `http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf`
3. R. Cardiff, (2007). Designing a Web Site for Young People: The Challenges of Appealing to a Diverse and Fickle Audience. In J. Trant and D. Bearman (eds). Museums and the Web 2007: Proceedings. Toronto: Archives & Museum Informatics, `http://www.archimuse.com/mw2007/papers/cardiff/cardiff.html`
4. P. Coppola, V. Della Mea, L. Di Gaspero, S. Mizzaro, I. Scagnetto, A. Selva, L. Vassena, and P. Zandegiacomo Riziò, (2005) Information Filtering and Retrieving of Context-Aware Applications Within the MoBe Framework. In Proceedings of CIR 2005 - International Workshop on Context-Based Information Retrieval, CONTEXT 2005, Paris, France.
5. P. Coppola, V. Della Mea, L. Di Gaspero, S. Mizzaro, I. Scagnetto, A. Selva, L. Vassena, and P. Zandegiacomo Riziò, (2005) MoBe: A Framework for Context-Aware Mobile Applications. In Proceedings of CAPS 2005 - Workshop on Context Awareness for Proactive Systems, Helsinki, Finland.
6. G. Crenn and G. Vidal, (2007). Les Musées Français et leurs publics a lâge du Web 2.0. Nouveaux usages du multimédia et transformations des rapports entre institutions et usagers? , in International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics, `http://www.archimuse.com/ichim07/papers/crenn/crenn.html`

7. T. De Jong, M. Specht and R. Koper, (2007). A reference model for mobile social software for learning. International Journal of Continuing Engineering Education and Life-Long Learning, 18(1), 118-138, http://hdl.handle.net/1820/996
8. B. Groen, (2007). Culture 2.0, Cultuur 2.0 Online PDF Publication, http://www.virtueelplatform.nl/download.php?id=5721
9. Y. Laurillau, and F. Paternò, (2004). Supporting museum co-visits using mobile devices. Proceedings of Mobile HCI 2004, Glasgow, Scotland, http://giove.cnuce.cnr.it/pdawebsite/publications/MobileHCI04.pdf
10. M. Middleton and J. Lee, (2007). Cultural institutions and Web 2.0. In Proceedings Fourth Seminar on Research Applications in Information and Library Studies (RAILS 4), RMIT University, Melbourne, http://eprints.qut.edu.au/archive/00010808/01/Cultural_Institutions_and_Web_2_0.pdf
11. S. Mizzaro, (2003). Quality Control in Scholarly Publishing: A New Proposal, Journal of the American Society for Information Science and Technology, 54(11):989-1005.
12. T. O'Reilly, (2005) What Is Web 2.0, Design Patterns and Business Models for the Next Generation of Software, http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
13. B. Shneiderman (2002). Leonardo's Laptop: Human Needs and the New Computing Technologies. MA: The MIT Press. Cambridge, http://mitpress.mit.edu/main/feature/leonardoslaptop/pdf/chapter5.pdf.
14. J. Trant, (2006). Exploring the potential for social tagging and folksonomy in art museums: proof of concept. New Review of Hypermedia & Multimedia, 12(1), 83-105, http://www.archimuse.com/papers/steve-nrhm-0605preprint.pdf

Managing conflicts between users in Wikipedia

Bernard Jacquemin[1], Aurelien Lauf[1], Celine Poudat[2], Martine Hurault-Plantet[1], and Nicolas Auray[2]

[1] LIMSI CNRS UPR 3251 — BP 133 – 91403 Orsay Cedex, France
{Bernard.Jacquemin,Aurelien.Lauf,Martine.Hurault-Plantet}@limsi.fr
[2] Sinequa — 12, rue d'Athènes – 75009 Paris, France
poudat@sinequa.com
[3] ENST Paris — 46, rue Barrault – 75013 Paris, France auray@enst.fr

Summary. Wikipedia is nowadays a widely used encyclopedia, and one of the most visible sites on the Internet. Its strong principle of collaborative work and free editing sometimes generates disputes due to disagreements between users. In this article we study how the wikipedian community resolves the conflicts and which roles do wikipedian choose in this process. We observed the users behavior both in the article talk pages, and in the Arbitration Committee pages specifically dedicated to serious disputes. We first set up a users typology according to their involvement in conflicts and their publishing and management activity in the encyclopedia. We then used those user types to describe users behavior in contributing to articles that are tagged by the wikipedian community as being in conflict with the official guidelines of Wikipedia, or conversely as being well featured.

Key words: Social network, Wikipedia, Web community, Conflict, Collaborative work

1 Introduction

The Wikipedia encyclopedia project has become a reference informational resource, and one of most visible sites on the Internet. Amazing and far removed from the Enlightenments spirit – where the expert and his signature constitute the text quality guarantee –, Wikipedia is based on a very different editorial process.

The whole project is based on a few strong ideological principles, also called *pillars*, *official guidelines* or *fundamental principles* in Wikipedia. First, the goal is clearly to be a generalist encyclopedia project with several linguistic instances that are independently managed. Then, the Wikipedia contents also have to be objective. Wikipedians reckon that the best way to grant the objectivity is to set out a *neutral point of view* (NPOV)[1]. Moreover, texts are freely edited and redistributed, and the encyclopedia has been developed with free and open

[1] The articulation between both is performed as follow: "What people believe is a matter of objective fact, and we can present that quite easily from the neutral point of view." (Jimbo Wales, co-founder of Wikipedia, http://en.wikipedia.org/wiki/

source software. The entire editorial process, from the writing articles to the macrostructure organization, is collectively managed. Finally, the wikipedians have to respect elementary good manners. So, even if the Wikipedia editorial process totally differs from the traditional encyclopedia one, the goals of encyclopedic relevance and objectivity are in fact very close [5, 7].

Several formal and informal ways to regulate and control the encyclopedia have progressively been introduced by the wikipedian community in order to obey and to make users obey the *pillars*. The common wikipedian philosophy makes it possible to gather together a large population of users writing about an unlimited number of themes or domains, to share their incomplete knowledge, to represent the various ways of thinking, and to delete errors thanks to successive users rectifications [15, 3]. However, this philosophy also generates disputes and conflicts linked to inevitable disagreements between contributors. What processes does the wikipedian community use to resolve the conflicts, and what roles do the wikipedians choose in this process?

In this article, by analyzing the contributors behavior in places where conflicts are resolved, we provide elements to help answer these questions. The users behavior is observed both in the articles that are tagged as being in particular accordance (*good* or *featured articles*), or conversely not in accordance, with the main guidelines of Wikipedia (*relevance dispute articles*, *NPOV dispute articles*...), and in pages specifically dedicated to serious personal conflicts, the *Arbitration Committee* [16, 13]. As a result, we present the following contributions:

First, we make a users typology according to parameters that bring to light their involvement in conflicts and their publishing and management activity in the encyclopedia. In particular, we establish relationships between the number of appearances before the Arbitration Committee, the initiation of a request to the Arbitration Committee, and the numbers of contribution to articles and talk pages of Wikipedia. We show that major contributors are often involved in arbitration, and mostly as the initiating party.

Then, we analyse the distribution of those types of users among the contributors to articles that do not respect a neutral point of view, given that it is one of the most important principles of Wikipedia. We find that all the major contributors who take their conflict before the Arbitration Committee are also contributors to NPOV articles, against only one half for the minor contributors.

Finally, by analysing the distribution of those wikipedians involved in serious disputes, among the contributors to tagged articles, we find that major contributors who are often involved in arbitration, are much more frequently contributing to protected articles (subject to disputes or vandalism), than to featured articles.

`Wikipedia_talk:Attribution/Role_of_truth`). Thus the Wikipedia's aim at the objectivity is only performed at an *opinion* inventory level, despite their uneven quality on the same page [8].

2 Related work

A number of authors study conflicts in Wikipedia in relation with coordination and cooperation underlying collaborative work. For instance, [9] develop quantitative measures of the costs involved by collaborative work, using the concepts of direct (i.e. writing article) and indirect work (i.e. discussion or anti-vandalism). At the article level, the history of the revisions is often used to model and identify conflict or coordination periods [9, 14]. The aim of the present study is rather to analyse the behavior of wikipedians, who are involved in conflicts, faced with the main tools wikipedians use to resolve conflicts.

Studies of conflict management and social control in virtual communities show that such social systems have the same kind of problems as real social systems. In particular, [10] show that the *social dilemma* between individual and collective interest in the problem of cooperation remains, even if it takes other forms. Furthermore, [4] observes that methods using both mediation and arbitration better manage conflicts than power strategies of social control, as it does in the real world. Indeed, the way a community manages its conflicts reveals its governance mode [2, 9, 14]. In the French Wikipedia, mediation takes place in talk pages of articles which have a template message at the top of the page, and arbitration takes place in the Arbitration Commitee pages.

In fact, template messages at the top of article pages are strongly linked to the official guidelines of Wikipedia. Indeed, these principles play an important role in the management and resolution of conflicts. [15] analysed the content of the article talk pages, and found that 7.9% of the activity in those pages consists in references to Wikipedia official guidelines.

The behavior of wikipedians has been studied either from their motivations point of view [11], either considering the type [12] or the evolution of their participation [3]. Our analysis of the behavior of wikipedians is based on quantitative data as well as in [12], but is restricted to those wikipedians who are involved in conflicts.

3 Corpus

Wikipedia is a generic term for the free multilingual and collaborative online encyclopedia[2] as well as a reference to every instance of this encyclopedia. Each instance refers to a different country and/or language. The instance we are interested in for this article is the French version of Wikipedia[3]. The corpus we used was extracted from the Wikipedia backup of 2006/04/02: more than 600,000 pages including 370,000 article pages and 40,000 talk pages (according to Wikipedia's internal architecture, each article page can be linked to a talk

[2] Available at `http://www.wikipedia.org/`.
[3] Available at `http://fr.wikipedia.org/`.

page). A tool called Wiki2Tei[4] was then used in order to convert the wikitext syntax to a TEI-compliant XML syntax (TEI standing for *Text Encoding Initiative*).

The articles of Wikipedia are written by voluntary contributors working with each other via a wiki. Since anyone can freely edit any article, many virtual places are provided to avoid or settle conflicts that may arise in the process. First of all, each article is linked to a discussion page where contributors can exchange and justify their assertions, and thus reach compromises according to Wikipedia's netiquette and neutrality policy. Furthermore, users can insert specific tags[5] on top of articles which do not respect Wikipedia's official guidelines (such as neutrality or relevance dispute) [11, 6] or, on the contrary, to reward an exemplary article (called *featured* or *good articles*[6]). Theses tags are used to highlight for the community the fact that some articles need improvement and thus can be used as points of reference for users. Finally, when disputes degenerate into personal conflicts and get out of hand, each user can register a complaint to the Arbitration Committee. The Arbitration Committee is a group composed of seven contributors to Wikipedia, elected by the rest of the community for six months. Deliberations and votes of the Arbitration Committee are public and usually tend to reach unanimity, which implies consensus, as it is the rule for the articles. The role of Arbitrators is not to express an opinion about the scientific rightness or the editorial policy of an article but to ensure that Wikipedia's official guidelines are respected: neutral point of view (NPOV), the need to cite general sources, netiquette (called *wikilove* by the wikipedian French community), the respect of the law, etc. They have the right to impose sanctions on users such as temporary or definitive article probation (meaning that the user cannot contribute anymore to one or more articles) or, less often, general restriction (meaning that the user is literally banned from all Wikipedia).

Thus, there are three virtual places to manage a conflict, in order of seriousness: the discussion pages linked to an article, the discussion pages linked to an NPOV dispute article and the pages of the Arbitration Committee. We focus on the last two because they correspond to open conflicts.

The first corpus we collected is composed of about 1,000 articles that have (or have had) the NPOV tag. Each article is associated, when possible, to its discussion page (some articles are not linked to a discussion page because the discussion may have started after we extracted the corpus). About 1,600 contributors intervened in these pages. We automatically added semantic tags to this corpus in order to extract each contribution and its size, who wrote it and when – which tells us which contributions were written during the conflict and which were not – and, when possible, to whom it answers. However, it is impossible to

[4] Open software available at `http://wiki2tei.sourceforge.net/` and freely distributed according to the terms of the BSD license (`http://www.opensource.org/licences/bsd-license.php`).

[5] Defined in Wikipedia as "a frame type in articles indicating a piece of information or a link" `http://en.wikipedia.org/wiki/Wikipedia:Template`.

[6] `http://en.wikipedia.org/wiki/Wikipedia:Good_articles`.

know who wrote a contribution when users do not sign it, deliberately or not. This is the reason why between 2% and 5% of the contributions may have been improperly tagged.

The second corpus is composed of about 80 pages from the Arbitration Committee. These pages are relatively well formed and homogeneous, allowing us again to automatically tag them so as to clearly make their essential architecture stand out: the conflict description, who registered the complaint and when, the parties involved, if the complaint is admissible or not, and the verdict of the arbitrators. Furthermore, each user is associated to his messages, and each arbitrator to his contributions and, of course, his vote. Finally, the verdict is composed of at least one verdict proposal and a vote; there are as many counterproposals and votes as needed until the arbitrators are able to reach an agreement. Each proposal is clearly identified and associated to the right arbitrator and each vote is associated to its arbitrator and to the proposition it refers to.

4 Typology of users in conflict

The Arbitration Committee is therefore a formal place for the resolution of conflicts. Though rather rare – only about one hundred users among 31 000 wikipedians were implied in an arbitration within a 5-year period –, arbitrations represent an important tool for Wikipedia governance. Indeed, elected arbitrators can impose penalties against Wikipedia users who transgressed the *pillars*. For instance, penalty may consist in blocking a user in order to keep the user from writing within articles during a certain period of time. It therefore gives strong means for controlling publication.

Among the hundred arbitrations which took place from the beginning of Wikipedia-France to 2006 april, some user names appear more often, either as the *initiating party*, or as the *other involved party*. Those two topics, frequency of appearance and role in the complaint, allow us to draw up an initial typology of users engaged in a dispute. We first distinguished three kind of protagonists depending on the frequency of their appearances: *very regular ones* who have between 3 et 14 appearances[7], *regular ones* who have two appearances, and *occasional ones* who have only one appearance. Concerning their role in the complaint, we then distinguished three categories, the *initiating party*, that is to say those who are most often the initiator of the complaints, the *other involved party*, and finally those who appear in a more balanced way, sometimes as initiating party and sometimes as other involved party. We can see on Table 1 that among the wikipedians who often appear, the *very regular ones*, are the initiating party for most part, even though *occasional ones*, who appeared only once, are mainly other involved party. We also note that most of those who appeared twice took once the initiating party position, and once the other involving party position.

[7] 14 is anyway a sort of record, then there are two of them having 7, another having 4, the other ones having 3 appearances

Table 1. Appearances before the Arbitration Committee

Appearances	Users	Initiating party	Other party	Both
3–14 (very regular ones)	10	50%	30%	20%
2 (regular ones)	17	12%	29%	59%
1 (occasional ones)	74	30%	70%	0%

We then added to that typology the way users contribute to Wikipedia. We considered the number of their contributions in editing articles, either in article pages, or in the discussion pages, because it is mainly in this place that conflicts begin[8]. Concerning this point, we noted big differences between users. We drew up four categories, the *major contributors* whose number of contributions extends from about 12,000 to 40,000 during the studied period, the *Large contributors*, between 2,800 and 12,000 contributions, the *middle contributors* between 600 and 2,800, and the *minor contributors*, between one and 600 contributions. Finally, we considered the type of their contributions according to whether they contribute to article pages or discussion pages. We therefore distinguished three categories according to whether they contribute more often to articles or to discussions, or to both of them in a balanced way.

Table 2. The contributions of the protagonists before the Arbitration Committee

Contributions	Users	Article orient.	Discussion orient.	Both
12,000–40,000 (Major contrib.)	7	100%	0%	0%
2,800–12 000 (Large contrib.)	23	96%	0%	4%
600–2,800 (Middle contrib.)	31	81%	0%	19%
1–600 (Minor contrib.)	40	70%	5%	25%

Table 2 shows that users who get involved in disputes in Wikipedia contribute more to articles than to the associated talk pages, despite their conflicts. Nevertheless, it also shows that the less they contribute to articles, the more they have a tendency to discuss.

Comparing the number of contributions and the frequency of appearances (Table 3), we realize that parties of the Arbitration Committee who are *very regular* are for the most part *big contributors*, while *occasional* ones are more often *small contributors*.

Table 3. Categories of contributors in complaints

Appearances	Contributors	Major	Large	Middle	Minor
3–14 (very regular ones)	10	20%	50%	30%	0%
2 (regular ones)	17	13%	29%	29%	29%
1 (occasional ones)	74	4%	18%	31%	47%

[8] We did not consider for instance contributions in the *bistrots* of Wikipedia.

Comparing the number of contributions and the role in the complaint (Table 4), we note that the *big contributors* are more often the initiating party and that the *small contributors* are more often the other involved party. Indeed we note an increase of the proportion of *other involved party* and a decrease of the proportion of *initiating party* as the number of contributions decreases. Part of protagonists who are sometimes the initiating party and sometimes the other involved party is marginal for each category of contribution size.

Table 4. Role in the complaint by size of contribution

Contributions	Users	Initiating party	Other party	Both
12,000–40,000 (Major contrib.)	7	57%	29%	14%
2,800–12,000 (Large contrib.)	23	39%	44%	17%
600–2,800 (Middle contrib.)	31	32%	58%	10%
1–600 (Minor contrib.)	40	15%	75%	10%

The analysis of these tables evokes that the big contributors assimilated the *pillars* of Wikipedia, and really care about enforcing them [1, 6]. Indeed, the emerging trend is that the more they contribute to articles, the more they carry out publication control at the same time. They exercise this control in the framework of the Arbitration Committee through their role as initiating party. They exercise this control mainly over *middle* and *small* contributors.

In the following section, we study whether we can complete this typology of contributors before the Arbitration Committee with the types of article they contribute to, involving the *pillars* of Wikipedia. Indeed, we saw that users put different tags within articles in order to warn other users about breaches of the rules of Wikipedia. We used those tags to categorize articles as *featured articles*, *NPOV dispute articles*, *relevance dispute articles*, and *protected articles*.

5 Users in conflict and *pillars* of Wikipedia

The NPOV dispute tag is the first tangible evidence of a disagreement between wikipedians. Thus we studied characteristics of contributors who participated in articles with the NPOV tag, and particularly the ones who are also parties of arbitration by the Arbitration Committee. This analysis reveals several behavior trends. In Table 5, we study the behavior of the contributors, shared out in categories following the number of their contributions. We compare contributors in articles with a NPOV tag to all the contributors in Wikipedia. The second column indicates for each section the number of contributors in NPOV articles. The third column shows the number of appearances before the Arbitration Committee for the contributors in NPOV articles in comparison with all the protagonists before the Arbitration Committee, for each category (see Table 2). In Table 6, we study the behavior of the contributors who appear before the Arbitration Committee, considering on the one hand the appearance frequency, and on the other

hand their role in the complaint. The second column indicates, for each category of frequency and of role, the number of contributors in Wikipedia who appear before the Arbitration Committee. The third column indicates for each category the number of contributors in NPOV articles who appear before the Arbitration Committee, and the proportion of these contributors to all the contributors of the same category who appear before the Arbitration Committee. Table 5 shows that 77% of the protagonists before the Arbitration Committee appear among the 1600 contributors participating to at least one article with the NPOV tag. It suggests that a lot of conflicts arise from an objectivity controversy.

Table 5. Protagonists who appear before the Arbitration Committee (AC) among the contributors in NPOV articles, by contributions size

Contributors categories	# NPOV contributors	NPOV contributors before the AC
Major contributors	30	7 (100% of 7)
Large contributors	151	21 (91% of 23)
Middle contributors	335	27 (84% of 31)
Minor contributors	1121	23 (57% of 40)
Total	1637	78 (77% of 101)

Table 6. Protagonists who appear before the Arbitration Committee (AC) among the contributors in NPOV articles, by appearances type

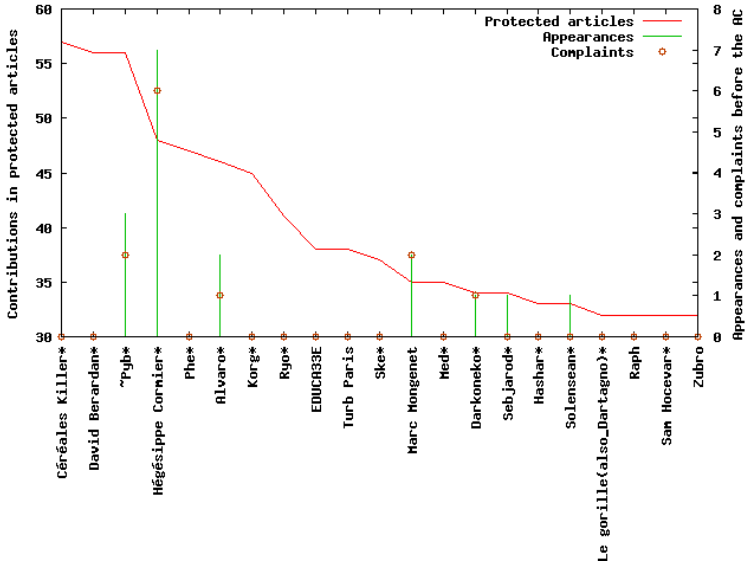
Protagonists categories	Before the AC	In NPOV articles
Very regular	10	10 (100%)
Regular	17	12 (70%)
Occasional	74	56 (76%)
Initiating party	29	26 (90%)
Other party	60	44 (73%)
Both	12	8 (67%)

We also notice (Table 5) a very marked presence of the protagonists who appear before the Arbitration Committee among the most verbose contributors of our sample. We also note (Table 6) that the *very regular protagonists* before the Arbitration Committee and the initiating parties contribute more in NPOV pages than *regular* and *occasional protagonists*, or than other involved parties. The *very regular protagonists* and initiating parties are particularly present in NPOV discussions.

In order to study further the behavior of the contributors in conflict, we now consider their participation in other articles with a particular tag, indicating either a breach of relevance or objectivity principles, or a particular ageement with the official guidelines of Wikipedia. These tags are the neutral point of view (NPOV) dispute tag, the relevance dispute tag and the protected article tag, that takes place when the controversy degenerates into conflict in order to prevent

the article from being modified, and the featured article tag, that indicates its particular quality, according to the *pillars*.

Fig. 1. Contributors in protected articles and protagonists



In Figures 1, 2, 3 and 4, the sample comprises only contributors in NPOV articles, who sometimes also contribute in articles with another tag. The curves in these figures present in descending order the number of contributions for the 20 most verbose contributors, respectively in protected articles, in featured articles, in NPOV articles and in non-relevant articles. For each contributor, the number of his appearances before the Arbitration Committee (vertical line) and the number of his complaints (small circle) are also indicated, corresponding to the right scale.

We observe several interesting differences in these figures. In particular, among the 20 most verbose contributors in protected articles (Figure 1), 7 are protagonists before the Arbitration Committee, namely 35% of the major contributors on these articles. Furthermore, their behavior before the Arbitration Committee is disparate: some of them initiate the procedure and the others are other involved parties, some are very regular or regular protagonists and the others are occasional ones. On the other hand, Figure 2 shows that, among the most verbose contributors in featured articles, only 3 appeared before the Arbitration Committee, all of them as initiating parties. Nonetheless their apparent aggressiveness must be put into perspective: as none of these protagonists is a regular one, the complaints are few.

Fig. 2. Contributors in featured articles and protagonists

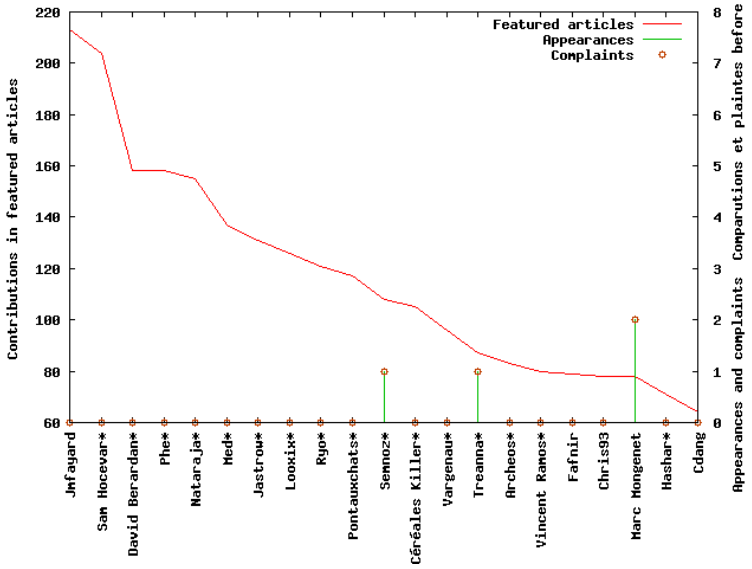
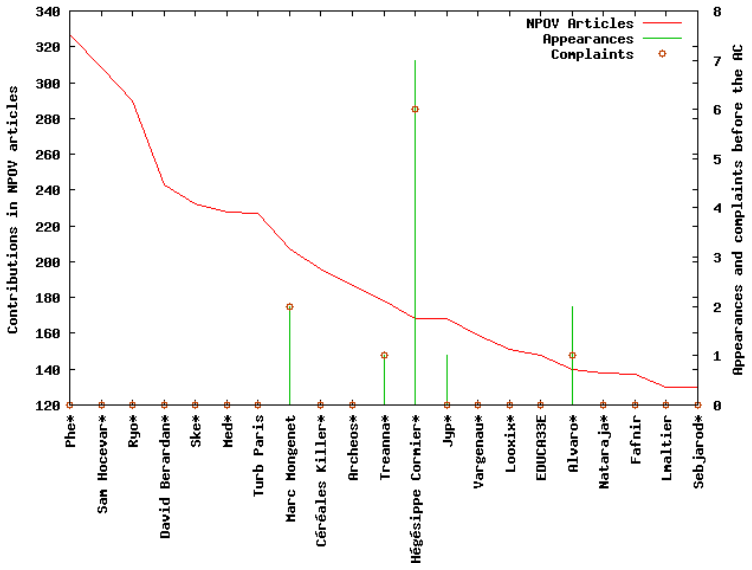


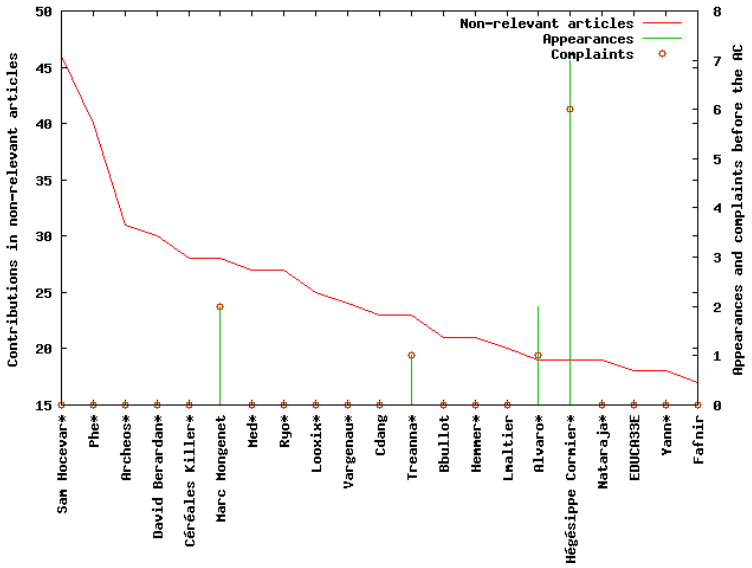
Fig. 3. Contributors in NPOV articles and protagonists



The behavior of the major contributors in NPOV and relevance dispute articles is between these two trends. Among the 20 most prolific contributors in NPOV articles indeed (Figure 3), 25% appeared before the Arbitration Com-

mittee. And 4 of the 20 major contributors in non-relevant articles, ie 20%, also appeared in arbitrations (Figure 4).

Fig. 4. Contributors in non-relevant articles and protagonists



In all these figures, the wikipedians with a particular status[9] are starred (*). It is interesting that most of the major contributors in the considered articles have also a particular status.

This observation confirms the previously mentioned correlation between a strong involvement of the contributors in the Wikipedia project, denoted both by the number of contributions and by the particular status [1, 6], and their intervention where and when the official guidelines need to be protected.

6 Conclusion

The Wikipedia encyclopedia is mainly based on collaborative work. This official guideline yields to cooperation patterns, including discussions and information sharing in order to realize the common goal. But such an extended collaboration also engenders conflicts. Disagreements which degenerate into serious personal disputes, with possible insults or systematic reverts, are finally not so frequent.

[9] Some particular status exists in the wikipedian community, e.g. administrator, steward, arbitrator, bureaucrat... Such a status is conferred by the community to a contributor through an election process. This status grants him/her extended rights in prospect of managing the encyclopedia.

They only involved one hundred users among 30,000 wikipedians over a period of five years. Official guidelines, the Wikipedia *pillars*, are clear, and there are not many of them. They constitute strong bases for conflict resolution. Tools and procedures have been developed step by step in order to enforce those principles.

We studied conflict evolution through the behavior of users who appear before the Arbitration Committee, and through their contributions to those articles that are tagged such as *featured articles*, *NPOV articles*, *non-relevant articles*, and *protected articles*. As expected, users appearing before the Arbitration Committee are more numerous on articles subject to a NPOV or relevance controversy, and much more on protected articles, than on featured articles.

The presence of involved parties before an Arbitration Committee has different meanings depending on whether one is the initiating party or the other involved party. We note that major and large contributors, also often involved as Wikipedia administrators, do most of the job of publication control. They are more often the ones who initiate arbitrations, and moreover the ones who contribute the most to featured articles. Tables 2, 3, 4 of Section 4 clearly show the evolution of the relative sizes respectively between initiating parties and other involved parties, between contribution to articles and contribution to discussions, between regular and occasional involved parties before Arbitration Committee, according to the size of contributions.

As a result, we may say that conflicts in Wikipedia are resolved both by means of a strong commitment to clear official guidelines, through specific places devoted to managing them, and by interventions of some attentive users.

Acknowledgements The research reported here was supported by a grant from the French National Research Agency (ANR), within the framework of the Autograph project ANR-05-RNRT-03002 (S0604108 W).

References

1. Anthony, D., Smith, S., Williamson, T.: Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia. Dartmouth College, Hanover (2005)
2. Auray, N., Poudat, C., Pons, P.: Democratizing scientific vulgarization. The balance between cooperation and conflict in French Wikipedia. Observatorio 3, 185–199 (2007)
3. Bryant, S.L., Forte, A., Bruckman, A.: Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In: ACM SIGGROUP Conference on Supporting Group Work, pp. 1–10. ACM Press, New York (2005)
4. DuVal Smith, A.: Problems of Conflict Management in Virtual Communities. In: Smith, M., Kollock, P. (eds.) Communities in Cyberspace, pp. 134–166. Routledge, London (1999)
5. Endrezzi L.: La communauté comme auteur et éditeur: l'exemple de Wikipédia. In: Jounée d'étude des URFIST "Évaluation et validation de l'information sur Internet" (2007)

6. Forte, A., Bruckman, A.: Why Do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing. In: GROUP 05 Workshop: Sustaining Community: The Role and Design of Incentive Mechanisms in Online Systems (2005)
7. Giles, J.: Internet encyclopaedias go head to head. Nature 438(7070), 900–901 (2005)
8. Gourdain, P., O'Kelly, F., Roman-Amat, B., Soulas, D., von Droste zu Hülshoff, T.: La Révolution Wikipédia. Les encyclopédies vont-elles mourir? Mille et Une Nuits, Paris (2007)
9. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, She Says: Conflict and Coordination in Wikipedia. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 453–462, ACM Press, New York (2007)
10. Kollock, P., Smith, M.,: Managing the virtual commons: cooperation and conflict in Computer communities. In: Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives, Susan Herring (ed.), pp. 109–128. John Benjamins, Amsterdam (1996)
11. Kuznetsov, S.: Motivations of contributors to Wikipedia. ACM SIGCAS Computers and Society 36(2), 1–7 (2006)
12. Ortega, F., Gonzalez-Barahona, J.M.: Quantitative Analysis of the Wikipedia Community of Users. In: WikiSym'07, pp. 75–86, Montreal, Canada (2007)
13. Stvilia, B., Twidale, M., Gasser, L., Smith, L.: Information Quality Discussions in Wikipedia. Technical Report, University of Illinois at Urbana-Champaign (2005)
14. Viégas, F.B., Wattenberg, M., Dave, K.: Studying Cooperation and Conflict between Authors with history flow Visualizations. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 575–582. ACM Press, New York (2004)
15. Viégas, F.B., Wattenberg, M., Kriss, J., Van Ham, F.: Talk Before You Type: Coordination in Wikipedia. In: 40th Hawaii International Conference on System Sciences (2007)
16. Zlatic, V., Bozicevic, M., Stefancic, H., Domazet, M.: Wikipedias: Collaborative web-based encyclopedias as complex networks. Physical Review E, 74(1) 6–11 (2006)

Investigating Weblogs in Small and Medium Enterprises: An Exploratory Case Study

Alexander Stocker[1], Klaus Tochtermann[123]

[1]Know-Center, Inffeldgasse 21, 8010 Graz
[2]Knowledge Management Institute, Inffeldgasse 21, 8010 Graz
[3]Institute for Networked Media, Elisabethstraße 20, 8010 Graz
{astocker, ktochter}@know-center.at

Abstract: Contrary to a Wiki where the opinion of the individual user disappears in favor of a more impartial 'collective intelligence', a weblog is author-centered, expressing the author's subjective point of view. This particular property of weblogs played a fundamental role for the popularity weblogs gained for making implicit knowledge explicit in an unsolicited, self-organized way. However, empirical studies from academia exploring internal corporate weblogs remain scarce, especially when they focus on small and medium enterprises (SMEs) which make up the majority of all enterprises worldwide. To counteract this lack of research, we investigate an internal corporate weblog in an ICT SME from a knowledge management perspective. We derive both research questions and hypotheses to test within future studies. Furthermore, we consider already gained findings from corporate weblog research and investigate their immediate applicability in the context of SMEs.

Keywords: Weblogs, Small and Medium Enterprises, Knowledge Management

1 Introduction

Not just because of the Web 2.0 hype (O'Reilly, 2004), weblogs enjoy a great popularity along with Wikis establishing a well-known source of user generated content on the Web. Being a 'log of the web', the term weblog, attributed to Jorn Barger refers to websites on which entries are commonly presented in reverse chronological order (Paquet, 2003). Termed with Enterprise 2.0 (McAfee, 2006) or Corporate Web 2.0 (Stocker et al, 2007), companies have identified an untapped potential in weblogs contributing to their business goals.

As a socio-technical object of investigation weblogs frame a broad area for interdisciplinary research. They continuously became a new form of 'mainstream personal communication' (Rosenbloom, 2004) for millions of people publishing and exchanging knowledge, thereby connecting to like minded people, establishing networks of relationships. Weblogs seem ideal for experts broadcasting their expertise to a large audience, but they are also suited for 'ordinary' people who want to share

stories with a small group (Wagner and Bolloju, 2005). Exploring the motivation of bloggers on the web, Nardi et al (2004) found that blogging is an unusually versatile medium, used for everything from spontaneously releasing emotion to supporting collaboration and community. However, there is also evidence that bloggers value sharing of their presented thoughts without getting the intensive feedback associated to other forms of communication (Nardi et al, 2004). Gumbrecht (2004) and Herring et al (2002) characterized blogs as a medium having limited interactivity, compared to e.g. listserv. Herring et al (2002) even found the modal number of comments in individual blogs to be zero, indicating the low level of interaction within the majority of weblogs.

In a corporate context, weblogs enjoy popularity in the form of organizational blogs. Those are (1) maintained by people who post in an official or semi-official capacity at an organization, (2) endorsed explicitly or implicitly by that organization, and (3) posted by a person perceived by the audience to be clearly affiliated with the organization (Kelleher and Miller, 2006). Employees are increasingly diffusing information about their experiences and progress at work to the public (Efimova, 2004). From a corporate view, utilization of weblogs has even been heralded a paradigm shift in the way, companies are interacting with their customers. They provide the ability of restoring a human face to a company's self-presentation with respect to information technology extending the customer relationship (Dwyer, 2007). Aiming towards a categorization of corporate weblogs, Zerfaß (2005) created a taxonomy describing fields of applications and upcoming challenges for weblogs.

In an Enterprise 2.0 movement (McAfee, 2006), companies started to adopt Wikis and weblogs, supporting knowledge transfer between their various actors and aiming in facilitation and improvement of their employees' knowledge work. Both tools entail the potential of making the practices of knowledge work and their output more visible and graspable. Corporate weblogs may contribute to codification and personalization of organizational knowledge (Kaiser and Müller-Seitz, 2005). While examining internal weblogs in project management within Microsoft, Grudin (2006) was referring to the request of further empirical studies on the topic of internal corporate weblogs.

With reference to Puntschart and Tochtermann (2004), knowledge transfer is the uni-directional targeted transfer of knowledge from a sender to a recipient. Knowledge sharing is an extension to knowledge transfer, where knowledge flows in both directions, from the sender to the recipient and vice versa.

After a brief literature review with explicit focus on internal weblogs within large-scale enterprises, we will address the need for empirical inquiries concerning the adoption of weblogs within small and medium enterprises (SMEs), which constitute the majority of all enterprises worldwide. Our presented findings are based on an exploratory case study conducted in an Austria SME settled in the ICT industry and employing 50 knowledge workers. We comprehensively analyze structure and properties of this internal weblog and explicitly probe its impact on knowledge management. Finally, we conclude with a summary and a discussion on the limitation of our research.

2 Related Work

Compared to the number of scientific publications on the topic of weblogs in total, those focusing on internal weblogs in corporate settings are scarce. A significant reason may be grounded in the fact that it is more challenging for researchers to investigate a weblog within a corporate context: Because of the access to critical business information published in the weblog, a close relationship of the researcher towards the enterprise is an inevitable precondition.

The four reviewed publications focus on a single case within a big multinational enterprise having a large set of weblogs. Such a weblog network already owns structures and properties similar to the Blogosphere, a collective term for the population of weblogs on the Web (Shi et al, 2007). Solely through examining electronic traces created by weblog users, interesting findings about weblogs can be gained.

To learn more about structures and properties of internal weblogs within organizations, Kolari et al (2007) investigated the internal Blogosphere of IBM. The weblog network was visualized as a social graph based on electronic traces, where bloggers and commentators constituted the nodes while the edges symbolized the relationships between them in terms of comments and trackbacks. The authors claimed to be the first to comprehensively characterize a social network expressed by weblogs within an enterprise. They presented new techniques to model the impact of a weblog post based on its range within an organizational hierarchy using mathematical operations but leaving an empirical inquiry open.

Jackson et al (2007) explored the social aspects of blogging within an unstated large-scale enterprise using empirical methods of research. They analyzed both motivation of blogging individuals and their practices of using weblogs. Pivotal for their analysis was the observed phenomenon that busy bloggers published almost twice as much comments within weblogs they visited than posts in their own. The authors brought to light that weblogs are able to strengthen the weak ties between bloggers. Furthermore weblogs enabled an informal mechanism to encourage disparate and widespread departments to go for a constructive contact. Weblogs provided good means for employees to establish and maintain personal networks. Busy bloggers did not only create value for themselves but also for the medium weblog users.

The growing several thousand both internal and external weblogs covering network of weblogs at Microsoft was investigated by Efimova and Grudin (2007). They probed where, how and why employees blogged, how personal the writing was in work related blogs and what happened when blogging became a formal work objective. While Microsoft valued external customer-oriented weblogs, a lot of skepticism existed towards internal weblogs to which no clear business purpose could be attributed. Contrariwise to external weblogs, internal ones were not formally supported by the company. Employees were free to determine whether, when and for what reason they blogged. A lot of bloggers described blogging as a way of sharing passion for their work and communicating directly with others inside and outside the company. Many described blogging as a desire to reveal the human side of a company, while others used weblogs purely for documentation and organization purposes.

Kosonen et al (2007) discussed roles and challenges of weblogs in internal communication in a large-scale ICT enterprise. They identified a two-dimensional framework based on the type of internal blogs and the related modes of communication. Blogs are employed in internal communication to fulfill strategy implementation goals and to foster informal interactions. Corporate climate and corporate culture determine the success of weblog adoption. Finding a balance between formal guidance and self-efficacy seems to be inevitable. Blogs offer an effective means for sharing knowledge in organizations in an informal manner.

3 Research framework

The goal of our research was to probe an internal manager weblog evolving in an Austrian ICT SME employing 50 knowledge workers. The European Union provides a recommendation for classifying SMEs: SMEs are enterprises which employ less than 250 persons and have a maximum annual turnover of 50 million EUR or 43 million EUR balance sheet total. Due to the different basic conditions in SMEs compared to those in large scale enterprises, we also assume different properties and structures of internal corporate weblogs. Our research was motivated by the lack of qualitative studies of weblogs in the context of SMEs. Taken into account that SMEs comprise the majority of all enterprises worldwide, we accentuate the relevance of our study.

We chose case study research as our preferred research technique, because the researched phenomenon, the weblog, can not be separated from its context, supporting knowledge transfer. According to Yin (1984) 'a case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident'. According to the principle 'use multiple sources of evidence' (Yin, 1984) different sources of information had been taken into account allowing us to address a broader range of historical, attitudinal and behavioral issues. Any findings such a case study generates are likely to be more convincing and accurate. Following Patton's recommendations (Patton, 1990), we chose an information-rich case providing many opportunities for learning.

We started investigating the weblog with respect to its property to facilitate the knowledge transfer between manager and employees. A comparison between content of e-mails sent by the manager to all employees and the weblog content is included. Furthermore, we had the chance to interview the manager talking about his intentions. We even received a certain amount of control over the weblog, shutting down the weblog for a short period of time. Finally we carried out a survey obtaining another set of findings. Using multiple sources of evidence enabled us to derive hypotheses with more accuracy and conviction in contrast to using just a single source of data.

Bearing in mind the extensive desktop research executed before, we were able to derive the following research scope covering techniques, previous findings and impact of weblogs on knowledge transfer:

- We showed why a weblog was used in this particular organization and how it affected knowledge transfer. Furthermore we addressed the question of weblog success in terms of popularity and how to raise it.
- From a researchers' perspective, we probed, whether present techniques from internal weblog research are applicable to weblog research in the context of SMEs.
- Researching weblogs in business settings is still lacking a strong theoretical body. Hence, the overall goal of this exploratory case-study was to develop some body of theory describing the adoption of weblogs in SMEs, which we will postulate in the form of hypotheses to be tested in further studies.

4 Conducting the exploratory case-study

4.1 Exploring the artifact

We began our exploration by investigating the weblog's history of creation: During a critical project meeting, the manager was reporting to all employees hourly, thereby adopting a very personal writing style. After the meeting was finished he expressed the desire to obtain a weblog for future coverage of relevant information.

An instance of Wordpress (www.wordpress.org) (licensed under the GNU General Public License) had been installed on the Web server of the company. Wordpress provides many features, but most of them remained unused within this case: A blogroll including other weblogs or web-sites which are regularly visited by the author was missing. The manager did neither insert hyperlinks to point to interesting internal or external resources, nor post multimedia-enriched content. Communicating confidential information, this weblog was accessible from the intranet only.

We explored the weblog content from both a qualitative perspective (i.e. what did the manager communicate to employees) and a quantitative perspective (i.e. how often did the manager inform the employees). From a quantitative perspective we measured operational figures in terms of number and frequency of posts and comments. Besides communicating via the weblog, the manager used e-mail as a supplemental channel. In the case of the investigated weblog, the reader group could be limited to the crowd i.e. 'ordinary employees', while the management was managed closely personally.

The manager mainly used the weblog to depict tasks accomplished on behalf of the represented organization. Thereby he adopted a subjective informal writing style (Kelleher and Miller, 2006). The communicated information was of both strategic nature, e.g. including knowledge about contracts, challenges, partner-acquisition or presentation of decisions from strategic meetings, and operative nature, e.g. including reports from business trips and stories about the participation at various events. While information being of relevance for all employees was diffused via the weblog, time-critical information being of particular interest to a limited group of employees was transported by personal talks, telephone calls or e-mails. Time-critical information relevant for everybody was still communicated via internal e-mails to assure the information transported reaches all receivers in time.

Table 1. Quantitative analysis of the weblog

month	number posts	number comments	min gap between posts (in days)	max gap between posts (in days)	avg gap between posts (in days)
May	8		0	5	1,1
June	5	1	2	14	5,6
July	9		0	7	3,7
August	3		2	21	10,3
September	2		8	18	13,0
October	1		19	19	19
November	2		5	24	15
	number posts (in total)		avg. total gap		
	30		5,8		

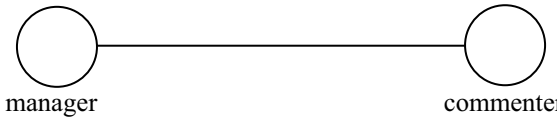
From studying the electronic traces we detected (1) a strong decrease of published posts and (2) a rise in the average gaps of posts. Furthermore we observed the phenomenon of only one comment being posted. We will give an explanation in the following sections, after extending the research scope.

4.2 Extending the research scope

The analysis of internal weblogs in large-scale corporate settings can be based upon extensive network data that is electronic traces of e.g. relations between a large set of internal corporate weblogs constituted by comments, trackbacks and blogrolls. Unfortunately, techniques that can be successfully applied in large enterprises (Kolari et al, 2007), including network theory and social network analysis based on electronic traces, can not be applied in the same way in SMEs. In the context of SMEs, there is often only a single or a small set of weblogs involved, which renders typical research measures of network approaches (Newman, 2003) such as degree or centrality of weblog networks impractical or even meaningless. Instead, it becomes more interesting how a weblog interferes and interfaces with nodes (actors) that are offline - such as the different stakeholders in an organization communicating with the weblog author. Our situation required extending the scope of analyzing purely electronic traces as done in many studies of weblogs in large scale enterprises or in the Blogosphere to including offline traces of actors, reading or interacting without authoring a weblog themselves.

In this paper we argue that especially for small and medium enterprises - though we expect the same argument to hold for large enterprises as well - traditional means of social network analysis are insufficient, due to the exclusive focus on electronic traces. Analyzing weblogs in SMEs requires methods that include the offline context. There may not be enough electronic traces to accurately understand the structure and properties of weblogs and how they may be embedded into SMEs. Therefore, phenomena which are investigated purely on the basis of electronic traces might turn out to be obvious, biased or simply wrong. Our investigated case evolved just one internal weblog.

Fig. 1. Social graph of the manager weblog



This social graph depicts the 'internal Blogosphere' as a very simple construct. We expected commenting practices to play an important indicator for the success of a weblog in terms of popularity. By considering only one posted comment, we first argued for a very low interest of the particular weblog within its possible audience. However we wanted to learn more about the respective weblog and therefore extended our investigation to the offline actors.

4.3 Conducting an experiment

Contrary to the approach from Kolari (Kolari et al, 2007) and our criticism expressed in the prior section we emphasized that it is very gainful to experience the impact of the weblog on nodes (actors) which are offline, not owning weblogs by themselves. We asked ourselves the subsequent questions:

- How did different actors perceive this weblog in the context of knowledge transfer?
- What were the benefits for employees reading this weblog? Did employees ignore this weblog as a source of information?
- What was the rationale of just one comment being published during the time of investigation?

We setup an experiment: First we deactivated the weblog exactly seven days after a post was created. By sending an e-mail to every of the 50 employees, we asked whether they had read the recent post and were able to recall the content. Our request was repeated once to receive a higher rate of return.

14 employees in total (28%) replied to our request. 11 employees (22%) were able to basically recite the content of the past weblog post. One employee expressed that he did not read the post. Two more employees provided us with an explanation of their rationale being a nonreader. They typically read weblogs within web-based feed readers, but the respective RSS feed could not be subscribed to in this way. Therefore they simply denied reading the weblog. This fact clearly depicted a goal conflict between manager and employees. Referring to Strohmaier et al (2007), we assumed further goal conflicts to be a reason for weakening the intended knowledge transfer.

Analyzing the findings gained by our experiment, we were able to derive the following hypotheses:

- Few comments in SMEs' weblogs do not necessarily equate few readers.
- Specific weblog configurations are able to counteract personal weblog practices, reducing the ability of the particular weblog to facilitate knowledge transfer.
- Studies of weblogs purely based on electronic traces may lead to biased or wrong findings. Having just a single or a small set of weblogs, it is more interesting to examine the impact of the weblog on offline nodes (actors).

4.4 Conducting a survey

Our first findings dealing with the actual reading behavior accentuated the need for a more detailed survey. The goal of this survey was to increase the accuracy of our findings regarding motivation of weblog readers and nonreaders. Additionally we intended to probe to what extent the goal of the manager – using the weblog to facilitate knowledge transfer towards the employees – was achieved.

All employees who were able to remember the last weblog post during our experiment were requested via e-mail to answer six questions concerning their weblog reading practices. This respective crowd formed group A – weblog readers. All employees refusing to reply in the experiment were surveyed using a different questionnaire including four questions. We probed their rationale of not reading the weblog, especially referring to conditions under which they would change their mind. Because we were not able to eliminate the possibility of also addressing readers, we also attached the questionnaire for group A to that e-mail. All non readers were finally added to group B.

Receiving 40 replies (80%) of 50 possible represented a very satisfactory response rate. Altogether 20 replies were written by members of group A (readers), and another 20 by those of group B (nonreaders).

Following, questions stated to and answers given by group A will be presented. The aim of questions 1-3 was to examine the motivation of employees reading the weblog. From an organizational perspective, further relevance is paid to what extent the manager's goal of informing the employees (a) had been achieved and (b) was in fact achievable by selecting a weblog as a knowledge transfer facilitator.

Q1: I read the weblog, because… Almost all replying employees clearly stated their interest in the tasks their manager was carrying out. One third also stated a general interest in what was happening within and in the periphery of their organization.

Q2: How and from where do you read the weblog? Ten employees used an ordinary Web browser, explicitly mentioning Internet Explorer and Mozilla Firefox. Six employees used a RSS-feed-reader, while two employees went for an RSS plugin for Microsoft Outlook. 16 employees read the weblog solely within the office. Three employees explicitly addressed the access restriction, which we were also pointed at in our experiment.

Q3: How often do you read the weblog? Half of the employees browsed the weblog for newly created posts at least once a week, while five employees visited the weblog more infrequently in broader intervals. From these findings we assumed reading this particular weblog is more like a scan for newly created posts. Only a minority group subscribed to the RSS feed, being notified instantly after a post was published.

The following question was aimed at exploring the reason of only one comment being posted during the time of investigation.

Q4: From your point of view, is commenting to a weblog post reasonable? Eight employees positively answered this question and quoted to mention different points of view to the author including additional information and aspects which had not been taken into consideration yet. Six employees clearly answered with 'no': The weblog

was purely perceived as an information portal, not a place for knowledge sharing. The remaining employees argued that reasons both for and against comments exist. We found this question to be stated in an ambiguous way, therefore failing to deliver an answer according to our intention exploring the rationale of non-commenting within *this* particular weblog. With respect to Nonnecke and Preece (2001) the observed behavior can be termed with 'lurking', when only a marginal fraction of members in virtual communities actively posts content. Cabrera and Cabrera (2002) provide a socio-economical explanation, investigating the employees' rationale of denying the sharing of ideas within an organizational context.

Approximately half of the employees were reading the weblog. The goal of the next question was to probe barriers when adopting internal weblogs in the context of SMEs.

Q5: To what extent is the manager able to improve the weblog from a technical, an organizational, and a content perspective? The most substantial argument given by the employees dealt with the perceived low frequency of posts. Nine employees explicitly requested a higher number of posts and three employees accentuated a call for a higher frequency of comments, too. By achieving the second, more employees would be encouraged to add comments on their own facilitating knowledge sharing. Two employees argued for making the weblog available from places outside the office. Merely three employees wanted the weblog to remain unaffected.

The substantial goal of the manager was to improve knowledge transfer towards the employees. The closing question for group A addressed whether the weblog had contributed to achieve that goal.

Q6: Has the knowledge transfer from manager to employees been improved by the weblog compared to prior (yes, rather yes, rather no, no)? Nine employees answered 'yes', seven employees 'rather yes'. The weblog constituted a new information channel towards employees, and the information communicated was of sufficient relevance to read the weblog. Three employees stated 'rather no' reasoning with the low frequency of posts, while one employee answered 'no'.

Subsequent, the results of the surveyed group B are displayed. Questions 1-2 dealt with the rationale of employees not reading the weblog.

Q1: I do not read the weblog because… The majority consisting of ten employees denied reading because they simply forgot either existence or URL of the weblog. Since its introduction as a new information portal, only one e-mail had been written by the author. Three employees criticized the weblogs's lacking ability to be read via web-based feed readers. Two employees did not read weblogs at all and one employee argued a lack of time for reading activities beside the work tasks.

Q2: I will read the weblog if... Four employees indicated to read the weblog if they received an e-mail notification for every new post created. They favored solutions based on push-mechanisms over those based on pull-mechanisms. With respect to the literature, McAfee (2006) also described knowledge workers preferring channels over portals. Three employees stated to read the weblog, if it was accessible from the web allowing subscription with web-based feed readers. Five employees did not see any relevance in the published content with respect to their personal work tasks. Two employees used different channels to obtain requested information and the weblog did not provide any new insights to them. Due to the fact that the author of the weblog

conducted little promotion, new employees did not know of its existence. However, three employees were not able to provide a rationale for their non-reading behavior and promised to read the weblog in the future.

Questions 3-3.1 addressed whether a weblog is perceived as an instrument for knowledge transfer by the nonreaders at all. Besides that we wanted to examine the preferred knowledge management activities from an employee's perspective.

Q3: From your point of view, which particular activities are able to improve the knowledge transfer from manager to employees? Prior to this survey, we assumed that nonreaders do not perceive the weblog as an instrument for knowledge transfer. Surprising to us, eight employees in fact did perceive the weblog as an instrument to facilitate knowledge transfer. Besides that, e-mail, newsletter, meetings and personal talks were named. Six employees placed importance on personal meetings between manager and employees.

Q3.1: Do weblogs account for that? 14 employees acknowledged weblogs as facilitators of knowledge transfer, explicitly naming asynchrony, ease of transporting information, little effort for operation and the informal narrative style as essential criteria. Five employees answered 'no', reasoning with the huge effort of retrieving relevant information. Notifications of new posts were not provided either. In addition, in this particular SME informal information channels were available in a manageable number, easily accessible by anybody rendering information communicated via the weblog unnecessary. Furthermore, information relevant for daily work assignments was not published.

Analyzing the findings, we derived the following hypotheses for validation in further studies:

- Weblogs will be read, if they provide sufficiently interesting content that is not available from alternative sources.
- The frequency of posts illustrates a key factor for weblog success in terms of popularity. A low frequency constitutes a barrier to accept the weblog as a knowledge transfer instrument.
- Commenting to weblog posts may lead to a change of the knowledge workers' perception of the weblog as a pure information portal, hence facilitating knowledge sharing.
- Access restrictions regarding tools and/or location will conflict with weblog reading practices.
- Lacking skills and personal accent for an ineffective utilization of the weblog in terms of knowledge transfer, e.g when employees demand notification features that are available but unknown to them.
- Employees will have limited desire to read the weblog if they perceive the relevance of the published content too low with respect to their daily work assignments.
- Weblogs have to be promoted by the authors to turn them into facilitators of knowledge transfer.
- Internal weblogs in SMEs are able to improve knowledge transfer in principle, however, as long as only one weblog with no comments exists, they seem inappropriate for knowledge sharing.

5 Limitation of research and future work

The motivation for our single-case study was based on the fact that known preliminary academic case-studies focused on large-scale enterprises, but most of the enterprises worldwide are made up of SMEs. We intended to advance weblog research to the SMEs context referring to their population.

However, one limitation of the findings generated by our study is noteworthy: First of all, data for deriving our hypotheses was generated by only one weblog in one SME. Single-case studies provide limited feasibility of generalizing to theory. However, unlike surveys, case studies do not make inferences about a population (or universe) on the basis of empirical data collected about a sample (Yin, 1984). In contrast to methods based on statistical generalization, case studies do not reason the selected cases as being sampling units. Individual cases are to be selected as a laboratory investigator selects the topic of a new experiment (Yin, 1984). A single case study can therefore be considered like a simple experiment. Hence findings from a single case-study can be reasoned like findings from a single experiment. If we had conducted a multiple-case study, the developed theory would have a stronger basis, allowing replication of findings. Keeping that in mind, we will test the hypotheses derived within further case studies to investigate whether replication may be achieved.

6 Conclusion

This exploratory case study aimed at generating findings about internal Weblogs in SMEs from a knowledge management perspective. We state out findings as hypotheses to be validated within further case-studies.

In conclusion, we outline our contribution to organizational weblog research in a nutshell: Weblogs are no fast-selling items. Promotion constitutes an important precondition for weblog success. Techniques from weblog research which are based on electronic traces may lead to wrong findings if applied in the context of SMEs that have only a single or a small set of weblogs. Employees will prefer weblogs providing information that is of sufficient interest or relevance for their work assignments and not available from other channels. A high frequency of posts constitutes a key factor for weblog success in terms of popularity. However, a low number of comments does not automatically equate a low number of readers. A specific weblog configuration will establish barriers, colliding with the reading practices of employees. Constituting information portals, weblogs are based on pull mechanism. However, in an organizational context, employees may favor technologies based on push principle. If adopted effectively, weblogs provide good means to facilitate knowledge transfer, but seem inappropriate for knowledge sharing.

Acknowledgements

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

(Nardi et al, 2004) Nardi, Bonnie A., Schiano, Diane J., Gumbrecht, Michelle, Swartz, Luke; Why We Blog, in Communication of the ACM, Volume 47, Issue 12, 2004.

(Cabrera and Cabrera, 2002) Cabrera, Angel; Cabrera Elizabeth, Knowledge Sharing Dilemmas, in Organization Studies, 2002.

(Dwyer, 2007) Dwyer, Paul, Building Trust with Corporate Blogs, in Proceedings of International Conference on Weblogs and Social Media (ICWSM '07), Boulder, Colorado, 2007.

(Efimova, 2004) Efimova, Lilika, Discovering the iceberg of knowledge work: A weblog case, in: Proceedings of Organizational Knowledge, Learning and Capabilities (OKLC '04), Innsbruck, 2004.

(Efimova and Grudin, 2007) Efimova, Lilia; Grudin, Jonathan, Crossing Boundaries: A Case Study of Employee Blogging, in Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40), 2007.

(Gumbrecht, 2004) Gumbrecht, Michelle, Blogs as 'Protected Space', Presented at the Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics: WWW-Confernce 2004. New York: ACM Press, 2004.

(Grudin, 2006) Grudin, Jonathan, Enterprise Knowledge Management and Emerging Technologies, in Proceedings of the 39th Hawaii International Conference on System Science (HICSS-39), 2006.

(Herring et al, 2002) Herring, Susan; Scheidt, Anne L.; Bonus, Sabrina; Wright, Elijah, Bridging the Gap: A Genre Analysis of Weblogs, in Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37), 2002.

(Jackson et al, 2007) Jackson, Anne; Yates, JoAnne; Orlikowski, Wanda, Corporate Blogging: Building Community through persistent digital talk, in Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40), 2007.

(Kaiser and Müller-Seitz, 2005) Kaiser, Stephan; Müller-Seitz Gordon, Knowledge Management via a Novel Information Technology – The Case of Corporate weblogs, in Proceedings of International Conference on Knowledge Management (I-KNOW' 05), Graz, 2005.

(Kelleher and Miller, 2006) Kelleher, Tom; Miller, Barbara, Organizational blogs and the human voice: Relational strategies and relational outcomes, in Journal of Computer-Mediated Communication, 11 (2), http://jcmc.indiana.edu/vol11 /issue2/kelleher.html, 2006, accessed on 28 01 2008.

(Kolari et al, 2007) Kolari, Pranam; Finin, Tim; Lyons, Kelly; Yesha, Yelena; Yesha; Yaacov; Perelgut, Stephen; Hawking, Jen, On the Structure, Properties and Utility of Internal Corporate weblogs, in Proceedings of International Conference on Weblogs and Social Media (ICWSM '07), Boulder, Colorado, 2007.

(Kosonen et al, 2007) Kosonen, Miia; Hentonen, Kaisa; Ellonen, Hanna-Kaisa, Weblogs and internal communication in a corporate environment: a case from the ICT industry, in International Journal of Knowledge and Learning, Vol. 3, No. 4/5, 2007.
(McAfee, 2006) McAfee, Andrew, Enterprise 2.0: The Dawn of Emergent Collaboration, MIT Sloan Management Review, Spring 2006, Vol.47 No. 3, 2006.
(Newman 2003) Newman, M.E.J, The Structure and Function of Complex Networks, in SIAM Review, 2003.
(Nonnecke and Preece, 2001) Nonnecke, Blair; Preece, Jenny, Why Lurkers Lurk, in Proceedings of American Conference on Information Systems (AMCIS '01), Boston, 2001.
(O'Reilly, 2004) O'Reilly, Tim, What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software, www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, 2005, accessed on 28 01 2008.
(Patton, 1990) Patton, Michael, Qualitative Evaluation and Research Methods, 2nd edition, Newbury Park, Sage Publications, 1990.
(Paquet, 2003) Paquet, Sebastian, Personal knowledge publishing and its uses in research, http://www.knowledgeboard.com/item/253, 2003, accessed on 29 01 2008.
(Puntschart and Tochtermann, 2004) Puntschart, Ines; Tochtermann, Klaus, Online Communities and the 'un'-importance of e-Moderators, in Proceedings of Fifth International Conference on Networked Learning (NLC '06), Lancaster, 2006.
(Rosenbloom, 2004) Rosenbloom, Andrew, The Blogosphere, in Communications of the ACM 17.12, 2004.
(Shi et al, 2007) Shi, Xiaolin; Tseng, Belle; Adamic, Lada, Looking at the Blogosphere Topology through different Lenses, in Proceedings of International Conference on Weblogs and Social Media (ICWSM '07), Boulder, Colorado, 2007.
(Stocker et al, 2007) Stocker, Alexander; Us Saaed, Anwar; Dösinger, Gisela; Wagner, Claudia, The Three Pillars of Corporate Web 2.0: A Model for Definition, in Proceedings of International Conference on New Media Technologies (I-MEDIA '07), Graz, 2007.
(Strohmaier et al, 2007) Strohmaier, Markus; Yu, Eric; Horkoff, Jennifer; Aranda, Jorge; Easterbrook, Steve, Analyzing Knowledge Transfer Effectiveness – An Agent Oriented Approach, in Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40), 2007.
(Wagner and Bolloju, 2005) Wagner, Christian; Bolloju, Narasimha, Supporting Knowledge Management in Organizations with Conversational Technologies: Discussion Forums, Weblogs, and Wikis, in Journal of Database Management 16, No. 2, 2005.
(Zerfaß, 2005) Zerfass, Ansgar, Corporate Blogs: Einsatzmöglichkeiten und Herausforderungen, http://www.zerfass.de/Corporateblogs-AZ-270105.pdf, 2005, accessed 29 01 2008.
(Yin 1984) Yin, Robert, Case Study Research. Design and Methods, Sage Publications, 1984.

Transforming Exchange-based Job Boards into Lasting Career Communities

Elfi Ettinger[1], Celeste Wilderom[1], and Rolf Van Dick[2]

[1]Department of Information Systems & Change Management,
University of Twente, the Netherlands
{e.ettinger, c.p.m.wilderom}@utwente.nl
[2]Department of Social Psychology, University of Frankfurt, Germany
{van.dick@psych.uni-frankfurt.de}

Abstract. In this paper, we use a qualitative approach to explore which design aspects an e-recruiting platform requires so as to achieve active long-term participation of its users. Our study is based on a case study of an e-recruiting platform for Austrian engineers. We use interviews and ethnographic methods. Although, e-recruiting trends point towards niche e-recruiting and career community approaches, our results show that a mere niche approach is not enough to maintain an actively participating user base. Our findings suggest that users are more inclined to re-use the same e-recruiting platform throughout their career if the userbase is comprised of many other users who share a similar social identity and who had already developed offline ties with each other before registration. Hence, integrated online community and social network applications for specified user segments will enable users to maintain and transform their existing offline contacts for career purposes. The paper concludes with recommendations for e-recruiting research and practice.

Keywords: e-Recruiting, Career Communities, User Participation, Social Identity.

1 Introduction

Despite the vibrancy of e-recruiting services, large numbers of them fail (Feldman and Klaas, 2002; Lin and Stasinskaya, 2002; Bishop, 2006). One challenge e-recruiting services are facing is keeping the applicants' profiles up-to-date. This is especially important if recruiters can search in the applicants' pools. E-recruiting services also need to show their success through indicators such as high numbers of applicants visiting their sites and clicking on ads, high page impressions, the accuracy of the matching between job ads and resumes, and by their ability to quickly suggest appropriate candidates to recruiters (Smith and Rupp, 2004; Zhao et al., 2007). The e-recruiting market is still dominated by traditional job boards such as monster.com or hotjobs.com in most countries. Although, e-recruiting trends suggest that the future of e-recruiting belongs to niche recruiting portals (see for instance: beyond.com; topjobsites.com) and career networks (linkedIn.com, xing.com), it is difficult for e-recruiting companies to design socio-technical features that generate ongoing

participation from a larger fraction of its initially perhaps instrumentally oriented users (Szmigin et al., 2005; Bishop, 2007).

In this paper, we study what design aspects an e-recruiting platform would need to meet in order to achieve active long-term participation of its registered users. To guide us in answering this question, we connect the theoretical streams of e-recruiting and community identity. Based on a case study with in-depth interviews with users of a niche e-recruiting platform for engineers, we present the conditions required for users to eagerly participate in online career services. Interestingly, our findings indicate that it is not enough to simply create niche portals for specific applicants (such as engineers, lawyers) or branch segments (such as chemistry, pharmacy). What seems to have an influence on users' active long-term participation in niche portals is rather the opportunity to maintain communication with other professionals (in this study's case: engineers) who already know each other from their offline network or via friends of friends (FoF). Based on this paper's findings as well as our experiences in previous studies on e-recruiting (e.g., Khapova and Wilderom, 2006; Khapova, et al., 2007), we conclude with recommendations for e-recruiting research and practice.

2 Theory

The usefulness of networked online services has already been discussed in reference to e-recruiting. Butler (2001), for instance, describes that online career services may be seen as virtual social communities rather than only as instrumental career-move services. In this regard, Khapova (2006) argues that studies of online career services need to include the design principles of a traditional community as well as the incorporation of social network research, so as to understand the various ways people make use of online career services. Innovative online career services require more input on user participation (von Hippel, 2007; Khapova, 2006). Sustaining online services depend on people visiting the sites, participating in social interactions, and most importantly, enhancing the loyalty of users (Kim et al., 2004; Bishop, 2007). Although networks and communities are conceptualized and studied in many diverse ways (Knoke and Kuklinski, 1982; Castells, 2000; Van Dijk, 2005), many researchers agree that networked communities are defined on the basis of shared identity, interests, and commonality among their members (Turkle, 1995; Preece, 2000; Castells, 2004; Plickert et al., 2007). Parker et al. (2000) define career communities as self-organizing, member-defined social structures through which people draw career support. The career scholar Michael Arthur (Arthur et al., 1995) notes that "intelligent careers" in the knowledge economy would need to be reflected in such communities.

Some researchers suggest that people use communities by merely adding internet contact to existing telephone and face-to-face contact, or by shifting their means of communication to the internet[1] (Wellman et al., 2002). Internet users have been found to join online communities for more than efficiency reasons. Ridings and Gefen

[1] Examples are facebook (www.facebook.com), myspace (www.myspace.com) or friendster (www.friendster.com).

(2004) identified friendship, social support, information exchange, and recreational as reasons for participation in online communities. According to social presence theorists (Biocca et al., 2003), the presence of other members (which can be complemented by offline interactions) may foster the ties of users to a specific online service. Hence, determinants of long-term sustainability of online career services may need to range from understanding how users judge online features, such as the quality of a career site's service, its system, and the provided information (DeLone and Mclean, 2003) to understanding offline features, such as the offline activities of users (Kim, 2000). Offline activities have been found to increase the solidarity and cohesiveness of virtual communities and strengthen the ties between members (Wellman and Haythornthwaite, 2002). A better understanding of the match between what is being offered (the supply) and (potential) users' interests will promote a stronger desire to participate and interact with other members, leading to shared feelings of belonging, responsibility, and commitment to the community (Kim, 2000). Academic studies investigating user needs in regard to offline and online activities of users' within niche e-recruiting services are rare. Thus, this paper may help in exploring users' view on effective design aspects of such online portals for the purpose of achieving long-term participation of its users.

3 Method

We conducted a qualitative case study of an e-recruiting platform for engineers. Since its establishment in 2005, the company has developed many partnerships with schools and companies across Austria and has obtained research grants for developing Web 2.0 technologies for e-recruiting services. The company actively collaborates with its users and customers in an effort to capitalize and distribute knowledge for system design improvements (e.g., von Hippel, 2007).

Given the exploratory nature of our study, we adopted a case-study design (Yin, 2003) with research methods that combine ethnography and in-depth interviewing. A case study is suitable when researching a contemporary phenomenon within its real-life context with multiple sources of evidence. Ethnography is a method that is frequently used to study the culture of users sharing a similar social identity (Boyd and Ellison, 2007). Hence, in order to better understand the users' needs in regard to such an online career network, the first author intensely emerged herself into the engineers' world which included numerous informal conversations with registered users and potential future users.

3.1 Sample and Data Collection

We randomly selected one registered user from each Higher Technical School (HTL) in Austria. He or she had a minimum of 3 years of work experience, and was interviewed by telephone. We also held numerous informal conversations with engineering students and teachers, especially during our visits to 7 career fairs throughout the year 2007. Also, agendas, copies of presentations and minutes from several staff meetings were compiled from the company developing the e-recruiting

platform. Email correspondence between registered users, system designers and service personnel was collected over a period of 6 months in 2007. We supervised 2 graduate students' projects related to this topic and we conducted a workshop with the system designers for the purpose of discussing how they create and utilize user-generated knowledge (Nonaka, 2007; von Hippel, 2007). The conversations with users were aimed at exploring collectively desired online design aspects as well as collecting their ideas for making use of online interactions with their fellow classmates to supplement their offline interactions. Each interview started by exploring how they found out about the e-recruiting platform, their evaluation of the webpage, and the appropriateness of questions asked in the resume forms, as well as their general online and offline job search behavior. A second set of questions were aimed at identifying users' shared needs and interests for active and long-term usage of the platform. While tracking, observing, and asking questions, we kept a record of field notes that enhanced the quality of later in-depth analysis. We also paid attention to Chatman's (1984) advice of establishing rapport with our informants so that they were more open and felt comfortable during our interactions.

3.2 Data Analysis

We first listened to all audio tapes of the 60 interviews and read all collected written documents. Then, we compiled narratives of the interviews and compared them with the content of the field notes, meeting minutes, email communications and presentation slides. Next, we identified broad themes in the data and reduced them to more precise categories (Miles and Huberman, 1994; Yin, 2003). We coded the collected data according to Riding's and Gefen's (2004) identified typology of reasons for joining online communities: information exchange, social support, friendship, and recreation. This first deductive analysis seemed functional as its four broad categories for joining online communities are based on reasons collected among a large number of different online communities representing many different segments. By systematically comparing the data (Strauss and Corbin, 1988), we noted patterns.

4 Results

Our findings suggest that users intend to use e-recruiting services on a continuing basis throughout their career if those services are complemented by community and social network applications aimed at specifically connecting users who share a similar social identity. We noted that engineers were very open in interacting with their fellow engineers: online as well as offline. These insights offer a fairly new network opportunity for transforming classical "job boards" to sustainable "career communities." Engineers identified a wide range of ideas that they want to share and exchange online with other fellow engineers. They predominantly intend to communicate online with offline known-fellow acquaintances from their schools or via extended networks (friends of friends). Interestingly, the interviewed engineers didn't seem to be keen on developing or maintaining a strong network with fully

unknown registered engineers. We found that most users found their jobs via personal relationships (such as knowing someone already employed by a certain company).

Engineers' impressions of the webpage and resume forms were largely positively evaluated and regarded as meeting users' needs. Some engineers had minor concerns with the length of resume forms or the support of uploading bigger file sizes. It appeared very important that the career site be clearly structured into sections for applicants and companies. Further, simplicity, easy navigation, quick loading of the pages, and perceived usefulness of the applications are important factors for re-using the service (see also David, 1989; DeLone and McLean, 2003). The possibility to adjust their own career profile to status of active or a passive job-seeker has been supported by many users. Importantly, system designers are challenged to create private (for friends) and public (for HR recruiters) spaces of the users' applicant profile so that trust is built by warding off the fear that personalized resume data might be abused.

As one engineer described: *"How can you make sure that my boss will not find my profile in the database?"* and *"Sure, I want my profile for friends to look different than my applicant profile."*

Fifty-six engineers reported that they would use an online career service in the long run if specifically targeted at engineers' needs. However, most of the interviewed users are not inclined to sign up at a general online job board that attracts many different job seekers. Jobs boards are seen as exchange-based career tools for finding a job when needed, but among them it is not desirable to connect online in such job boards with unknown users or to maintain active long-term membership in such portals. It also appeared that engineers prefer to fill out resume forms of corporate recruiting pages to general online job boards. Corporate pages are perceived as being more trustworthy than general job boards.

Integrated social network and community features within career services have been frequently found to make it easier to keep in touch with and follow their fellow classmates' career paths.

As one engineer noted: "*When I want to apply for a job in a different location in Austria, then it would be great to have some online search option to find out if any of my schools' graduates already work in the same area or company."*

Another engineer describes: *"It would be neat to see who of my friends have friends who work at BMW; you get a much more realistic picture of the company when getting advice from fellow-colleagues than trying to find out everything yourself.*"

It appears that much social interaction in such portals may occur among pre-existing social ties. School or university ties seem to offer the foundation for continued online interaction in the engineering e-recruiting portal. Moreover, when people discover that they have similar problems, opinions and experiences, they may feel closer and have more trust (Preece et al., 2004). Consequently, if users have more trust, they are more likely to share sensitive personalized information that e-recruiting

services require from their users to enhance applicant search quality. It has been found that identifying users' shared interests for collective usage and sense of belonging is important for enhancing users' participation (Preece, 2001). We classified engineers shared interests into Riding's and Gefen's (2004) typology of reasons for participating in online communities. Besides information exchange (career info page on how to find jobs, how to prepare application documents, interview tips and info on training programs, links to continuing education, links to companies, sector info page, salary calculator etc.) we found that applications that support friendship (personal page with contact info, pictures, friends, hobbies, city groups, civil service, army groups) are important so as to sustain users' participation. Also, social support (sharing reports, sharing problems with each other, sharing music, games etc.) and recreation (work-climate index in different companies, events and activities across Austria, sports info) were identified as important for long-term active participation.

5 Conclusion

This paper addresses which design aspects e-recruiting services are required so as to achieve active long-term participation of its users. We interviewed registered users in regard to their expectations, shared needs and interests as well as collectively reported requirements for long-term use of the platform. Our findings indicate that users are willing to bring their offline ties with them online into niche-based career communities. They are also willing to maintain these online ties if they have been properly connected within groups sharing a similar social identity offline (before entering the online career community). Hence, we support Boyd's findings (2008) that users of online communities are not necessarily looking to meet new people. Instead, the interviewed engineers primarily expressed interest in communicating with people who are already a part of their extended offline social network[2].

On a practical level, niche career platforms are advised to complement their traditional job posting services with social network and community applications so that users can find and connect each other. Our results suggest that e-recruiting portals require not only useful information sections on careers and continuing education, but should also encourage friendship and social activities of its users. The future is likely to belong to those providers that best understand their users' shared social identity and succeed in providing semantic technologies so as to enhance the users' online experiences. Finally, niche-providers are well-advised to stay in close touch with the potentially shifting needs of their active and most innovative (lead) users (von Hippel, 2007).

[2] Also, Ellison et al. (2007) suggest that www.facebook.com is used to maintain existing offline relationships or solidify offline connections, as opposed to meeting new people.
Similarly, according to recent research, 91% of U.S. teens who use social network sites do so to connect with offline or known friends (Lenhart and Madden, 2007).

References

Arthur, M. B., Claman, P. H., DeFillippi, R. J.: Intelligent enterprise, intelligent career. Academy of Management Executive, 9(4), 7-20 (1995)

Biocca, F., Harms, C., Burgoon, J. K.: Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. Presence, 12(5), 456-480 (2003)

Bishop, J.: Increasing participation in online communitites: A framework for human-computer interaction. Computers in Human Behavior, (23)4, 1881-1893 (2007)

boyd, d., Ellison, N. B.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), article 11 (2007)

boyd, d.: Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), Youth, Identity, and Digital Media (119-142). MIT Press, Cambridge, MA (2008)

Butler, B. S.: Membership size, communication activity, and sustainability: A resource-based model of online social structures. Information Systems Research, 12(4), 346-362 (2001)

Castells, M.: The Rise of the Network Society. Second Edition. Blackwell, Cambridge, MA, USA (2000)

Castells, M.: The Power of Identity, Blackwell, Oxford (2004)

Chatman, E. A.: Field research: Methodological themes. Library and Information Science Research, 6, 425-38 (1984)

Davis, F. D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 13(3), 319-340 (1989)

DeLone, W. H., McLean, E. R.: The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4), 9-30 (2003)

Ellison, N., Steinfield, C., Lampe, C.: The benefits of Facebook "friends": Exploring the relationship between college students' use of online social networks and social capital. Journal of Computer-Mediated Communication, 12 (3), article 1 (2007)

Feldman, D. C., & Klass, B. S.: Internet job hunting: A field study of applicant experiences with online recruiting. Human Resource Management, 41(2), 175-192 (2002)

Khapova, S.: Careers in the knowledge economy and web-based career support: New challenges and opportunities. PhD thesis, Print Partners Ipskamp B.V., Enschede (2006)

Khapova, S. N.,Wilderom, C. P. M.: Computer-based career support. In J. H. Greenhaus G. A. Callanan (Eds.), Encyclopaedia of Career Development (191-193). Sage, Thousand Oaks, CA (2006)

Khapova, S. N., Arthur, M. B., Wilderom, C. P. M., Svensson, J. S.: Professional identity as the key to career change intention. Career Development International, 12(7), 584-595 (2007)

Knoke, D., Kuklinski, J. H.: Network Analysis. Sage, Beverly Hills, CA (1982)

Lakhani, K. R., von Hippel, E. How open source software works: “free” user-to user assistance. Research Policy, 32(6), 923-943 (2003)

Lin, B., Stasinskaya V. S.: Data warehousing management issues in online recruiting. Human Systems Management, 21(1), 1-8 (2002)

Lenhart, A., Madden, M.: Teens, privacy & online social networks. Pew Internet and American Life Project Report. Retrieved Jan 31, 2008 from http://www.pewinternet.org/pdfs/PIP_Teens_Privacy_SNS_Report_Final.pdf (2007, April 18)
Nonaka, I.: The Knowledge-Creating Company. Harvard Business Review, 85(7/8), 162-171 (2007)
Parker, P.: Career communities. Doctoral thesis, University of Auckland, New Zealand (2000)
Parker, P., Arthur, M. B., Inkson, K.: Career communities: a preliminary exploration of member-defined career support structures. Journal of Organizational Behavior, 25, 489-514 (2004)
Plickert, G., Côté, R. R., Wellman, B.: It's not who you know, it's how you know them: Who exchanges what with whom? Social Networks, 29(3), 405-429 (2007)
Preece, J.: Online Communities. John Wiley & Sons Inc., New York (2000)
Preece, J.: Nonnecke B., Andrews, D. The top five reasons for lurking: Improving community experiences for everyone. Computers in Human Behavior, 20(2), 201-223 (2004)
Ridings, C., Gefen, D.: Virtual Community Attraction: Why People Hang Out Online. Journal of Computer-Mediated Communication, 10(1), article 4 (2004)
Smith, A. D., Rupp, W.T.: Managerial challenges of e-recruiting: extending the life cycle of the new economy employees, Online Information Review, 28(1), 61-74 (2004)
Strauss, A., Corbin, J. Basics of qualitative research: Techniques and procedures for developing grounded theory. (2nd Ed.) Sage, Thousand Oaks, CA (1998)
Szmigin, I., Canning, L., Reppel, A. E.: Online community: Enhancing the relationship marketing concept through customer bonding. International Journal of Service Industry Management, 16(5), 480-496 (2005)
Turkle, S.: Life on the screen: Identity in the age of the Internet. Simon & Schuster, New York (1995)
van Dijk, J.: The Network Society: Social Aspects of New Media. Sage, Thousand Oaks, CA (2005)
von Hippel, E.: Horizontal innovation networks—by and for users. Industrial and Corporate Change, 16(2), 293-315 (2007)
Wallace, J. E.: Organizational and professional commitment in professional and non-professional organizations. Administrative Science Quarterly, 40, 228-255. (1995)
Wellman, B., Haythornthwaite, C.: The Internet in everyday life. (Eds.) Blackwell, Oxford (2002)
Wellman, B., Boase, J., Chen, W.: The networked nature of community on and off the Internet. IT and Society, 1(1), 151-165 (2002)
Yin, R. K.: Case Study Research: Design and Methods. Third Edition, Sage, Thousand Oaks, CA (2003)
Zhao, D., Rosson, M. B., Purao, S.: The future of work: what does online community have to do with it?. Proceedings of the 40th Annual Hawaii International Conference on System Science, IEEE Computer Society Press (2007)

1st Workshop on Advances in Accessing Deep Web (ADW 2008)

May 7th, 2008,
Innsbruck, Austria
http://www.integror.net/adw/

Workshop Co-Chairs

Dominik Flejter, Poznan University of Economics, Poland
Tomasz Kaczmarek, Poznan University of Economics, Poland
Marek Kowalkiewicz, SAP Research Brisbane, Australia

Workshop Program Committee

Manuel Alvarez, University of A Coruna, Spain
Irene Celino, CEFRIEL - Politecnico di Milano, Italy
Terence Critchlow, Pacific Northwest National Laboratory, the USA
Emanuele Della Valle, CEFRIEL - Politecnico di Milano, Italy
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Denis Shestakov, University of Turku, Finland
Altigran Soares da Silva, Universidade Federal do Amazonas, Brazil

Determining Relevant Deep Web Sites by Query Context Identification

Zsolt T. Kardkovács[1] and Domonkos Tikk[1]

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics,
{kardkovacs,tikk}@tmit.bme.hu

Abstract. Deep web search requires a transformation between search keywords and semantically described and well-formed data structures. We approached this problem in our "In the Web of Words" (WoW) project by allowing natural language sentence queries and by a context identification method that connects the queries and deep web sites via database information. In this paper we propose a novel SQL based approach that can identify the focus of input questions if the information is represented in a database. We propose a new relational database design technique called normalized natural database (NNDB) to capture the meaning of data structures. We show that a proper NNDB is a context database, and it can serve as the basis of context identification combining the template based techniques and the world model encoded in the database.

Key words: deep web search, context identification

1 Introduction

The online accessible information organized and stored in structured databases is the content of *deep web*. The content of such databases is presented to the user as dynamic web pages being created to answer user's query, and thus standard search engines can hardly index and find them [1, 2]. Therefore, searching on the Internet today can be compared to dragging a net across the surface of the ocean. While a great deal may be caught in the net, there is still a wealth of information that is deep, and missed. Internet content is considerably more diverse and the volume certainly much larger than commonly understood [2, 3, 4].

In the WoW project, our purpose was to create a prototypical search service that could integrate surface web and deep web search, and for the latter case, could allow users to formulate their search expressions in the form of Hungarian natural language questions. The overview of the system is depicted on Figure 1. The core of the system consists of the natural language interface, the context identifier and the deep web search engine (DWSE). These components transforms simple interrogative sentences of a given language into SQL queries in several steps (see [5]). The important issue in this processing chain is the determination of the context (topic) of the question, and databases related to the topic, which

is the topic of this paper. At the end of the search the user is directed the answer pages provided by the deep web sites.

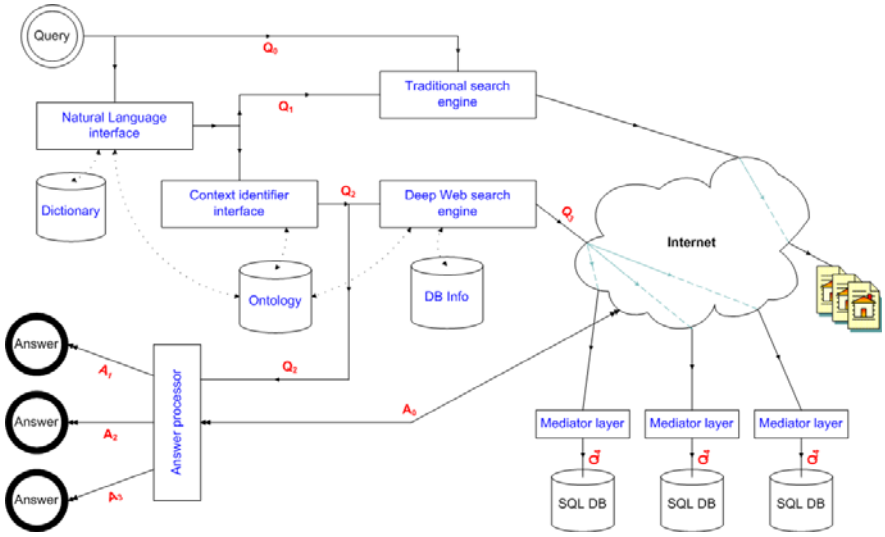


Fig. 1. The complete overview of the WoW system

The deep web search facility of WoW is restricted to the database content of contracted partners available online, i.e. no automatic exploration of deep web sites is performed. Therefore throughout the paper, we assume that schema information of deep web site are available.

The deep web search model of WoW is language independent, however its implementation for a specific application requires natural language processing (NLP) tools. In our experiments we implemented the model of WoW for Hungarian, but it can be easily adopted for other languages having the necessary NLP tools available. We have reported on this issue for Hungarian in our previous works [6, 7].

In this paper, we primarily focus on the context identification and the deep web search engine architecture, and particularly we present some important improvements made to these components of the original model in the WoW system [5]. For context identification, we propose a relational theory based mathematical foundation and we outline its implementation. Using this new approach, we can substitute the formerly used intermediate languages (ER-models, ontologies, or language Q_2 [5]) with standard SQL expressions. This change simplifies the model by omitting the transformation from the intermediate languages to SQL, which was formerly performed by the DWSE component. In order to make this change, we also exploit NLP tools (such as morphological and syntactic parsers, named entity recognizers) and natural language interfaces to databases.

In our proposed framework, the operation of DWSE exclusively relies on the result of context identification. Indeed, if there is available (meta)information on the data that is stored at the deep web site, and the internal representation language is SQL, then the DWSE engine can select the deep web sites relevant in answering the user query by determining the attributes and schemas appearing in the transformed SQL form of the query.

The paper is organized as follows. We briefly review related works on database theory and deep web search in Section 2. Section 3 introduces the most important concept of relational theory used in this article. Section 4 defines the new normalized natural database (NNDB) structure that stores language processing related information. In Section 5, we present a context identification algorithm based on NNDB. Section 6 discussed the role of DWSE in our framework, and Section 7 gives a few illustrative examples on context identification, and briefly presents some test results.

2 Related works

This paper concerns two different though partly connected research areas for which we summarize the literature separately.

Deep web search have been received attention from the late 90's [4, 3], but the first works only intended to characterize the scale of deep web from various aspects. As shown recently in [2], traditional crawl-and-index techniques may not be appropriate to provide a good coverage over the information content of the hidden and dynamic deep web site. In contrast to our cooperative model, another direction in discover the deep web is called the discover-and-forward access model. In this model, search engine having already discovered *query interface* of databases of the web would forward users automatically to the appropriate database to submit or refine the query. There are a few projects, such as MetaQuerier [8] and WISE-Integrator [9] which exploring this particular research direction. We remark that this solution also requires context identification to select the relevant databases for the user query. Our concept provide smaller coverage since it requires active cooperation from the deep web sites but ensures higher quality at answering, which can draw the attention of deep web sites in longer term. Another alterative method for the integration of deep web sites is highly accurate semantic matches between the attributes of the source query interfaces. However, due to the pervasive lack of data present at the query interface the matching accuracy is often questionable. WebIQ [10] solves this problem by discovering instances for interface attributes from the surface and deep web.

Traditional database management systems fail to capture the intentional semantics between data structure elements. In order to overcome this deficiency—which is a motivation of our semantic database approach—Sagiv's [11] and Honeyman's [12] works introduced the notion of intentional database in the early 80's which aimed at capturing the meaning of the represented relations in the database. Their path was followed by Chan end Mendelson [13] on the concept of separable databases. Their works intend to determine conditions on the

separability of schema structure of the database which enables that the modification of a particular relation does not affect other relations, and also eases the introduction of new schemas end relations into the database.

3 Preliminaries

Throughout the paper, we focus on relational databases and relational theory, although our results can be easily adopted to other well-known data models as well. We use the traditional conventions, terms and notations of relational data model theory.

- r denotes a relation; t denotes an element of a relation (aka n-tuple or record) t; R, S, T denote schemas; and A, B, C denote attributes.
- Capitals from the end of the alphabet (like X, Y, Z) denote attribute sets.
- The expression $r(R)$ denotes a relation on schema R.
- The expression $t[X]$ denotes the value of a record over the attribute set X.
- $t \in r$ means the record t is the element of relation r. $A \in R$ means that attribute A is an element of schema R. For simplicity, schemas are considered to be attribute sets.
- $X \rightarrow Y$ denotes a *functional dependency* defined on the schema $X \cup Y \subseteq R$. The functional dependency is true iff all records t_1, t_2 of any relation $r(R)$ over the schema R satisfy: $t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y]$. If $X \subseteq R$ and $X \rightarrow R$ hold then X is a *superkey*. If additionally there exist no $X' \subset X$ for which $X' \rightarrow R$ is true then X is the *key* of schema R. The set of functional dependencies is denoted by $\mathbf{F}$.
- The closure of an attribute set X over a dependency set $\mathbf{F}$ is defined as the maximal set having the property $X^+(\mathbf{F}) = \{A | \mathbf{F} \models X \rightarrow A\}$.
- The closure of dependency set $\mathbf{F}$ is defined as the maximal set of elements $\mathbf{F}^+ = \{X \rightarrow Y | \mathbf{F} \models X \rightarrow Y\}$. The dependency set $\mathbf{F}$ is a *coverage*, if there exists no such dependency $X \rightarrow Y$ that satisfies
  1. $(\mathbf{F} \setminus \{X \rightarrow Y\})^+ = \mathbf{F}^+$;
  2. $\nexists X' \subset X$, for which $\mathbf{F} \models X' \rightarrow Y$;
  3. Y consists of a single attribute.
- Let $X \dashrightarrow Y$ denote the *inclusion dependency*, which is true iff the domain of X and the domain of Y are identical, and for all relations r_1, r_2 over schemas R_1 and R_2, where $X \subseteq R_1, Y \subseteq R_2$ satisfy: $r_1(X) \subseteq r_2(Y)$. The set of inclusion dependency is denoted with $\mathbf{I}$.
- Let the relational database be denoted with $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$, where $\mathbf{R}$ is the finite set of schemas in the database, $\mathbf{r}$ is the finite set of relations over $\mathbf{R}$[1], and $\Sigma = \mathbf{F} \cup \mathbf{I}$ is the finite set of (functional and inclusion) dependencies of the schemas of the database.
- We say that a database $\mathcal{DB}$ complies a dependency set Σ, denoted by $\mathcal{DB} \models \Sigma$, if $\forall r \in \mathbf{r}\ r \models \Sigma$ is fulfilled.

[1] Obviously, there is exactly one relation over a given schema.

Relational databases do not specify how to define sensible operations between relations. As a consequence, one can join two relations if data are in appropriate formats along any two attributes. For example, one can join tables about cities and books on attributes population and title, respectively. Therefore database management systems can not determine the intentional semantics between elements of the data structure. In order to capture the notion "self-descriptiveness" we introduce a new concept called semantic database.

Definition 1 (Semantic databases). *Let* $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ *be a semantic database if for any attributes and schemas of* $\mathcal{DB}$ *, the following conditions hold regarding a reference function* $\lambda : \mathbf{A} \rightarrow \mathbf{A}$ *(*$\mathbf{A}$ *is the set of all attributes), and a binary (is-a) relation* $\Xi : \mathbf{R} \times \mathbf{R}$*:*

1. *for any* $S \in \mathbf{R}$ *schema with key* A*,* $\lambda(A) = A$*,*
2. *for any attribute* A *there is a schema* $S \in \mathbf{R}$ *and a simple key* $B \rightarrow S$ *such that* $\lambda(A) = B$*,*
3. $\Sigma \models A \dashrightarrow \lambda(A)$ *holds for any attribute* A*,*
4. $\Xi(R, S)$ *is true for any* $R, S \in \mathbf{R}$ *schemas if and only if* $\Sigma \models X \dashrightarrow Y$ *for some appropriate* $X \rightarrow R, Y \rightarrow S$ *attribute sets.*

For example: $\lambda(\mathsf{wife}) = \mathsf{name}$, $\lambda(\mathsf{name}) = \mathsf{name}$, and $\Xi(\mathsf{actor}, \mathsf{person})$. The definition extends the traditional relational database model by the reference function and an is-a relation in order to represent which attributes and schemas are related, and how.

Definition 2 (Valid relationship). *Let* $X \subseteq R_i, Y \subseteq R_j$ *be two attribute sets of the not necessarily distinct schemas* $R_i, R_j \in \mathbf{R}$ *in a database* $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$*. A valid relationship between* X *and* Y *exists if* $\Sigma \models \{X \dashrightarrow Y, Y \rightarrow R_j\}$*, and is denoted by* $\varepsilon(X, Y)$*.*

Proposition 1. $\varepsilon(A, \lambda(A))$ *for any attribute* A *of the semantic database* $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$*.*

Proposition 2. *Let* $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ *be a semantic database. If* X, Y *are keys of schemas* $R, S \in \mathbf{R}$ *respectively that make the relation* $\Xi(R, S)$ *valid, then* $\Xi(R, S) \Rightarrow \varepsilon(X, Y)$*.*

4 Normalized Natural Databases

According to the traditional database design concept, schemas are partitioned until a normal form is reached to keep them clear of redundancy. It must be noted that the model being designed depicts a closed system from a well-defined perspective. In such cases a schema may still contain many attributes, which makes further decomposition necessary if a new view appears. Thus to provide portability the number of attributes must be minimized in order to avoid further decomposition.

Naturally, for an arbitrary decomposition a schema must contain one attribute as something must identify the entities in the relation. It is easy to show that some specific attributes without the key property may have to remain in the schema too. For example, a name used in a natural language to identify an object in the world is inseparable from the object, though it is not a key, because different objects may have the same name. Let us call such an attribute *natural key* introduced by [14, 15]. The natural key of a schema S is denoted by $\kappa(S)$.

In other words, values of natural keys are the (compound) nouns or noun phrases in sentences, which are exact, unambiguous identifiers of an entity of the real world in the given language and context. For example, $\kappa(\mathsf{book}) = \mathsf{title}$, or $\kappa(\mathsf{person}) = \mathsf{family\_name}$, but $\kappa(\mathsf{Book}) \neq \mathsf{author}$, though author is an attribute of the book schema.

Definition 3 ([15]). *A semantic database is called natural, if any of the following conditions are fulfilled:*

1. *If the reference function of the database is λ, then $\lambda(A)$ is a natural key for all attributes A.*
2. *If $\varepsilon(A, B)$ is true for some attributes A, B of the database, then B is a natural key.*
3. *Each key is a natural key.*

Proposition 3. *The three statements in the Definition 3 are equivalent.*

Proof. $1 \Leftrightarrow 3$: According to definition 1, $\lambda(A)$ is always a key. From statement 1 $\lambda(A)$ is a natural key also which implies 3. Conversely, because of statement 3, each key is a natural key, which also holds for $\lambda(A)$, thus implying 1.

$2 \Leftrightarrow 3$: According to the definition of $\varepsilon(A, B)$, B is always a key. Therefore statement 2 implies 3, since B is a natural key. Conversely, the natural key property holds for key B in $\varepsilon(A, B)$ from statement 3, because B is a key in $\varepsilon(A, B)$.

Definition 4 (Normalized natural database). *A natural database $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \mathbf{F} \cup \mathbf{I}, \lambda, \Xi>$ is normalized (abbreviated by NNDB) if the following statements hold (here $S, R \in \mathbf{R}$):*

1. *Any attribute name appears in a single schema.*
2. *Any R schema with a natural key contains only one element. Such schemas are called primary schemas.*
3. *Any R schema without natural keys contains at least two elements, and for all attributes $A \in R$ there is an $S \neq R$ schema such that $\lambda(A) = \kappa(S)$. Here R is termed a secondary schema.*
4. *For any two different R, S schemas $R \cap S = \emptyset$.*
5. *There are no distinct secondary R, S schemas such that $R^+(\mathbf{F}) \supseteq S$.*
6. *For any R secondary schema there are either no valid non-trivial dependencies on it or there is an embedded dependency $\mathbf{F} \models X \rightarrow A$ defined on it such that $\mathbf{F} \models X \rightarrow R$, and $\nexists X' : X' \subset X$ for which $\mathbf{F} \models X' \rightarrow A$ holds.*
7. *For any inclusion dependency $X \dashrightarrow Y$, Y is a natural key.*

Proposition 4. *Let* $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \mathbf{F} \cup \mathbf{I}, \lambda, \Xi>$ *be a NNDB. The function* λ *implies an equivalence relation on attributes of* $\mathcal{DB}$ *in the following way: the equivalence class of attribute* A *consists of those attributes* A_j *for which* $\lambda(A_j) = \lambda(A)$*. We denote by* $\|A\|$ *the equivalence class of the attribute* A*.*

Due to the lack of space we omit the proof here (see [16]).

5 Context Identification in NNDBs

All primary schemas have a single attribute: a natural key, and named entities are its instances. For all other attributes, only references must be stored, i.e. references to the attributes they are referring to in a role. A pragmatic property of NNDBs is that they can be easily described by a few metaschemas. That is, NNDBs are extensible and portable to various topic areas. Hence, the context identification procedure is very similar to the idea first proposed in [17]. However, the proposed implementation uses purely relational semantics, it can be decomposed by any database schemas, therefore there is no need for re-engineering techniques. The proposed implementation also solves the identification of the question focus.

The search model of WoW uses NLP tools, namely morphological and syntactic parsers, and named entity recognizer. In our implementation for Hungarian queries we used the morphological parser Hunmorph [18], and our own syntactic parser [6] and named entity recognizer [7].

The context identification algorithm (or the disambiguation) based on a relational NNDB (see Figure 2) is composed of the following steps:

- Determine entities and phrase structures in the interrogative sentence of the natural language. Use a state-of-the-art named entity recognizer to determine with high accuracy the idiomatic expressions, names, labels (i.e. dates, currencies, etc.)—commonly referred to as *entities*—,and a syntactic parser to obtain the beginnings and endings of the phrases.
- Map the identified entities into NNDB elements using the Dictionary and KnowledgeBase. The mapping is essential for handling multilingual issues and synonymity. Naturally, a single entity in the real world may have more than one name, e.g. the Virgin of Orleans, Joan of Arc, Jeanne d'Arc, Joan the Saint refer to the very same person. The NNDB itself does not deal with this kind of ambiguity. For handling multilingual naming, synonymity, and other naming conventions a dictionary layer is necessary on top of the NNDB, which provides the real mapping between entities in questions and elements of the NNDB. Entities are mapped into primary schemas, attributes, and attribute values of natural keys [15].
  However, elements of NNDBs can still be ambiguous in the sense that the different entities may share the same representation forms. For example, the name Charles de Gaulle could stand for a historical personage, a name of street, an airport, a railway station, a national research project, and so on.

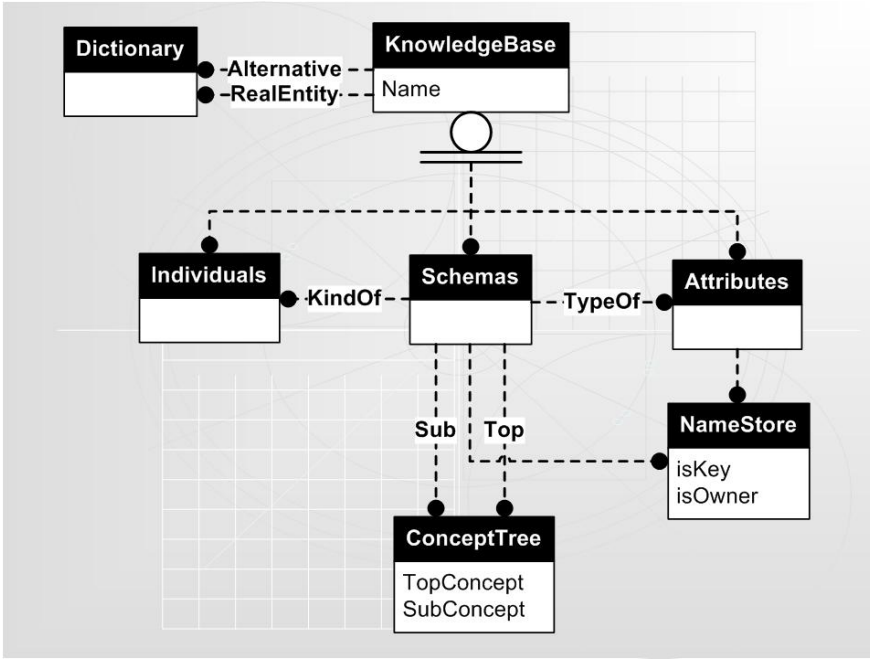


Fig. 2. NNDB example database (IDEF1X notation).

– Identify noun phrases in sentences using a state-of-the-art syntactic parser. Noun phrases describe a possibly non-empty set of entities that the question is all about. In other words, noun phrases contain the key data that identify the context of the sentence.
– Let $t_1, t_2, \ldots t_n$ be the representations of noun phrase heads in a $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \mathbf{F} \cup \mathbf{I}, \lambda, \Xi>$. If noun phrase heads are all described by $\mathcal{DB}$ then find the common schema which connects all possible meanings of noun phrase heads. In order to do that, one must distinguish among three different cases depending on whether the noun phrase head t_i is a schema, an attribute or an attribute value in $\mathcal{DB}$:

$$\gamma(t_i) = \begin{cases} \|\kappa(t_i)\| & \text{if } t_i \text{ is a schema} \\ \|t_i\| & \text{if } t_i \text{ is an attribute} \\ \{\|t_j\| | t_i \in \mathrm{DOM}(t_j)\} & \text{if } t_i \text{ is a value.} \end{cases}$$

The common schema which connects all noun phrase heads is

$$\Gamma = \bigcap_{i=1}^{n} \gamma(t_i).$$

Note that ambiguities are partially eliminated, therefore meaningless expressions result in an empty set. On the other hand, the disambiguation is not fully

done at this point since alternatives with common schemas are still present after this step. Two typical cases are interrogative sentences that contain a single noun phrase, and sentences that are ambiguously stated by the user. In either case, common schemas are all returned to the user as alternatives for topic restriction.
If noun phrase heads are not represented in databases, the algorithm terminates with an error. Unfortunately, information on purposes, usage, causes, accidental events, experiments, are typically can not be represented and therefore not stored in databases. Moreover, the outlined method is unable to properly resolve cultural idioms (e.g. The Voice as Frank Sinatra), common sense expressions (e.g. Washington as U.S. government), symbolic, associative expressions (e.g. Mecca of movies), etc.

– Noun phrases identify contexts not by naming a common schema but by naming the focus of the question. The information needs of a user executing a search can be represented by a triplet: the question tag, the head of the verb phrase, and by the head of the first noun phrase after the question tag or the verb phrase excluding pronouns if the former is missing. This observation led us to apply a template based focus identification algorithm, i.e., template triplets determine a set of schemas Δ.
– The context of question is $\Gamma \cap \Delta$.

What about ambiguities in NNDBs? Natural language ambiguities can be resolved only by context similar to the way they are resolved in natural conversations. If $\Gamma \cap \Delta$ contains more than a single schema then user is prompted to choose between these options.

6 Deep Web Search Engine in Work

Once the context is identified, one needs to find the databases that may contain information about the input question. It is easy to see that deep web sites might have a correct answer for the input question if and only if all their schemas and attributes can be mapped into the NNDB. Unfortunately, such a mapping is not easy to find algorithmically, hence schema and attribute names, and their semantics might differ from the ones used in NNDBs. This is why WoW requires their partners to declare which schemas and attribute elements are present in their databases, and how they are stored. With these information mediator layers can make way for the proper transformation from NNDB queries to URLs, forms, or deep web database queries.

The role of DWSE is twofold. First, it serves to maintain information on all connected deep web sites by storing their metaschemas and data structures, and second, it determines by simple mathematical relations which deep web sites may be relevant to answer the user's question.

7 An Illustrative Example

One stores information on books, movies, locations, and cultural events. Let $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \mathbf{F} \cup \mathbf{I}, \lambda, \Xi>$ be defined in the following way:

- $\mathbf{R}$ = {movie, cinema, person, director, actor, location, datetime, shows, acts, directs, lives} where the first 7 schemas are primary schemas, and others are secondary ones.
- For primary schemas: κ(movie) = title, κ(cinema) = cname, κ(person) = pname, κ(actor) = aname, κ(director) = dname, κ(location) = place, κ(datetime) = date.
- For secondary schemas: shows = {movie, datetime, location}, acts = {amovie, aactor}, directs = {dmovie, ddirector}, lives = {person, address}. λ(amovie) = λ(dmovie) = title, λ(datetime) = date, λ(address) = place, λ(aactor) = aname, λ(director) = dname, and λ(person) = pname.
- Ξ(director, person), Ξ(actor, person), Ξ(cinema, location).
- There are no other dependencies.

In the following examples we skip both the named entity recognition and the NNDB mapping step for the sake of simplicity.

For the question "Where did Churchill live?" the context identification procedure finds the single phrase head Churchill. Churchill as a name could stand for either a person, a place or a movie, i.e. $\Gamma = \mathbf{R}$. We have a template for <where, live, ...> where "..." means arbitrary first phrase head. The template maps this triplet into Δ = {location}, therefore the context of this sentence must be $\Delta \cap \Gamma$ = {location}. Based on these information, deep web algorithm transforms user question into the following SQL statement.

```
SELECT place FROM lives WHERE person = 'Winston Churchill'
```

For basic ideas on this transformation see [15, 16]. Due to lack of space, full details will be published in the near future.

The algorithm processes the question "In which movie did Quentin Tarantino play?" in the following way. It first identifies two simple noun phrases: movie and Quentin Tarantino. Since Quentin Tarantino is both a director and an actor $\Gamma = \gamma$(Quentin Tarantino) $\cap\, \gamma$(movie) = {acts, directs}. The focus identification process determines Δ = {acts} using the template <which, play, movie>, that is, the result must be $\Gamma \cap \Delta$ = acts. During the next phase, the next SQL statement is generated:

```
SELECT movie FROM acts WHERE aactor = 'Quentin Tarantino'.
```

When ambiguity has to be resolved, the system prompts the user to clarify the question "Where can I see Pulp Fiction?". The solution produced by the algorithm contains two possibilities: the user either asked about the cinema or the place where the movie will be shown. The algorithm does not prioritize, therefore the user interaction is unavoidable.

In the current implementation our knowledge base contains 18000 terms (∼110 schemas, ∼230 attributes) on 23 topic areas, e.g. movies, books, restaurants, locations, cultural events, and related institutions, groups, people, etc (sources are port.hu, National Széchényi Library, eszemiszom.hu). Our template database consists of cca. 1000 templates. We found that this approach has a 83% precision on a cca. 1000 sentence corpus extracted from Szeged Treebank [19], however, in most cases (cca. 72%) it finds the question to be incomplete. Obviously, this approach also has its limitations: it cannot retrieve information that does not fit into the database model (e.g. questions about reasons, causality, subjectivity). We excluded these types of questions from testing.

8 Concluding Remarks

In this paper, we presented a new relational database design technique called NNDB. We pointed out that a proper NNDB is a context database, and it could serve as the basis of context identification combining the template based techniques and using the world model encoded in the database design. The database structure can be easily extracted from any relational database, and needs no re-engineering technique. Moreover, it has a well-formed mathematical background based on relational theory.

Acknowledgement

Our work was sponsored by Mobile Innovation Center through the Asbóth Oszkár Programme of the Hungarian State.

References

1. Winkler, H.: Suchmaschinen. Metamedien im Internet? In Becker, B., Paetau, M., eds.: Virtualisierung des Sozialen, Frankfurt/NY (1997) 185–202 (In German; English translation: `http://www.uni-paderborn.de/~timwinkler/suchm_e.html`).
2. He, B., Patel, M., Zhang, Z., Chang, K.: Accessing the Deep Web: A Survey. Communications of the ACM **50** (2007) 94–101
3. Bergman, M.K.: The deep web: surfacing hidden value. Journal of Electronic Publishing **7** (2001) `http://www.press.umich.edu/jep/07-01/bergman.html`.
4. Lawrence, S., Giles, C.: Accessibility of information on the web. Nature **400** (1999) 107–109
5. Tikk, D., Kardkovcs, Z.T., Magyar, G.: Searching the deep web: the WOW project. In: New Methods and Practice for the Networked Society. Volume 2 of Advances in Information Systems Development., New York, Springer (2007) 493–504
6. Tikk, D., Kardkovács, Z.T., Andriska, Z., Magyar, G., Babarczy, A., Szakadát, I.: Natural language question processing for hungarian deep web searcher. In Elmenreich, W., Haidinger, W., Machado, T., eds.: 2^{nd} IEEE International Conference on Computational Cybernetics, ICCC 2004, Vienna, Austria (2004) 303–309

7. Tikk, D., Szidarovszky, F.P., Kardkovács, Z.T., Magyar, G.: Entity recognizer in Hungarian question processing. In Bandini, S., Manzoni, S., eds.: AI*IA 2005: Advances in Artificial Intelligence. Number 3673 in Lecture Notes in Artificial Intelligence. Springer, Berlin–Heidelberg–New York (2005) 535–546
8. Chang, K.C.C., He, B., Zhang, Z.: Toward large scale integration: Building a MetaQuerier over databases on the web. In: Proceedings of the 2^{nd} Biennial Conference on Innovative Data Systems Research (CIDR 2005), Asilomar, CA, USA (2005) 44–55
9. He, H., Meng, W., Yu, C., Wu, Z.: Wise-integrator: an automatic integrator of web search interfaces for e-commerce. In: Proceedings of the 29^{th} international conference on Very large data bases. Volume 29., VLDB Endowment (2003) 357–368
10. Wu, W., Doan, A., Yu, C.: WebIQ: learning from the web to match deep-web query interfaces. In: Proc. of the 22^{nd} Int. Conf. on Data Engineering (ICDE'06), IEEE Computer Society (2006) 44–53
11. Sagiv, Y.: A characterization of globally consistent databases and their correct access paths. ACM Transactions on Database Systems **8** (1983) 266–286
12. Honeyman, P.: Testing satisfaction of functional dependencies. Journal of the ACM **29** (1982) 668–677
13. Chan, E.P.F., Mendelzon, A.O.: Independent and separable database schemes. In: PODS'83: Proceeding of the 2^{nd} ACM SIGACT-SIGMOD symposium on Principles of database systems, New York, NY, USA, ACM Press (1983) 288–296
14. Kardkovács, Z.T.: On the transformation of sentences with genitive phrases to SQL statements. In: Proc. of the 10^{th} NLDB. Volume 3513 of Lecture Notes in Computer Science., Alicante, Spain, Springer Verlag (2005) 10–20
15. Kardkovács, Z.T., Tikk, D.: On the transformation of sentences with genitive relations to SQL queries. Data & Knowledge Engineering **61** (2007) 406–416
16. Kardkovács, Z.T.: Querying Heterogeneous Databases Based on Semantic Processing of Well-Formed Language Phrases. PhD thesis, Budapest University of Technology and Economics, Budapest, Hungary (2007) In Hungarian.
17. Meng, X., Wang, S., Wong, K.F., Lum, V.: A chinese query language chiql: Design and evaluation. In: SEEP'98: Proceedings of International Conference on Software Engineering: Education and Practice, New Zealand, IEEE Press (1999) 190–197
18. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., Varga, D.: Hunmorph: open source word analysis. In: Proc. of ACL. (2005)
19. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proc. of the 8^{th} TSD. Volume 3658 of Lecture Notes in Computer Science., Karlovy Vary, Czech Republic, Springer Verlag (2005) 123–131

Deep Web Navigation by Example

Yang Wang and Thomas Hornung

Institute of Computer Science, Albert-Ludwigs University Freiburg, Germany
{wangy,hornungt}@informatik.uni-freiburg.de

Abstract. Large portions of the Web are buried behind user-oriented interfaces, which can only be accessed by filling out forms. To make the therein contained information accessible to automatic processing, one of the major hurdles is to navigate to the actual result page. In this paper we present a framework for navigating these so-called Deep Web sites based on the page-keyword-action paradigm: the system fills out forms with provided input parameters and then submits the form. Afterwards it checks if it has already found a result page by looking for pre-specified keyword patterns in the current page. Based on the outcome either further actions to reach a result page are executed or the resulting URL is returned.

Key words: Form Analysis, Deep Web Navigation by Page-Keyword-Actions

1 Introduction

A recent study by He et. al [5] has found an exponential growth and great subject diversity of Deep Web [2] sites. Taking into account the vast amount of high-quality data, which is geared towards human visitors, it is not surprising that many different research questions are actively pursued in this area at the moment, e.g. vertical search engines [4].
In this paper we present a framework which bridges the gap between the front page and the desired result page which actually contains the relevant data. In a user-assisted acquisition step we first analyze the relevant form fields on the Web page we are interested in and then build a navigation model based on the page-keyword-action paradigm. The main idea is twofold: first, the user has to identify and label the relevant input form fields. For these we pre-compute and store the dependencies in a database so that we can check for illegal combinations offline. Second, we use the user-provided input values at runtime to fill out the appropriate form fields and then check after submittal if we have already reached the result page. The check is based on a keyword sequence, which gives us a hint if we are on an intermediate, or bridge, page. If so, a series of actions, which are associated with this bridge page, is performed, which yields us to a new page. Here again, we check if we have reached a result page. If this is the case, we return the URL, otherwise we (again) perform the associated actions or return an error message.
This framework has been developed for use in the FireSearch project [6] whose

aim it is to organize Deep Web sources in a mashup graph, where it is used in conjunction with the ViPER [13] wrapper tool to convert Deep Web sources into machine-processable query interfaces. However, as it has been implemented in JavaScript and Java as a Firefox plugin it could be used with minor modifications in other projects, e.g. for a domain-specific Meta Search engine, where the relevant Deep Web sources could be integrated by an interested community, as well.
The paper is structured as follows: we start with a description of the two main components of our framework, namely the analysis of form fields in Section 2 and the navigation model in Section 3. In Section 4 we present an evaluation of our system and in Section 5 we discuss related work. Finally, we conclude in Section 6 with an outlook on future work.

2 Form Analysis

Initially for each new Web page we store all occurring forms in a database for later analysis. Afterwards the user can load the desired form field and label the desired input elements[1], e.g. in Figure 1 the maximum desired price the visitor is willing to pay for a used car has been labeled *Price-To*. Overall she has labeled six input elements, e.g. the desired brand and the make of the car. Now we check for each labeled input element, if they are static or if there are any dynamic dependencies, which might be due to Ajax interactions with the server. Note, that only these input elements of the form can be used later on for querying that have been labeled in this stage. Our running example is the analysis of a Web search engine for used cars[2], where each car model depends on its car make. The other input elements are static, i.e. they do not change if one of the other input elements is changing. The dynamic and static combinations are determined automatically after the user has finished labeling the desired input elements based on the following idea: modify the first dropdown menu[3] and check all other labeled dropdown menus, if the available options have changed. If this is the case, then modify the dependent dropdown menu to uncover layered dependencies and mark the dependent menu as dynamic. After all dropdown menus have been checked, we mark all menus that are not dynamic as static. To avoid loops, we only check possible dropdown menus that have not participated in a dependency in the current analysis cycle before, e.g. in the example shown in Figure 1 the car model would not be considered if we check for further dynamic dependencies for the car make input element.
Figure 2 and Figure 3 show the resulting static and dynamic dependencies for our running example. After the *frontend* analysis is finished, we continue with

[1] In the context of this paper we refer to all elements in the form field that can be provided with a value, e.g. checkboxes, as input elements.
[2] `http://www.autoscout24.de`
[3] Only dropdown menus are currently considered as candidates for dynamic elements, all other input element types are assumed to be static by default.

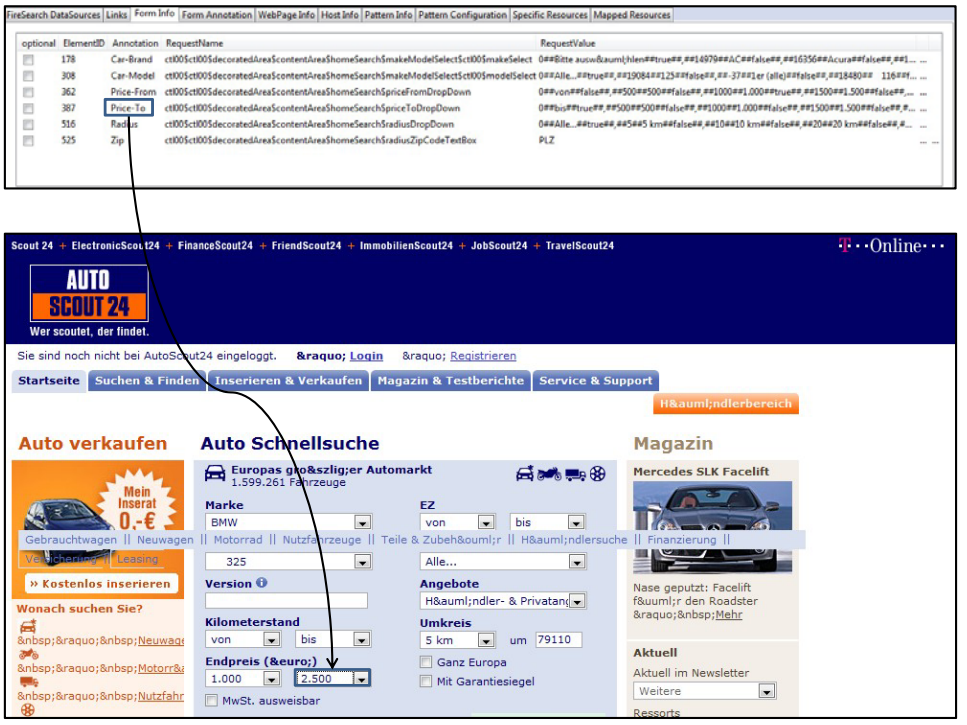


Fig. 1. Annotation of form attributes

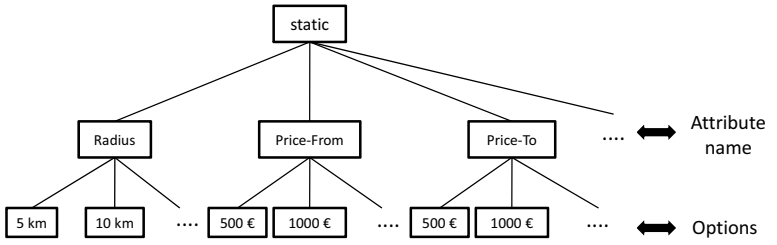


Fig. 2. Relation tree for static input elements for `http://www.autoscout24.de`

the analysis of the possible navigation patterns for this page described in the next section.

3 Deep Web Navigation

The navigation model is a crucial part of our system. Based on the model the system can anytime determine, if it has already reached the result page or if it is on an intermediate page. Additionally the model determines the actions, which should be perfomed for a specific intermediate page, e.g. to click on a link or fill

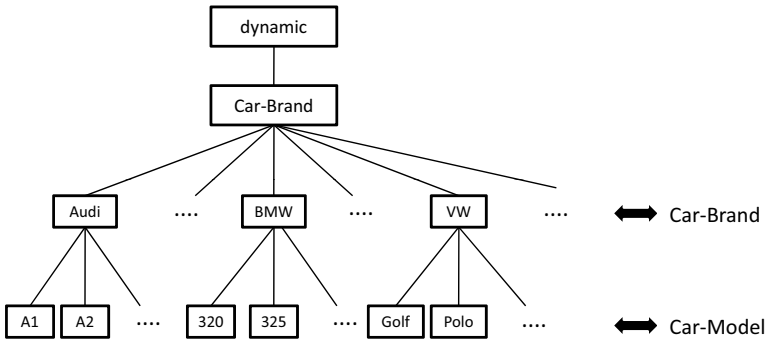


Fig. 3. Relation tree for dynamic input elements for `http://www.autoscout24.de`

out a new form field. The key idea of our *Page-Keyword-Action* paradigm is that the system first determines its location (intermediate vs. result page) based on a *page keyword* and then invokes a series of associated *actions* if appropriate.

3.1 Deep Web Navigation

The overall navigation process is illustrated in Figure 4: the user provides the system with a value map that contains for each desired input element label/value combinations. If the form field contains dynamic input elements for which she has provided input label/value combinations we check if they are legal. If so, we subsequently fill out and submit the form field with these combinations, which yields a new Web page[4]. For this Web page we check, if we can find one of our defined keywords (cf. Section 3.2). If so, we perform the associated actions which result in a new Web page and check again if we are on a intermediate page. The cycle continues as long as we can find keywords on the Web page. To avoid an infinite loop, the user can specify an upper bound on the number of possible intermediate pages, after which an error message is returned. If we cannot find a keyword on the current Web page, we have found the goal page and return its URL.

3.2 Intermediate Page Keyword

Deep Web pages are typically created dynamically, i.e. data from a background database is filled into a predefined presentation template. Therefore, we can usually identify fixed elements, which are part of the template, which are almost identical between different manifestations. After the form analysis is finished the user can iteratively submit the form with different options. If an input value combination leads her to an intermediate page, she can identify the relevant

[4] Additionally, we use the information obtained during form analysis for directly generating the request POST/GET URL. Thereby we can offline mimic the behavior of the form field.

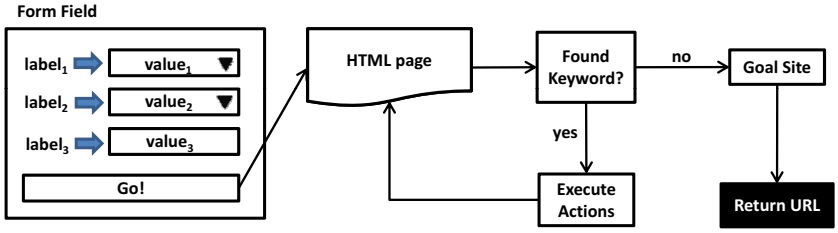


Fig. 4. Navigation process

keyword as described in the following. If she has already reached a result page for a value combination no further user interactions are required. Note that as long as she is in the context of the currently active form field, she can also access a series of intermediate pages and for each page specify a series of actions.
For the identification of a specific intermediate page we opted for a static text field. The reason is that it can be included in many HTML elements, e.g. the `div`, `h2`, or the `span` tag and given our template assumption they serve as a sufficient discriminatory factor. Other more advanced techniques based on visual markers on the page or more IR-related techniques, such as text classification approaches [10], could be used in this context as well and are planned as future work. In Figure 5 we have marked potential candidates for keywords with a rectangle. The most likely candidates which are most characteristic are encircled with an ellipse, e.g. the error message for the car search service shown on the left. After the user has identified the keyword in the page, she can now specify actions that should be performed in order to reach the result page.

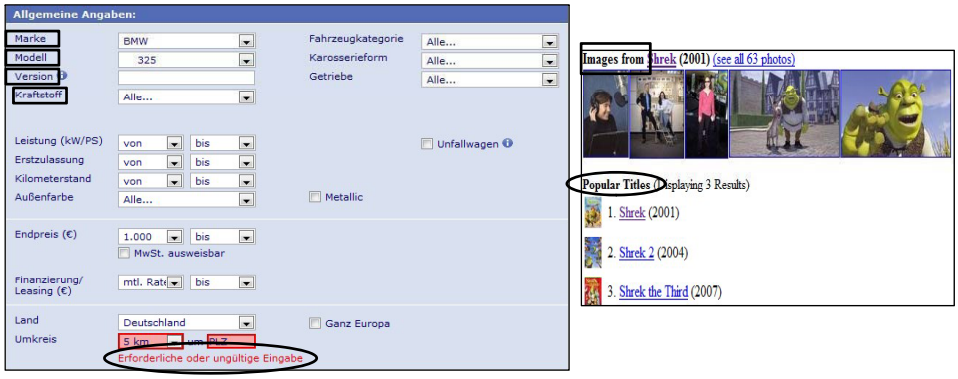


Fig. 5. Intermediate pages for `http://www.autoscout24.de` (left) and `http://www.imdb.com` (right)

3.3 Intermediate Page Actions

The above specified keywords can be used to identify intermediate pages. However, our ultimate goal is to find a result page given a set of input value combinations for the initial form field. Therefore some actions, such as clicking on a link or filling out and submitting a new (intermediate) form, have to be perfomed to access the next - preferrably result - page.
In order to uniquely identify the appropriate HTML elements on which the stored actions should be executed, we defined a path addressing language called *KApath*, which is a semantic subset of XPath [16]. In order to access the appropriate action element, the system first finds the common ancestor of the keyword element and the action element and then descends downwards in the action element branch. Afterwards, the registered actions are executed for the found action element. Thus, KApath supports the following path expressions:

- **/Node[@aname$_1$=avalue$_1$] ... [@aname$_n$=avalue$_n$]:** The element in the DOM tree that matches the specified attribute name-value combinations of type *Node*,
- **/P:** Immediate parent node of current node,
- **/P::P:** All (transitive) parent nodes of current node,
- **/P::P/Node[@aname$_1$=avalue$_1$] ... [@aname$_n$=avalue$_n$]:** The first found parent node in the DOM tree that matches the specified attribute name-value combinations starting from the current node and is of type *Node*,
- **/Child:** Immediate child nodes of current node,
- **/Child::Child:** All (transitive) child nodes of the current node,
- **/Child::Child/Node[@aname$_1$=avalue$_1$] ... [@aname$_n$=avalue$_n$]:** The first found child node in the DOM tree that matches the specified attribute name-value combinations starting from the current node and is of type *Node*.

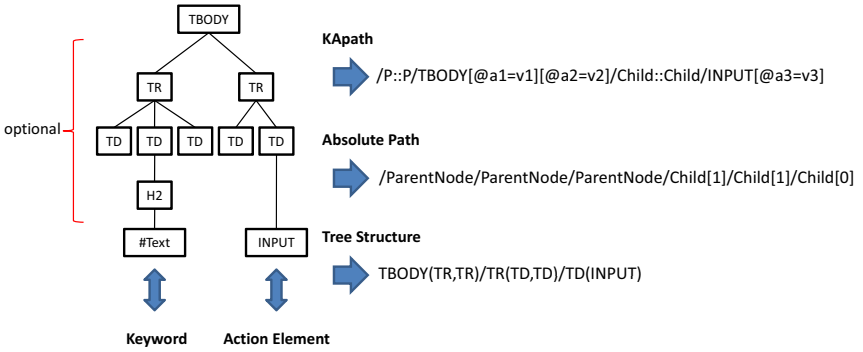


Fig. 6. Example KApath expression that allows optional HTML elements in the intermediate page

Figure 6 shows an example how the associated action element in a page can be referenced with respect to the page keyword with a KApath expression. Here,

the TBODY node is the first common parent node for both (keyword and action) elements. Therefore the system automatically generates a KApath expression which allows optional intermediate elements between the keyword and the first common parent node. For finding the correct action element it is crucial to consider its attributes as well. However, it can still happen that the desired action elements have no (e.g. links) or dynamic attributes (e.g. visibility). For these cases we additionally store the absolute path from keyword to action element and the tree structure starting from the common parent. Another situation where we can make use of the absolute path is when the HTML page structure has changed and the common parent node is still on the same level in the DOM tree but in another branch. The tree structure is helpful if there are changes on the way downwards from the common parent node. Together, keyword, KApath, absolute path and the tree structure form the navigation model for this intermediate page (cf. Figure 6).
Based on the user's browsing behavior, the system can generate the complete navigation model. First, she identifies the keyword for an intermediate page by clicking on the relevant text in the Web page. Then, the system determines the closest surrounding HTML element and stores the relevant context information. Afterwards, the system monitors the user behavior and stores each action she performs until she reaches a new page. Based on this action log, the system can automatically determine the paths and tree structures for each action.
The following type of actions are supported by our system:

- Clicking on links,
- Entering text in input fields,
- Selecting options from a dropdown or checkbox menu, and
- Submitting forms.

4 Evaluation

In our experiments, we evaluated the following aspects for our two major components: accuracy and runtime. For this, we selected 100 Deep Web sites from different domains, e.g. car search and video search. 60 of them were directly adopted from the website table in [2], because they contain a large amount of data. The others were selected by a focused search on Google on Deep Web repositories. For a full list of the tested Web sites we refer the interested reader to [14].

4.1 Experimental Results

Frontend Analysis For 99% of the tested Web sites the frontend analysis was successful, finding the correct static and dynamic dependencies. Depending on the number of items in the dropdown menus of the form fields, the time needed for analysis took from 0.5 to 30 seconds, i.e. 4.28 seconds on average. Since this analysis has only to be performed once, we feel that performance optimizations

# Int. Pages	# Web Sites	Page Load	1 Model	6 Models
0	58	2.25	2.26	2.31
1	22	-	4.60	4.66
2	14	-	6.47	6.55
3	4	-	8.12	8.23
4	1	-	9.70	9.83
5	1	-	11.06	11.22

Table 1. Time (in seconds) for navigation experiments

for this analysis are of limited benefit, because our major focus is on correctly identifying hidden dependencies between the dropdown menus.

Deep Web Navigation For 96% of the tested Web sites we were able to successfully find a keyword and to navigate to the desired result page. The navigation process took from 2.26 to 11.22 seconds, i.e. 3.79 seconds on average. As shown in Table 1 most of the time was spend for loading pages, i.e. 2.25 seconds on average. The columns labeled *1 Model* and *6 Models* indicate the number of registered navigation models for each page. As can be seen, the overhead for checking multiple models was marginal in contrast to the time spent for loading pages. This is due to the fact that the execution of the actions is performed by the browser on the client side and since no computationally intensive algorithm is required to identify intermediate pages.

4.2 Open Issues

Our evaluation revealed the following open issues of our system.

Frontend Analysis

– Delayed AJAX interactions: For one Web site we were unable to correctly detect the dynamic dependencies because the server took longer than our specified threshold to change the items in the respective dropdown menu.

This could be remedied by increasing our threshold value to some extent, but further investigation is needed to find a general solution for this problem.

Deep Web Navigation

– Dynamic request URLs: Usually, different request URLs only differ in the searchpart[5], due to different variable bindings, which are transferred to the server. Two Web sites in our test bed used different paths as well, which our system converts into illegal request URLs.
– Hidden form elements: Since the user can only drag labels to visible form elements, values in hidden form elements that have to be correlated with visible elements cannot be detected by our system.

[5] The part of the URL after the ?.

– Session IDs: Session IDs are often used to track user interactions with Web pages and are only valid for a certain period. Because we are not able to produce a new (fake) session ID for each service, the offline generated URL becomes invalid over time.

All of the abovementioned issues could be solved by filling out the frontend form at runtime and skipping the offline generation of the URL for such resources.

– Static URLs: Our system determines, if a new Web page has been loaded based on the current URL. If the URL does not change after a form has been submitted, we are not able to initiate the navigation process.

This can be solved by using another metric for determining if a new Web page has been loaded, e.g. a checksum of the Web page.

5 Related Work

A number of navigation concepts have been proposed for accessing Deep Web sources. [3] and [9] proposed process-oriented navigation maps, which describe a set of paths from a start page to a result page. But these maps rely on consecutive state transitions and fixed interactions between them. In [7] the user actions from a specified start page over possibly multiple intermediate pages to an end page are recorded in a navigation map. The actions that link two adjacent pages are strongly connected as well. A sophisticated Deep Web navigation strategy based on the branched navigation model is proposed in [1]. The navigation is represented as a sequence of pages, with envisioned future support for standard process-flow languages such as BPEL [15]. In [12] a navigation sequence was specified in NESQL [11]. The NESQL expression contains several informations about action elements, for instance, their specified names and types. Each expression will be interpreted based on these element properties. By storing historical information from previous accesses of a Deep Web resource and utilizing browser pools, their system tries to reuse the current state of a browser.
Our framework is not dependent on a rigid sequence of intermediate pages, because for each new page all keyword patterns are checked and therefore the previous state of the system is not important for our page-oriented navigation model. Besides, we do not need a complex navigation algebra or calculus for the navigation process because we just save the above described navigation model for each intermediate page. For instance, the framework proposed by [3] relies on a subset of serial-Horn Transaction F-Logic [8]. As discussed in Section 3.3, the saved action sequences are just macro procedures, which are interpreted by our JavaScript macro engine.

6 Conclusion and Future Work

In this paper we presented a framework for bridging the gap between the start and goal page of Deep Web pages. We have proposed a simple, but efficient, Deep

Web navigation strategy, which we have found to be very effective thus far. The main idea is to change a heavy-weight navigation calculus for an intermediate page identification procedure and a set of actions that navigate to the next page. Our experiments suggest that the determination of a suitable keyword is crucial for the successful identification of an intermediate page, and that for some cases it might be better to skip the offline generation of the start URL.
For future work we plan to investigate how to automatically suggest meaningful and discriminatory keywords to the user and the use of more elaborate techniques to identify intermediate pages, such as the visual appearance of the Web page.

References

1. Baumgartner, R., Ceresna, M., Ledermüller, G.: Deep Web Navigation in Web Data Extraction. In: CIMCA/IAWTIC, pp. 698–703. (2005)
2. Bergman, M. K.: The Deep Web: Surfacing Hidden Value. White Paper, http://www.brightplanet.com/images/stories/pdf/deepwebwhitepaper.pdf (2001)
3. Davulcu, H., Freire, J., Kifer, M., Ramakrishnan, I. V.: A Layered Architecture for Querying Dynamic Web Content. In: SIGMOD, pp. 491–502. (1999)
4. He, H., Meng, W., Yu, C. T., Wu, Z.: WISE-Integrator: A System for Extracting and Integrating Complex Web Search Interfaces of the Deep Web. In: VLDB, pp. 1314–1317. (2005)
5. He, B., Patel, M., Zhang, Z., Chang, K. C.-C.: Accessing the Deep Web. In: Commun. ACM, 50(5), pp. 94–101. (2007)
6. Hornung, T., Simon, K., Lausen, G.: Mashing Up the Deep Web - Research in Progress. In: WEBIST 2008. (2008)
7. Julasana, N., Khandelwal, A., Lolage, A., Singh, P., Vasudevan, P., Davulcu, H., Ramakrishnan, I. V.: WinAgent: A System for Creating and Executing Personal Information Assistants Using a Web Browser. In: IUI, pp. 356–357. (2004)
8. Kifer, M.: Deductive and Object-oriented Data Languages: A Quest for Integration. In: DOOD, pp. 187–212. (1995)
9. Lage, J. P., da Silva, A. S., Golgher, P. B., Laender, A. H. F.: Collecting Hidden Web Pages for Data Extraction. In: WIDM, pp. 69–75. (2002)
10. Nigam, K., McCallum, A., Thrun, S., Mitchell, T. M.: Learning to Classify Text from Labeled and Unlabeled Documents. In: AAAI/IAAI, pp. 792–799. (1998)
11. Pan, A., Raposo, J., Álvarez, M., Hidalgo, J., Viña, Á.: Semi-Automatic Wrapper Generation for Commercial Web Sources. In: EISIC, pp. 265–283. (2002)
12. Raposo, J., Álvarez, M., Losada, J., Pan, A. : Maintaining Web Navigation Flows for Wrappers. In: DEECS, pp. 100–114. (2006)
13. Simon, K., Lausen, G.: ViPER: Augmenting Automatic Information Extraction with Visual Perception. In: CIKM, pp. 381–388. (2005)
14. Wang, Y.: Deep Web Navigation by Example. In: Master's Thesis, Institute of Computer Science, Albert-Ludwigs University Freiburg. (2008)
15. Web Services Business Process Execution Language Version 2.0, `http://www-128.ibm.com/developerworks/library/specification/ws-bpel/`. (2007)
16. XML Path Language (XPath) Version 1.0, `http://www.w3.org/TR/xpath`. (1999)

Fuzzy Constraint-based Schema Matching Formulation

Alsayed Algergawy, Eike Schallehn, and Gunter Saake

Department of Computer Science,
Otto-von-Guericke University,
39106 Magdeburg, Germany
{alshahat,eike,saake@iti.cs.uni-magdeburg.de}

Abstract. The deep Web has many challenges to be solved. Among them is schema matching. In this paper, we build a conceptual connection between the schema matching problem *SMP* and the *fuzzy constraint optimization problem FCOP*. In particular, we propose the use of the *fuzzy constraint optimization problem* as a framework to model and formalize the schema matching problem. By formalizing the *SMP* as a *FCOP*, we gain many benefits. First, we could express it as a combinatorial optimization problem with a set of soft constraints which are able to cope with uncertainty in schema matching. Second, the actual algorithm solution becomes independent of the concrete graph model, allowing us to change the model without affecting the algorithm by introducing a new level of abstraction. Moreover, we could discover complex matches easily. Finally, we could make a trade-off between schema matching performance aspects.

Key words: Schema matching, Constraint programming, Fuzzy constraints, Objective functions.

1 Introduction

The number of deep Web sources has increased rapidly [3]. To open the deep Web to users software systems are needed to enable users to explore and integrate deep Web sources. Schema matching is the core task of these systems.

Schema matching is the task of identifying semantic correspondences among elements of two or more schemas. It plays a central role in many data application scenarios [12]: in *data integration*, to identify and characterize inter-schema relationships between multiple (heterogeneous) schemas; in *data warehousing*, to map data sources to a warehouse schema; in *E-business*, to help to map messages between different XML formats; in the *Semantic Web*, to establish semantic correspondences between concepts of different web sites ontologies; and in *data migration*, to migrate legacy data from multiple sources into a new one [9].

Due to the complexity of schema matching, it was mostly performed manually by a human expert. However, manual reconciliation tends to be a slow and

inefficient process especially in large-scale and dynamic environments. Therefore, the need for automatic schema matching has become essential. Consequently, many schema matching systems have been developed for automating the process, such as Cupid [12], COMA [5], LSD [6], BTreeMatch [10], and Spicy [2]. Manual semantic matching overcomes mismatches which exist in element names and also differentiates between differences of domains. Hence, we could assume that manual matching is a perfect process. On the other hand, automatic matching may carry with it a degree of uncertainty, as it is based on syntactic, rather than semantic, means. Furthermore, recently, there has been renewed interest in building database systems that handle uncertain data in a principled way. Hence a short rant about the relationship between databases that manage uncertainty and data integration systems appears. Therefore, we should surf for a suitable model which is able to meet the above requirements.

A first step in discovering an effective and efficient way to solve any difficult problem such as schema matching is to construct a complete problem specification. A suitable and precise definition of schema matching is essential for investigating approaches to solve it. Schema matching has been extensively researched, and many matching systems have been developed. Some of these systems are rule-based [5, 12, 14] and others are machine learning-based [6, 7]. However, formal specifications of problems being solved by these systems do not exist, or are partial. Little work is done towards schema matching problem formulation e.g. in [18, 16].

In the rule-based approaches, a graph is used to describe the state of a modeled system at a given time, and graph rules are used to describe the operations on the system's state. As a consequence in practice, using graph rules has a worst case complexity which is exponential to the size of the graph. Of course, an algorithm of exponential time complexity is unacceptable for serious system implementation. In general, to achieve acceptable performance it is inevitable to consequently exploit the special properties of both schemas to be matched. Beside that, there is a striking commonality in all rule-based approaches; they are all based on *backtracking paradigms*. Knowing that the overwhelming majority of theoretical as well as empirical studies on the optimization of backtracking algorithms is based on the context of *constraint problem (CP)*, it is near to hand to open this knowledge base for schema matching algorithms by reformulating the schema matching problem as a CP [17, 13, 4].

In this paper, we build a conceptual connection between the schema matching problem *(SMP)*and the *fuzzy constraint optimization problem (FCOP)*. On one hand, we consider schema matching as a new application of fuzzy constraints; on the other hand, we propose the use of the fuzzy constraint satisfaction problem as a new approach for schema matching. In particular, in this paper, we propose the use of the *FCOP* to formulate the *SMP*. However, our approach should be generic, i.e. have the ability to cope with different data models and be used for different application domains. Therefore, we first transform schemas to be matched into a common data model called rooted labeled graphs. Then we reformulate the graph matching problem as a constraint problem. There are many

benefits behind this formulation. First, we gain direct access to the rich research findings in the *CP* area; instead of inventing new algorithms for graph matching from scratch. Second, the actual algorithm solution becomes independent of the concrete graph model, allowing us to change the model without affecting the algorithm by introducing a new level of abstraction. Third, formalizing the *SMP* as a *FCOP* facilitates handling uncertainty in the schema matching process. Finally, we could simply deal with simple and complex mappings.

The paper is organized as follows: Section 2 introduces necessary preliminaries. Our framework to unify schema matching is presented in Sect. 3 to show the scope of this paper. Section 4 shows how to formulate the schema matching problem as a constraint problem. The concluding remarks and ongoing future work are presented in Sect. 5.

2 Preliminaries

This paper is based mainly on two existing bodies of research, namely *graph theory* [1] and *constraint programming* [4, 13]. To keep this paper self-contained, we briefly present in this section the basic concepts of them.

2.1 Graph Model

In this subsection we present formally rooted (multi-)labeled graphs used to represent schemas to be matched. More formally, we can define the labeled graph as follows:

Definition 1. *A Rooted Labeled Graph G is a 6-tuple* $G = (N_G, E_G, Lab_G, src, tar, l)$ *where:* $N_G = \{n_{root}, n_2, ..., n_n\}$ *is a finite set of nodes, each of them is uniquely identified by an object identifier (OID), where* n_{root} *is the graph root.* $E_G = \{(n_i, n_j) | n_i, n_j \in N_G\}$ *is a finite set of edges.* $Lab_G = \{ Lab_{NG}, Lab_{EG} \}$ *is a finite set of node labels* Lab_{NG} *, and a finite set of edge labels* Lab_{EG}*. These labels are strings for describing the properties of nodes and edges. src and tar:* $E_G \mapsto N_G$ *are two mappings (source and target), assigning a source and a target node to each edge. And* $l : N_G \cup E_G \mapsto Lab_G$ *is a mapping label assigning a label from the given* Lab_G *to each node and each edge.*

2.2 Constraint Programming

Many problems in computer science, most notably in artificial intelligence, can be interpreted as special cases of constraint problems. *Semantic schema matching is also an intelligent process which aims at mimicking the behavior of humans in finding semantic correspondences between two schemas' elements. Therefore, constraint programming is a suitable scheme to represent the SMP.*

Constraint programming is a generic framework for declarative description and effective solving for large, particulary combinatorial, problems. Not only it

is based on a strong theoretical foundation but also it is attracting widespread commercial interest as well, in particular, in areas of modeling heterogeneous optimization and satisfaction problems. We, here, concentrate only on constraint satisfaction problems *(CSPs)* and present definitions for *CSPs*, constraints, and solutions for the *CSPs*.

Definition 2. *A Constraint Satisfaction Problem* $\boldsymbol{P}$ *is defined by a 3-tuple* $\boldsymbol{P}$*=(X, D,C) where,* $X = \{x_1, x_2, ..., x_n\}$ *is a finite set of variables,* $D = \{D_1, D_2, ..., D_n\}$ *is a collection of finite domains. Each domain* D_i *is the set containing the possible values for the corresponding variable* $x_i \in X$*, and* $C = \{C_1, C_2, ..., C_m\}$ *is a nonempty, finite set of constraints on the variables of X.*

Definition 3. *A Constraint* C_s *on a set of variables* $S = \{x_1, x_2, ...x_r\}$ *is a pair* $C_s = (S, R_s)$*, where* R_s *is a subset on the product of these variables' domains:* $R_s \subseteq D_1 \times ... \times D_r \rightarrow \{0, 1\}$.

The number r of variables a constraint is defined upon is called arity of the constraint. The simplest type is the *unary constraint*, which restricts the value of a single variable. Of special interest are the constraints of arity two, called *binary constraints*. A constraint that is defined on more than two variables is called a *global constraint*.

Example 1. (Map Coloring:) Let us assume we have a map comprising n countries. We want to color each country using one of four colors: *red*, *green*, *white*, or *blue* in a way that no two adjacent countries have the same color. This problem could be formulated as CSP $\boldsymbol{P}$*=(X,D, C)* where: $X = \{x_1, x_2, ..., x_n\}$ represents n countries, $D = \{D_1, D_2, ..., D_n\}$ represents the domains of the variables such that $D_1 = D_2 = ... = D_n = \{red, green, blue, white\}$, and Crepresents constraints which should be satisfied such that $C_{(x_i,x_j)} = \{(v_i, v_j) \in D_i \times D_j | v_i \neq v_j\}$.

Solving a *CSP* is finding assignments of values from the respective domains to the variables so that *all constraints* are satisfied. However, in the schema matching field, we do not need to find any solution but the best solution. The quality of a solution is usually measured by an application dependent function called objective function. The goal is to find such a solution that satisfies all the constraints and minimizes or maximizes the objective function respectively. Such problems are referred to as *Constraint Optimization Problems (COP).*

Definition 4. *A Constraint Optimization Problem* $\boldsymbol{Q}$ *is defined by couple* $\boldsymbol{Q}$ *=(P,g) such that* $\boldsymbol{P}$ *is a CSP and* $g : D_1 \times ... \times D_n \rightarrow [0, 1]$ *is an objective function that maps each solution tuple into a value.*

While powerful, both CSP and COP present some limitations. In particular, all constraints are considered mandatory. In many real problems, there are constraints that could be violated in solutions without causing such solutions to be unacceptable. If these constraints are treated as mandatory, this often causes

problems to be unsolved. If these constraints are ignored, solutions of bad quality are found. This is a motivation to extend the CSP schema and make use of *soft constraints*. A way to circumvent inconsistent constraints problems is to make them fuzzy. The idea is to associate fuzzy values with the elements of the constraints, and combine them in a reasonable way.

A constrain, as defined before, is usually defined as a pair consisting of a set of variables and a relation on these variables. This definition gives us the availability to model different types of uncertainty in schema matching. In [8], authors identify different sources for uncertainty in data integration. Uncertainty in semantic mappings between data sources can be modeled by exploiting fuzzy relations while other sources of uncertainty can be modeled by making the variable set a fuzzy set. In this paper, we take the first one into account while the other sources are left for our ongoing work.

Definition 5. *A Fuzzy Constraint C_μ on a set of variables $S = \{x_1, x_2, ...x_r\}$ is a pair $C_\mu = (S, R_\mu)$, where the fuzzy relation R_μ, defined by $\mu_R : \prod_{x_i \in var(C)} D_i \rightarrow [0, 1]$ where μ_R is the membership function indicating to what extent a tuple v satisfies C_μ. $\mu_R(v) = 1$ means v totaly satisfies C_μ, $\mu_R(v) = 0$ means v totaly violates C_μ, while $0 < \mu_R(v) < 1$ means v partially satisfies C_μ.*

Definition 6. *A Fuzzy Constraint Optimization Problem $\boldsymbol{Q}_\mu$ is a 4-tuple $\boldsymbol{Q}_\mu$= (X, D, Cμ, g) where X is a list of variables, D is a list of domains of possible values for the variables, C_μ is a list of fuzzy constraints each of them referring to some of the given variables, and g is an objective function to be optimized.*

In the following section we shed light on our schema matching framework to determine the scope of schema matching understanding.

3 A Unified Schema Matching Framework

Most of existing schema matching systems deal with the schema matching problem from its point of view, but we need a generic framework that unifies the solution of this intricate problem independent of the domain of schemas to be matched and independent of the model representations. To this end, we propose the following general phases to address the schema matching problem. Figure 1 shows these phases with the main scope of this paper. In the following subsection we introduce a framework for defining different data models and how to transform them into schema graphs.

3.1 Schema Graph

To make the matching process a more generic process, schemas to be matched should be represented internally by a common representation. This uniform representation reduces the complexity of the matching process by not having to cope with different representations. By developing such import tools, schema match

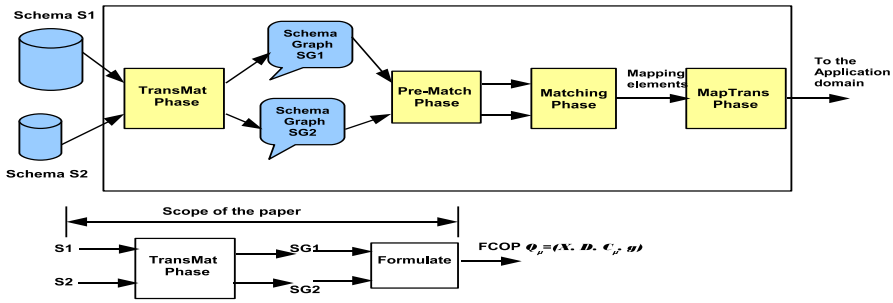


Fig. 1: Matching Process Phases

implementation can be applied to schemas of any data model such as *SQL*, *XML*, *UML*, and etc. Therefore, the first step in our approach is to transform schemas to be matched to a common model in order to apply matching algorithms. We make use of labeled graphs as the internal model. We call this phase *TransMat*; Transformation for Matching process.

In general, to represent schemas and data instances, starting from the root, the schema is partitioned into relations and further down into attributes and instances. In particular, to represent relational schemas, XML schemas, etc. as rooted labeled graphs, independently of the specific source format, we benefit from the rules found in [18, 15, 11]. These rules are rewritten as follows:

- Every prepared matching object in a schema such as the schema, relations, elements, attributes etc. is represented by a node, such that the schema itself is represented by the root node. Let schema S consist of m elements *(elem)*, then

$$\forall\ elem \in S\ \exists\ n_i \in N_G \wedge S \mapsto n_{root},\ s.t.\ 1 \leq i \leq m$$

- The features of the prepared matching object are represented by node labels Lab_{NG}. Let features *(featS)* be the property set of an element *(elem)*, then

$$\forall\ feat \in featS\ \exists\ Lab \in Lab_{NG}$$

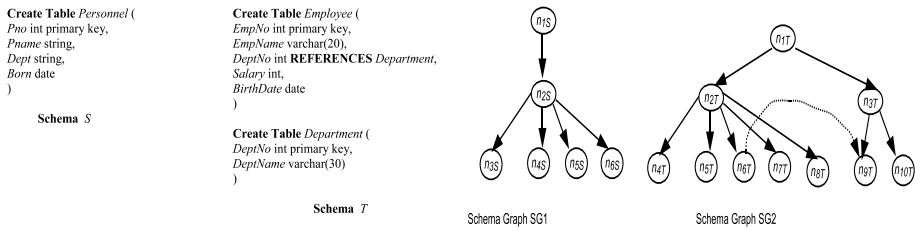
- The relationship between two prepared matching objects is represented by an edge. Let the relationships between schema elements be *(relS)*, then

$$\forall\ rel \in relS\ \exists\ e(n_i,\ n_j) \in E_G \text{ s. t. } src(e) = n_i \in N_G \wedge tar(e) = n_j \in N_G$$

- The properties of the relationship between prepared objects are represented by edge labels Lab_{EG}. Let features *rfeatS* be the property set of a relationship *rel*, then,

$$\forall\ rfeat \in rfeatS\ \exists\ Lab \in Lab_{EG}$$

Example 2. (Relational Database Schemas) Consider schemas S and T depicted in Fig. 2(a) (from [14]). The elements of S and T are tables and attributes. Applying the above rules, the two schemas *Schema S* and *Schema T* are represented



(a) Two Relational Schemas (b) Schema Graphs

Fig. 2: Two Relational Schemas & their Schema Graphs (without labels)

by *SG1* and *SG2* respectively, such that $SG1 = (N_{GS}, E_{GS}, Lab_{GS}, src_S, tar_S, l_S)$ where $N_{GS} = \{n_{1S}, n_{2S}, n_{3S}, n_{4S}, n_{5S}, n_{6S}\}$, $E_{GS} = \{e_{1-2}, e_{2-3}, e_{2-4}, e_{2-5}, e_{2-6}\}$, $Lab_{GS} = Lab_{NS} \cup Lab_{ES} = \{name, type, datatype\} \cup \{part-of, associate\}$. src_S, tar_S, l_S are mappings such that $src_S(e_{1-2}) = n_{1S}$, $tar_S(e_{2-3}) = n_{3S}$ and $l_S(e_{1-2}) = part-of$. Figures 2(b) shows only the nodes and edges of the schema graphs (SG2 can be defined similarly).

In this example, we exploit different features of matching objects such as *name, datatype,* and *type*. These features are represented as nodes' labels. These features shall be the input parameters to the next phase. For example, the *name* of a matching object in SG1 will be used to measure linguistic similarity between it and another matching object from SG2, its *datatype* is to measure datatype compatibility, and its *type* is used to determine semantic relationships. However, our approach is flexible in the sense that it is able to exploit more features as needed. Moreover, in this example, we exploit one structural feature *"part-of"* to represent structural relationships between nodes at different levels. Other structural features e.g. association relationship, that is a structural relationship specifying both nodes are conceptually at the same level, are represented between keys. One association relationship is represented in Fig. 2(b) between the nodes n_{6T} and n_{9T} to specify a key/foreign key relation. Visually, association edges are represented as dashed lines.

So far, recent schema matching systems directly determine semantic correspondences between schemas' elements as a graph matching. In this paper, we extend the internal representation, schema graphs, and reformulate the graph matching problem as a constraint problem.

4 Schema Matching as a FCOP

4.1 Schema Matching as Graph Matching

Schemas to be matched are transformed into rooted labeled graphs and, hence, the schema matching problem is converted into graph matching. Two types of

graph matching exist *isomorphism* and *homomorphism*. In general, a match of one graph into another is given by a *graph morphism*, which is a mapping of one graph's object sets into the other's, with some restrictions to preserve the graph's structure and its typing information.

Definition 7. *A Graph Morphism* $\phi : SG1 \rightarrow SG2$ *between two schema graphs* $SG1 = (N_{GS}, E_{GS}, Lab_{GS}, src_S, tar_S, l_S)$ *and* $SG2 = (N_{GT}, E_{GT}, Lab_{GT}, src_T, tar_T, l_T)$ *is a pair of mappings* $\phi = (\phi_N, \phi_E)$ *such that* $\phi_N : N_{GS} \rightarrow N_{GT}$ *(ϕ_N is a node mapping function) and* $\phi_E : E_{GS} \rightarrow E_{GT}$ *(ϕ_E is an edge mapping function) and the following restrictions apply:*

1. $\forall n \in N_{GS}\ \exists\ l_S(n) = l_T(\phi_N(n))$
2. $\forall e \in E_{GS}\ \exists\ l_S(e) = l_T(\phi_E(e))$
3. $\forall e \in E_{GS}\ \exists$ *a path* $p' \in N_{GT} \times E_{GT}$ *such that* $p' = \phi_E(e)$ *and* $\phi_N(src_S(e)) = src_T(\phi_E(e)) \wedge \phi_N(tar_S(e)) = tar_T(\phi_E(e))$.

The first two conditions preserve both nodes and edges labeling information, while the third condition preserves graph's structure.

Graph matching is an isomorphic matching problem when $|N_{GS}| = |N_{GT}|$ otherwise it is homomorphic. Obviously, the schema matching problem is a homomorphic problem.

Example 3. For the two relational schemas depicted in Fig. 2(a) and its associated schema graphs shown in Fig. 2(b), the schema matching problem between schema S and schema T is converted into a homomorphic graph matching problem between *SG1* and *SG2*.

Graph matching is considered to be one of the most complex problems in computer science. Its complexity is due to two major problems. The first problem is the computational complexity of graph matching. The time required by backtracking in a search tree algorithm may in the worst case become exponential in the size of the graph. The second problem is the fact that all of the algorithms for graph matching mentioned so far can only be applied to two graphs at a time. Therefore, if there is more than two schemas that must be matched, then the conventional graph matching algorithms must be applied to each pair sequentially. For applications dealing with large databases, this may be prohibitive. Hence, choosing graph matching as a platform to solve the schema matching problem may be effective process but inefficient. Therefore, we propose transforming graph homomorphism into a *FCOP*.

4.2 Graph Matching as a FCOP

In the schema matching problem, we are trying to find a mapping among the elements of two schemas. Multiple conditions should be applied to make these mappings valid solutions to the matching problem, and some objective functions are to be optimized to select the best mappings among matching result. The analogy to the constraint problem is quite obvious: here we make a mapping

between two sets, namely between a set of variables and a set of domains, where some conditions should be satisfied to a certain extent. In order to obtain an equivalent constraint problem CP for a given schema matching problem (assuming that schemas to be matched are transformed into schema graphs) we utilize the followings rules:

1. take objects of one schema graph to be matched as the CP's set of variables,
2. take objects of other schema graphs to be matched as the variables' domain,
3. find a proper translation of the conditions that apply to schema matching into a set of fuzzy constraints, and
4. form objective functions to be optimized.

We have defined the schema matching problem as a graph matching homomorphism ϕ. We now proceed by formalizing the problem ϕ as a FCOP problem $Q_\mu = (X, D, C_\mu, g)$. To construct a FCOP out of this problem, we follow the above rules. Through these rules, we take the two relational database schemas shown in Fig. 2(a) and its associated schema graphs shown in Fig. 2(b) as an example, taking into account that $|N_{GS}|(= 6) < |N_{GT}|(= 10)$

– The set of variables X is given by $X = N_{GS} \cup E_{GS}$ where the variables from N_{GS} are called *node variables* X_N and from E_{GE} are called *edge variables* X_E

$$X = X_N \cup X_E$$
$$= \{x_{n1}, x_{n2}, x_{n3}, x_{n4}, x_{n5}, x_{n6}\} \cup \{x_{e_{1-2}}, x_{e_{2-3}}, x_{e_{2-4}}, x_{e_{2-5}}, x_{e_{2-6}}\}$$

– The set of domain D is given by $D = N_{GT} \cup E_{GT}$, where the domains from N_{GT} are called *node domains* D_N and from E_{GT} are called *edge domains* D_E,

$$D = D_N \cup D_E =$$
$$\{D_{n1}, D_{n2}, D_{n3}, D_{n4}, D_{n5}, D_{n6}\} \cup \{D_{e_{1-2}}, D_{e_{2-3}}, D_{e_{2-4}}, D_{e_{2-5}}, D_{e_{2-6}}\}$$

where $D_{n1} = D_{n2} = D_{n3} = D_{n4} = D_{n5} = D_{n6} = \{n_{1T}, n_{2T}, n_{3T}, n_{4T}, n_{5T}, n_{6T}, n_{7T}, n_{8T}, n_{9T}, n_{10T}\}$ (i.e. *the node domain* contains all the second schema graph nodes) and $D_{e_{1-2}} = D_{e_{2-3}} = D_{e_{2-4}} = D_{e_{2-5}} = D_{e_{2-6}} = \{e_{1-2T}, e_{1-3T}, e_{2-4T}, ..., p_{1-2-4T}, ...\}$ (i.e. *the edge domain* contains all the available edges and paths in the second schema graph)(the edge e_{1-2} reads the edge extends between the two nodes n_1 and n_2 such that $e_{1-2} = e(n_1, n_2)$).

Using this formalization enables us to deal with holistic matching. This can be achieved by taking the objects of one schema as the variable set, while the objects of other schemas are the variable's domain. Let we have n schemas which are transformed into schema graphs *SG1, SG2,...,SGn* then $X = X_N \cup X_E$, $D_N = \sum_{i=2}^{n} D_{Ni}$, $D_E = \sum_{i=2}^{n} D_{Ei}$. Another benefit behind this approach is that our approach is able to discover complex matchings of types *1:n* and *n:1* very easily.

In the following subsections, we demonstrate how to construct both constraints and objective functions to obtain a complete problem definition.

4.3 Constraints Construction

The exploited constraints should reflect the goals of schema matching. Schema matching based only on schema element properties has been attempted. However, it does not provide any facility to optimize matching. Furthermore, additional constraint information, such as semantic relationships and other domain constraints is not included, and schemas may not completely capture the semantics of data they describe. Therefore, in order to improve performance and correctness of matching, additional information should be included. In this paper, we are concerned with both syntactic and semantic matching. Therefore, we shall classify constraints that should be incorporated in the *CP* model into: *syntactic constraints* and *semantic constraints*.

Syntactic Constraints

1. Domain Constraint: It states that a node variable must be assigned a value (or a set of values) from a node domain, and an edge variable must be assigned a value from the edge domain. That is $\forall x_{ni} \in X_N$ and $x_{ej} \in X_E \exists$ a unary constraint $C^{dom}_{\mu(x_{ni})}$ and $C^{dom}_{\mu(x_{ei})}$ ensuring domain consistency of the match, where
$$C^{dom}_{\mu(x_{ni})} = \{d_i \in D_{Ni}\},\ C^{dom}_{\mu(x_{ei})} = \{d_i \in D_{Ei}\}$$
2. Structural Constraints: There are many structural relationships between nodes in schema graphs such as:
   - Edge Constraint: It states that if an edge exists between two variable nodes, then an edge (or path) should exist between their corresponding images. That is, $\forall x_{ei} \in X_E$ and its source and target nodes are x_{ns} and x_{nt} $\exists$ two binary constraints $C^{src}_{\mu(x_{ei},x_{ns})}$, $C^{tar}_{\mu(x_{ei},x_{nt})}$ representing the structural behavior of matching, where:
   $$C^{src}_{\mu(x_{ei},x_{ns})} = \{(d_i,d_j) \in D_E \times D_N |\ src(d_i) = d_j\ \}$$
   $$C^{tar}_{(x_{ei},x_{nt})} = \{(d_i,d_j) \in D_E \times D_N |\ tar(d_i) = d_j\}$$
   - $\forall$ two variable nodes x_{ni} and $x_{nj} \in X_N\ \exists$ a set of binary constraints as follows:
     a) Parent Constraint $C^{parent}_{\mu(x_{ni},x_{nj})}$ representing the structural behavior of parent relationship, where
     $$C^{parent}_{\mu(x_{ni},x_{nj})} = \{(d_i,d_j) \in D_N \times D_N |\ \exists\ e\ (d_i,\ d_j)\ s.t.\ src(e) = d_i\ \}$$
     b) Child Constraint $C^{child}_{\mu(x_{ni},x_{nj})}$ representing the structural behavior of child relationship, where
     $$C^{child}_{\mu(x_{ni},x_{nj})} = \{(d_i,d_j) \in D_N \times D_N |\ \exists\ e\ (d_i,\ d_j)\ s.t.\ tar(e) = d_j\ \}$$
     c) Sibling Constraint $C^{sibl}_{\mu(x_{ni},x_{nj})}$ representing the structural behavior of parent relationship, where
     $$C^{sibl}_{\mu(x_{ni},x_{nj})} = \{(d_i,d_j) \in D_N \times D_N |\ \exists\ d_n\ s.t.\ parent(d_n,\ d_i) \wedge paren(d_n,\ d_j)\ \}$$

Semantic Constraints

1. Labeled Constraints: $\forall x_i \in X \exists$ a unary constraint $C^{Lab}_{\mu(x_i)}$ ensuring the semantics of the predicates in the schema such that: if $x_i \in X_N$ and if $x_i \in X_E$:

$$C^{Lab}_{\mu(x_i)} = \{d_j \in D_N |\ lsim(l_S(x_i), l_T(d_j)) \geq t\ \}$$
$$C^{Lab}_{\mu(x_i)} = \{d_j \in D_E |\ lsim(l_S(x_i), l_T(d_j)) \geq t\},$$

   where *lsim* is a similarity function determining the semantics similarity between nodes/edges labels such name and t is a predefined threshold.

The above syntactic and semantic constraints are by no means the contextual relationships between elements. Other kinds of domain knowledge can also be represented through constraints. Moreover, each constraint is associated with a membership function $\mu(v) \in [0, 1]$ to indicate to what extent the constraint should be satisfied. If $\mu(v) = 0$, this means v totally violates the constraint and $\mu(v) = 1$ means v totally satisfies it. Constraints restrict the search space for the matching problem so may benefit the efficiency of the search process. On the other hand, if too complex, constraints introduce additional computational complexity to the problem solver.

4.4 Objective Function Construction

The objective function is the function associated with an optimization process which determines how good a solution is and depends on the object parameters. The objective function constitutes the implementation of the problem to be solved. The input parameters are the object parameters. The output is the objective value representing the evaluation/quality of the individual. In the schema matching problem, the objective function simulates human reasoning on similarity between schema graph objects.

In this framework, we should consider two function components which constitute the objective function. The first is called *cost function* f_{cost} which determines the cost of a set constraint over variables. The second is called *energy function* f_{energy} which maps every possible variable assignment to a cost. Then, the objective function could be expressed as follows:

$$g = min|max(\sum_{setofconstraint} f_{cost} + \sum_{setofassignment} f_{energy})$$

5 Summary and Future Work

In this paper, we have introduced a fuzzy constraint-based framework to model the schema matching problem. Our approach is able to handle uncertainty in schema matching by exploiting fuzzy constraints. Moreover, our framework is generic which has the feature to deal with different schema representations by transforming the schema matching problem into graph matching. Instead of solving the graph matching problem which has been proven to be an NP-complete

problem, we reformulate it as a constraint problem. We have identified two types of constraints syntactic and semantic to ensure match semantics. We also shed light on how to construct objective functions.

The main benefit of this approach is that we gain direct access to the rich research findings in the CP area; instead of inventing new algorithms for graph matching from scratch. Another important advantage is that the actual algorithm solution becomes independent of the concrete graph model, allowing us to change the model without affecting the algorithm by introducing a new level of abstraction.

Understanding the schema matching problem is considered the first step towards an effective and efficient solution for the problem. In our ongoing work, we will exploit constraint solver algorithms to reach our goal.

References

1. R. Babakrishnan and K.Ranganathan. *A textbook of graph theory*. Spring Verlag, 1999.
2. A. Bonifati, G. Mecca, A. Pappalardo, and S. Raunich. The spicy project: A new approach to data matching. In *SEBD*. Turkey, 2006.
3. S. C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured databases on the web: Observations and implications. *SIGMOD Record*, 33(3):61–70, 2004.
4. R. Dechter. *Constraint Processing*. Morgan Kaufmann, 2003.
5. H. H. Do and E. Rahm. COMA- a system for flexible combination of schema matching approaches. In *VLDB 2002*, pages 610–621, 2002.
6. A. Doan. Learning to map between structured representations of datag. In *Ph.D Thesis*. Washington University, 2002.
7. A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. *SIGMOD*, pages 509–520, May 2001.
8. X. Dong, A. Halevy, and C. Yu. Data integration with uncertainty. In *VLDB'07*, pages 687–698, 2007.
9. C. Drumm, M. Schmitt, H.-H. Do, and E. Rahm. Quickmig - automatic schema matching for data migration projects. In *Proc. ACM CIKM07*. Portugal, 2007.
10. F. Duchateau, Z. Bellahsene, and M. Roche. An indexing structure for automatic schema matching. In *SMDB Workshop*. Turkey, 2007.
11. F. Giunchiglia and P. Shvaiko. Semantic matching. *KER Journal*, 18(3), 2003.
12. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In *VLDB 2001*, pages 49–58. Roma, Italy, 2001.
13. K. Marriott and P. Stuckey. *Programming with Constraints: An Introduction*. MIT Press, 1998.
14. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE'02*, 2002.
15. L. Palopoli, D. Rossaci, G. Terracina, and D. Ursino. A graph-based approach for extracting terminological properties from information sources with heterogeneous formats. *Knowledge and Information Systems*, 8:462–497, 2005.
16. M. Smiljanic. *XML Schema Matching Balancing Efficiency and Effectiveness by means of Clustering*. PhD thesis, Twente University, 2006.
17. E. Tsang. *Foundations of Constraint Satisfaction*. Acadimic Press, 1993.
18. Z. Zhang, H. Che, P. Shi, Y. Sun, and J. Gu. Formulation schema matching problem for combinatorial optimization problem. *IBIS*, 1(1):33–60, 2006.

Workshop on E-Learning for Business Needs

May 7th, 2008,
Innsbruck, Austria

Workshop Co-Chairs

Slawomir Grzonkowski, DERI, NUI Galway, Ireland
Tadhg Nagle, DERI, NUI Galway, Ireland
Jonny Parkes, Enterprise Ireland, Ireland

Workshop Program Committee

Juri Luca De Coi, L3S and University of Hannover, Germany
Bill McDaniel, DERI, NUI Galway, Ireland
Katarzyna Stankiewicz, Gdansk University of Technology, ZIE, Poland
Tony Hall, NUI Galway, Ireland
Cristina Chuva, Universidade de Coimbra, Portugal
Giuseppe Bux, Italy

Web2Train: a Design Model for Corporate e-Learning Systems

Katerina Papanikolaou and Stephanos Mavromoustakos

European University Cyprus, Department of Computer Science,
1516 Nicosia, Cyprus
{k.papanikolaou, s.mavromoustakos}@euc.ac.cy

Abstract. Web2.0 has revolutionized the way we use the Web by opening the doors of collaborative learning and direct communication and making the web an open source for learning and exchanging ideas. The aim of this paper is to give Web2.0 its prominent aspect into e-Learning. We present a design model for e-Learning corporate environments that incorporates the social and collaborative aspect of the knowledge transfer process, the quality peculiarities and the training requirements. The introduction of the use of social networks in e-Learning will help improve the effectiveness of e-Learning in reaching its training objectives something that is currently lacking.

Keywords: Web2Train, e-Learning system design, Web2.0 e-Learning, design model.

1 Introduction

In the past few years we have experienced the ever increasing use of e-Learning platforms for business training purposes. The use of e-Learning for business training offers numerous advantages such as ease to set-up, better use of employee time, cost savings, cross-country collaborations, directness and efficiency.

As e-Learning technology is progressing, so should the effectiveness achieved by its use. The virtual environment in order to be effective should find new ways and methods in order to achieve the training objectives. The new medium lacks in certain aspects compared to the traditional training room. It lacks in perceived degree of interactivity, it lacks in communication means, it lacks in creating a sense of community and communication among learners. The employee usually feels isolated behind a screen where hiding and avoiding communication is easy. Learners often lose motivation and self-discipline resulting in lagging behind in their training and failure.

In this paper, we present a design model for corporate e-Learning environments, namely Web2Train. Web2Train incorporates Web 2.0 Tools and is based on three axes; the social and collaborative aspect of the knowledge transfer process, the quality peculiarities and the training requirements.

Our design model is based on our previous work on the inclusion of socio-cultural differentiation in quality-based design and implementation of e-Learning platforms [1, 2].

The use of Web 2.0 tools, such as Blogs and Wikis allow users to express their opinions, communicate and learn from one another in a separate channel than the official e-Learning platform. The learners can use the collectiveness of these tools in order to share information, exchange experiences, monitor theirs and their co-workers progress and address issues of their field. The formulation, use and implementation of these tools follow the quality standards and are diversified according to the learner requirements of age, gender, socio-cultural factors, educational background and intended training objectives. The effectiveness of the tools is evaluated for both trainers and learners at the evaluation phase in terms of achieving the set goals. The learners as part of the learning process should be involved in all phases contributing to the final formulation of the platform. The process is engineered so as to control and vary the degree of involvement of the partners of this learning process according to the intended objective and the skills and competences of the learners.

The rest of the paper is organized as follows: Section 2 describes common Web 2.0 Tools, while section 3 explains the Web2Train model. Finally, section 4 draws the concluding remarks and presents some future work.

2 Web2.0 Tools

One of key factors affecting e-Learning effectiveness was identified to be the lack of interactivity [3, 4, 5]. The isolation of the learner behind a screen leads to declining motivation, loss of interest and failure. In parallel the same need for user interactivity has its effects on the Internet and its use. Internet users realized the asymmetrical flow of information most of them were content consumers rather than content providers. But the Internet can inherently provide access to users both as content consumers and as content creators. The realization of this fact leaded to the establishment of Web2.0. Web2.0 is harnessing the Web in a more interactive and collaborative manner, emphasizing peers' social interaction and collective intelligence, and presents new opportunities for leveraging the web and engaging its users more effectively. Within the last three years, Web2.0 ignited by successful Web2.0 based social applications such as wikis and blogs and application specific software such as my MySpace, Flickr and YouTube, has been forging new applications that were previously unimaginable. In the next section, we present the basic applications enabling social networking and in e-Learning interactivity (Fig. 1).

Weblogs
Wikis

Mashups
Podcasts

Fig. 1. Corporate e-Learning environment

2.1 Weblogs

The Web offered the perfect medium for immediate press releases and quick dissemination of news and information. The publishing on the Web of an individual's diary, with the thoughts, views, comments and positions created the revolution of Web-Logging or Blogging. The term was initially coined by Jorn Barger in 1997 and in its simplest form is a website with data entries, presented in reverse chronological order [6]. This is the outcome of a common need for the sharing and expression of thoughts, criticism and experiences by individuals and was from the beginning one of the strongest tools of the Internet. Blogger is the owner of the Blog and contributing to the Blog is blogging. Each Weblog is part of the Blog-o-sphere. The number of existing blogs is rapidly growing and there seems to be no end in the near future. The new found "democracy" has many supporters but there are concerns regarding the extend of the freedom of speech. Although, the concept of an on-line diary is far from new, despite this their popularity is increasing rapidly. Two are the main reasons for their success as also identified by other researchers [11]:

1. Personalization: the Blog is personal with the authors' views and ideas but others can contribute too. Directness of communication and the ground for discussions, exchange of ideas.
2. Usability: The crucial factor for the success of Web2.0 applications is the ease of use. They are not addressed to people with technical computer programming skills. Everyone is able to contribute to the WWW and become a content creator by clicking on his/her weblog, register and writing with the help of a WYSIWYG-Editor.

The amount of information trafficking in Blogs can be enormous, to avoid this a personal overflow RSS (Really Simple Syndication) technology is used. With the help of XML structure, so called RSS-Reader can provide feeds of subscribed Blogs or other applications. The big advantage is that new information can be read without opening a site. Further, the possibility of using Aggregators and Search functions help to make the information consumption more efficient. The popularity of blogs has raised concerns and legal liabilities regarding the release of confidential information, use of language etc.

2.2 Wikis

A *wiki* is a simple yet powerful Web-based collaborative-authoring (or content-management) system for creating and editing content was introduced by Bo Leuf and Ward Cunningham in 1995 [7]. It lets anyone add a new article or revise an existing article through a Web browser. Users can also track changes made to an article. The term wiki is derived from the Hawaiian word *wikiwiki*, which means fast or quick. The

user-generated online encyclopaedia Wikipedia (http://en.wikipedia.org) is a wiki. Wiki features include:

- *A wiki markup language.* "Wikitext" provides a shorthand way of formatting text and linking external documents and contents.
- *Simple site structure and navigation.* Contributors can create new pages and easily link one page to another. Because a blog site's hierarchy and structure is flat, the navigation is simple.
- *Simple templating.* When a page of wikitext is requested, wiki software converts the wiki markup to HTML and creates links between pages, and wraps this converted content in a template to provide a consistent look to all pages in the wiki.
- *Support for multiple users.* Hyperlinks to pages within the wiki are created automatically. Wiki software makes links based on the page's title, so the author doesn't need to use, remember, or type long URLs to link one page to another within a wiki.
- *Simple workflow.* You can write or edit and publish without editorial oversight or approval. Content in a wiki is managed through change monitoring and the wiki's ability to roll back to a previous version and prevent spam. You can also control user access and privileges, if required.
- *A built-in search feature.* You can search for specific information or topic within a wiki using associated keywords.

Wikis facilitate collaborative work and this is their main difference from Blogs. Due to this collaborative ability wikis can significantly enhance the learning environment.

2.3 Mashups

A Web mashup is a Web page or Web site that combines information and services from multiple sources on the Web. Web mashups can combine information and/or complementary functionality from multiple Web sites or Web applications. A Web mashup server lets you connect, collect, and mash up anything on the Web as well as data on some backend systems. Seven are the major categories: mapping, search, mobile, messaging, sports, shopping, and movies. More than 40 percent of mashups are mapping mashups [8] Several other new-breed Web applications similarly integrate multiple services under a rich user interface.

Typical applications are HousingMaps (http://www.housingmaps.com), that display sales and rental information from a classsifed ads Web site into Google Maps. The Users can view the map enhanced with information on what property is available for rent or sale in the area. Another example is, Fishing Solutions (http://www.fishingsolutions.com.au) that uses Google Maps and information from anglers to help users find fish.

It is easier and quicker to create a mashup than to code an application from scratch in a traditional way. This capability is one of Web2.0's most important and valuable features.

2.4 Podcast

A Podcast as defined in Wikipedia is: “A Podcast is a multimedia file that is distributed by subscription (paid or unpaid) over the Internet using syndication feeds, for playback on mobile devices and personal computers”. At the beginning the multimedia files were equal to audio files (.mp3). Nowadays also Video Files are distributed via Podcasts. Similar to Weblogs the technology behind is rather simple. With the help of RSS, the easy production of my own Podcast and the widespread bandwidth of the internet connection (which make bigger downloads possible) together with the availability of mobile devices Podcasts get their popularity. Some examples that describe the use of Podcast in Education can be found in [9, 10]. It seems that this technology is gaining in popularity.

The need for including social and collaborative elements in e-learning academic environment has already been identified. An example includes the University of Technology of Graz [11].

3 Web2Train Framework

The advent of Web2.0 and the interactivity introduced by its tools cannot be ignored by e-Learning and will be the medium to make on-line learning an efficient and productive process reaching its training goals. The lack of interactivity and the learner isolation this entails has been identified as one of the key reasons for e-Learning failing to reach its objectives. The learning process does not only depend on the instructor delivering the material to the trainee, there is also a social and collaborative element where learners exchange ideas, share resources even help one another that is vital to the success of the process. Up to now is has been very hard to transfer this collaborative element to the electronic environment due the immaturity and lack of user-friendliness of the existing technology. Web2.0 technology comes to fill the gap with user-friendly applications promoting collaborative corporate training.

This lead us to propose a novel design model for e-Learning corporate environments that incorporates the social and collaborative aspect of the knowledge transfer process, the quality issues as identified by ISO9126 and W3C and training requirements.

3.1 Social and Collaborative Aspects

The importance of the development of e-Learning methods and resources addressing these aspects is proving more and more important every day. A number of e-Learning systems are currently available serving the purpose of knowledge transfer and dissemination. These systems however, have been mainly developed and produced for the Anglo-Saxon corporate environment. Large corporations’ personnel usually include employees from different backgrounds and nationalities, while in the same respect these companies may have offices around the world. Human, social and cultural factors such as the learning background, the training needs, the availability and acceptability of the use of various resources change from country to country even

from area to area within the same country and make the need for differentiation even for the same application imperative if we are to support a successful corporate e-Learning environment.

An e-Learning application must be tailored-made for each country, regions in the same country and groups of countries located in the same geographical area. In requirements analysis phase the emphasis should be placed on the specific characteristics of the countries targeted by the e-Learning application. These characteristics include [12]:

- Demographics - It is well known that human behavior varies according to gender and age. Therefore, these issues can significantly affect system design and performance. The Web engineer or project manager must specify and design the e-Learning application based on the targeted population.
- Social characteristics - The analyst/developer must examine the educational system, the literacy level, as well as the languages spoken within the population, in order for the e-Learning application to be designed in such a way that will accommodate diverged features.
- Technical characteristics - Identifying the technology level of each targeted country will help the Web engineer to decide on the type of technology and resources to use. Countries with advanced technologies and high Web usage are excellent candidates for an e-Learning application utilizing the full potential of the technology. On the other hand, countries new to the Internet arena with primitive or basic technologies may need to design e-Learning systems, for low bandwidth networks and reduced communication capabilities.

3.2 Quality Components

Quality factors such as usability, functionality, efficiency, reliability and maintainability as defined in the ISO 9126 standard [13] together with W3C's recommendations and other web engineering quality components as presented in the research arena [14, 15] need to be addressed and incorporated into the framework proposed leading to the successful design and development of quality corporate e-learning systems. Each component is decomposed into several features that must be separately addressed to fulfill specific user needs:

- Usability - Issues like understandability, learnability, friendliness, operability, and ethics are vital design factors that Web engineers cannot afford to miss. The system must be implemented in such a way to allow for easy understanding of its functioning and behavior even by the non-expert Internet employees. Aesthetics of user-interface, consistency and ease-of-use are attributes of easy-to-learn systems with rapid learning curve. E-Learning corporate systems, by keeping a user profile and taking into consideration human emotions, can provide related messages to the user, whether this is a welcome message or a trainee customization page, thus enhancing the friendliness of the system. These training systems must reflect useful knowledge looking at human interactions and decisions.
- Functionality - The system must include all the necessary features to accomplish the required task(s). Accuracy, suitability, compliance, interoperability and privacy are issues that must be investigated in designing an e-Learning corporate system to

ensure that the system will perform as it is expected to. The system must have all the capabilities encountered in the traditional learning process enhanced by the latest high technology features.

- System Reliability - Producing a reliable system involves understanding issues such as fault tolerance, crash frequency, recoverability and maturity. The system must maintain a specified level of performance in case of software faults with the minimum crashes possible. It must also have the ability to re-establish its level of performance. A system must consistently produce the same results, and meet or even exceed users' expectations. The e-Learning corporate system must have correct link recognition, user input validation and recovery mechanisms.
- Efficiency – Trainees expect the system to run in an efficient manner when utilizing an e-Learning environment. System response-time performance, as well as page and graphics generation speed, must be high enough to satisfy learner's needs. Fast access to information must be examined also throughout the system life to ensure that user requirements are continuously met on one hand, and that the system remains competitive and useful on the other.
- Maintainability - Some crucial features related to maintaining such a training system are its analyzability, changeability, stability, and testability. The primary target here is to collect data that will assist designers to conceive the overall system in its best architectural and modular form, for a future maintenance point of view. With the rapid technological changes especially in the area of Web engineering, as well as the rigorous user requirements for continuous Web site updates, easy system modifications and enhancements, both in content and in the way this content is presented, are also success factors for the development and improvement of such system.

3.3 Training Requirements

Transferring the dynamic nature of learning to the new e-Learning environment, maintaining learners' individuality and differentiation according to personal preferences and abilities, as well as motivating and inspiring learners are key factors for the acceptance of the new learning environment [16, 17]. The key factors are identified as follows:

- The identification of learners' needs – The e-Learning environment should be shaped according to the predefined learners' needs and course required pedagogical outcome.
- The structuring of the learning material – The material should be constructed in a way that facilitates the successful transfer of the required knowledge.
- The enhancement of the e-Learning environment – The e-Learning environment can be used either complimentary or in parallel to the real training environment. In either case the e-Learning environment should adhere to the basic mechanisms and functions of the real environment. In the pure distance learning case this enhancement is even more imperative.
- The motivation for trainees' participation – The transferring to the virtual environment is not always straight forward and easy. Trainees are not always willing to use the virtual environment for a number of reasons, such as the

difficulty of the e-Learning tool, the non-intuitive nature of the environment, the provision of reduced interactivity, etc.

- The ability of the e-Learning environment to answer and solve questions and problems. – The e-Learning environment should be able to offer learners a basic problem solving mechanism. Mechanisms such as on-line tutorials, contact with the instructor, reference to useful resources and even access to a technical helpdesk would offer learners support and help.
- The establishment of collaborative mechanisms among trainees – In the virtual environment the trainee can be easily isolated and separated from the rest of the class. This is usually avoided in the real classroom and should be avoided in the virtual classroom too, by organizing and operating in a collaborative basis so that learners can interact and communicate.
- The utilization of the relevant tools (e.g. Web 2.0) for the support of any specific solution – Depending on the targeted audience and the required learning outcome the appropriate tools should be implemented and differentiated accordingly. Tools and components can be utilized to enhance the e-Learning environment more efficiently.
- The right mix of the learning processes implemented – The most important learning processes are identified as follows: analysis, synthesis, reasoning, judging, problem solving, collaboration, simulation, evaluation, presentation and relation. These processes should be used dynamically for constructing the learning scene for each course and trainee.

4 Conclusions

The proposed design model, namely Web2Train incorporates Web 2.0 Tools and is based on three axes; the social and collaborative aspect of the knowledge transfer process, the quality peculiarities and the training requirements. The use of Web 2.0 tools are incorporated in the model in order to facilitate learner-to-learner interaction as well as learner-to-instructor interaction achieving the learning objectives, through collaborative learning. The use these tools will also improve the effectiveness of e-Learning by distilling real classroom practices in the electronic environment. The learners can achieve a sense of belonging, creating a collaborative environment and facilitating the exchange of information and ideas. The learners are also involved in cross-discussion groups to cross-fertilize ideas and build relationships; the idea is to facilitate interactivity and reduce isolation of the learner that is often associated with e-learning.

Future work includes the development of a quality e-learning corporate environment based on the Web2Train model. The system will be implemented in different corporate environments to proof its effectiveness in reaching training objectives.

References

1. Mavromoustakos, S., Papanikolaou, K., Andreou, A. S.: A methodology for developing Quality E-Learning Systems. In: e-Challenges e-2004, (2004)
2. Mavromoustakos, S., Papanikolaou, K., Leonidou, C., Andreou, A. S.: The Development of a Quality e-Learning Environment based on Human, Social, and Cultural factors. In: IEEE 1st International Conference on Information and Communication Technologies, (2004)
3. Elvheim, M.: Supporting group work in distance education. In: The Third International Conference on Extreme Programming and Flexible Processes in Software Engineering (XP2002), pp. 204--205, (2002)
4. Khalifa, M., Lam, R.: Web-based learning: Effects on learning process and outcome. In: IEEE Transactions on Education, 45(2), pp. 350--356, (2002)
5. Mehlenbacher, B. Miller, C. R., Covington, D., Larsen, J. S.: Active and interactive learningonline: A comparison of web-based and conventional writing classes. In: IEEE Transactions on Professional Communication, 43(2), pp. 166– 184, (2000)
6. Paquet, S.: Personal Knowledge publishing and its uses in research, Knowledge Board, (2003)
7. Leuf, B., Cunningham, W.: The Wiki Way. Quick Collaboration on the Web. Addison-Wesley, (2001)
8. Vlist, E.: Professional Web 2.0 Programming,Wrox, (2006)
9. Campell, G. There's something in the Air – Podcasting in Education. Educause review, November/December, pp. 33-46, (2005)
10. Towned, N.: Podcasting in Education, Viewfinder, Nr. 61, i-ii, (2005)
11. Ebner, M.: E-Learning 2.0 = e-Learning 1.0 + Web 2.0?. In: The Second International Conference on Availability, Reliability and Security, pp. 1235--1239, (2007)
12. Andreou, A. S., Mavromoustakos, S. M., Schizas, C. N.: Enhancing the Analysis Phase of E-Commerce Systems Development with Human, Social, Cultural and Organizational Factors. In: 2nd International Interdisciplinary Conference on Electronic Commerce ECOM-02, (2002)
13. International Standards Organization, IEC 9126-1, Software Engineering – Product Quality – Part 1: Quality model, (2001)
14. Olsina, L., Godoy, D., Lafuente, G., Rossi, G.: Specifying Quality Characteristics and Attributes for Websites. In: Proceedings ICSE´99 Web Engineering Workshop, Los Angeles, USA, (1999).
15. Norton, K.: Applying Cross Functional Evolutionary Methodologies to Web Development. In: Proceedings of the First ICSE Workshop on Web Engineering, ACM, (1999)
16. Lytras, M. D., Doukidis, G. I., Skagou, T. N.: Value Dimension of the E-Learning Concept: Components and Metrics. In: Proceedings of the 20th ICDE World Conference on Open Learning and Distance Education, (2001)
17. Lytras, M. D., Pouloudi, A.: E-learning: Just a waste of time. In: Proceedings of the Seventh Americas Conference on Information Systems, (AMCIS 2001), pp. 216--222, (2001)

Evaluate – An Innovative Service for Learning Performance Monitoring in Businesses

Bernd Simon, Kasra Seirafi, Åsmund Realfsen, Mark Strembeck, Gustaf Neumann

Vienna University of Economics and Business Administration,
Institute for Information Systems & New Media

{bernd.simon, kasra.seirafi, asmund.realfsen, mark.strembeck, neumann}@wu-wien.ac.at

Abstract. In this paper we present Evaluate, a platform for learning performance monitoring. Evaluate manages a number of artefacts that can be used to monitor learning performance, like metrics, measures, surveys, questionnaires, or reports. The portal serves various types of users such as business process leaders, monitoring project leaders, learners, and instructors. Based on a powerful modular component framework the processes supported include formative and summative course evaluation as well as sharing of survey instruments. Evaluate's business model is based on a number of advantages, such as a reduced effort for setting up learning performance monitoring projects, low costs for collecting empirical data, and support for benchmarking.

Keywords: Training evaluation, e-learning, learning management, software-as-a-service, performance monitoring

1 Introduction

In today's knowledge-driven society, human resources are increasingly considered as a crucial input factor for high performance. As a result, organisations have started to - implicitly or explicitly - identify *competency objectives* for their key processes. Based on these competency objectives existing employees are trained or new employees are recruited. Therefore, thoroughly planed learning processes and learning management have become important factors to generate competitive advantages. In this context, several studies have recognised corporate learning as an effective way to increase an organisation's overall performance [1].

Today, the predominant technology serving personnel development processes are learning management systems (LMS). An LMS supports an organisation in the administration of learning courses, the registration procedures for learners, and in the distribution of learning materials. In large organisations, LMS are frequently accompanied by specific modules of enterprise resource planning (ERP) solutions providing support for related processes such as performance appraisals.

However, such infrastructures often suffer from specific drawbacks. For example, in most cases LMS are solely focused on managing centralized corporate learning processes, but largely ignore the business processes they are supposed to support. In other words, LMS mainly focus on learning delivery, and lack support of processes that focus on the identification of learning needs, or the subsequent assessment of

learning transfer and performance improvement. Thus, the evaluation of learning processes demands for open systems that support the exchange of standardized measurement items as well as corresponding benchmark data. Moreover, support for the identification and definition of learning metrics and measures that can be used to collect data on intangible assets are also rarely found in current state-of-the-art tools.

In addition, deploying an in-house learning technology infrastructure is cost-intensive. For example, Brandon Hall recently reported costs ranging from $72,370.-(500 users) over 349,414.-(10,000 users) to $ 601,358.- (25.000 users) for installed implementations [2]. Together with the internal resource requirements for implementing LMS, these costs constitute a significant obstacle for professional personnel development. This is especially true for small and medium-sized enterprises which represent the majority of European businesses.

In this paper, we present a new learning technology infrastructure that aims to address the above mentioned shortcomings. The Evaluate platform provides a number of services for learning management and performance monitoring.

The remainder of the paper is structured as follows: In Section 2, we outline a methodological framework for learning and performance monitoring. In Section 3, an example case is introduced which motivates the application of Evaluate in a business setup. Section 4 describes Evaluate in more detail, following the design space framework for learning media. Based on the example case presented in Section 3, we illustrate a concrete implementation of Evaluate in Section 5. After giving an overview of the technological architecture in Section 6, Section 7 concludes the paper.

2 Methodological Framework for Performance Monitoring in Learning Environments

Evaluate provides different components for monitoring the performance of learning activities, the transfer of learning to the work environment, and the subsequent impact of learning on the corresponding business processes.

The PROLIX Methodological Framework for Competency Evidence Elicitation and Performance Monitoring distinguishes between the following five phases [3]: (1) learning process monitoring, (2) learning outcome monitoring, (3) competency monitoring, (4) process performance monitoring, and (5) business performance monitoring.

At the time of writing Evaluate focuses on learning process monitoring: “Learning Performance Monitoring is concerned with tracking critical success factors of learning arrangements such as quality of learning material, empathy of instructors, or service quality. Learning process monitoring enables organisations to influence learning activities and the management of those, so that they produce better learning outcomes and enhanced competencies. Learning process monitoring can be performed at all levels ranging from informal evaluations of small learning activities (e.g. a tutoring session on a specific aspect of a software tool) over training and training programme evaluations, to a corporation-wide assessment of the effectiveness of learning management.” [3]

Monitoring of l*earning processes* can be done for example by performing course evaluations. In Evaluate the "Course Evaluation" service covers formative as well as summative evaluation [4]. *Formative evaluation* is performed in order to influence a learning experience while it is delivered. For instance, a formative evaluation helps a learner to reflect her learning goals before, during, and after a learning activity in order to improve the achievement of learning objectives. To thoroughly support formative evaluation, Evaluate provides a survey tool for carrying out expectation analysis, satisfaction analysis, and transfer analysis. The corresponding questionnaires mainly consist of open questions that help the learner (employee) to reflect on her goals. A s*ummative evaluation* is designed to assess the results of a learning process. In case of a summative evaluation, questionnaires based on standardised measures with closed questions are predominant. Such an instrument supports the collection of data that serves as a basis for target achievement and benchmarking.

3 An Example Case

This section discusses an example case where the adoption of both summative and formative approaches could yield benefits for the respective company. "Soft Solutions Ltd." is an SME providing customized ERP solutions in the print industry. Over the last five years, the company has quickly expanded in Central and Eastern Europe. The 2,000 employee company recruited up to 100 software developers a year, who had to be trained to build special purpose programming skills. In particular, Soft Solutions' Chief Technology Officer, Frida Smith, has identified the need of teaching the company's predominant development process, an approach based on the principles of "Extreme Programming", to new developers.

Those trainings are of paramount importance for the company's effectiveness. As a consequence, Frida decided to install a quality management process for learning activities. Together with the head of personnel development, trainers, and managers of her software development department, the following objectives were identified:

- Inform and track *learning transfer*, since this constitutes the ultimate goal behind the investment
- Measure the *usefulness* of different learning activities, since literature revealed that usefulness (especially in new media environments) constitutes a powerful key performance indicator for corporate learning [5]
- Observe *satisfaction* with learning offerings, since the learning activities are the first deep contact between Soft Solutions and its new employees, and employee satisfaction is an important factor to Soft Solutions' top management.
- Track *performance of instructors* since instructors are considered as a main driver for learning transfer and satisfaction.
- Gather data the *quality of learning materials*, since content quality constitutes a key influence factor for learning success, especially in learning environments where content-based learning is a predominant form of knowledge transfer – like it is the case with the learning offerings of Soft Solution.

In addition, the results should enable follow-up measures. For example, in case a certain maximum threshold is reached (e.g. 80% of course participants agree that

learning service was satisfactory), the team arranging and delivering the learning service is rewarded. On the other hand, measures are required to be taken (e.g. train the trainer activities, improved transfer support) in case the evaluation results for a learning arrangement are below a certain threshold.

4 The Web Portal Evaluate

In this section, we will describe Evaluate in more detail following the design space methodology. Before each of the instantiations of the four design spaces is described, the methodology itself is briefly described.

4.1 Design Spaces of Learning Media

As illustrated in Fig. 1., design choices are grouped into the following four design spaces: business model design space, organizational design space, artefacts design space, and agents design space [6]. A design space includes design issues in a system component that incorporates both a socio-economic as well as a technological perspective of the system.

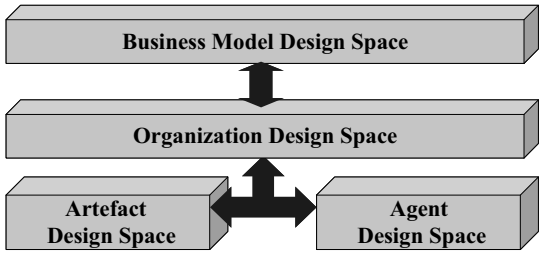


Fig. 1. Design Spaces of Learning Media [6]

Here, a *business model* represents a high-level architecture for product, service, and information flows, including a rough description of the various actors and their roles. In the business model design space, decisions concerning the learning media's position on the educational value chain [7] are taken. Therefore, the target audience needs to be specified (e.g. corporate learners vs. independent learners, or high school teachers vs. faculty of higher education).

Decisions taken in the business model design space will be the basis for the definition of hierarchies and processes in the *organization design space*. In addition, the organizational integration of the learning media and the different institutions has to be defined. In the organization design space, objectives for learning tasks are outlined. Based on these objectives, organizational competences are defined and workflow processes are designed. The resulting processes link learning agents and learning artefacts together.

In the *artefact design space*, decisions about the description of artefacts are made which influence the flexibility of the overall system. For example, the requirements

for the data model of a learning media are determined by the use cases it aims to support, and by the types of learning objects exchanged in learning processes. Selecting an appropriate set of attributes and attribute values for the description of artefacts has a significant impact on the ease of use of the learning media.

The *agent design space* defines the user roles supported by the learning environment. A learning environment can support roles such as learner, course instructor, teaching assistant, evaluator, administrator, etc. Registration and authentication processes also need to be defined here.

4.2 Artefacts

In Evaluate everything is centred on performance monitoring projects. A performance monitoring project represents a real world monitoring activity that has a beginning and an end time. It uses surveys in order to generate reports that provide the data for subsequent performance improvement measures, for example in the context of a course on software development methodologies.

Reports consist of multiple report items. For example, a report item can represent a certain metric that aggregates data from a particular data series. A data series is collected via surveys. When aggregated to a report item it can be displayed in multiple formats (e.g. via absolute values, or percentages).

Metrics and measures are key artefacts of Evaluate. A metric can result in different types of values, e.g. actual values, benchmark values, target values, or alarm values. Metrics based on empirical data sources constitute so-called qualitative metrics, like satisfaction with trainer, or usefulness of a learning activity for example. Quantitative metrics, on the other hand, refer to figures gathered by observing a process either financially or operationally - examples are total inbound costs, time-to-market, training budget, or number of courses attended.

In order to empirically capture data about qualitative metrics, validated measures are needed. Those measures consist of different measure items, which - in aggregated form - satisfy a particular information need manifested in the qualitative metric. Such measures need to be well defined in order to ensure precise measurement and comparability. A measure usually consists of an assertion (e.g. “I was able to transfer the knowledge gained in the course into my work environment”) and a predefined scale for answering (e.g. a 5-item Likert scale ranging from strongly agree to strongly disagree).

In Evaluate, questionnaires are used to collect measures. In addition to measure items, a questionnaire of a survey may also include additional questionnaire items. For example, open questions might be used for allowing respondents to express their opinion with a maximum amount of freedom. Other questionnaire items might be used to capture demographic information (e.g. sex, age) or other data to filter the results according to different target groups.

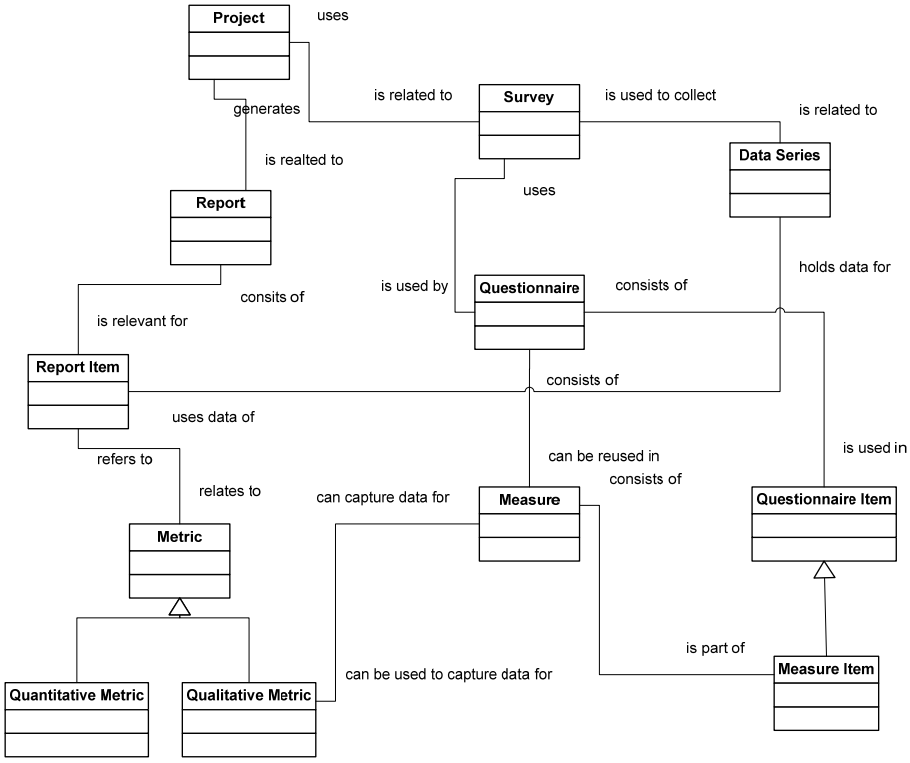


Fig. 2. Key Artefacts of an Evaluate Service

4.3 Users and Roles

From a user-perspective, Evaluate provides a Web-based portal which is managed by a portal management and is used by companies. From our experiences we assume that most services will use questionnaires to collect information from users (like employees, course participants, etc.). Thus, those "Survey Participants" also have to be represented in the model.

Portal Management
The *Portal Administrator* is responsible for general user administration and system operation. The *Portal Quality Manager* is responsible for the availability of quality metrics and measures which are used by the different Evaluate services.

Business Users
The *Business Process Leader* is responsible for ("technical") output and ("financial") outcome of the learning process. Usually she is the head of the business unit, in which the monitoring project is implemented. She has to coordinate the monitoring process

with the monitoring project leader. Additionally she has to ensure that all kinds of infrastructure (e.g. office space, hardware, IT tools, knowledge) are available to the team in order to achieve the best possible performance.

The *Monitoring Project Leader* can be an internal or external person supporting the deployment and maintenance of an Evaluate monitoring service. A person taking the role of monitoring project leader advises the business process leader in selecting metrics, choosing measures, designing reports, and creating follow-up action plans.

The *Learning Employee* - or Learner in short - is an actor in the business process that aims at acquiring knowledge, skills or a change in attitudes by getting involved in formal or non-formal learning activities.

The *Instructor* is the process leader of a particular learning activity. The instructor stimulates learning and hereby changes attitudes, abilities, or behavior of the learning employee.

Survey Participant

An *Anonymous Survey Participant (SP)* is able to fill in a questionnaire without any need for authentication. An *Authentication SP* is a registered and authenticated user. A *Self-Registered SP* is a mixture of the two types above. Here a SP has to create an account before filling in a survey. As the login data are not authenticated, the user can still remain anonymous.

4.4 Processes and Hierarchies

Formative and Summative Course Evaluation

The formative evaluation workflow of Evaluate is defined as follows: Once a learning employee books a learning activity, she first fills in an expectation analysis questionnaire, where she is asked to express her transfer intentions, for instance. Other potential questionnaire items are: related organisational and individual goals, or motivation to participate in the training. Subsequently, this information is forwarded to the respective instructor in order to properly adapt the corresponding training activity.

In case of an electronic learning environment, this information can be used to personalize the learning experience. In addition, the collected expectations are also forwarded to the instructors involved in the learning activity. After the learning activity is completed a learning wrap-up questionnaire is presented to the employee, where transfer intentions are again reflected. After a while - usually between 4 to 20 weeks - the learning transfer is evaluated by performing a transfer evaluation. The results of this process are documented in a report. This report is again forwarded to the respective manager and to the respective tutor, e.g. in case follow-up sessions are planned.

The “learning wrap-up” and “transfer evaluation” phases of the process sketched above can also be combined with a summative evaluation. A summative evaluation is designed to assess the results of a learning process. An ex-post assessment of learning activities evaluates the effectiveness of a certain learning activity based on the

identified metrics and corresponding measures. Kirkpatrick [8, 9] suggests to evaluate training on four levels: learner's satisfaction (reaction), learning outcome (change of attitudes, skills, knowledge), change in behaviour (transfer), and business impact (results).

The collected data is aggregated in a report and forwarded to the stakeholders. Company-internal and external benchmarks visualized in the scorecard help to interpret the benchmarks. Follow-up actions are defined in case the learning activity has not produced satisfactory results.

Sharing of Measures

Many Evaluate services rely on validated measures for capturing data for qualitative metrics. Driven by this demand Evaluate offers a set of standard measures that are accessible for all participating companies on a "Public Measure Space". In addition to the standard measures, companies have the possibility to create and manage their own customized measures, for example by adapting standard measures to their specific needs. Such customizations create a measure in a closed "Company Spaces". In general, this results in two different options:

- A company creates new measure which can be added to surveys of monitoring projects.
- A company "imports" a measure from the public space and modifies it. Afterwards, the measure is added to surveys of monitoring projects.

In addition to importing measures, Evaluate enables users to "publish" user-generated measures from the company space to the public space. This way, new user-generated measures can be made accessible to other users. A motivation for sharing measures is the possibility of using public measures in related performance monitoring projects for benchmarking.

In order to ensure that only high quality measures are distributed via Evaluate, published measures have to undergo a quality check before being published. Therefore, Evaluate divides public spaces in a user-generated part and a standard part. Published measures are first stored in the user-generated measures section. The portal quality manager then approves and regulates the possible incorporation of a specific user-generated measure into the standard section.

4.5 Business Model

Evaluate provides a web-based interface that enables companies to perform high quality learning and competency monitoring. In particular, Evaluate aims to address shortcomings of current state-of-the-art learning management solutions. For example, Evaluate aims to provide:

- reduced effort for setting up learning performance monitoring projects through reuse of measures and questionnaires;
- reduced effort for collecting relevant data;
- increased data quality through validated measures;
- straightforward interpretation of reports as they can be enriched with benchmark data

- an hosted web-based service that is instantly available for "play" (no set-up costs)

At the time of writing we foresee a revenue model that is based on advertising and service fees. Service fees are charged per performance monitoring project. Advertising is foreseen for the publicly accessible measures that can be used for free.

5 How Evaluate Can Serve Soft Solutions Ltd.

Following the example case presented in Section 3, our CTO Frida asked her assistant to use Evaluate's "Course Evaluation" (CE) service for the monitoring process of the software development courses ("Extreme Programming for Beginners"). Every time a new group of learners is assigned to the introductory courses, a new CE-monitoring project is initiated.

Participants of the respective courses are then invited to fill in different questionnaires: one addressing expectations for the course (survey 1: before the course), a second regarding the learner's satisfaction with the course (survey 2: directly after the course) and a third concerning transfer into the workplace (survey 3: eight weeks after the course). Below, we address these surveys in more detail:

(1) *Survey* "Expectation Analysis"
uses a *Questionnaire* to collect data using the following items:
"Expectations about learning content" (free text), and
"Intensions to transfer learning to workplace" (free text)

Please note, that the questionnaire items of the first survey do not capture data for specific metrics. With these questions we mainly want to make course participants reflect on the training and to provide information for the trainer.

(2) *Survey* "Satisfaction Analysis"
uses a *Questionnaire* to collect data on
"Instructor competency" (5-item measure),
"Instructor learning techniques" (3-item measure),
➔ Capture data for m*etric* "performance of instructor"
"Learning material" (5-item measure),
➔ Capture data for *Metric* "quality of learning material"
"Service quality" (3-time measure),
➔ Capture data for m*etric* "satisfaction with learning offerings"

(3) *Survey* "Transfer Analysis"
uses a q*uestionnaire* to collect data on*:*
"Applicability of learning" (3-item measure)
➔ Capture data for m*etric* "usefulness"
"Actual transfer of learning" (5-item measure)
➔ Capture data for m*etric* "learning transfer"

During and after the evaluation process, the Course Evaluation Service will then provide Frida's team with a wide range of possibilities, for example:

- Trainers of a course receive the results of the expectation analysis in order to improve course preparation;
- Reports on transfer success as well as usefulness and learner satisfaction;
- Reflection for participants when filling out the questionnaires;
- Internal benchmarking: all courses of a year are compared and high performing courses are identified as best practises.

6 Architecture

Evaluate is based on the open source Web application framework OpenACS[1]. Despite being an open source framework, OpenACS is developed and maintained by a community that primarily consists of professional software developers [10]. Figure 3 shows the conceptual structure of the Evaluate platform.

Evaluate uses AOLserver[2], a high performance Web server developed by NaviSoft, later acquired by America Online (AOL). In addition to commercial sites, AOLserver is used as the Web server for a number of non-commercial projects, as the DotLRN LMS framework [11] for example. OpenACS provides native Tcl [12] support for server side scripting.

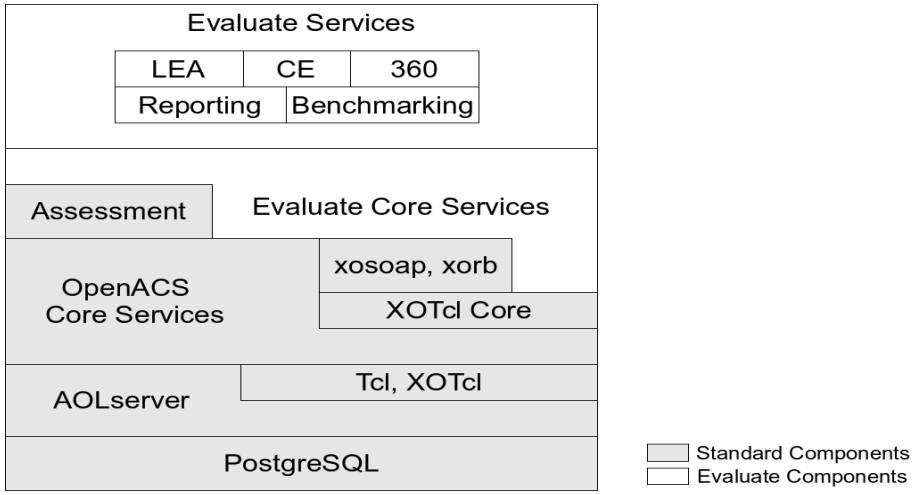


Fig. 3. Conceptual structure of Evaluate

The different Evaluate components are implemented in XOTcl (Extended Object Tcl). XOTcl is a fully dynamic object-oriented programming language [13, 14] that can be loaded in every Tcl compatible environment. Moreover, XOTcl can directly be integrated with arbitrary software components providing C or Tcl linkage, as

[1] http://openacs.org/
[2] http://aolserver.com/

AOLServer for example. The xosoap and xorb components provide functions that enable the communication with other applications that provide SOAP bindings.

The core services of Evaluate extend the functionality of the underlying components. Information about different types of Evaluate entities, e.g. companies, metrics, roles, or permissions, is captured via special purpose XOTcl objects. Each of the different Evaluate services, as "Course Evaluation" for example, is implemented as an own customizable software component, and each of these services seamlessly integrates with other Evaluate services.

The assessment component, shown in Figure 3, is an OpenACS component which can be used to perform surveys or tests of anonymous as well as registered users. The assessment component can be integrated with Evaluate core services to enable the administration of metrics.

As the Evaluate core services are based on OpenACS, each company can autonomously maintain its own company space, and enable or disable Evaluate services within its company space. Moreover, customization of the user interface is supported via a set of predefined templates.

7 Conclusion and Outlook

In this paper we presented the Evaluate platform for learning performance monitoring. Evaluate aims to reduce the gap between corporate learning offerings and knowledge transfer at the work place. We sketched the methodological framework behind Evaluate and described Evaluate along the different dimensions of the "design spaces for learning media" approach. Moreover, we motivated the application of Evaluate in a business context on an example case which uses the "Course Evaluation" service of the Evaluate platform.

Beyond "Course Evaluation", Evaluate supports a number of other performance monitoring methods. For example, we are currently working on the Learning Environment Assessment and the 360-Degree Assessment component. At the same time we continue to develop additional measures and extend the collection of corresponding benchmark data. In our future work, we plan to investigate critical success factors for the design of performance monitoring services as well as their impact on the adoption of such services. The benchmarking support foreseen in Evaluate shall help us to investigate novel methodologies of learning performance monitoring, such as control groups.

8 References

1. Zwick, T.: Weiterbildungsintensität und betriebliche Produktivität. Zeitschrift für Betriebswirtschaft 74 (2003) 651-668
2. Brandon Hall Research: Pricing Trends in Learning Management Systems, http://www.brandonhallnews.com/promos/5feb8.htm (2008)
3. Simon, B., Ackema, R.: Extended Methodological Framework for Competency Evidence Elicitation and Performance Monitoring (D7.3), Vienna, Antwerp (2007)
4. Caffarella, R.S.: Planning Programs for Adult Learners - A Pratical Guide for Educators, Trainers, and Staff Developers. John Wiley & Sons, San Francisco, CA (2002)
5. Alliger, G.M., Tannenbaum, S.I., Bennett, W.J., Traver, H., Shotland, A.: A Meta-Analysis of the Relations among Training Criteria. Personnel Psychology 50 (1997) 341-357
6. Guth, S., Neumann, G., Simon, B.: UNIVERSAL - Design Spaces for Learning Media. In: Sprague, R.H. (ed.): Proceedings of the 34th Hawaii International Conference on System Sciences. IEEE, Maui, USA (2001)
7. Oblinger, D., Kidwell, J.: Distance Learning - Are We Being Realistic? Educause Review 35 (2000) 31-38
8. Kirkpatrick, D.L.: Techniques for evaluating training programs. Journal of ASTD 13 (1959) 3-9
9. Kirkpatrick, D.L., Kirkpatrick, J.D.: Evaluating Training Programs: The Four Levels. Berrett-Koehler Publisher, Berkley, USA (2005)
10. Demetriou, N., Koch, S., Neumann, G.: The Development of the OpenACS Community. In: Lytra, M., Naeve, A. (eds.): Open Source for Knowledge and Learning Management: Strategies Beyond Tools. Idea Group Publishing (2006)
11. Alberer, G., Alberer, P., Enzi, T., Ernst, G., Mayrhofer, K., Neumann, G., Rieder, R., Simon, B.: The Learn@WU Learning Environment. In: Uhr, W., Esswein, W., Schoop, E. (eds.): Tagungsband Wirtschaftsinformatik 2003, Vol. 1. Physica Verlag, Dresden, Germany (2003) 593-612
12. Ousterhout, J.K.: Tcl and the Tk Toolkit (1994)
13. Neumann, G., Zdun, U.: XOTcl, an Object-Oriented Scripting Language. Proceedings of the 7th USENIX Tcl/Tk Conference, Austin, Texas, USA (2000)
14. Neumann, G., Zdun, U.: XOTcl - Extended Object Tcl, http://media.wu-wien.ac.at/ (2008)

VIEW: A Framework for Interactive eLearning in a Virtual World

Kamal Bijlani, P. Manoj and P. Venkat Rangan

Amrita eLearning Research Center
Amrita University, India
{kamal,manoj,venkat}@amrita.edu

Abstract. In the field of eLearning systems, several systems have been implemented in order to create a virtual university or virtual training center. By connecting several virtual universities together, we extend the concept of an online university setting to a virtual world where there is richer interaction between the players. In this virtual world, several universities can interact, the best instructors deliver the lectures, and after the class, the interaction and learning continue in a social online chat room where members from all the universities can participate. We propose a generic eLearning framework called A-VIEW (Amrita Virtual Interactive eLearning World) for educational, corporate, and other applications. The A-VIEW project is a joint venture of Amrita University, ISRO (Indian Space Research Organization), TIFAC India, and the Indo-US eLearning Initiative.

Keywords: eLearning, Virtual University, A-VIEW, Interactive Lectures, Knowledge Café, Knowledge Library

1 Introduction

"…Through the sharing of knowledge and experience, new horizons will open up in the highest realms of science, technology, and the corporate world…"
— Chancellor, Amrita University

In this technological era, eLearning is an emerging technique with great potential. The main reasons for the need for eLearning are: rapid changes in technology; high demand for skilled practitioners; and the dearth of experts in several fields of research and knowledge. The availability, scalability, and inter-operability of information technologies make eLearning a viable alternative to traditional teaching methodologies.

The basic idea of a virtual university is that a particular course is taught with an online system. There have been some definitions and models [1], [2], [3] of a virtual university. [14], [15] show that several universities have employed this model and are offering a variety of courses.

We have been exploring, experimenting, and applying eLearning in various educational courses. Initially, we were providing some of the features of what is

known as a virtual university. As part of our "Indo-US Initiative in eLearning" program, various courses were conducted by seasoned faculty from the USA in India. These courses were transmitted simultaneously to multiple universities. During this time, we updated some of our initial ideas of a virtual university. We found that by networking a set of virtual universities and providing them with tools and paradigms, we could create an interactive virtual world. This allows the students of all the participating universities to get a high-value technical education irrespective of the location of their campus.

In this paper, we present a framework called A-VIEW (Amrita Virtual Interactive eLearning World) which can be used to provide a rich interactive social environment for eLearning in educational, corporate, and other applications. We are using A-VIEW in various contexts: for example, teaching undergraduate courses in various Amrita campuses at the same time; imparting simultaneous corporate training to the branch offices of a company; teaching yoga to various centers, academia, and corporations; broadcasting a live series of seminars from one place to others. In this paper, we focus on the general architecture and applications of the A-VIEW framework; and not on the technical details.

In the A-VIEW model, one expert instructor can teach several groups of students. The expert instructor can teach from one location, and the teaching assistants provide support at the other locations. As a result, the leverage of skilled resources is increased.

2 A-VIEW Framework

We are presenting the A-VIEW framework as a generic paradigm that provides Live Interactive Distance Learning to several centers and a supporting Knowledge Environment for the student to continue the learning at a self-controlled pace. In the A-VIEW framework, we target a set of nodes or places that are physically distant to be connected by an eLearning network. The A-VIEW framework can be roughly divided into two parts. The first part consists of a set of tools that are provided for Live Lectures. The assumption is that the class can be transmitted simultaneously to a set of nodes. Several tools are provided for the instructor to use in a live lecture. The instructor has the ability to interact with the students, and there are tools for testing the awareness and basic understanding of the students during the class. During the live lecture, each receiving node becomes a live virtual university. The instructor can interact with each location, and they can share resources. The instructor node and all the student nodes together form the virtual world during the live lecture.

As an example, Figure 1 shows A-VIEW running two separate live classes on two subjects: Topic A and Topic B. For Topic A, the main instructor is at Location 1, and the Teaching Assistant (TA) provides support at the other locations. Similarly for Topic B, the main instructor is at Location 3. In this way, there can be any number of live classes that are being conducted. For each class, there is one main instructor, and for each distant location, there is a teaching assistant.

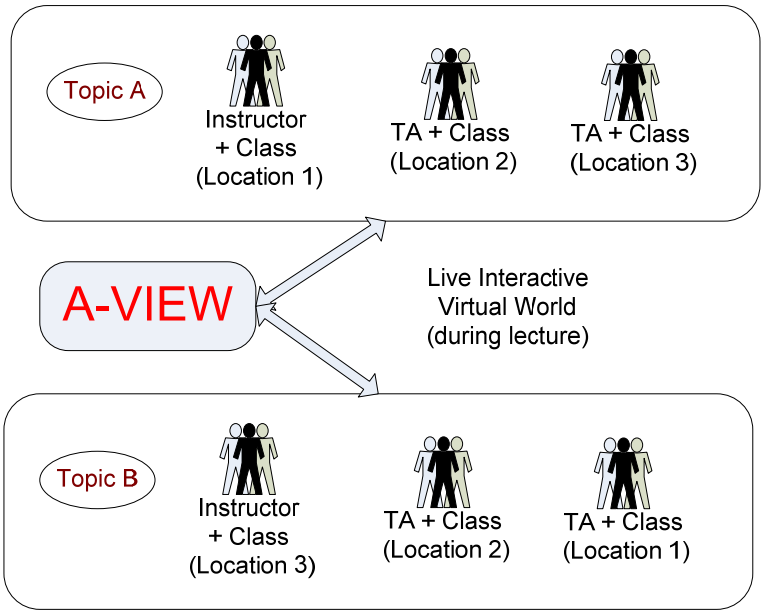


Fig. 1. A-VIEW: Live Lecture with various topics

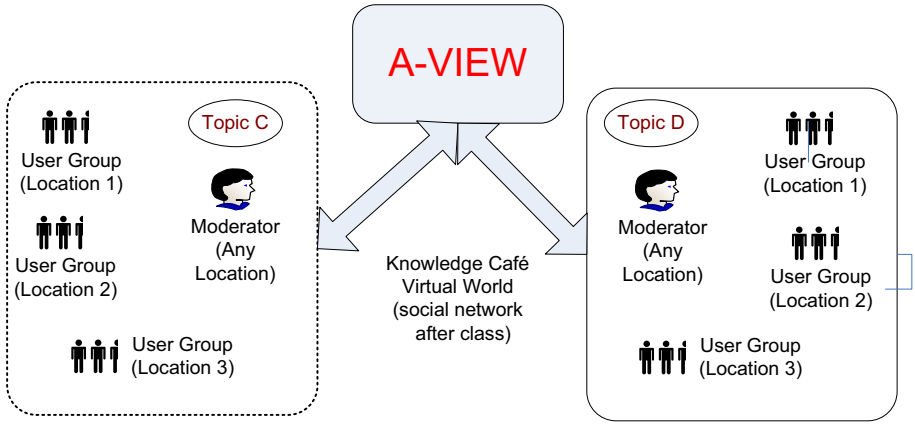


Fig. 2. A-VIEW: Knowledge Café with various topics

After the live lecture is over, the second part of A-VIEW provides an online social environment with various tools for the students to interact, learn, and disseminate information to each other. This part is called the Knowledge Café and Knowledge Library. Here members of all ranks from all centers can log into a common

environment and share resources. This creates a virtual world of all the participating centers. As an example, Figure 2 shows that groups of users from various locations are logged in. Moderators of the Knowledge Café can work in shifts, and can be from any location. However, for the live lectures the main instructor is usually fixed at one particular location.

2.1 Live Lectures to Multiple Centers

As instructors move toward the online learning environment enabled by the various communication networks, they seek tools, processes and teaching methods that are equal to or better than those they use in the traditional teaching environment. Although the tools used in the online system are different, they have to accomplish the same functionalities. A student in any location should able to see and hear the instructor properly, and also be able to review any documents, applications or graphic explanations being offered by the instructor at the physical teaching environment. Even though there may be multiple distant classes and one main instructor, any student in any class should be able to ask questions of the instructor.

Based on our previous work, we have designed the basic video and audio of the instructor [4], [5]. In addition, A-VIEW provides a variety of tools for the instructor to convey information to the student. These include the following:

Whiteboard: Similar to the blackboard in a traditional classroom, the instructor can write on a board, or a tablet, etc. The students can see the instructor's display and also modify, if given control by the instructor. Multiple devices are supported.

Chat: The multi-user chat window is normally used by the students to post questions to the instructor, without disturbing the flow of the class. In this way, the instructor can choose to answer the questions at an appropriate time.

Document Sharing: The instructor can open various types of documents like PowerPoint, PDF, HTML, JPG, etc. and these can also be seen by all the student nodes. A compressed form of this document is sent to the student node. As the instructor moves around in a document, the same part of the document is displayed for all the student nodes. This mode has very little impact on the bandwidth, since only a marker in the document is transferred over the network.

Window Sharing: In this mode, whatever application is running in a certain window on the instructor side can be shown in a window on the student side. For example, the instructor could be teaching VLSI, Marketing or Graphics. The instructor is able to run an application, and the students see the changes as they happen. The entire contents of the instructor's application window are sent "live" to the student nodes. Depending upon the settings used for the transmission of this window on the instructor side, this activity can be bandwidth intensive.

Application Sharing: The same application is open on the instructor side and the student side. The instructor can pass control for editing or running this application over to the student side.

Fig. 3. Large Multiple Displays

Large Multiple Displays: All the above facilities can be shown on large LCD displays at the same time. In this way, the students have access to all the facets of a live lecture. The students can pay attention to the part of the display that they need to in order to understand the subject matter.

Mobile Lectures: The instructor can teach the class using a laptop with an attached device to use the Whiteboard facility. Several devices are supported. So, the instructor could use a tablet PC or attach an external device like a Wacom tablet.

2.2 Live Lectures with Interactivity

Research shows that individuals vary greatly in how they learn, and their learning styles depend on various demographic and psychological factors. In an eLearning environment, the instructor/trainer is in one location and the students in another. The students may be having problems listening or understanding the subject material. The instructor may not be fully aware of the problems of the student environment. For example, in one case study we found that students had difficulty in understanding the accent of the instructor. However, the students did not complain. In such a situation, it is important to have some feedback system so that the instructor-student communication mechanism can be evaluated [7], [8].

For A-VIEW, we have designed a set of tests and quizzes that are automatically presented to all the students in each class. It is assumed that all students can log into the class. The quiz has several questions to test a wide array of variables including: Verbal Communication, Basic Comprehension, Concept Insight, and Grasp of Previous Lectures. The answers are quickly presented to the instructor. In this way, the instructor is able to get an idea of the performance of the students, and acute problem situations can be diagnosed. Our investigations indicate that students who have quizzes in the class have better overall performance than students who do not. By keeping a history of the regular quizzes, the student's participation in the course, and the final grades of the students, we are working to determine if there are any underlying correlations between these variables.

2.3 Knowledge Café

For some time, the students had only live lectures. In this situation, many students felt that they did not have a way to connect to the original instructors to continue their study, or access to the relevant materials to go deeper into the particular subject [1],

[6]. To alleviate these issues, we introduced the concept of the Knowledge Café and Library. The students, teaching assistants, and instructors typically enter the Knowledge Café after the live class is over. For a specific course, the Knowledge Café consists of a 24-hour Online Chat Room. Here the members of one center can interact with other members from all other centers/campuses.

In the Knowledge Café, the teaching assistants act as the moderators of the chat room. The moderator is on duty for some fixed times every day. Since the moderator can be from any center/campus, the load on any one center is reduced. It is the duty of the moderator to check the chat history for each day, and to present the useful points to the instructor of the course. The instructor in turn reviews the notes, edits them, and can add them to the Knowledge Library. The members also have access to a discussion forum which is used to converse about various topics associated with the particular course. Figure 4 shows a screen shot of the Knowledge Café. We can see that there are members logged in from various universities. In this way, the Knowledge Café can also be used for joint collaboration on any project.

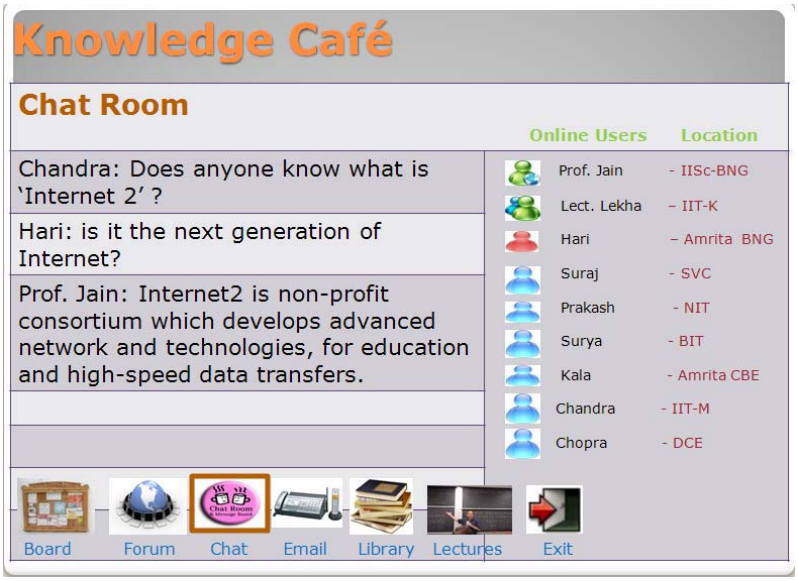


Fig. 4. Screen Shot of Online Chat Room in Knowledge Café

2.4 Knowledge Library

The Knowledge Library consists of the various lecture archives for the specific course and related materials. All the live lectures are recorded and are available to the students. The lectures can be searched by various criteria like the instructor, date, school, etc. Also, the lectures can be searched by specifying a topic keyword.

The members have access to various resources for the course like FAQs, Related Documents, Relevant Sites, Tutorials, etc. and a search facility. We are also keeping an audit trail of how the members use the system, and what patterns, if any, they exhibit. For example, we are using Artificial Intelligence pattern-matching techniques to analyze the common problems faced by the students. We hope that by identifying the main common problems that the students face in a course, we will find some associated patterns, and thereby take appropriate steps to improve the future classes. Automatic emails and reminders are sent to the members about their courses.

3 Application of A-VIEW to Universities

3.1 Teaching at Amrita Campuses

As shown in Figure 5, the Amrita Campuses at different locations are connected via VSAT satellite network. This network from ISRO connects the teaching node and the student nodes in the network. The A-VIEW system is implemented using this network. A-VIEW creates an interactive, multi-disciplinary, multimedia, virtual world without geographical limitations. Now the students of Amrita can not only attend, but also interact during lectures that are taking place at any of the Amrita campuses. We have delivered various courses through A-VIEW in a number of significant fields including engineering, medicine, management, etc. In addition to Amrita campuses, A-VIEW is also being utilized for community services like Village Resource Centers and is being extended to remote schools.

Fig. 5. Satellite Network (VSAT) in Amrita

3.2 EDUSAT connecting Indian Universities

Figure 6 shows that about 40 universities in India are connected via the EDUSAT satellite network from ISRO. EDUSAT is the first Indian satellite exclusively used for serving the educational sector. The figure shows the participating universities. Through EDUSAT, A-VIEW is already being used to teach various courses to engineering colleges across the country.

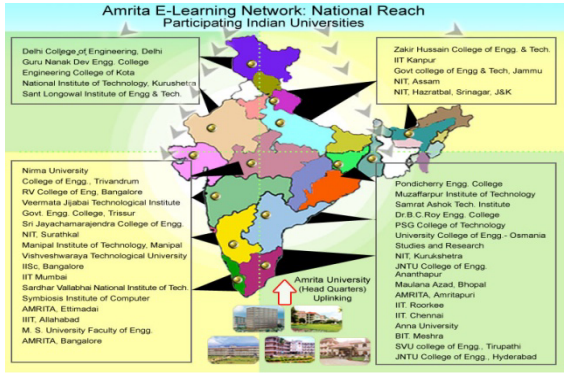


Fig. 6. Indian Universities connected by Satellite (EDUSAT)

Professor name	University	Topic
Dr. Eric A Brewer	Prof. at UC Berkeley, Founder of Inktomi Corp. & Federal Search Foundation.	Technologies for Emerging Regions in Societal Transformation
Dr. N.Narayana Rao	Prof. University of Illinois at Urbana-Champaign	Engineering Electromagnetism
Dr.Ponisseril Somasundaran	Director, NSF/IUCR, Prof. Columbia University	Challenges and opportunities for Nano technology
Dr. Ashok Agrawala	Prof. of CS, Director of the MIND Lab, University of Maryland	The Emerging Technologies at the MIND Lab

Fig. 7. Lectures on EDUSAT

Under the "Indo-US E-learning initiative" program, several experts have come to India and taught classes and led research programs. 26 U.S.-based Universities including Harvard, Yale, Princeton, etc. are participating in the initiative, as well as 41 Universities from India including IITs, NITs, Amrita, etc. This program is helping to improve the quality of the students and also the instructors in India. Hundreds of

lectures have been broadcast (http://amritauniversity.info/). Some of them are listed in Figure 7.

3.3 Corporate World

In the corporate world, we are using A-VIEW for training. In principle, this is similar to teaching at the university [11], [12], [13]. We have found that depending upon the applications, the emphasis and usage of the tools varies to some extent. For example, we are finding that the Corporate Training applications use the 'Shared Window' tool to exhibit various demonstrations. Although this increases the required bandwidth, the system is able to show more videos and training materials.

Fig. 8. Lecture on Mobile Phone

Employees from separate branch offices are able to talk to each other using the Knowledge Café. Compared to the university students, there is much more communication between the employees, and an ability to assist each other. Usage and communication is better.

The role of the mobile phone is especially important in the corporate world [9], [10]. In the university, laptops are common, but in the corporate world, the executives typically carry mobiles, PDAs, etc. Figure 8 shows that the video of the class can be seen on the mobile phone. Due to the limit of the screen resolution on the mobile, some of the features and functions are provided in a simpler manner on the mobile. For the mobile, it is an engineering challenge to design the user screens so that they provide the necessary functionality and still are simple and easy to use. However, the role of mobiles is critical in the corporate world, and for executives it is necessary to be able to continue their eLearning training whenever they have time.

3.4 Telemedicine

Amrita Telemedicine is a fully integrated telemedicine solution for linking various clinics and hospitals located in various remote areas of India with Amrita Institute of Medical Sciences and Research Centre (AIMS) at Cochin. A-VIEW helps the patients at different locations to get consultation with the doctors. It also helps the doctors at remote areas to seek expert opinions on treatment, medicine, etc. from expert doctors at AIMS. It also allows doctors to share data with one another.

3.5 OLPC (One Laptop Per Child)

Amrita is doing a joint project with University of Texas, Austin on the OLPC. This laptop computer designed for kids will be used by village children in India. A-VIEW will be used by instructors to teach children in villages in India.

4 Conclusion

In summary, the goal of the A-VIEW framework is to create a virtual world across multiple centers for the student to receive live interactive lectures and to provide a complementary online social environment where the students can continue learning at their own pace. In this setup, we are able to leverage the knowledge and teaching skills of the best instructors. In general, the A-VIEW framework aims to provide sufficient power and flexibility for a variety of applications. The classroom interactive quizzes keep the instructor and the students alert and aware of the actual level of communication that is taking place. Used properly, the Knowledge Café provides technical, moral, and social support—particularly for the weaker students.

At the same time, we continue to monitor the various applications, setups and variables. We continue to perform research into the individual and group behaviors, satisfaction criteria, and the performance of the students in these eLearning laboratories

References

1. Onay, P., Yalabik, N., Koksal, G.: e-Kampus-IS: Information System for a Virtual University Consortium: ITHET 2005, pp. S3A/1 - S3A/6, (2005)
2. Barjis, J.: An overview of Virtual University Studies: Issues, Concepts and Trends from Virtual Education Cases in Learning & Teaching Technologies, (2003)
3. Xinyou Zhao, Yan Zhang: An Instructor-Oriented Prototype System for Virtual Classroom: IEEE Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06): 0-7695-2632-2/06, (2006)
4. P.Venkat Rangan and Harrick M Vin: Multimedia conference as a universal paradigm for collaboration, Proceedings of Euro graphic workshop on multimedia system, applications, and Interaction, Stockholm Sweden, (1991)

5. P.Venkat Rangan and D.C.Swinehart: Software architecture for integration of video services in the ether phone environment, IEEE journal on selected areas in communication, 9(9), pp. 1395-1404, (1991)
6. Pantovic, V., Lazovic, N., Starcevic, D.: Improved indexing for distributed virtual university, FIE 2000, 30th Annual, Vol. 2, pp. F3D/12, (2000)
7. Yazdani, M, Bligh, D.: Cooperative learning in a virtual university, Cognitive Technology, Humanizing the Information Age, Second International Conference, pp. 251 – 255, (1997)
8. Beuschel, W.: Ubiquitous e-learning: are we there yet? Advanced Learning Technologies, The 3rd IEEE International Conference, pp. 414 – 415, (2003)
9. Ogino, S., Sakauchi, M.: Mobile applications on virtual university, Research Challenges, Academia/Industry Working Conference, pp. 243 – 248, (2000)
10. Schlageter, Gunter: E-learning in Distance Education - Towards Supporting the Mobile Learner, ITHET '06, 7th International Conference, pp. 338 – 342, (2006)
11. Vaida, M.-F., Vescan, L.N.: Managing Web-based learning for distance education, ITI 2002. Proceedings of the 24th International Conference, pp.165 – 169, vol.1, (2002)
12. Rabenstein, R.: SYSTOOL - an online learning tool for signals and systems, ICASSP '02, IEEE International Conference, vol. 4: pp. IV-4128 - IV-4131, (2002)
13. Chen, M.: A corporate insider's view about virtual universities, COMPSAC 2000, The 24th Annual International, pp. 286–287, (2000)
14. Koper, R.: Use of the Semantic Web to Solve Some Basic Problems in Education, Journal of Interactive Media in Education, vol. 6, Special Issue on the Educational Semantic Web, (2004)
15. Hee-do, Chun: Innovation of Regional Universities through e-Learning, Analysis of the New University for Regional Innovation (NURI) Project Performance, KERIS@Vol.2 No.115, (2006)

PRELIMINARY CALL FOR PAPERS

3rd Workshop on Social Aspects of the Web (SAW 2009)

in conjunction with

12th International Conference on Business Information Systems (BIS 2009)

Poznań, Poland
27, 28 or 29 April 2009
http://www.integror.net/saw/

Deadline for submissions (preliminary): December, 1st 2008

In recent years, the Web has moved from a simple one-way communication channel extending traditional media, to a complex "peer-to-peer" communication space with a blurred author/audience distinction and new ways to create, share, and use knowledge in a social way.

This change of paradigm is currently profoundly transforming most areas of our life: our interactions with other people, our relationships, ways of gathering information, ways of developing social norms, opinions, attitudes and even legal aspects as well as ways of working and doing business.

It also raises a strong need for theoretical, empirical and applied studies related to how people may interact on the Web, how they actually do so and what new possibilities and challenges are emerging in the social, business and technology dimensions.

Following the two previous editions, the goal of this workshop is to bring researchers and practitioners together to explore the issues and challenges related to social aspects of the Web.

TOPICS OF INTEREST

We want to facilitate discussion on theoretical, empirical and applied studies related to:

- *Users in the social Web*
  - *User identity/identities on the Web*
  - *Activity patterns*
  - *Privacy / intimacy in the social Web*
  - *Psychological aspects of acting in the social Web*
  - *Analysis and reduction of the socio-technical gap in social software*
- *Communities on the Web*
  - *User roles, leadership and interactions*
  - *Conflicts and their resolution*
  - *Social norms and their enforcement*
  - *Trust and reputation in communities*
  - *Relations of on-line and off-line communities*
  - *Social discourse and decision-taking on the Web*
- *Large-scale social Web mining and empirical studies*

- *Social network analysis*
- *Associations mining from social network*
- *Large-scale behavior patterns and anomalies' mining*
- *Moods' / opinions' / social problems' analysis*
- *Experts finding on the social Web*
- *Mining formal semantics from social sources*
- *Methodologies of Web-based social macro and micro studies*

- *Social Web and business*
  - *Social Web as a source of business information*
  - *Social Web as a business communication channel*
  - *Business models for social software and services*
  - *Specific types of social software on the Web (bookmarking, social networks etc.)*
  - *Use cases and best practices*
- *Applications of Web-based social software*
  - *Social software architectures*
  - *Social software on the Semantic Web*
  - *Strategies for bootstrapping social software / bypassing the critical mass problem*
  - *Social software in information processing and retrieval*
  - *Social software in collaborative maintenance of content and data*

SUBMISSION

The following types of papers can be submitted to SAW 2009:

- *Long papers: max. 12 LNBIP pages*
- *Work-in-progress report: max. 6 LNBIP pages*
- *Position papers: max. 6 LNBIP pages*
- *Demo papers: max. 6 LNBIP pages*

SAW 2009 proceedings will be published by Springer as a volume in Lecture Notes in Business Information Processing series, together with proceedings of BIS 2009.

CHAIRS

- *Dominik Flejter, Poznan University of Economics, Poland*
- *Tomasz Kaczmarek, Poznan University of Economics, Poland*
- *Marek Kowalkiewicz, SAP Research Brisbane, Australia*

PROGRAM COMMITTEE (PRELIMINARY, TO BE EXTENDED)

- *Tanguy Coenen, Vrije Universiteit Brussel, Belgium*
- *Davide Eynard, Politecnico di Milano, Italy*
- *Marcin Paprzycki, Polish Academy of Science, Poland*
- *Katharina Siorpaes, STI, University of Innsbruck, Austria*
- *Jie Tang, Tshingua University, China*
- *Celine van Damme, Vrije Universiteit Brussel, Belgium*
- *Valentin Zacharias, FZI Karlsruhe, Germany*

PRELIMINARY CALL FOR PAPERS

2nd Workshop on Advances in Accessing Deep Web (ADW 2009)

in conjunction with

12th International Conference on Business Information Systems (BIS 2009)

Poznań, Poland
27, 28 or 29 April 2009
http://www.integror.net/adw/

Deadline for submissions (preliminary): December, 1st 2008

The main way of accessing content on contemporary Web is by means of general purpose search engines. However, for reasons such as: password protection, FORM based interfaces and usage of dynamic client-side technologies (JavaScript, AJAX, Flash, Adobe Air, and others), a significant portion of modern Web content cannot be indexed and thus is unavailable to the majority of Web users.

In many cases these information sources that cannot be indexed, known altogether under the names of Deep Web, Hidden Web or Invisible Web, are better structured and of better quality than indexable surface Web sources. First attempts to index deep Web sources are proving that the task is not trivial.

Started recently, Deep Web research combines challenges from several active research areas including information retrieval, information extraction, hypertext, Web engineering, data integration, database technologies and the Semantic Web.

The goal of this workshop is to bring researchers and practitioners together to explore the issues and challenges related to domain dependent and independent Deep Web empirical studies, methodologies and techniques of accessing and processing Deep Web content, as well as their real life applications.

TOPICS OF INTEREST

In the workshop we would like to present and discuss research, business cases and working prototypes or applications in areas including, but not limited to:

- *Modeling and Describing Deep Web*
  - *Models of Deep Web Navigation*
  - *Hidden Web Data and the Semantic Web*
  - *Description of Deep Web Sources Contents*
  - *Description of Deep Web Sources Querying Capabilities*
  - *Addressing Content and Data in Deep Web Sources*
  - *Semantic Annotation of Deep Web Sources*
- *Empirical Studies of Deep Web*
  - *Research on Deep Web Size and Topicality*

- *Deep Web Content and Quality Studies*
- *Domain-Specific Deep Web Sources Studies*
- *Comparative Studies of Deep Web and Surface Web Content*

- *Working with Deep Web Sources*
  - *Probing Deep Web Sources*
  - *Classification and Clustering of Deep Web Sources*
  - *Handling Client-Side Technologies (JavaScript/dHTML/AJAX) for Deep Web Access*
  - *Extraction of Data from Deep Web*
  - *Deep Web Sources Discovery*
  - *Archiving and Preservation of Deep Web Content*
  - *Methods of Deep Web Indexing*
- *Data Integration from Deep Web Sources*
  - *Deep Web Sources Selection*
  - *Rewriting Queries for Deep Web Sources*
  - *Matching and Mapping of Deep Web Sources Schemas*
  - *Methods of Deep Web Meta-Search*
- *Applications of Deep Web Research*
  - *Monitoring Hidden Web Data*
  - *Cases and Best Practices of Deep Web Data Usage*
  - *Business Models of Deep Web Data Integration*
  - *Legal and Ethical Aspects of Using Deep Web Content*
  - *Detecting and Disabling Deep Web Harvesting*

SUBMISSION

The following types of papers can be submitted to ADW 2009:

- *Long papers: max. 12 LNBIP pages*
- *Work-in-progress report: max. 6 LNBIP pages*
- *Position papers: max. 6 LNBIP pages*
- *Demo papers: max. 6 LNBIP pages*

ADW 2009 proceedings will be published by Springer as a volume in Lecture Notes in Business Information Processing series, together with proceedings of BIS 2009.

CHAIRS

- *Dominik Flejter, Poznan University of Economics, Poland*
- *Tomasz Kaczmarek, Poznan University of Economics, Poland*
- *Marek Kowalkiewicz, SAP Research Brisbane, Australia*

PROGRAM COMMITTEE (PRELIMINARY, TO BE EXTENDED)

- *Manuel Alvarez, University of A Coruna, Spain*
- *Irene Celino, CEFRIEL - Politecnico di Milano, Italy*
- *Emanuele Della Valle, CEFRIEL - Politecnico di Milano, Italy*
- *Francesco Guerra, University of Modena and Reggio Emilia, Italy*
- *Altigran Soares da Silva, Universidade Federal do Amazonas, Brazil*

Author Index