How much semantics on the "wild" Web is enough for machines to help us?*

Mária Bieliková

Institute of Informatics and Software Engineering, Slovak University of Technology Ilkovicova 3, 842 16 Bratislava, Slovakia maria.bielikova@fiit.stuba.sk WWW home page: http://fiit.stuba.sk/~bielik

Abstract. The current Web is not only a place for the content available in any time and location. It is also a place where we actually spend time to perform our working tasks, a place where we look for not only interesting information, but also entertainment, and friends, a place where we spend part of our rest. The Web is also an infrastructure for applications which offer various services. There is so many aspects of the Web that this diverse organism is a subject of study of researchers from various disciplines. In this paper we concentrate on information retrieval aspect of the Web, which is still prevailing. How we can improve information retrieval, be it goal-driven or exploratory? To which extent we are able to give our machines means for helping us in information retrieval tasks? Is there any level of semantics, which we can supply for the Web in general, and it will help? We present some aspects of information acquisition by search on the "wild" Web together of examples of approaches to particular tasks towards the improvement of information search, which were proposed in last two years within the Institute of Informatics and Software Engineering at the Slovak University of Technology, especially within the PeWe (Personalized Web) research group.

1 Introduction

The Web is amazing by the amount of diversity of its stuff, by the conception of so much thoughts, discussions, opinions that all show in many cases wisdom and creativity of people. This is also the bottleneck of current web – it is its nature, which involves "web objects" of various type (text, multimedia, programs) representing conceptually different entities (the content, people, things, services) and constantly changing. Particular objects are not formally defined, e.g. the content is semistructured, which leads to the complexity considering machine processing.

Obvious sentences are expected here – how is the Web important for our lives (both work and private), how the Web grows, how it is dynamic and constantly changing, how it absorbs people with their opinions, ratings and tags¹. Especially its dynamic nature prevents us from a direct employing of the most methods developed for closed information worlds (even though big or actually present on the Web). And its size requires automatic (or semiautomatic) approaches for information acquisition from this large heterogeneous information space.

The Web is undergoing constant development with

- the Semantic Web initiative, which aims for a machine readable representation of the Web [3],
- the Adaptive Web initiative, which stresses the need for personalization and broader context adaptation on the Web [6],
- the Web 2.0 initiative called also the Social Web, which focuses on social and collaborative aspects of the Web [14].

Development in this area matures to the point where the Web is becoming so important and in fact still unknown phenomenon that is identified as a separate, original object of investigation, and there are even initiatives which want to establish the Web Science as a new scientific discipline [7].

Considering information retrieval based on search (be it goal-driven or exploratory) includes also effective means for expressing users' information needs – how should a user specify his query or a broader aim of the search (be it a concrete requirement for explanation of particular term or an abstract need for finding out what is interesting or new in some domain). The "effective" here means that the user gets what he expects, even if his expectations are not completely known – this is pretty similar to the software requirements specification, but within the "wild" Web we have so much and so diverse users with various needs that we are not able to do this manually as software engineers do with the software specification.

In general, user's information needs usually come into existence while the user solves a task. Information needs can be classified into three categories [5]:

^{*} This work was partially supported by the projects VEGA 1/0508/09, KEGA 028-025STU-4/2010, and it is the partial result of the Research & Development Operational Programme for the project SMART II, ITMS 26240120029, co-funded by the ERDF.

¹ We do not mention and elaborate further another important view on the Web as an infrastructure for services and software applications.

2 Mária Bieliková

- Informational. The user's intent is to get specific information assumed to be present on the Web. The only assumed interaction is reading.
- Navigational. The user's intent is to reach a particular web page. It is assumed that a user will "travel" through the Web space taking advantage of getting a starting point.
- *Transactional.* The user's intent is to perform some activity enabled by the Web, i.e. the use of a service offered by particular web page.

These categories cannot be directly inferred from the user's query. However, good search engine should consider various information needs as this implicates a move from a static information retrieval (first two categories) to the third category, which integrates not just data stored on the Web, but also services that can provide right information (e.g. planning a flight).

2 Web and semantics

I do not feel a need for putting here well-known arguments about the importance of semantics for automatic reasoning. Yes, it is important! This fact was stated already many times from its first publishing in [3] even though what we give a machine actually is not the semantics; for the machine it is only a syntax – formal description of a resource.

The question is not what we can do with the semantics when it is perfect, but how to acquire it. How much semantics we can acquire for constantly changing world of the Web, or what amount is already useful to such extent that we can report an improvement in fulfilling our information needs.

With the Web development several sources for the semantics come into existence. Except the

 web content as a fundamental source for the semantics,

there are other sources of the semantics that can be mined:

- web structure with the focus on links analysis, and
- usage logs with the focus on a user activity on the Web mainly by an analysis of clickstreams.

As a special case of the content source we consider

- web annotations,

when viewing the annotations as a layer above the content created either automatically [11] or manually (in particular by user interactions and social tagging). The web annotations can be viewed also as a result of the users' activity and as such considered as a source for the web usage mining.

2.1 Considering the web content

The content, or resources in general are basically described by metadata. Metadata were used by librarians already before the Web era. They typically recognize three categories of metadata: administrative, structural, and descriptive [21]. Considering the Web and it content we focus on descriptive metadata related to the content. Moreover, metadata for the Web comparing to libraries resources should conform the fact that we cannot predict all kinds of the Web objects and their evolution.

The semantics of the content can be expressed many ways ranging from

- the set of keywords (or tags) through
- the Resource Description Framework or topic maps as a general model for conceptual description of resources to
- ontologies with all power resultant from formal logic where the ontology consists of concepts, relations, attributes, data types, a concept hierarchy, and a relation hierarchy.

Having ontologies that cover (almost) "complete" semantics which we are presently able to specify seems to be a solution for the Semantic Web. But it is not, at least now. Considering the complexity of defining such semantics recalls the situation some more than 40 years back when people tried devise general solving machines. Even though they moved later to expert knowledge capturing, the results were still limited mainly due to the ability of people to specify knowledge explicitly. So the situation repeats in some sense.

Right after the Semantic Web establishment we have witnessed a boom of various approaches to representing semantics for specific domains and methods for reasoning including mapping ontologies. However, ontology-based semantics is spreading slowly because we obviously have solutions just for very specific and rather static domains. It is perfect way for the application architecture as knowledge bases were in 70ties. But it does not fit well with the "wild" Web.

Even if we would have formally represented knowledge that would be sufficient for the best part of our needs (knowledge representation problem in Artificial Intelligence), and would have strong reasoning mechanisms, it is not enough for the changing Web – we still miss a component for matching this knowledge to particular web objects. Moreover, the Web is evolving as we people evolve in unpredictable way. New information and knowledge is constantly added to the Web either as semistructured content or as services or applications running on the Web.

Web 2.0 brought or vitalized a role of people in the whole process. We witness the power of crowd and

3

its limitations. *Folksonomy* is simply a returning back to the most elemental way to enrich a resource with semantics employing a set of keywords. Fundamental difference lays in the process of keywords acquisition. Folksonomy is created by users through the process of social tagging [12]. The advantage is real power of users, so keywords attached to the resource by social tagging represent rather objective notation of a web page content. The problem is that folksonomies are coarse-grained, informal and flat.

Following this trend we proposed a model of *light-weight semantics* of the web content referred to as the resource *metadata* [17]. It is promising in the sense of its automatic acquisition for open corpus, or vast and dynamic domains. It provides a meaningful abstraction of the Web content, i.e. provides a metadata model, and a mapping between web pages and this metadata model.

The model consists of interlinked concepts and relationships connecting concepts to resources (subjects of the search) or concepts themselves (see Figure 1). Concepts feature domain knowledge elements (e.g., keywords or tags) related to the resource content (e.g., web pages or documents). Both the *resource-toconcept* and the *concept-to-concept* relationships are weighted. Weights determine the degree of concept relatedness to the resource or to other concept, respectively. Interlinked concepts result in a structure resembling lightweight ontology, and form a layer above the resources allowing an improvement of the search.



Fig. 1. Content model based on lightweight semantics.

The advantage of modeling domain knowledge as described above lies in its simplicity. Hence, it is possible to generate metadata enabling lightweight semantic search for a vast majority of resources on the Web. We have already performed several experiments of automatic metadata extraction with promising results [15]. This models conforms also with existing and evolving folksonomies that can supplement extracted metadata, and can be fully captured within the model.

We believe that proposed model can improve information search. Our confidence is supported by partial results achieved (some of them are briefly mentioned in the Section 3). There are still some issues related to the proposed model. As the most serious we consider:

- extracting the right terms (concepts);
- creating and typing relationships between concepts;
- multilingual and multicultural aspects as for example some terms can have completely different meaning in dependence of culture.

Especially term extraction is well developed field with term-indexing approaches and named entity resolution. Considering the model alone, the semantics is still rather low as we cannot recognize properly important terms for particular user in particular context. That is why there is the need to combine all sources for the semantics [13]. We mention here except the content also web users' activity (web structure and web annotation are out of the scope of this paper).

2.2 Considering web users' activity

Monitoring a user's activity can serve as important source for semantics. Utilizing an implicit user feedback we can recognize which web pages (or even their parts) are interesting in particular context, and thus adjust or enrich metadata related to that content. User related metadata (i.e., a user model) allow personalization. Considering the "wild" Web with its lightweight semantics the spreading the personalization to the whole Web becomes possible (to some extent).

Resource metadata model introduced above serves also as a bottom layer for an overlayed user model. As we operate in open corpus it is not possible to have either of the models in advance. We propose to represent user's interests (discovered via web usage mining) by the same means as the resource metadata, and provide constant mapping between these two models.

If we want to employ such models for the purpose of information retrieval on the "wild" Web, we need to acquire terms (keywords, tags, concepts) from the web pages visited by the users. Because the Web is an open information space, we need to track down and process every page the user has visited in order to update his model appropriately.

To achieve this, we developed an enhanced proxy server, which allows for realization of advanced operations on the top of requests flowing from a user with responses coming back from the web servers, all over the Internet [2]. Figure 2 depicts the schema how the



Fig. 2. Monitoring a user based on an enhanced proxy platform.

proxy server operates. When the web server sends the response to the required resource back to the user, the proxy server enriches the resource by a script able to capture the user activities (due evaluation of the user feedback). In parallel we run a process of extracting the meta-data and concepts from the web page. Together with the user feedback, these are stored in the user profile. Before the extraction phase based on various algorithms to semantic annotation and keywords, and category extraction, we realize main content detection (relevant textual part of the HTML document) and machine based translation into English, which is required by the extraction algorithms.

The aforementioned process gathers metadata for every requested web page, and creates a basic (evidence) layer of a user model. Naturally, as the time flows, the keywords which represent long-term user interests occur more often than the others. Therefore, by considering only *top* K most occurring keywords, we get a user model which can be further analyzed, and serves as a basis for personalization.

We deployed our enhanced proxy platform to determine the efficiency of the solution in real-world usage. The proxy solution can be, apart from user activity logging, used to improve user experience with ordinary web pages by adapting them according actual user needs. More, we provide users with a wordlebased visualization (Wordle tag cloud generator, http://www.wordle.net/) of their user profiles, and collected a precious feedback, which helped us to determine "web stop-words", i.e., words which occur often on web pages but do not make any sense from the user's interests point of view. An example of such a user profile of one of the proxy authors is displayed in Figure 3.

3 Examples

We present several examples of approaches to particular tasks towards the improvement of information search, which were proposed and evaluated in last two years within the Institute of Informatics and Software Engineering at the Slovak University of Technology in Bratislava, especially within the PeWe (Personalized Web) research group.

3.1 Gaming as a source of semantics

Computer games are potential sources of metadata that are hard to extract by machines. With game rules properly set and sufficient motivation, players can indirectly solve otherwise costly problems.

Little Google Game. We proposed a method for term relationship network extraction via analysis of the logs of unique web search game [19]. Our game called Little Google Game focuses on web search query guessing. Players have to formulate queries in a special format (using negative keywords) and minimize amount of results returned by the search engine (we use Google at the moment). Afterwards we mine the game logs and extract relationships of terms based on their frequent common occurrence in the Web.

3.2 Domain dependent approaches

In spite of domain independence of proposed models, knowing the domain allows for more accurate models. This is common approach also used by the most popular web search engines, which blend data from multiple sources in order to fulfil the user's need behind his query using the advantage when domain is known (e.g. flight planning or cooking a meal).



Fig. 3. Michal's tag cloud.

ALEF, Adaptive Learning Framework. We proposed a schema for adaptive web-based learning and based on it we developed ALEF (Adaptive LEarning Framework), a framework for creating adaptive and highly interactive web-based learning systems [16].

ALEF domain model follows the resource metadata model described above. The content includes learning objects that can be of three types: explanation, question and exercise. The domain model covers for every learning object: actual content (text and media), and additional metadata that contain information which is relevant for personalization services (concepts, tags, comments). Comparing to other existing approaches, the notion of metadata in ALEF is quite simplified, which allows for automatic construction of domain model, and on the other hand, it still provides a solid basis for reasoning resulting in advanced operations such as metadata-based personalized navigation.

News recommendation. We proposed content based news recommendation based on articles similarity. Considering high dynamic and large every day volume of news we devised and evaluated in real settings two representations for effective news recommendation:

- efficient vector comprising title, term frequency of title words in the article content, names and places, keywords, category and readability index [9],
- balanced tree built incrementally; it inserts articles based on the content similarity [23].

Different approach to news recommendation provided on the same e-news portal (www.sme.sk) is presented in [20]. It employs k-nearest neighbor collaborative filtering algorithm based on generic full text engine exploiting power-law distributions Important property of proposed algorithm is that it maintains linear scalability characteristics with respect to the dataset size. Adaptive faceted browser. We devised a faceted semantic exploratory browser taking advantage of adaptive and social web approaches to provide personalized visual query construction support and address guidance and information overload [22]. It works on semantically enriched information spaces (both data and metadata describing the information space structure are represented by ontologies). Our browser facilitates user interface generation using metadata describing the presented information spaces (e.g., photos).

3.3 User centric approaches

Monitoring users and implicit feedback is promising approach for the "wild" Web. Even though an explicit user feedback (filling forms by a user) is easy to implement, it has serious problems with credibility, disturbing the user and dependence on his will.

Query expansion by social context. We proposed a method which implicitly infers the context of search by leveraging a social network, and modifies the user's search query to include it [10]. The social network is built from the stream of user's activity on the Web, which is acquired by means of our enhanced proxy server.

User interest estimation. We proposed a method for adaptive link recommendation [8]. It is based on an analysis of the user navigational patterns and his behavior on the web pages while browsing through a web portal. We extract interesting information from the web portal and recommend it in the form of personalized calendar and additional personalized links.

Search history tree. We proposed an approach intended to reduce user effort required to retrieve and/or revisit previously discovered information by exploiting web search and navigation history [18]. It is based on collecting streams of user actions during search

6 Mária Bieliková

sessions. We provide the user with a history map – a scrutable graph of semantic terms and web resources with full-text search capability over individual history entries. It is constructed by merging individual session history trees and the associated web resources.

Discovering keyword relations from Crowd. We proposed an approach of determining keyword relations (mainly a parent-child relationship) by leveraging collective wisdom of the masses, which is present in data of collaborative (social) tagging systems on the Web [1]. We demonstrated the feasibility of our approach on the data coming from the social bookmarking systems delicious and CiteULike.

4 Conclusions

In this paper we described just particular aspects of the whole picture. It is not in any sense complete. It should be viewed as a discussion on certain aspects and possible partial solutions.

At the moment we have more questions as the answers. How the Web should be described? What properties are important? How to discover interesting information for particular individual? Is there any emergent phenomena? What we could do? How we can really connect people in such a way that it will be convenient and useful? Can we trust the Web? Is its infrastructure right?

One day maybe we people will discover silver bullet for the Web. Meantime we should be open for various small enhancement, try to understand the Web as much as possible, and try to integrate all particular successes.

Acknowledgements. Figures and parts of descriptions in the Section 3 are taken from published papers, which present particular examples, all mentioned in References.

The author wish to thank colleagues from the Institute of Informatics and Software Engineering and all students – members of PeWe group, pewe.fiit.stuba.sk for their invaluable contribution to the work presented in this invited lecture. The most current state of ongoing projects within the group is reported in [4].

References

- M. Barla, M. Bielikova: On deriving tagsonomies: keyword relations coming from the Crowd. LNCS 5796, ICCCI 2009, Springer, 2009, 309–320.
- M. Barla, M. Bielikova: Ordinary web pages as a source for metadata acquisition for open corpus user modeling. Proc. of IADIS WWW/Internet 2010, 2010.
- 3. T. Berners-Lee, J. Hendler, O. Lassila: *The semantic web*. Scientific American Magazin, May 2001.

- M. Bielikova, P. Navrat, (Eds): Workshop on the Web – Science, Technologies and Engineering, 2010. ISBN 978-80-227-3274-1. Available online at pewe.fiit.stuba.sk/ontoparty-2009-2010-spring/.
- A. Broder: A taxonomy of web search. t ACM SIGIR forum, 36 (2), 2002, 3–10.
- P. Brusilovsky et al.: *The adaptive web.* LNCS 4321, Springer, 2007, ISBN 978-3-540-72078-2, 763p.
- J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, D. Weitzner: Web science: an interdisciplinary approach to understanding the web. Commun. ACM, 51 (7), July 2008, 60–69.
- M. Holub, M. Bielikova: Estimation of user interest in visited web page. Proc. of Int. Conf. on World Wide Web, WWW 2010, ACM, 2010, 1111–1112.
- M. Kompan, M. Bielikova: Content-based news recommendation. LNBIP Series, E-Commerce and Web Technologies, Springer, 2010.
- T. Kramar, M. Barla, M. Bielikova: Disambiguating search by leveraging a social context based on the stream of users activity. LNCS 6075, UMAP 2010, Springer, 2010, 387–392.
- M. Laclavik et al.: Ontea: platform for pattern based automated semantic annotation. Computing and Informatics 28 (4), 2009, 555–579.
- P. Mika: Ontologies are us: a unified model of social networks and semantics. LNCS 3729, ISWC 2005, Springer, 2005, 522-536.
- P. Navrat, T. Taraba, A. Bou Ezzeddine, D. Chuda: Context search enhanced by readability index. IFIP WCC Series 276, Springer, 2008, 373–382.
- T. O'Reilly: What is Web 2.0. O'Reilly Network. 2005. [Accessed 2010-07-30] Available at http://oreilly.com/web2/archive/what-is-web-20.html
- M. Simko, M. Bielikova: Automated educational course metadata generation based on semantics discovery. LNCS 5794, EC TEL 2009, Springer, 2009, 99-105.
- M. Simko, M. Barla, M. Bielikova: ALEF: A framework for adaptive web-based learning 2.0. KCKS 2010, IFIP AICT 324, Springer, 2010, 367–378.
- M. Simko, M. Bielikova: Improving search results with lightweight semantic search. CEUR, 491, SemSearch 2009 at WWW 2009, 53–54.
- J. Simko, M. Tvarozek, M. Bielikova: Semantic history map: graphs aiding web revisitation support. Proc. of 9th Int. Workshop on Web Semantics, IEEE Computer Society, 2010.
- J. Simko, M. Tvarozek, M. Bielikova: Little google game: relationships term extraction by means of search game. Proc. of Datakon 2010.
- J. Suchal, P. Navrat: Full text search engine as scalable k-nearest neighbor recommendation system. AI 2010, IFIP AICT 331, Springer, 2010, 165–173.
- 21. A.G. Taylor: *The organization of information*. Libraries Unlimited: Englewood, USA. 1999, 300p.
- M. Tvarozek, M. Bielikova: Generating exploratory search interfaces for the semantic web. HCIS 2010, IFIP AICT 332, Springer, 2010, 175–186.
- D. Zelenik, M. Bielikova: Dynamics in hierarchical classification of news. Proc. of WIKT 2009, 83–87.