Proceedings of the 4th International Workshop

Social Data on the Web

Alexandre Passant, DERI NUI Galway, Ireland
Sergio Fernández, Fundación CTIC, Spain
John Breslin, DERI NUI Galway, Ireland
Uldis Bojārs, University of Latvia, Latvia

http://sdow.semanticweb.org/2011

Workshop at the 10th International Semantic Web Conference

ABSTRACT

The 4th international workshop Social Data on the Web (SDoW2011), co-located with the 10th International Semantic Web Conference (ISWC2011) , aims to bring together researchers, developers and practitioners involved in semantically-enhancing social media websites, as well as academics researching more formal aspect of these interactions between the Semantic Web and Social Web.

It is now widely agreed in the community that the Semantic Web and the Social Web can benefit from each other. One the one hand, the speed at which data is being created on the Social Web is growing at exponential rate. Recent statistics showed that about 100 million Tweets are created per day and that Facebook has now 500 million users. Yet, some issues still have to be tackled, such as how to efficiently make sense of all this data, how to ensure trust and privacy on the Social Web, how to interlink data from different systems, whether it is on the Web or in the enterprise, or more recently, how to link Social Network and sensor networks to enable Semantic Citizen Sensing.

Following the successful SDoW workshops at ISWC 2008, 2009 and 2010, this workshop will tackle these various topics and aims at bringing together researchers and practitioners, as in the 3 previous editions. We aim to bring together Semantic Web experts and Web 2.0 practitioners and users to discuss the application of semantic technologies to data from the Social Web. It is motivated by recent active developments in collaborative and social software and their Semantic Web counterparts, notably in the industry, such as FaceBook Open Graph Protocol.

Alexandre Passant, DERI NUI Galway, Ireland
Sergio Fernández, Fundación CTIC, Spain
John Breslin, DERI NUI Galway, Ireland
Uldis Bojārs, University of Latvia, Latvia

Program Committee

- Alessandra Toninelli, Research & Innovation Division Engineering Group, Italy
- Axel Ngonga, Universität Leipzig, Germany
- Chris Bizer, FUB, Germany
- Dan Brickley, FOAF project & Vrije Universiteit Amsterdam, The Netherlands
- Daniel Gayo-Avello, University of Oviedo, Spain
- Daniel Schwabe, PUC Rio, Brasil
- Diego Berrueta, Fundación CTIC, Spain
- Emanuele Della Valle, Politecnico di Milano, Italy
- Fabien Gandon, INRIA, France
- Gunnar Aastrand Grimnes, DFKI Knowledge Management Lab, Germany
- Harry Halpin, University of Edinburgh / W3C, UK
- Henry Story, Apache Software Foundation, France
- Irene Celino, CEFRIEL, Italy
- Jose E. Labra, University of Oviedo, Spain
- Libby Miller, BBC, UK
- Matthew Rowe, University of Sheffield, UK
- Michael Hausenblas, DERI, NUI Galway Ireland
- Mischa Tuffield, Garlik, UK
- Olaf Hartig, Humboldt-Universität zu Berlin, Germany
- Oscar Corcho, UPM, Spain
- Pablo López, Treelogic, Spain
- Pablo Mendes, Kno.e.sis, Wright State University, USA
- Richard Cyganiak, DERI, NUI Galway, Ireland
- Sebastian Tramp, Universität Leipzig, Germany
- Sheila Kinsella, DERI, NUI Galway
- Sofia Angeletou, KMi, The Open University, UK
- Steve Harris, Garlik, UK
- Yves Raimond, BBC, UK

Table of Contents

Pragmatic metadata matters: How data about the usage of data effects semantic user models

Claudia Wagner[1], Markus Strohmaier[2], and Yulan He[3]

[1] JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
claudia.wagner@joanneum.at
[2] Graz University of Technology and Know-Center
Inffeldgasse 21a, 8010 Graz, Austria
markus.strohmaier@tugraz.at
[3] The Open University, KMi
Walton Hall, Milton Keynes MK7 6AA, UK
yhe@open.ac.uk

Abstract. Online social media such as wikis, blogs or message boards enable large groups of users to generate and socialize around content. With increasing adoption of such media, the number of users interacting with user-generated content grows and as a result also the amount of *pragmatic metadata* - i.e. data about the usage of content - grows.
The aim of this work is to compare different methods for learning topical user profiles from Social Web data and to explore if and how pragmatic metadata has an effect on the quality of semantic user models. Since accurate topical user profiles are required by many applications such as recommender systems or expert search engines, learning such models by observing content and activities around content is an appealing idea.
To the best of our knowledge, this is the first work that demonstrates an effect between pragmatic metadata on one hand, and the quality of semantic user models based on user-generated content on the other. Our results suggest that *not all types of pragmatic metadata are equally useful* for acquiring *accurate* semantic user models, and some types of pragmatic metadata can even have detrimental effects.

Keywords: Semantic Analysis, Social Web, Topic Models, User Models

1 Introduction

Online social media such as Twitter, wikis, blogs or message boards enable large groups of users to create content and socialize around content. When a large group of users interact and socialize around content, *pragmatic metadata* is produced as a side product. While *semantic metadata* is often characterized as *data about the meaning of data*, we define *pragmatic metadata* as *data about the usage of data*. Thereby, pragmatic metadata captures how data/content is used

by individuals or groups of users - such as who authored a given message, who replied to messages, who "liked" a message, etc. Although the amount of pragmatic metadata is growing, we still know little about how these metadata can be exploited for understanding the topics users engage with.

Many applications, such as recommender systems or intelligent tutoring systems, require good user models, where "'good"' means that the model accurately reflects user's interest and behavior and is able to predict future content and activities of users. In this work we explore to what extent and how pragmatic metadata may contribute to semantic models of users and their content and compare different methods for learning topical user profiles from Social Web data.

To this end, we use data from an online message board. We incorporate different types of pragmatic metadata into different topic modeling algorithms and use them to learn topics and to annotate users with topics. We evaluate the quality of different semantic user models by comparing their predictive performance on future posts of user. Our evaluation is based on the assumption that "better" user models will be able to predict future content of users more accurately and will need less time and training data.

Generative probabilistic models are a state of the art technique for unsupervised learning. In such models, observed and latent variables are represented as random variables and probability calculus is used to describe the connections that are assumed to exist between these variables. Only if the assumptions made by the model are correct, Bayesian inference can be used to answer questions about the data. Generative probabilistic models have been successfully applied to large document collections (see e.g. [1]). Since for many documents one can also observe metadata, several generative probabilistic models have been developed which allow exploiting special types of metadata (see e.g., the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). However, previous research [10] has also shown that incorporating metadata into the topic modeling process may lead to model assumptions which are too strict and might overfit the data. This means that incorporating metadata does not necessarily lead to "better" topic models, where "better" means, for example, that the model is able to predict future user-generated content more accurately and needs less trainings data to fit the model.

Our work aims to advance our understanding about the effects of pragmatics on semantics emerging from user-generated content and specifically aims to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?
2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work, while Section 3 describes our experimental setup. In Section 4 we report our results, followed by a discussion of our findings in Section 5.

2 Related Work

From a machine learning perspective, social web applications such as Boards.ie provide a huge amount of unlabeled training data for which usually many types of metadata can be observed. Several generative probabilistic models have been developed which allow exploiting special types of metadata (such as the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). In contrast to previous work where researchers focused on creating new topic models for each type of metadata, [9] presents a new family of topic models, Dirichlet-Multinomial Regression (DMR) topic models, which allow incorporating arbitrary types of observed features . Our work builds on the DMR topic model and aims to explore the extent to which different types of pragmatic metadata contribute to learning topic models from user generated content.

In addition to research on advancing topic modeling algorithms, the usefulness of topic models has been studied in different contexts, including social media. For example, [5] explored different schemes for fitting topic models to Twitter data and compared these schemes by using the fitted topic model for two classification tasks. As we do in our work, they also point out that models trained with a "'User"' scheme (i.e., using post aggregations of users as documents) perform better than models trained with a "'Post"' scheme. However, in contrast to our work they only explore relatively simple topic models and do not take any pragmatic metadata (except authorship information) into account when learning their models.

In our own previous work, we have studied the relationship between pragmatics and semantics in the context of social tagging systems. We have found that, for example, the pragmatics of tagging (users' behavior and motivation in social tagging systems [11, 6, 4]) exert an influence on the usefulness of emergent semantic structures [7]. In social awareness streams, we have shown that different types of Twitter stream aggregations can significantly influence the result of semantic analysis of tweets [12]. In this paper, we extend this line of research by (i) applying general topic models and (ii) using a dataset that offers rich pragmatic metadata.

3 Experimental Setup

The aim of our experiments is to explore to what extent and how pragmatic metadata can be exploited when semantically analyzing user generated content.

3.1 Dataset

The dataset used for our experiments and analysis was provided by Boards.ie,[4] an Irish community message board that has been in existence since 1998. We used all messages published during the first week of February 2006 (02/01/2006 - 02/07/2006) and the last week of February 2006 (02/21/2006 - 02/28/2006). We only used messages authored by users who published more than 5 messages and replied to more than 5 messages during this week. While we performed our experiments on both datasets, the results are similar. Consequently, we focus on reporting results obtained on the first dataset which consists of 1401 users and 27525 posts which were authored by these users and got replies.

To assess the predictive performance of different topic models we estimate how well they are able to predict the content (i.e. the actual words) of future posts. We generated a test corpus of 4007 held out posts in the following way: for each of the 1401 user in our training corpus we crawled 3 future posts which were authored by them and to which at least one user of our training corpus has replied. From here on, we refer to this data has *hold-out* data.

3.2 Methodology

In this section we first introduce the topic modeling algorithms (LDA, AT-model and DMR topic model) on which our work is based and then proceed to describe the topic models which we fitted to our training data, their model assumptions and how we compared and evaluated them.

Latent Dirichlet Allocation (LDA) The idea behind LDA is to model documents as mixtures of topics and force documents to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms. That means the generation of a collection of documents is modeled as a three step process: First, for each document d a distribution over topics θ_d is sampled from a Dirichlet distribution α. Second, for each word w_d in the document d, a single topic z is chosen according to this distribution θ_d. Finally, each word w_d is sampled from a multinomial distribution over words ϕ_z which is specific for the sampled topic z.

The Author Topic (AT) model The Author Topic model [10] is an extension of LDA, which learns topics conditioned on the mixture of authors that composed the documents. The assumption of the AT model is that each document is generated from a topic distribution which is specific to the set of authors of the document. The observed set of variables are the words per document (similar as in LDA) and the authors per document. The latent variables which are learned by fitting the model, are the topic distribution per author (rather than the topic distribution per document as in LDA) and the word distribution per topic.

[4] `http://www.boards.ie/`

We implemented the AT-model based on Dirichlet-multinomial Regression (DMR) Models (explained in the next section). While the original AT-model uses multinomial distribution (which are all drawn from the same Dirichlet) to represent an author-specific topic distributions, the DMR-model based implementation uses a "fresh" Dirichlet prior for each author from which then the topic distribution is drawn.

Dirichlet-multinomial Regression (DMR) Models Dirichlet-multinomial regression (DMR) topic models [9] assume not only that documents are generated by a latent mixture of topics but also that mixtures of topics are influenced by an additional factor which is specific to each document. This factor is materialized via observed features (in our case pragmatic metadata such as authorship or reply user information) and induce some correlation across individual documents in the same group. This means that e.g. documents which have been authored by the same user (i.e., they belong to one group) are more likely to chose the same topics. Formally, the prior distribution over topics α is a function of observed document features, and is therefore specific to each distinct combination of feature values. In addition to the observed features we add a default feature to each document, to account for the mean value of each topic.

Fitting Topic Models In this section we describe the different topic models which we fitted to our training datasets (see table 1 and 2). Each topic model makes different assumptions on what a document is (see column 3), takes different types of pragmatic metadata into account (see column 4) and makes different assumptions on the document-specific topic distributions θ which generates each documents (see column 5).

For all models, we chose the standard hyperparameters which are optimized during the fitting process: $\alpha = 50/T$ (prior of the topic distributions), $\beta = 0.01$ (prior of the word distributions) and $\sigma^2 = 0.5$ (variance of the prior on the parameter values of the Dirichlet distribution α). For the default features $\sigma^2 = 10$. Based on the empirical findings of [13], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. All models share the assumption that the total number of topics used to describe all documents of our collection is limited and fixed (via hyperparameter T) and that each topic must favor few words (as denoted by hyperparameter β which defines the Dirichlet distribution from which the word distributions are drawn - the higher β the less distinct the drawn word distributions).

Following the model selection approach described in [3], we selected the optimal number of topics for our training corpus by evaluating the probability of held out data for various values of T (keeping $\beta = 0.01$ fixed). For both datasets (each represents one week boards.ie data), a model trained on the "'Post"' scheme (i.e., using each post as a document) gives on average (over 10 runs) the highest probability to held out documents if $T = 240$ and model trained on the "'User"' scheme (i.e., using all posts authored by one user as a document) gives on av-

erage (over 10 runs) the highest probability to held out documents if $T = 120$. We kept T fixed for all our experiments.

Evaluation of Topic Models To compare different topic models we use perplexity which is a standard measure for estimating the performance of a probabilistic model. Perplexity measures the ability of a model to predict words on held out documents. In our case a low perplexity score may indicate that a model is able to accurately predict the content of future posts authored by a user. The perplexity measure is defined as followed:

$$perplexity(d) = exp[-\frac{\sum_{i=0}^{N_d} lnP(w_i|\hat{\phi}, \alpha)}{N_d}] \quad (1)$$

In words, the perplexity of a held out post d is defined as the exponential of the negative normalized predictive likelihood of the words w_i of the held out post d (where N_d is the total number of words in d) conditioned on the fitted model.

ID	Alg	Doc	Metadata	Model Assumption
M1	LDA	Post	-	A post is generated by a mixture of topics and has to favor few topics.
M2	LDA	User	-	All posts of one user are generated by a mixture of topics and have to favor few topics.
M3	DMR	Post	author	A post is generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M4	DMR	User	author	All posts of one user are generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M5	DMR	Post	user who replied	A post is generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M6	DMR	User	user who replied	All posts of one user are generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M7	DMR	Post	related user	A post is generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

M8	DMR	User	related user	All posts of one user are generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

Table 1: Overview about different topic models which incorporate different types of pragmatic metadata.

ID	Alg	Doc	Metadata	Model Assumption
M9	DMR	Post	top 10 forums of author	A post is generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post.
M10	DMR	User	top 10 forums of author	All posts are generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post-aggregation.
M11	DMR	Post	top 10 communication partner of author	A post is generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post.
M12	DMR	User	top 10 communication partner of author	All posts are generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post-aggregation.

Table 2: Overview about different topic models which incorporate different types of smooth pragmatic metadata based on behavioral user similarities.

4 Experimental Results

Our experiments were set up to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?

To answer this question, we fit different models to our training corpus and tested their predictive performance on future posts authored by our trainings users.

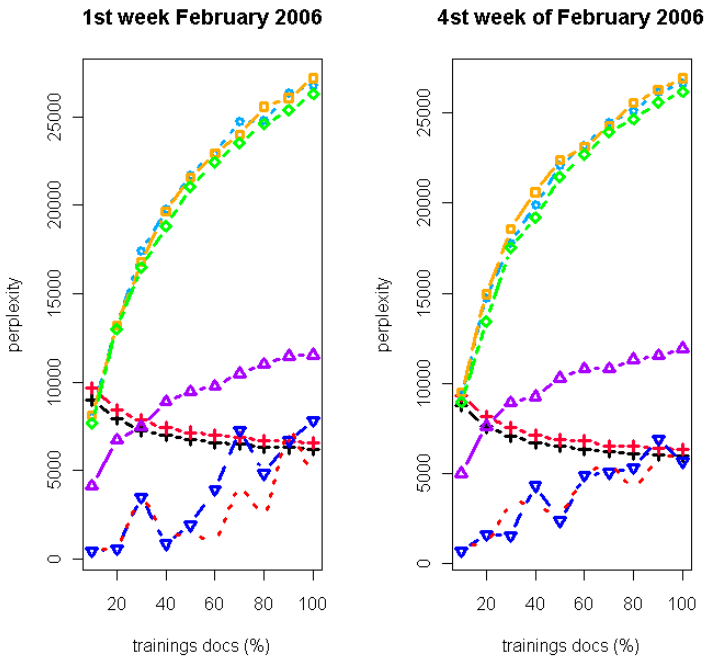


Fig. 1. Comparison of the predictive performance of different topic models on held out posts. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

Figure 1 shows that the predictive performance of semantic models of users which are either solely based on the users (i.e., aggregations of users‘ posts) to whom these users replied (M6) or which take in addition also the content authored by these users (M8) into account, is best. Therefore, our results suggest that it is beneficial to take user‘s reply behavior into account when learning topical user profiles from user generated content.

We also noted that all models which use the “User” training scheme (M4, M6 and M8) perform better than the models which use the “Post” training scheme (M3, M5 and M7). One possible explanation for this is the sparsity of posts which consist of only 66 tokens on average.

Since we were interested in how the predictive performance of different models change depending on the amount of data and time used for training, we split our training dataset randomly into smaller buckets and fitted the model on different proportions of the whole training corpus. One would expect that as the percentage of training data increases the predictive power of each model would improve as it adapts to the dataset. Figure 1 however shows that this is only true for our baseline models M1 and M2 which ignore all metadata of posts. The model M3 which corresponds to the Author Topic model exhibits a behavior that is similar to the behavior reported in [10]: When observing only few training data, M3 makes more accurate predictions on held-out posts than our baseline models. But the predictive performance of the model is limited by the strong assumptions that future posts of one author are about the same topics as past posts of the same author. Like M3, also M5 (and M7) seem to over-fit the data by making the assumptions that future posts of a user will be about the same topics as posts he replied to in the past (and posts he authored in the past).

To address these over-fitting problems we decided to incorporate smoother pragmatic metadata into the modeling process which we get by exploiting behavioral user similarities. The pragmatic metadata we used so far capture information about the usage behavior of individuals (e.g., who authored a document), while our smoother variants of pragmatic metadata capture information about the usage behavior of groups of users which share some common characteristics (e.g., what are the forums in which the author of this document is most active). Our intuition behind incorporating these smoother pragmatic metadata which are based on user similarities is that users which behave similar tend to talk about similar topics.

2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

From Figure 2 one can see that indeed all models which incorporate behavioral user similarity exhibit lower perplexity than our baseline models, especially if only few training samples are available. The model M12, which is based on the assumption that users who talk to the same users talk about the same topics, exhibits the lowest perplexity and outperforms our baseline models in terms of their predictive performance on held out posts.

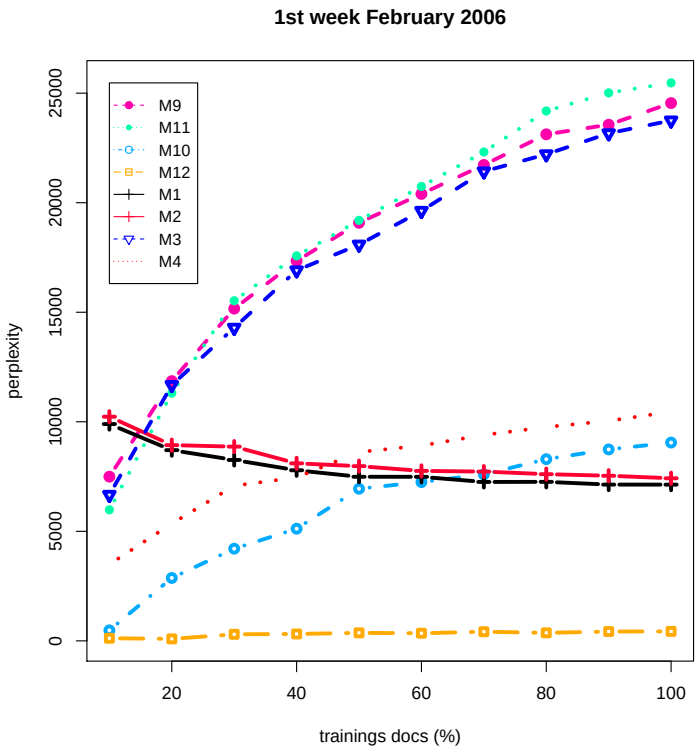


Fig. 2. Comparison of the predictive performance of topic models which take smooth pragmatic metadata into account by exploiting user similarities. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

For the model M10 which assumes that users who tend to post to the same forums talk about the same topics, we can only observe a lower perplexity than our baseline models when only few trainings data are available, but it still outperforms other state of the art topic models such as the Author topic model.

5 Discussion of Results and Conclusion

While it is intuitive to assume that incorporating metadata about the pragmatic nature of content leads to better learning algorithms, our results show that not all types of pragmatic metadata contribute in the same way. Our results confirm previous research which showed that topic models which incorporate pragmatic metadata such as the author topic model tend to over-fit data. That means incorporating metadata into a topic model can lead to model assumptions which are too strict and which yield the model to perform worse.

Summarizing, our results suggest that:

- **Pragmatics of content influence its semantics:** Integrating pragmatic metadata information into semantic user models influences the quality of resulting models.
- **Communication behavior matters:** Taking user's reply behavior into account when learning topical user profiles is beneficial. Content of users to which a user replied seems to be even more relevant for learning topical user profiles than content authored by a user.
- **Behavioral user similarities improve user models:** Smoother versions of metadata based topic models which take user similarity into account always seem to improve the models.
- **Communication behavior based similarities matter:** Different types of proxies for behavioral user similarity (e.g., number of forums they both posted to, number of shared communication partners) lead to different results. User who have a similar communication behavior seem to be more likely to talk about the same topics, than users who post to similar forums.

Acknowledgments. The authors want to thank Boards.ie for providing the dataset used in our experiments and Matthew Rowe for pre-processing the data. Furthermore we want to thank David Mimno for answering questions about the DMR topic model and Sofia Angelouta for fruitful discussions. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: International Conference on Machine Learning. pp. 233–240 (2007)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(Suppl. 1), 5228–5235 (April 2004)

4. Helic, D., Trattner, C., Strohmaier, M., Andrews, K.: On the navigability of social tagging systems. In: The 2nd IEEE International Conference on Social Computing (SocialCom 2010), Minneapolis, Minnesota, USA. pp. 161–168 (2010)
5. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics. pp. 80–88. SOMA '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1964858.1964870
6. Koerner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: 21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada, ACM. ACM, New York, NY, USA (June 2010)
7. Koerner, C., Benz, D., Strohmaier, M., Hotho, A., Stumme, G.: Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In: Proc. of the 19th International World Wide Web Conference (WWW 2010). ACM, Raleigh, NC, USA (Apr 2010), http://www.kde.cs.uni-kassel.de/benz/papers/2010/koerner2010thinking.pdf
8. Mccallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Tech. rep., UMass CS (December 2004)
9. Mimno, D., McCallum, A.: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08) (2008), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.6925
10. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487–494. UAI '04, AUAI Press, Arlington, Virginia, United States (2004), http://portal.acm.org/citation.cfm?id=1036843.1036902
11. Strohmaier, M., Koerner, C., Kern, R.: Why do users tag? Detecting users' motivation for tagging in social tagging systems. In: International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26. AAAI, Menlo Park, CA, USA (2010)
12. Wagner, C., Strohmaier, M.: Exploring the wisdom of the tweets: Knowledge acquisition from social awareness streams. In: Proceedings of the Semantic Search 2010 Workshop (SemSearch2010), in conjunction with the 19th International World Wide Web Conference (WWW2010), Raleigh, NC, USA, April 26-30, ACM (2010)
13. Wallach, H.M., Mimno, D., McCallum, A.: Rethinking LDA: Why priors matter. In: Proceedings of NIPS (2009), http://books.nips.cc/papers/files/nips22/NIPS2009_0929.pdf
14. Wang, X., Mohanty, N., Mccallum, A.: Group and topic discovery from relations and text. In: In Proc. 3rd international workshop on Link discovery. pp. 28–35. ACM (2005)

Linked Opinions: Describing Sentiments on the Structured Web of Data

Adam Westerski, Carlos A. Iglesias, and Fernando Tapia Rico

Universiadad Politecnica de Madrid,
Escuela Tecnica Superior de Ingenieros de Telecomunicacion,
Avenida Complutense 30, Ciudad Universitaria,
28040 Madrid, Spain
westerski@dit.upm.es,cif@gsi.dit.upm.es,
fernando.tapia.rico@alumnos.upm.es

Abstract. In the paper we report on the results of our experiments on the construction of the opinion ontology. Our aim is to show the benefits of publishing in the open, on the Web, the results of the opinion mining process in a structured form. On the road to achieving this, we attempt to answer the research question to what extent opinion information can be formalized in a unified way. Furthermore, as part of the evaluation, we experiment with the usage of Semantic Web technologies and show particular use cases that support our claims.

Keywords: structured data; ontology; appliance; knowledge management; idea management; opinion mining

1 Introduction

The rise of the Social Web has stimulated progress in many disciplines and gave birth to new trends. One of the research domains that noted especially big progress within recent years is opinion mining. From the information systems point of view, opinion mining aims to harness the flows of unstructured (or poorly structured) subjective user generated textual content that otherwise is hard to analyse, accurately categorise and reason upon. However, while in many cases opinion mining delivers satisfying results it should be aligned to the constantly evolving Web.

One of the problems that we would like to bring to attention is that many web systems (e.g Swotti[1]) that employ opinion mining after gaining understanding of the user generated content, process the extracted parameters (e.g. polarity, features) and publish the outcomes again in an unstructured form (i.e. HTML). On the other hand, others (e.g. Tweet Sentiment[2]) that allow to access the data via web services establish their own formats or languages due to lack of standards that would define clear rules for publishing such information.

[1] http://www.swotti.com/
[2] http://www.tweetsentiments.com/

In our research we aim to show what kind of benefits could it bring to establish a common web metadata schema that would enable to publish in a formalized manner the results of the opinion mining process. As we report on the research done, firstly we introduce the abstract data model - an ontology that formalizes all concepts derived from the opinion mining process (see Sec. 3). Further we propose the use of Semantic Web technologies to adapt that ontology for web use and show exactly what profits can that bring (see Sec. 4). Finally, we present the results of the evaluations run for large scope use cases as well as limited to particular web systems (see Sec. 5).

2 Motivation

Embedding opinion mining functionality for websites that are rich in user comments can aid to automatically rank comments and let users faster reach the types of opinions that they seek [17]. Furthermore, given the same data, opinion mining algorithms can be used to supply additional metrics to rate products and content [20]. However, all of this value is often limited only to the single site of origin that performed the opinion mining algorithm.

Based on the achievements and research done in the area of Semantic Web [7] and more specifically its evolution into proposal of Linked Data [6], we point to publishing opinion information using a universal metadata format that would extend the usability of such data. First and foremost, when having opinions described across the Internet in a unified way it is possible to compare them and perform an Internet wide search and statistics. At the moment it is possible to find opinions of desired polarity about selected product using contemporary Internet search engines, however the simple text based indexing is far less accurate and less flexible than what could be achieved with metadata indexing [11]. Furthermore, if the opinion mining data would be accompanied and linked with other metadata that describes the context of the subjective content, then the capabilities of search and browsing would rise even more (e.g. with regard to aggregation and data mashups [10]).

Finally if all of the above motivations seem fair but far away and hard to realize in practice, we would like to point to what currently seems to be the principal argument for content providers to publish metadata: improve viability on the web and in the search engines. Metadata can help to increase the precision and recall of search [4] but also the value of metadata becomes more visible as the search results in the leading Internet search engines start to contain data extracted form the metadata published along with HTML (e.g. Google Rich Snippets [3]), thus making particular search results more attractive in comparison to competitive links. Through annotation of opinions, exactly the same benefit could be delivered for the websites that provide opinion mining results over subjective content posted on them or remote sources (see Fig. 1).

[3] http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html

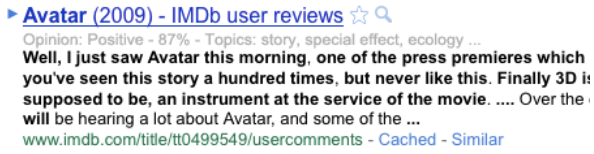


Fig. 1. A Google snippet modified with Greasemonkey script and enriched with data extracted from RDF

3 Marl: An Ontology for Opinion Mining

When designing the ontology our aim was to analyse the properties that characterize opinions expressed on the web or inside various IT systems. The final set of concepts that we propose (see Fig. 2) is a result of a two step process.

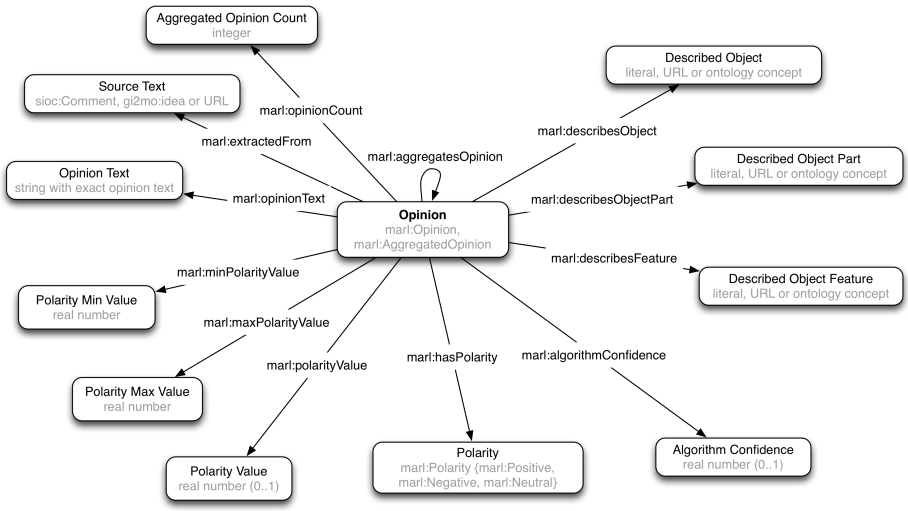


Fig. 2. Conceptual model for opinion and the proposed **Marl** ontology

First, we analysed different kinds of subjective data sources and produced a common model that was formalized as Marl Ontology v0.1. For this part we started with three common cases of opinions expressed on the Web: movie opinions, movie review opinions and products opinions. Later, in addition, we also analysed characteristics of opinions in enclosed communities and used an enterprise open innovation system as a case study. In the second phase, we evaluated the proposed ontology against live data and corrected the discovered

drawbacks in version 0.2 of the ontology (see Sec. 5). The description of particular properties and explanation of their meaning can be found in Table 1.

Table 1. Marl ontology: classes and properties breakdown

Name	**Description**
Opinion	Class that represents the opinion concept
extractedFrom	Indicates the source text from which the opinion has been extracted.
opinionText[1]	The exact string that contains the calculated sentiment.
hasPolarity	Points to either entity or literal that indicates if the opinion is positive/negative or neutral
polarityValue	A numerical representation of the polarity value.
maxPolarityValue	Maximal possible numerical value for the opinion
minPolarityValue	Lowest possible numerical value of the opinion
describesObject	Indicates the object that the opinion refers to
describesObjectPart	Indicates a particular element of the object that the opinion refers to (eg. laptop *battery*)
describesFeature	Points to a feature of an object that the opinion refers to (eg. laptop battery *life*)
algorithmConfidence	A number that describes how much the algorithm was confident with its assessment
AggregatedOpinion	Subclass of Opinion class that aggregates a number of opinions.
aggregatesOpinion	Points to Opinion instances that are aggregated
opinionCount[1]	Amount of opinions aggregated.
Polarity	Instances of this class represent the positive, neutral or negative polarity

In the particular model that we created we attempted to center all the data properties around a single opinion class. This and other ontology design choices that we made with Marl relate to one of the common problems of modelling ontologies for web use: the choice between modelling certain concepts fully as classes of domain ontologies, literals or simply URLs. While for using the full potential of Semantic Web it is best to model metadata concepts as entities described by particular ontologies the reality proves that this is far from being a practical solution. Therefore, we propose a model that accommodates both (see Fig. 3) assuring future extendibility yet facilitating more simple and practical use. In the next section we describe the benefits and applications of either of the cases.

4 Publishing and consuming opinion metadata on the web

Following the description of the opinion ontology we show its possible uses and the differences that various closed and open systems impose. Furthermore, to support the ontology design decisions described earlier, we expose the benefits and drawbacks of publishing opinion data in different forms and with a different level of detail using the Marl ontology.

[1] Properties added in Marl v0.2

```
(1) Example A: Entity referencing for describing contextual information
marl:describesObject <http://dbpedia.org/
                      resource/Avatar_%282009_film%29>
marl:describesObjectPart dbpedia-owl:director
marl:extractedFrom <http://www.gi2mo.org/
                    index.php?sioc_type=comment&sioc_id=157>
marl:polarity marl:Positive
marl:polarityValue "0.6"

(2) Example B: Using literal values to describe contextual information
marl:describesObject "Avatar"
marl:describesObjectPart "director"
marl:extractedFrom <http://www.gi2mo.org/
                    2010/09/introducing-marl/#comment-157>
marl:polarity marl:Positive
marl:polarityValue "0.6"
```

Fig. 3. Referencing entities (1) and using literals (2) with Marl ontology

4.1 Internet wide keyword search and comparison of opinion values

In the simplest case where opinion ontology would be used only with properties expressed with literals, the structure information (connection between opinion text, opinion value and the full body of text) can still be very useful. Even with the contemporary keyword search engines publishing opinion metadata could make a lot of sense. While the discovery of information remains impaired and inaccurate, once actually having found the desired textually expressed opinions, thanks to the metadata it is possible to compare them or transform in different ways. Furthermore, as the research on semantic metadata indexing [15] progresses it is already possible to utilize these simple relationships to make useful search queries on large data sets (see Fig. 4).

```
* <http://purl.org/marl/ns#extractedFrom> * AND
* <http://purl.org/marl/ns#hasPolarity> <http://purl.org/
                                        marl/ns#Positive> AND
* <http://purl.org/marl/ns#describesObject> "Avatar"
```

Fig. 4. Sindice Semantic Index [15] sample query for: "Search positive opinions about Avatar"

4.2 Internet wide entity based search and/or improved data discovery

One of the envisioned bold goals of Semantic Web is to provide entity based search. This would allow to point exact concepts that the user is referring to and eliminate ambiguity of user query present in the keyword search. Slowly this is becoming achievable much due to popularization of big linked data silos (e.g. DBpedia [2]) and wide adaptation of certain ontologies (e.g. GoodRelations [12]). In our research, we also considered using opinion metadata in such scenario. In comparison to the previous case, instead of using literals to describe opinion context Marl ontology properties could point to the exact concepts defined in one of the commonly refereed datasets. This, for example, would allow to formulate queries that distinguish opinions about "Avatar" movie by James Cameron from other meanings of this word (see Fig. 5).

```
* <http://purl.org/marl/ns#extractedFrom> * AND
* <http://purl.org/marl/ns#hasPolarity> <http://purl.org/
                                        marl/ns#Positive> AND
* <http://purl.org/marl/ns#describesObject> <http://dbpedia.org/
                                        resource/Avatar_%282009_film%29>
```

Fig. 5. Sindice Semantic Index [15] sample query for: "Search positive opinions about Avatar" using DBpedia Avatar entity for disambiguation

From a technical point of view, the establishment of such metadata infrastructure would physically link the opinions together with the Linked Data cloud and therefore each other as well via reference to similar topics. In turn, this would allow to traverse the distributed graph in many different ways for numerous use cases, such as aggregation of opinions (see Fig. 6).

4.3 Semantic search engines for dedicated systems

The large scale entity search engines still cope with a number of problems such as insufficient data, efficiency problems etc. even in aforementioned cases of vertical search (e.g. single topic or content type, like the movies). Nevertheless, we also would like to show that similar techniques, that expose the benefits of Marl ontology, can be very useful even if limited to very narrow systems or groups of heterogeneous systems where most of the problems of Internet wide search are eliminated (e.g. in an enterprise).

Following the example of movies that we used in previous cases, the local search could limit to a single website but thanks to the rich data descriptions with the ontological structure it would enable more precise queries than in text search. In this case Marl fills the gap for describing opinions in conjunction with complex taxonomy trees that enable to query for opinions related to particular elements in the class hierarchy that characterizes the given domain.

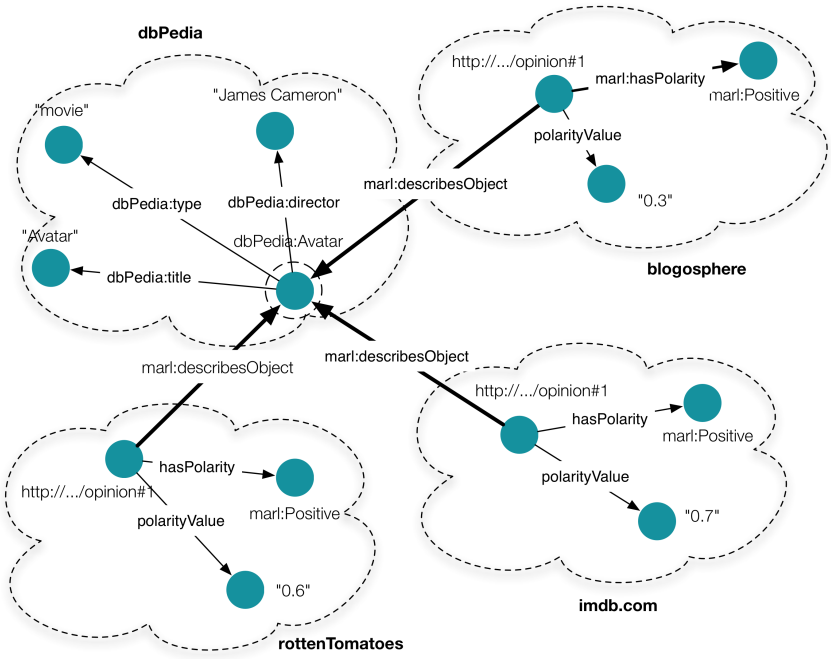


Fig. 6. Sample RDF graph with opinions linked indirectly via metadata references to common entities.

Finally, one can move away from the World Wide Web context to the enterprise environments or other closed systems. In such case the difference is the full control over created data and very strictly defined vocabularies that do not need to be aligned with Web publishing standards. In that case, Marl can be used to together with the verity of enterprise ontologies in the enterprise collaborative systems (e.g. Idea Management Systems or collaborative knowledge management systems). The opinions can be linked via products that they refer to, innovation proposals that are commented by employees, projects in which context the opinions are expressed etc.

5 Evaluation

In order to evaluate our proposal for annotation of opinions we did two experiments. In the first, the goal was to analyse the coverage of the proposed schema against different datasets. In the second experiment we wanted to test in practice how the linked opinion metadata would work with the capabilities of the contemporary search engines and semantic web query endpoints.

During the coverage experiments we analysed two kinds of data: (a) datasets created by other researchers and annotated with opinion mining data; (b) services available on-line that use opinion mining for various goals. The final list consisted of 5 research datasets and 8 services, for each we analysed the data that is exposed and provided Marl mappings. Next, we calculated the coverage as an amount of properties that were possible to describe with Marl over the total amount of data properties used in a dataset. In the first experiment we considered all the dataset fields and the average coverage we got was 64%. However, it has to be noted that the individual characteristics of the data sources varied a lot. According to ontology design goals presented by Noy et al. [14] one of the characteristics of good design is not to cover the very individual elements of datasets. Therefore, after removing the dataset fields that did not repeat at least once, we ran the experiment again and got the average coverage of 76%. The results of the experiments have been summarized in Table 2.

Table 2. Marl ontology coverage experiment results, considering all dataset fields (exp1) and after removing fields that did not repeat at least once (exp2).

Dataset/service name	**#covered/#total**		**coverage**	
	exp1	**exp2**	**exp1**	**exp2**
Congressional speech data [19]	7 / 12	7 / 7	58%	100%
Movie Review Data [16]	3 / 4	3 / 3	75%	100%
Customer Review Data [13]	5 / 9	5 / 6	56%	83%
French Newspaper Articles [8]	1 / 3	1 / 2	33%	50%
Multi-Domain Sentiment Dataset [9]	4 / 4	4 / 4	100%	100%
Swotti (www.swotti.com)	9 / 13	9 / 13	69%	69%
Tweetsentiments (www.tweetsentiments.com)	6 / 11	6 / 11	55%	55%
Mombo (www.mombo.com)	10 / 16	10 / 12	63%	83%
Opinion Crawl (www.opinioncrawl.com)	4 / 9	5 / 9	44%	44%
OPAL (www.gi2mo.org/apps/opal/)	8 / 11	8 / 11	73%	73%
OPfine (www.jane16.com)	6 / 6	6 / 6	100%	100%
Evri (www.evri.com)	3 / 5	3 / 5	60%	60%
Opendover (www.opendover.nl)	4 / 9	4 / 6	44%	67%
Average	**5 / 8**	**5 / 7**	**63%**	**76%**

In the second part of our experiments we tested the capabilities of Marl to be used in context of Semantic Web queries. We started with creating a list of competency questions and tested them against the ontology (a total of 20 query templates where created). Later, for a more practical approach, we extracted small parts of datasets mapped in the previous experiment and used them to check with software prototypes if the queries involving Marl deliver anticipated results with different kinds of search. On this stage the problem that we encountered in most cases was insufficient data to create rich links to expose true power of Marl. Ultimately, for Internet wide data, we did our tests in the context of movie reviews and filtering opinions by polarity from different sites such as Tweetsentiments, IMDB (via Cornell dataset [16]) and Swotti in a single query. We repeated this both for references to movies expressed as

literals and for the entity search (with DBpedia entity references). In both cases we used Sindice search engine as back-end for the demonstration. Finally, for tests of metadata search in closed private environments we have setup a local SPARQL endpoint and used the OPAL opinion mining module in conjunction with technologies from Gi2MO project [21] to extract opinions from independent Idea Management Systems and visualise them together. The additional challenge was that the two systems had data in different languages: one Spanish and the other English. As an outcome, the opinion mining algorithm enabled us to leverage the multilingual instances to the same level but ultimately the Marl ontology in conjunction with other Semantic Web vocabularies worked as an enabler to integrate the systems and run queries over the data to aggregate all information in a single view (e.g. show all ideas with community opinions and compare aggregated opinion scores, or compare the amount of positively received ideas by idea categories etc.). All together the query experiments proved that the ontology is capable in answering all the formulated questions in test scenarios of: movie opinions, product opinions, Idea Management Systems. A common problem, that confirmed the test results of coverage experiment, was that many queries expected the direct link to text fragment of the opinion - which is not facilitated by Marl. An example of a query constructed for data serialized with Marl v0.1 during our experiments can be seen at figure 7.

```
PREFIX gi2mo: <http://purl.org/gi2mo/ns#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX marl: <http://purl.org/marl/ns#>
SELECT ?idea_uri
       COUNT(?negative_opinion_uri) AS ?negative_opinions
       COUNT(?positive_opinion_uri) AS ?positive_opinions
FROM <http://etsit.gi2mo.org/etsit_ideas_en.rdf>
WHERE {
{
    ?idea_uri a gi2mo:Idea .
    ?idea_uri gi2mo:hasComment ?comment_uri .
            ?positive_opinion_uri marl:extractedFrom ?comment_uri .
            ?positive_opinion_uri marl:hasPolarity marl:Positive .
}
UNION {
    ?idea_uri a gi2mo:Idea .
    ?idea_uri gi2mo:hasComment ?comment_uri .
            ?negative_opinion_uri marl:extractedFrom ?comment_uri .
            ?negative_opinion_uri marl:hasPolarity marl:Negative .
} } GROUP BY ?idea_uri
```

Fig. 7. A sample SPARQL query for "Show amount of positive and negative opinions for all ideas submitted into the Idea Management System". The source data was serialized using Marl v0.1 therefore aggregation operator was used to go around the lack of opinion count information.

Concluding both of the experiments, we used the acquired knowledge to produce a second iteration of the ontology (Marl 0.2) and included the new properties that according to our tests were uncovered and often used by other datasets; or were expected as output for search queries. After repeating the coverage experiments with the new version of the ontology we got 79% coverage for experiment 1 (all dataset fields considered) and 94% coverage for experiment 2 (dataset fields that did not repeat at least one time across different sources ignored).

6 Related Work

The research presented in this paper is primary focused on developing a universal model for describing and comparing opinions on the World Wide Web. As such, it is tied to efforts of the Semantic Web research community, which goals have been outlined by Sir Tim Berners-Lee [5]. Furthermore, as much as we are interested in reasoning and giving birth to the intelligent web, our research is focused to a much more extent on the sole goal of publishing and consuming data. Therefore, we have aligned our investigation with the efforts undertaken by the Linking Open Data project[4] - an attempt to build an interlinked Web of Data using Semantic Web technologies.

In terms of related research conducted in those areas, to our knowledge, there has been only one attempt to achieve a similar goal as our. Softic et al. [18] has proposed an opinion ontology and performed a number of experiments to show its use. However, as authors claim themselves the ontology is unfinished and missing the key element of opinion formalization leaving it for later research which has not done yet. In our work we aimed to use the opinion mining as a tool in our main research area of Innovation Management, therefore we needed a full solution for metadata publishing that could be applied in practice. In comparison to Softic et al. we propose a different conceptual model for the opinion ontology, deliver new properties that describe not only a generic concept but enable to publish the numerical values from the opinion mining process (which is impossible using Softic et al. opinion ontology). Furthermore, with our research we propose a different evaluation framework and test our solution in different cases, which in the end delivers new conclusions and opens new possibilities (see Sec. 7).

Within commercial services related to the area of opinion mining there are different data serialization methods used for APIs but all use own vocabularies. In relation to our work, a standing out service by Opendover moves towards the Semantic Web technologies but the vocabularies used refer only to individual sentiments (thus being more similar to a dictionary) rather than full opinions like in case of Marl ontology.

On the other hand, not related to opinion mining, we recognize that for a practical solution, opinions could be conceptually modelled as reviews. Therefore, in terms of related work we also considered vocabularies created for describing online reviews. Among those, the most popular are: hReviews [1], the RDF

[4] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects /LinkingOpenData

mapping of hReview [3], Google's RDF vocabulary for reviews[5] and Schema.org Review vocabulary[6]. In comparison to our work the existing review formalization vocabularies are much more generic and conceptually describe less referring to the entire review body, whereas we see that the opinion ontology needs to describe particular elements of the review and features discussed in the review (e.g. one might imagine a query using both concepts "show all sci-fi movie **reviews** that contain positive **opinions** about director"). Furthermore, we see reviews as judgement based on factual information and comprehensive knowledge whereas opinions are less formal, smaller pieces of information. For those reasons we believe there is a need for making a distinction between the two concepts in terms of metadata and web search.

7 Conclusions and Future Work

In the paper we have presented a solution for describing opinions on the web with well known and widespread metadata standards of Semantic Web. Furthermore, we have shown how adapting the available metadata specification can help to link opinions with other concepts on the web and lead to better search capabilities and improved exposure of data. Whereas, the full potential of the solution depends on the adoption of W3C recommendations such as RDF or RDFa, we have proven that even with the minimal use of entity search, the publishing of metadata about opinions can be very beneficial. In terms of future work, our aim is very much related to more specific domain research and usage of the Marl ontology in synergy with dedicated ontologies to provide complex search facilities in enclosed systems, very much in a manner as described in the article when referring to vertical search engines and search engines for dedicated systems.

Acknowledgements

This research has been partly funded by the Spanish Ministry of Industry, Tourism and Trade through the project RESULTA (TSI-020301-2009-31) and Spanish CENIT project THOFU. We express our gratitude to Paradigma and Atos Origin R&D for their support and assistance as well as providing us access to their software.

References

1. Allsopp, J.: Review and resume microformats: hreview and hresume. In: Microformats: Empowering Your Markup for Web 2.0 (2007)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: In 6th Int'l Semantic Web Conference, Busan, Korea (2007)

[5] http://www.google.com/support/webmasters/bin/answer.py?answer=146645
[6] http://schema.org/Review

3. Ayers, D., Heath, T.: RDF Review Vocabulary Specification. http://hyperdata.org/xmlns/rev/ (2007)
4. Beall, J.: How google uses metadata to improve search results. The Serials Librarian 59(1) (2010)
5. Berners-Lee, T.: Semantic web road map. http://www.w3.org/DesignIssues/ Semantic.html (1998)
6. Berners-Lee, T.: Linked data. http://www.w3.org/DesignIssues/LinkedData (2006)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American pp. 29–37 (May 2001)
8. Bestgen, Y., Fairon, C., Kerves, L.: Un baromètre affectif effectif: Corpus de référence et méthode pour déterminer la valence affective de phrases. In: Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (2004)
9. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL) (2007)
10. Catasta, M., Cyganiak, R., Tummarello, G.: Towards ecsse: live web of data search and integration. In: Semantic Search 2009 Workshop (SemSearch2009) (2009)
11. Hemminger, B.M., Saelim, B., Sullivan, P.F., Vision, T.J.: Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. J. Am. Soc. Inf. Sci. Technol. 58(14) (2007)
12. Hepp, M.: Goodrelations: An ontology for describing products and services offers on the web. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008). vol. 5268, pp. 332–347. Springer LNCS, Acitrezza, Italy (2008)
13. Hu, M., Liu, B.: Mining and summerizing customer reviews. In: Proceeding of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2004)
14. Noy, N.F., McGuinness, D.L.: Ontology development 101: A guide to creating your first ontology. Tech. rep., Stanford Medical Informatics, Stanford (2001)
15. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Tummarello, G.: Sindice.com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies (2008)
16. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL (2004)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2) (2008)
18. Softic, S., Hausenblas, M.: Towards opinion mining through tracing discussions on the web. In: Social Data on the Web (SDoW 2008) Workshop at the 7th International Semantic Web Conference (2008)
19. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of EMNLP (2006)
20. Tian, P., Liu, Y., Liu, M., Zhu, S.: Research of product ranking technology based on opinion mining. In: Second International Conference on Intelligent Computation Technology and Automation (2009)
21. Westerski, A., Iglesias, C.A., Rico, F.T.: A model for integration and interlinking of idea management. In: Metadata and Semantic Research: 4th International Conference, MTSR 2010. Alcalá de Henares, Spain (2010)

Semantic Annotation from Social Data

Geir Solskinnsbakk and Jon Atle Gulla

Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
{geirsols, jag}@idi.ntnu.no

Abstract. Folksonomies can be viewed as large sources of informal semantics. Folksonomy tags can be interpreted as concepts that can be extracted from the social data and used as a basis for creating semantic structures. In the folksonomy the connection between these concepts and the tagged resources are explicit. However, to effectively use the extracted conceptual structures it is important to be able to find connections between the concepts and not only the already tagged documents, but also new documents that have not previously been seen. Thus, we present in this paper an automatic approach for annotating documents with concepts extracted from social data. This is based on representing each tag's semantics with a *tag signature*. The tag signature is then used to generate the annotations of documents. We present an evaluation of the approach which shows promising results towards automatic annotation of textual documents.

1 Introduction

The last years we have seen a growing amount of social services on the web. Amongst these are a wide range of collaborative services that offer users the possibility of tagging a multitude of resources. These resources can be anything on the web, ranging from images, videos to documents. These services can aid the user in organizing information by letting the user attach tags to the resources for easy access at a later time. In addition, the social aspect lets users share resources and tags, so that others can also take advantage of the effort each individual user puts into tagging. There exist many tagging systems, like Flickr (http://www.flickr.com) which lets users share and tag images, Delicious (http://www.delicious.com) which lets users tag and share any resource specified with a URL, Bibsonomy (http://www.bibsonomy.org) which lets users tag and share literature references. Users are free to choose which tags to apply to resources with no centralized control of the vocabulary. The networked data structure resulting from such systems are often referred to as Folksonomies [1].

Tags in folksonomies can be seen as a basis for concept extraction for semantic data structures, which can also be seen in several publications lately [2, 3]. The conceptual structures are one side of the story, however, it is also an interesting problem to connect the concepts (tags) with documents on the web.

This is especially interesting for applications that require search and browsing of the structure and documents. On one hand, we already have a mass of manual annotators (the users of the folksonomy) who generate annotations. Unfortunately, the users have not tagged every single document. This means that there is a huge amount of documents that have not yet been annotated by folksonomy users. Although the documents have not been tagged by users, the documents may be interesting for a browsing facility. Determining the correct annotation of a document automatically is thus the problem we are targeting in this paper. As a solution towards this problem we propose an approach towards fully automatic annotation of documents that have never been seen by the system (i.e. documents that have not yet been tagged by any user). Since we are working on folksonomy data we will use the terms tag and tagging rather than concept and annotation, respectively, for the remainder of the paper. Tags on their own carry only limited semantics. However, we can exploit that the folksonomy can be seen as a large repository of informal semantics to extend the semantics of the tags. This is done by associating each tag with a *tag signature*. The signature takes the form of a vector of semantically related terms, which are weighted to describe the strength of the relations between the tag and the terms in its vector. The tag signature is constructed based on the (textual) resources that have previously been tagged by the users of the folksonomy. By utilizing the tag signatures for suggesting tags to documents, we are using the content (or topic) of the document and tag signature to suggest tags. Thus our approach is not only able to suggest tags to resources that have been tagged before, but also to resources which are new to the system. The approach is evaluated (using training and test data) based on a data set crawled from Delicious. The results of the evaluation are promising in terms of automatically assigning tags to documents.

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work, while Section 3 gives an overview of tag signatures and the approach for automatic tag suggestion. Section 4 describes the evaluation and results, followed by a discussion of our findings in Section 5. Finally the paper is concluded in Section 6.

2 Related Work

The related work for this paper is directed at tag recommender systems, since these systems essentially provide some of the same functionality that we are targeting.

Mishne [4] presents an approach for suggesting tags for weblog posts. This is done by first finding similar weblog posts using information retrieval techniques. The tags used on the most similar posts are retrieved and ranked before being presented to the user. Another system for tagging of blog posts is described by Qu et al. [5]. The system uses key phrase extraction applied to the blog content to find tags which can be applied to the blog post. The system described by Baruzzo et al. [6] also uses key phrase extraction for generating tag recommendations to the user. The keyphrases are extracted from the text and mapped

to domain ontology. Spreading activation is employed in the ontology to locate common ancestors which are presented to the user as new tag recommendations. In [7], Lipczak et al. present an approach based on a combination of extracting candidate tags from the resource and using information found in the folksonomy. Candidate tags are found from the title and the URL of the resource, tags related to the resource, and tags related to the user.

Musto et al. [8] apply a combination of content-based and collaborative-based approaches to generate tag recommendations. The content-based approach analyzes the resource to tag, and extracts candidate tags from the URL, the HTML title and meta tags. The candidates are scored by taking into account type of source (URL, title etc.) and the occurrence frequency within each source type. The collaborative approach searches an underlying corpus of users, resources, and tags to find candidate tags. Finally the user is presented with tags from one or both of the candidate tag sets based on some strategy. Jäschke et al. [9] present two different algorithms for tag recommendation based on folksonomy data. The first is based on collaborative filtering, and the second is based on the FolkRank algorithm. Gemmell et al. [10] describe an approach for tag recommendation based on adapting the k-nearest neighbor algorithm to folksonomy data.

Most current methods use either the content of the resource (key phrase extraction), or the data found in the folksonomy as a source of tags to recommend to the user. Our approach to automatic tagging is based on a combination (even though we do not extract tags from the content). We use the information in the folksonomy (the mapping from tag to resource) and the content of the resource to build a semantic representation of each tag. In this way our approach is able to suggest tags (that are used in the folksonomy) to documents that have not been seen before. Systems that purely use the graph structure of the folksonomy to recommend tags, will suffer when trying to recommend tags to a resource not previously seen. On the other hand, systems that purely rely on extracting tags from the content may lead to an increase in the tag vocabulary. Hence, reusing tags that already exist in the folksonomy will ensure that the vocabulary in the folksonomy is consolidated.

3 Tag Signatures

Users that contribute within a community to tag and share resources on the web generate what is often referred to as a folksonomy [1]. Folksonomies consist mainly of three entities; (1) users; (2) tags; and (3) resources. Bookmarking is the action of a user attaching one or more tags to a specific resource, and the combined data is called a bookmark.

Heymann views this data as *triples*[11] {user, tag, URL}. The interpretation of the triple is that *user* has applied *tag* to the resource identified by *URL*. As the user has actively engaged in applying the tag(s) to the resource we make a basic assumption that the tag(s) make up a description of the documents' content. In terms of the user, the tag(s) applied signal the semantics of the resource and

should be representative for the resource's content, so it is later easy to find (both for the user himself and others in the community).

The assumption made above is used as a basis for generating an extended semantic representation of the tags using the contents of documents to which a tag has been applied. This representation associates each tag in the folksonomy with a vector of semantically related terms. Each term is given a weight that reflects the importance of the term with respect to the tag. This means that a term can be connected to several different tags, but with different weighting, signaling that the term has a different importance with respect to each tag. We refer to our semantic representation as a *Tag Signature*. Two different considerations are made when deciding how to weigh the terms in each tag signature. The first is that the weight should reflect the internal semantics of the tag. This means that we want to give a high weight to terms that are good at characterizing important aspects of the tag. The second is that we want the weight to reflect the external semantics of the tag. This in essence means that we want the term to be good for discriminating this tag from others. Thus we apply the $tf \cdot idf$ [12] measure for weighting the terms in the signatures. The collection of terms and their weights collectively represent the semantic content of the tag, and we thus refer to the tag signature as an extended semantic representation of the tag, which greatly extends the pure syntactic representation of the tag. The tag signature materializes as a vector. The definition is given in [13] (in [13] we use the term Tag Vector), but we repeat it here for convenience as Definition 1.Details of the construction of the tag signature can be found in [13].

Definition 1. *Tag Signature. Let V be the set of n terms (vocabulary) in the collection of tagged resources. $t_i \in V$ denotes term i in the set of terms. The tag signature for tag j is defined as the vector $T_j = [w_1, w_2, \ldots, w_n]$ where each w_i denotes the semantic relatedness weight for each term t_i with respect to tag j.*

3.1 Unsupervised Tagging Approach

Unsupervised tagging can be used in many application areas such as tag recommendation, automatic tagging of a set of documents, document classification, etc. Our approach to automatic tagging takes as input an untagged document and returns a ranked list of tags. The similarity between the document content and the tag is based on the tag signature. Since the tag signature is represented as a vector of weighted terms, and similarly the document can be viewed as a vector of weighted terms, we propose to use the cosine measure to calculate the similarity between the two. The calculation is shown as Equation 1 [12], where $w_{i,d}$ is the weight of term t_i in the document, $w_{i,j}$ is the weight of term t_i in T_j, and n is the number of terms. In our implementation, we have stored all tag signatures in a tag signature index, and use the document as a large query into the tag signature index. The list of tags returned can be cut off at top m tags, or at a threshold for the similarity score.

$$sim(d, T_j) = \frac{\sum_{i=1}^{n} w_{i,d} \times w_{i,j}}{\sqrt{\sum_{i=1}^{n} w_{i,d}^2 \times w_{i,j}^2}} \tag{1}$$

Our approach does not increase the tag vocabulary (as for instance key word extraction techniques might do by proposing new tags). This is a benefit since the document will be tagged according to the already used tags. This means that we can classify the documents according to tags that are already used and are found in the semantic structure. However, if the coverage of the tags is not sufficient, it may be the case that new tags have to be introduced. In such cases the system could have as fallback strategy to implement one of the content based tag suggestion algorithms found in the literature. Another benefit is that the extended semantic representation of the tags allows us to adapt the semantics of a tag to the way it has been used by the users. This implies that a tag may have a different tag signature in different communities, since the tags may be used in slightly different contexts. However, this also means that there will be domain restrictions to the approach. For automatic tagging of good quality we are reliant upon a good coverage of the domain.

4 Evaluation

The experiment is performed on a data set from Delicious that we crawled between December 2009 and January 2010. We only kept bookmarks pointing at resources under "http://en.wikipedia.org/wiki/", the English section of Wikipedia. The crawl resulted in 228536 bookmarks created by 51296 users, 72420 unique tags, and 65922 unique URLs. We kept only English Wikipedia documents so that we could map the documents to a dump of Wikipedia (from June 2008) which has been cleaned and Part of speech (POS) tagged [14]. We performed some simple filtering of the crawled data, removing bookmarks pointing at certain document classes. All bookmarks pointing at documents prefixed with *category:*, *user:*, *image:*, etc. were removed from the delicious data set. This filtered 14162 bookmarks. We were able to map the URLs in 91.2% of the remaining bookmarks to the Wikipedia dump, leaving us with a total of 195471 bookmarks. Mapping failures may have been due to encoding problems, articles that have moved, etc. Next, we filtered the bookmarks based on tags. This was done by lowercasing tags and removing all tags that had not been used by at least 5 users and in 25 bookmarks. This is to ensure that we remove some of the noisy tags found in folksonomies, and assure that the tags have been sufficiently used. The final tag set consisted of 2988 tags (used to tag 59610 documents).

The data set has been randomly split into two parts based on the documents, one for generating the *tag signatures* (training set) and one for the evaluation (test set). The training set consists of 29845 documents while the test set consists of 29765 documents. The *tag signatures* have been constructed according to the description given in Section 3. Further we have performed the evaluation using both standard preprocessing and by extracting terms based on POS tags in the Wikipedia collection. The POS based pre-processing is based on extracting only noun phrases from the text, splitting phrases and stemming individual terms.

The first part of our evaluation is designed to find how well the tag assignments made by our approach corresponds with the tags assigned to the docu-

ments by the folksonomy users. This is done by constructing the tag signatures based on the training set and comparing the tag assignments generated by our approach in the test set with the original tag assignments in the bookmarks of the test set. As a simple base line, we have chosen to use keyword search (named KW Tags). The keyword search is performed by using each tag in the folksonomy (same tag set as we use for tag signatures) as a keyword query matched against the document and generates for each document a ranked list of tags which we compare our method to. All indexing and search has been implemented using Lucene[1].

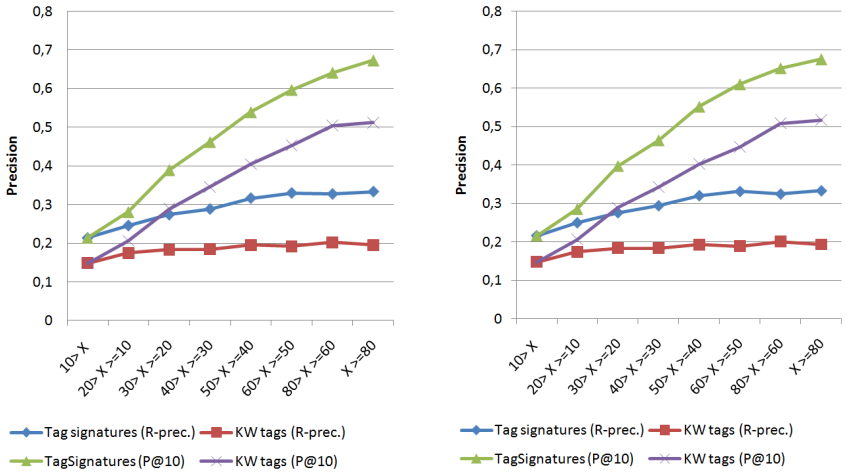
The second part of the evaluation, the user evaluation, has been performed by presenting a group of 6 persons (including one of the authors) with 15 randomly selected documents. For each document the user has been presented with the top 10 ranked KW Tags, and the top 10 Tag Signature based tags (in random order). Tags that have been used to tag each document in the original folksonomy data set have been removed from the evaluation set. Thus we are evaluating only new tag assignments. This is done to learn more about the quality of the tags that are suggested but that have not previously been used to describe the documents. In case of overlap between the two result sets, the list of tags has been padded with extra tags so that the user always is presented with 20 tags. The evaluators used a 5 point scale in which 1 meant that the tag was not appropriate to describe whole or parts of the document content, while 5 meant that the tag was highly descriptive of whole or parts of the document.

4.1 Results

In the first part of the evaluation, we investigate how well our results compare to the tag assignments made by users in the folksonomy. We have used the training set to generate the tag signatures and the test set for evaluation. This means that the text of the documents we evaluate is not incorporated in the training phase. Consequently, the set of bookmarks has been split in two, one for the training set and one for the test set. We have calculated two different measures, the R-precision, and Precision @ 10 (P@10). R-precision for the tag assignments of a single document is calculated by taking all tags assigned to the document by the users (of the folksonomy; the original tag assignments) in the test set as the relevant set of tags, R, with $|R|$ elements. Next we take the top $|R|$ results from KW Tags and our method and calculate the precision in these sets. We also check the precision in the top 10 tags (ranked by the cosine measure) as these tags are the most interesting to suggest to users. We have grouped the results according to the number of unique tags assigned to the documents (Figure 1), the number of times a user has tagged the document (Figure 2(a)), and the size of the documents after preprocessing (Figure 2(b)).

The average R-precision value calculated over all documents in the test set is 0.224 for our approach and 0.155 for the keyword based approach. The average P@10 is 0.238 for our approach and 0.168 for the keyword based approach.

[1] http://lucene.apache.org



(a) Standard preprocessing

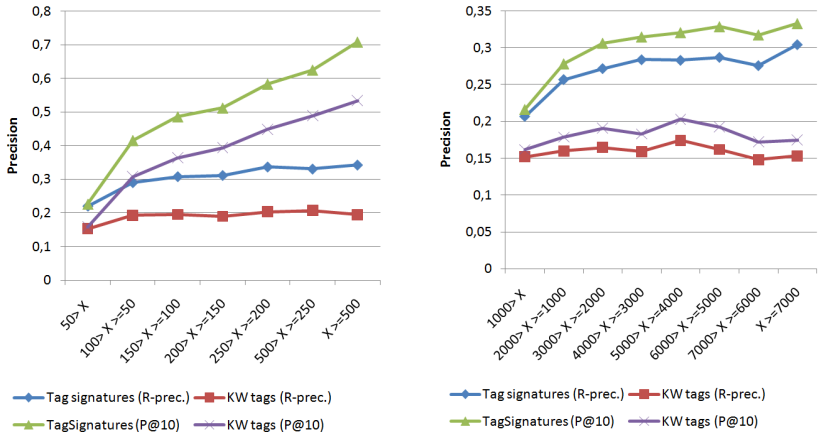
(b) POS based preprocessing

Fig. 1. Results grouped by number of unique tags (X) assigned to each document.

Figure 1(a) and 1(b) show the results of KW Tags and our method based on standard preprocessing and POS tag based preprocessing, respectively. The results show that the quality of the two approaches seem quite comparable, thus using the POS information does not improve the quality of the results significantly. Next we can note from the figure that our results are consistently significantly better through all groups than using the pure keyword based approach. Manual examination of the results also shows that our approach is able to find tags that are not present in the document text.

In Figure 2(a) we have grouped the results according to the number of tag assignments to each document. These results show the same trends as the previous graph, as should be expected, since there is a correlation between the number of unique tags assigned to a document and the total number of tag assignments to a document. As the number of tags assigned by users to a document increases, so does the probability of being able to suggest one of these tags. The increase in the experiment metrics with increasing number of unique tags/tag assignments should thus cater for at least parts of this effect.

Figure 2(b) shows the results grouped by document size (after preprocessing). From the graph we can see that our approach scores consistently higher than KW Tags for both measures. From the figure we can see that the results from the KW Tags seems quite stable with only small changes as the size of the documents increase. The approach based on the tag signature on the other hand seems to increase, but with a lower rate as the document size increases as in a logarithmic function. This is an quite interesting result. Since the tag signatures have the form of a vector, we should expect that the number of tags that mach a given document increase as the document size increases (the number of



(a) Number of user tag assignments (b) Document size in number of terms

Fig. 2. The figures 2(a), and 2(b) show the results grouped by the number of user tag assignments and the document size in number of terms, respectively.

potential keywords to match increase). This should also be visible for the KW tags case. However, the results do not show this kind of effect, rather a decrease in the evaluation metric as the number of document terms passes 5000. Thus we interpret this as a result pointing towards that the added semantics of the vectors are able to generate better suggestions.

Figure 3 shows the results from the user evaluation. The data series named *Tag Signatures* is based on the top 10 tags suggested by our approach, while the data series *KW Tags* is based on the top 10 tags suggested by using the existing tags in the system as keyword queries into the documents. Tags that have been used to tag these documents in the folksonomy data set have not been evaluated. Thus the tags evaluated are "new" to each of these documents. The evaluation is performed to check the quality of the remaining tags from the first part of the evaluation, i.e. tag assignments from our system that are not present in the form of bookmarks in the data set collected. The graphs show that the quality of the tags were assessed by the evaluators to be, on average, of higher quality for the *Tag Signature* data series in 10 out of 15 documents. The average value was found to be 3.18 for tags suggested by our approach and 2.91 for tag suggested by the keyword based approach. Although not statistically significant results, we see this as a positive tendency. Manual examination of the documents and tag evaluations showed that there was some disagreement (as can also be seen from Table 1 which shows the standard deviation of the user evaluation scores). This seems to point towards that it is hard to understand the mechanisms that lie behind tagging. It seems that one tag may be valuable to one user, while it is not that valuable to others. The users' intention when tagging (or evaluating a tag in our case) seems to be very important. Some users would like to tag based on the general topic of the document, while others may want to tag based

on certain details in the document. This makes it hard to evaluate tagging on single documents, and our approach seems to be more appropriate when we take a large sample of documents into consideration. Two types of tags our approach seems to not handle satisfactory are subjective tags and very general tags (like interesting, history, etc.). Subjective tags are hard to handle in general and will be discussed further in the next section. Very general or broad terms may cover a very wide topic (like history, which can be used to tag documents about World War II and music history in the 60's). This can however be viewed as a variation on tag ambiguity which we address in the next section.

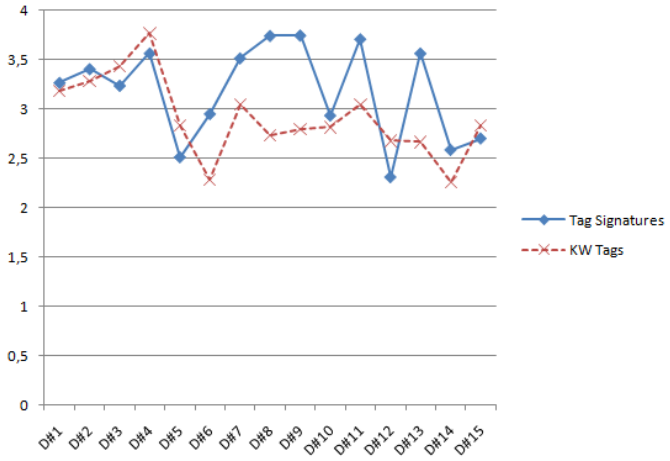


Fig. 3. The results from the user evaluation. Based on 15 randomly chosen documents.

Table 1. The standard deviation of the results from the user evaluation.

Exp./Doc.	D#1	D#2	D#3	D#4	D#5	D#6	D#7	D#8	D#9	D#10	D#11	D#12	D#13	D#14	D#15
$\sigma_{TagSign.}$	1.550	1.379	1.546	1.544	1.601	1.502	1.334	1.198	1.385	1.469	1.160	1.380	1.395	1.544	1.476
σ_{KWTags}	1.525	1.427	1.358	1.280	1.703	1.379	1.455	1.388	1.527	1.266	1.443	1.479	1.481	1.469	1.510

Table 2 shows the tag assignments given by our approach and by the keyword based approach for the document "Comparison of layout engines (HTML5)" (based on the 2008 Wikipedia dump). The results show that the two approaches have an overlap of two tags, *firefox* and *xhtml*. If we polarize tag suggestions as being either good or bad and define good tag suggestions as those with an average score above 3, we see that in our approach 9 out of 10 suggested tags qualify, while for the keyword based approach only 5 out of 10 qualify.

Table 2. Example set of tags for the document "Comparison of layout engines (HTML5)" (2008 version).

Tag signatures		KW Tags	
Tag	**Score**	**Tag**	**Score**
ie	4.5	firefox	4.2
firefox	4.2	xhtml	4.2
xhtml	4.2	engine	3.3
compare	3.7	emulation	2
mozilla	3.8	values	1.8
xforms	3.8	xml	4
webstandards	4.2	input	2
png	1.8	property	1.7
css	4	experimental	1.2
xslt	3.3	internet	3.7

5 Discussion

The results described in the previous section show that our approach using tag signatures for automatic assignment of tags to documents previously not seen by the system has quite good performance. However, when looking at P@10 (average 0.238), we see that we are not able to find all tags applied to the documents by users of the folksonomy. What about the quality of the remaining tags suggested? The second part of the evaluation was supposed to give us an answer to this question, but due to disagreement among the users, it is hard to give a conclusive answer. In fact, the disagreement among the evaluators highlights the problem of evaluating tag assignments. The intention of a user is highly relevant as discussed in the previous section. We found that on average, the score given to tags suggested by our system (3.18) seems to indicate that tags suggested by our system have some positive aspects. Thus although we do not have any conclusive evidence, P@10 would most likely be higher since our system suggests tags that, even though not applied to the document by users in the folksonomy, seem to make sense among the evaluators. For a definite answer to the question of the overall quality of the tag suggestion, we would need to perform a larger evaluation.

One of the strengths of our approach is, in our opinion, that it is able to assign tags to documents that have not been seen by the system previously. We are thus not as bound as methods that adhere to strictly using approaches based on collaborative filtering. The tags suggested are based on the content of documents previously tagged with each tag, and the terms are weighted based on balancing the internal and external representation of the tag. Thus we might say that our approach is a combination of content based and folksonomy based tagging. Further the positive results we have achieved, tell us that the quality of the tag signatures seems reasonable, they are able to describe the characteristics of the tags in terms of a weighted vector of terms.

Tag disambiguation is a concern that we have not addressed in the current phase of our research. Tags have a tendency to be ambiguous (polysemy, homonymy etc.), which is also described in the literature (e.g. in Heymann et al. [15]). Take for instance the tag *apple*. Apple can be used in the computer company sense or in the fruit sense. In our case, tag ambiguity may cause the tag signatures to be imprecise, meaning that they span two or more specific topics (causing drift of the signature). This may have affected our results negatively, by suggesting inappropriate tags to documents. Tag ambiguity can however be reduced by applying one of several measures found in the literature for tag disambiguation (e.g. in Garcia-Silva et al. [16] or Angeletou et al. [17]). In our approach tag disambiguation could be applied during tag signature construction, and would generate several tag signatures (one for each sense) for ambiguous tags. Subjective tags would also give rise to some degree of ambiguity. How do you quantify what *cool* or *interesting* means? These types of tags are hard to deal with in automatic systems, since what one person finds interesting may be uninteresting to another. Thus these types of tags are rather useless to apply in automatic systems, and these kinds of systems should focus on the topic of the document.

6 Conclusions

In this paper we have presented an approach for automatically annotating documents with folksonomy tags using tag signatures. The signatures are materialized as a vector of weighted terms, in which the weights reflect the semantic relatedness of the term with respect to the tag. Our evaluation shows that our approach beats naive tagging, using a direct match between tag and document. We found that we are able to annotate documents in which the tag is not present using the tag signature as a semantic connection. Further, the annotations are not made purely based on what a document has been tagged with in the folksonomy, but takes into account the content of the document as well. The evaluation is based on presenting annotations to documents that have not been seen by the system before and interpret this as evidence that our tag signatures carry more semantics that the tag on its own.

Acknowledgment. This research was carried out as part of the IS_A project, project no. 176755, funded by the Norwegian Research Council under the VERDIKT program.

References

1. Thomas Vander Wal. Folksonomy coinage and definition. http://vanderwal.net/folksonomy.html, Accessed February 8. 2011.
2. Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *The Semantic Web - ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer Berlin / Heidelberg, 2005.

3. Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, Raleigh, NC, USA, 2010.
4. Gilad Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of 15th International Confernece on World Wide Web (WWW)*, pages 953–954. ACM Press, 2006.
5. Lizhen Qu, Christof Müller, and Iryna Gurevych. Using tag semantic network for keyphrase extraction in blogs. In *Proceedings of 17th Conference on Information and Knowledge Management*, pages 1381–1382. ACM, 2008.
6. Andrea Baruzzo, Antonina Dattolo, Nirmal Pudota, and Carlo Tasso. Recommending new tags using domain-ontologies. In *WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 409–412. IEEE, 2009.
7. Marek Lipczak, Yeming Hu, Yael Kollet, and Evangelos Milios. Tag sources for recommendation in collaborative tagging systems. In *Proceedings of the ECML/PKDD 2009 Discovery Challenge Workshop*, pages 157–172, 2009.
8. Cataldo Musto, Fedelucio Narducci, Pasquale Lops, and Marco de Gemmis. Combining collaborative and content-based techniques for tag recommendation. In *Proceedings of 11th International Conference on E-Commerce and Web Technologies (EC-Web)*, volume 61 of *LNBIP*, pages 13–23. Springer, 2010.
9. Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 4702 of *LNAI*, pages 506–514. Springer, 2007.
10. Jonathan Gemmell, Thomas Schimoler, Maryam Ramezani, and Bamshad Mobasher. Adapting K-Nearest Neighbor for Tag Recommendation in Folksonomies. In *Proceedings of the 7th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems*, 2009.
11. P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *First ACM International Conference on Web Search and Data Mining (WSDM'08)*.
12. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.
13. Geir Solskinnsbakk and Jon Gulla. A hybrid approach to constructing tag hierarchies. In *On the Move to Meaningful Internet Systems, OTM 2010*, volume 6427 of *Lecture Notes in Computer Science*, pages 975–982. Springer, 2010.
14. J. Artiles and S. Sekine. *Tagged and Cleaned Wikipedia*. Available from http://nlp.cs.nyu.edu/wikipedia-data/, Accessed December 2009.
15. Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008.
16. A. Garcia-Silva, M. Szomszor, H. Alani, and O. Corcho. Preliminary results in tag disambiguation using dbpedia. In *CKCaR'09: Proceedings of the 1st International Workshop on Collective Knowledge Capturing and Representation at K-CAP 2009*, 2009.
17. Sofia Angeletou, Marta Sabou, and Enrico Motta. Semantically enriching folksonomies with flor. In *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at ESWC*, 2008.

Using SKOS to Integrate Social Networking Sites with Scholarly Information Portals

Arnim Bleier, Benjamin Zapilko, Mark Thamm and Peter Mutschke

GESIS-Leibniz Institute for the Social Sciences
{arnim.bleier, benjamin.zapilko, mark.thamm,
peter.mutschke}@gesis.org

Abstract: Web 2.0 platforms have become a ubiquitous way of information exchange, but are seldom integrated with the Web of Data. To overcome this situation we propose the usage of SKOS thesauri acting as back-of-the-book index providing domain-specific axes transcending applications. We illustrate this concept with a use-case in the social sciences domain but applications in other domains are possible.

Keywords: User-Generated Content, Digital Libraries, Semantic Web, SKOS, Social Sciences

1 Introduction

The Social Web represents the vision of user-friendly online platforms fostering the generation of content in a collaborative manner. In the case of emerging scholarly online platforms it is the hope that innovative patterns of knowledge creation and dissemination enhance collective intelligence by overcoming the role-asymmetry between producer and consumer known from the Web 1.0. Members of such communities can form virtual networks around topics of their academic interest propelling the exchange of ideas along the social graph. However promising this might be, the concept breaks with the borders of the application and leaves a gap left to fill for enabling linked and open Web 3.0 information systems.

In the remainder we suggest domain specific LOD traversal axes crosscutting application boundaries. We begin with the problem statement. Next we sketch out the process of linking heterogeneous academic data sets via SKOS thesauri and later extend this idea to unstructured user-generated content found on Social Networking Sites (SNS). We conclude with having a look into challenges still faced as well the current status of a prototype.

2 Problem Statement

Leaving the distinction between content provider and user with the advent of the so called Web 2.0 was the empowerment leading to a flood of content with only light-

weight explicit semantics such as mentions of other users or tagging. But as Schmidt rightfully argues [1] this new "produser" role suffers another distinction, this time towards the ontology manager and in the case of scholarly Web 3.0 platforms prohibiting a rich semantic encoding of new ideas in the first place.

The W3C Incubator report on a "Standards-based, Open and Privacy-aware Social Web"[2] provided valuable visions and bottom up projects (e.g. GNU social or Diaspora) have targeted cross-application interoperability, but few of these initiatives focused on the empowerment of users when it comes to the semantic dimension of their content to provide a linkage to rest of the Web of Data. A related situation is faced on the side of heterogeneous LOD services; this time however highly specialized application specific vocabularies prevent a coherent picture to emerge and make it difficult to traverse (scientific) data along domain concepts.

We argue that both cases are related in requiring a domain-specific representation that is precise enough to capture existing concepts, but also leaves flexibility to express new ideas.

3 Integrating Heterogeneous Data on the Web of Data

In this section we will have a closer look at the process of building up LOD services at GESIS combining fine-grained vocabularies for individual data sets and more coarse-grained representations for the thematic traversal of the social sciences domain. As the leading German social sciences infrastructure facility GESIS publishes large amounts of scientific information in form of library references, survey studies and corresponding statistical data sets on several sites (i.e. Sowiport[1], SOFIS[2] or ZACAT[3]) addressing different use cases and user categories. The development of such targeted applications scenarios yielded services enjoying high usage. However, leaving many information sources unconnected proved disadvantageous for growing beyond the originally foreseen use cases. Addressing these shortcomings with a complete rebuild of applications was not a choice, but a integrative approach was needed. The Simple Knowledge Organization System (SKOS) proved as a useful choice, allowing classic knowledge representations to be encoded, in terms of a high-level thesaurus, for the Web of Data. With the RDF representation of the Thesaurus for the Social Sciences [3] (TheSoz) a formal multilingual representation of the social sciences domain has been developed. This TheSoz[4] acts as a back-of-the-book index for the social sciences and glues together data items belonging to various application domains; now we are looking into ways extending this concept to third-party applications such as academic Social Networking Sides.

[1] http://www.gesis.org/sowiport/
[2] http://www.gesis.org/en/services/research/sofis-social-science-research-information-system/
[3] http://zacat.gesis.org/webview/
[4] http://lod.gesis.org

4 Connecting Social Networking Sites

As we have addressed GESIS is providing different kinds of data sets and the Thesaurus for the Social Sciences provides the glue making it possible to traverse them along domain-specific axes. A similar situation is faced in case of integrating Social Networking Sites. If one agrees that applications centered on user-generated content should be aligned with the Web of Data a two-way mechanism would be needed to support ingoing as well as outgoing links to and from further LOD resources. While the subject of supporting application/rdf+xml request types on a partnering SNS is an open issue, progress on supporting outgoing links and requests to further LOD resources has been made. Users of our prototype can either manually select TheSoz concepts tags they think are suitable to their contribution(s) or use an automatic suggestion service recommending appropriate thesaurus concepts. The usage of an automatic suggestion service proves in particular useful since it requires only little user knowledge of the vocabulary itself and makes adoption of the service more likely. While these "TheSoz tags" can act just as traditional tags with a human readable label integrating seamless into the expected user experience, they are in fact a smart resource. Since the thesaurus is multilingual, literal forms of labels provide translations and semantic relations with other thesaurus concepts provide refinement and inclusion [4] in the tag-space. Moreover the semantic machine- and human-readable meaning of these tags does not end at artificial application boundaries but provides connections to other applications for the traversal of information along axes of user interests.

5 Current Status and Challenges

The discussed concept has still a long way to go. A particular challenge to get initial user involvement is the optimization of the multi-label classifier used in the TheSoz concept recommendation and consequently we are considering ways to integrate multi-modal data (e.g. mentions or the structure of discussion threads) into the feature space of the classifier. Most importantly to us, however, will be the user feedback to our prototype on the iversity[5] platform launching this fall.

References

1. Schmidt, J. and Pellegrini, T.: Das Social Semantic Web aus kommunikationssoziologischer Perspektive. In: Social Semantic Web, pp. 453—468. Springer (2009)
2. Harry, H., Tuffield, M.: A Standards-based, Open and Privacy-aware Social Web: W3C Incubator Group Report. W3C (2010)
3. Zapilko, B., Sure-Vetter, Y.: Converting the TheSoz to SKOS. GESIS Report (2009)
4. Miles, A. and Bechhofer, S.: SKOS simple knowledge organization system reference. W3C (2008)

[5] http://www.iversity.org

Social Semantic Web Access Control*

Serena Villata, Nicolas Delaforge, Fabien Gandon, Amelie Gyrard

INRIA Sophia Antipolis {firstname.lastname}@inria.fr

Abstract. In the Social Web, the users are invited to publish a lot of personal information. These information can be easily retrieved, and sometimes reused, without providing the users with fine-grained access control mechanisms able to restrict the access to their profiles, and resources. In this paper, we present an access control model for the Social Semantic Web. Our model is grounded on the Social Semantic SPARQL Security for Access Control Ontology. This ontology can be used by the users to define, thanks to an Access Control Manager, their own terms of access to the data. Moreover, the Access Control Manager allows to check, after a query, to which extent the data is available, depending on the user's profile. The evaluation of the access conditions is related to different features, such as *social tags*, contextual information, being part of a group, and relationships with the data provider.

1 Introduction

One of the key features of Social Web is the ability to publish, and thus find a lot of personal and professional information about people. With the advent of Social Semantic Web, this is more evident, as underlined by Breslin et al. [2]. This availability of personal data of the users has both positive and negative sides. On the one hand, this allows people to share their data, e.g., photos, videos, posts, with their friends and the persons they know. On the other hand, semantic forms of the users' profiles can be reused elsewhere, e.g., what happened with FOAF search engines and aggregators as Plink, or FoaFSpace. This leads to the need of mechanisms whereby users can restrict the access to their data.

In this paper, we address the research question: *How to define an access control model for the Social Semantic Web?* This question has to deal with different aspects that need to be taken into account when designing a model of access control for the Social Web. First of all, we avoid the usual access control lists, often maintained by a sole authority, because we cannot specify the access restrictions to any particular user, in a context where the user information is so dynamic as in the Social Web. Second, we rely on the social tags assigned to the users and their data. Moreover, the contextual information are considered in this model, i.e., time constraints, geo-localization information, maximum of accesses to a resource. Finally, the model supports a user friendly interface allowing both expert, and non expert users to define their own terms of access.

* The authors acknowledge support of the projects ISICIL ANR-08-CORD-011 and DataLift ANR-10-CORD-09 funded by the French National Research Agency.

We define the Social Semantic SPARQL Security for Access Control vocabulary (S4AC), a lightweight ontology which allows the users to specify fine-grained access control policies for their RDF data (Figure 1). At the core of S4AC is the Access Condition which is a SPARQL 1.1. `ASK` clause that specifies the condition to be satisfied in order to grant the access to a resource. Moreover, the users can define Access Conditions based on *tags* which restrain the conditions to the resources tagged with such tags, e.g., resources tagged "friends", "amici", "ami". The conditions can be bound to specific values to provide an Access Evaluation Context, e.g., `<''?user'', <http://myExample.net#sery>>` where the URI of the user is bound to `<http://myExample.net#sery>`. Finally, the Access Condition is associated with a temporal and spatial validity. The Access Privilege, instead, defines which kind of privilege is granted to the user satisfying the Access Conditions and the contextual constraints, e.g., `s4ac:Read` grants the user the privilege to read the requested data. Moreover, we introduce the Access Control Manager letting the users in the Social Semantic Web to (i) define the access conditions for their RDF data, e.g., their FOAF profile, and (ii) filter the RDF data depending on the access conditions the user who wants to access satisfies.

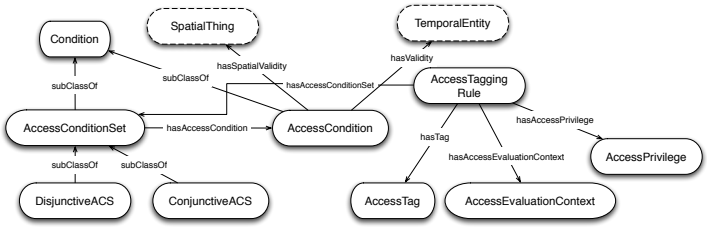


Fig. 1. An overview of the S4AC Ontology.

A key feature of our approach is to rely only on Semantic Web languages. As a consequence, our access control model is platform independent, and can be used by any kind of system based on those languages. In particular, the semantics of our policies is grounded in SPARQL 1.1[1] `ASK` queries. Relying on SPARQL semantics, our model allows the user to submit arbitrary queries while enforcing fine-grained access rules on the results he will receive. If the result of the `ASK` query is *true*, then the user is provided with the information he requires. If the result is $false$, then the model returns to the user a denial coupled with one or more rule labels explaining the reasons of the denial.

The reminder of the paper is organized as follows: Section 2 provides a description of the S4AC ontology and the kind of policies which can be defined using such ontology, and Section 3 presents the access control model and describes the developed prototype. Related work and conclusions end the paper.

[1] http://www.w3.org/TR/sparql11-query/

2 Social Semantic SPARQL Security for Access Control Ontology

The Social Semantic SPARQL Security for Access Control Ontology (S4AC), online at `http://ns.inria.fr/s4ac/v1#`, is detailed in Figure 2. One of the key features of our access control approach is to be integrated with the models adopted in the fields of the Social Web, and of the Web of Data. In particular, S4AC reuses concepts from SIOC[2], SCOT[3], NiceTag[4], WAC, TIME[5], GEO[6], and the access control model as a whole is grounded on further existing ontologies, as FOAF[7], Dublin Core[8], and RELATIONSHIPS[9].

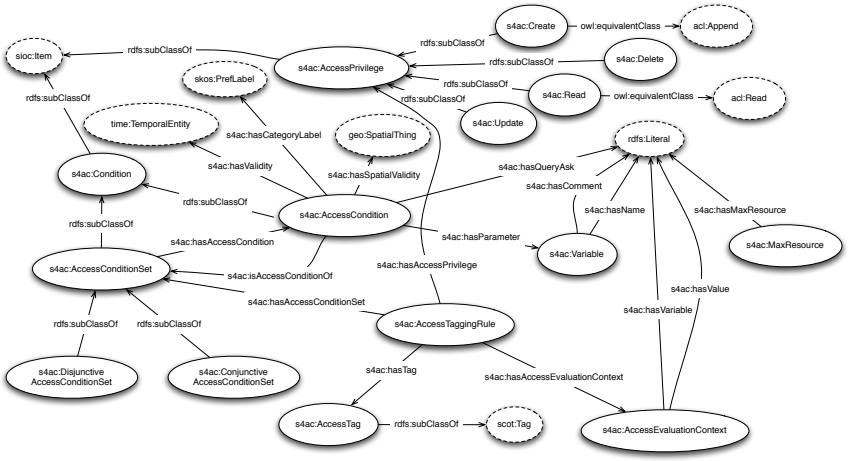


Fig. 2. The S4AC Ontology.

The main class of the S4AC ontology is the class *AccessCondition*, which is a subclass of the class *Condition*, itself a subclass of `sioc:Item`.

Definition 1. *An Access Condition (AC) is a SPARQL 1.1* `ASK` *query. If a solution exists, the* `ASK` *query returns true, and the Access Condition is said to be verified. If no solution exists the* `ASK` *query returns false, and the Access Condition is said not to be verified.*

[2] http://rdfs.org/sioc/spec/
[3] http://scot-project.net/
[4] http://ns.inria.fr/nicetag/2010/09/09/voc.html
[5] http://www.w3.org/TR/2006/WD-owl-time-20060927/
[6] http://www.w3.org/2003/01/geo/wgs84_pos
[7] http://xmlns.com/foaf/spec/
[8] http://dublincore.org/documents/dcmi-terms/
[9] http://vocab.org/relationship/.html

The *Access Condition* grants or restricts the access to the data. If the `ASK` returns *true*, the access is granted to the user. In order to return the user a more informative answer if the access is denied, we introduce the property *hasCategoryLabel*. This property allows to associate to each AC one or more natural language labels which "identify" the access condition, and they are returned to the user to provide him the reasons of the denial. We cannot return the user all the access conditions, because this would make him aware of the policies of the provider. If it is the case that only some results are filtered, it is a matter of the access control model whether to communicate or not, thanks to the *hasCategoryLabel* property, that an access restriction has been applied. The *AccessCondition* defines two properties of the access polices: *hasValidity*, and *hasSpatialValidity*. They allow to define the validity of an Access Condition. Thanks to the use of the concept `time:TemporalEntity`, the validity can be expressed in various ways: valid from/through a specific date/time, or valid in a specific time interval. *hasSpatialValidity*, instead, deals with the spatial localization of the user at the moment of trying to access the data. We use the concept `geo:SpatialThing` in order to express the spatial constraints. These properties are used to express policies in which not only the identity of the user requesting the data is checked, but also the contextual information related to the time and place in which the request is performed. A further class is *MaxResource* which defines the number of times the user can access all or a specified resource. We introduce also the property *hasParameter* which provides for each variable used in the ACs, a comment in natural language explaining the meaning of the variable. This is introduced with the aim to explain to the user how the variables are used in the access policies he is adopting, e.g., "?date" has the associated comment "the date of creation of the resource".

Definition 2. *An Access Evaluation Context (AEC) is a list L of predetermined bound variables of the form* $L = (\langle var_1, val_1\rangle, \langle var_2, val_2\rangle, \ldots, \langle var_n, val_n\rangle)$ *that is turned into a SPARQL 1.1 Binding Clause to constrain the* `ASK` *query evaluation when verifying the Access Conditions.*

The *AEC* is represented in the ontology as the class *AccessEvaluationContext* which has two properties, *hasVariable* and *hasValue*, which are respectively the variable, and the value to which the variable is bound. It is used to provide a standard evaluation context to the access conditions, e.g., requesting user, resource provider. Consider the following example: `L=(<''?resource'', ''<http://MyExample.net#doc>''>,<''?user'',''<http://MyExample.net#sery>''>)`. This list can be used to generate an additional SPARQL 1.1 Binding Clause for the access conditions of the form: `BINDINGS ?resource ?user {(<http://MyExample.net#doc>, <http://MyExample.net#sery>)}`.

Definition 3. *An Access Condition Set (ACS) is a set of Access Conditions.*

The *AccessConditionSet* class has a property *hasAccessCondition* which identifies which Access Conditions form the ACS. Two subclasses of *AccessConditionSet* are introduced: conjunctive, and disjunctive ACS.

Definition 4. *A Conjunctive Access Condition Set (CACS) is a logical conjunction of Access Conditions of the form* $CACS = AC_1 \wedge AC_2 \wedge \ldots \wedge AC_n$. *A CACS is verified if and only if every access conditions it contains is verified.*

Definition 5. *A Disjunctive Access Condition Set (DACS) is a logical disjunction of Access Conditions of the form* $DACS = AC_1 \vee AC_2 \vee \ldots \vee AC_n$. *A DACS is verified if and only if at least one of the access conditions it contains is verified.*

Definition 6. *An Access Tagging Rule (ATR) is a triple* $R = \langle ACS, TagSet, Bindings \rangle$ *where ACS is an Access Condition Set, TagSet is a set of tags* $\{tag_1, tag_2, \ldots, tag_m\}$, *and Bindings is an Access Evaluation Context. An ATR is verified for a resource tagged with one or more tags from TagSet if and only if the ACS is verified for that resource. The ACS may be reduced to a single access condition. In this case, the ATR is said to be verified if and only if the single access condition is verified. The TagSet may be empty, in which case the ATR applies to any named graph.*

An ATR declares that the access conditions in the ACS applies to any RDF graph tagged with one or more tags from *TagSet*. The class *AccessTaggingRule* has three properties: *hasAccessConditionSet*, associating an ACS to the ATR, *hasTag*, providing a set of tags to the ATR, and *hasAccessEvaluationContext*, associating to the ATR the AEC, i.e., the bindings applied to the rule. Moreover, it has the property *hasAccessPrivilege* which defines the access privilege the user is granted to: *Read*, *Create*, *Update*, *Delete*. We expand the `acl:Write` class, which is used for every kind of modification on the content, and we allow fine-grained access control privileges. The class *AccessTag*, used to define the set of tags, is a sub-class of `scot:Tag`.

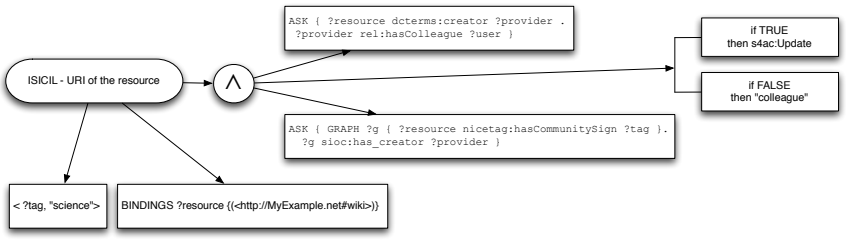


Fig. 3. An example of access policy.

We show now in detail which kind of access control policies are enabled by the proposed access control model. Consider the policy defined below: the data provider defines an access policy such that only his resources tagged with tag "family" are constrained by the access condition which grants the access to those users which have a *hasParent* relationship with the data provider, i.e.,

the parents of the provider. The Access Condition Set is composed only by one access condition, thus this is the only one which needs to be evaluated. The access privilege is `Read`. Thus, given a `SELECT` query of the user, if he is granted with the access, then he is allowed to `Read` the requested data. The use can access the data from December 31th at 23:59. If the user is not granted with the access then the label the system returns him together with the failure message is "parents", to explain that the reasons of the failure have to be associated to the fact that the user is not a parent of the provider; we choose not to send any message if some results are filtered.

```
<http://MyExample.net/expolicies>
 a s4ac:AccessTaggingRule;
  s4ac:hasAccessConditionSet [
   s4ac:hasAccessCondition [
     s4ac:hasValidity [
       time:hasBeginning [
         time:inXSDDateTime 2011-12-31T23:59:00
       ];
     ];
     s4ac:hasCategoryLabel ''parents''@en;
     s4ac:hasQueryAsk ''
       ASK { ?resource dcterms:creator ?provider .
       ?provider rel:hasParent ?user }''
   ];
  ];
  s4ac:hasAccessPrivilege s4ac:Read;
  s4ac:hasTag ''family''@en.
```

The table below presents some examples of the `ASK` queries which may be associated with the access conditions. *Cond1* grants the access to those users who have a relationship of kind "colleagues" with the provider. *Cond2* grants the access to the friends of the provider, and *Cond3* extends this access condition also to the friends of friends. *Cond4* is more complicated[10]. It grants the access to those users that are marked with a specified tag. For specifying the tag, we use the NiceTag ontology which allows to specify the relationship among the resources and the tags for each tagging action. Also negative access conditions are allowed, where we specify which specific user cannot access the data. This is expressed, as shown in *Cond5*, by means of the `FILTER` clause, and the access is granted to every user except *sery*. *Cond6* expresses an access condition where the user can access the data only if he is a minimum lucky, e.g., one chance out of two. *Cond7* provides a positive exception where only a specific user can access the data, it is the contrary of *Cond5*. *Cond8* grants the access to those users who are members of a particular group, to which the provides belongs too. Finally, *Cond9* ensures the access to all the resources tagged with *?tag*.

An example of conjunctive ACS is as follows: $CACS_{friends-but-sery} = Cond_2 \wedge Cond_5$, where the access is granted to the users who are friends of the provider, but the user `<http:MyExample.net#sery>`, even if she is a friend of the provider, cannot access the data. An example of disjunctive ACS is

[10] The GRAPH keyword is used to match patterns against named graphs.

cond1	ASK { ?resource dcterms:creator ?provider . ?provider rel:hasColleague ?user }
cond2	ASK { ?resource dcterms:creator ?provider . ?provider rel:hasFriend ?user }
cond3	ASK { ?resource dcterms:creator ?provider . ?provider rel:hasFriend{1,2} ?user }
cond4	ASK { ?resource dcterms:creator ?provider . ?provider dcterms:creator ?g . GRAPH ?g { ?user nicetag:hasCommunitySign ?tag }}
cond5	ASK { FILTER(! (?user= <http://MyExample.net#sery>))}
cond6	ASK { FILTER(random()>0.5) }
cond7	ASK { FILTER(?user= <http://MyExample.net#sery>)}
cond8	ASK { ?resource dcterms:creator ?provider . ?provider sioc:member_of ?g . ?user sioc:member_of ?g }
cond9	ASK { GRAPH ?g { ?resource nicetag:hasCommunitySign ?tag } ?g sioc:has_creator ?provider }

$DACS_{colleagues-or-friends} = Cond_1 \vee Cond_2$, where it is ensured that the users who are colleagues or friends of the provider are allowed to access the data.

The ATR detailed above can be constrained to a wider set of tags such as $ATR_{parents} = \langle Cond, \{''parent'','' parents'','' family'','' relatives''\}, \emptyset\rangle$ where no AEC is provided. Further examples of ATRs are: (i) $ATR_{friends} = \langle Cond_2, \{''friends'','' amici'','' ami''\}, \emptyset\rangle$ where the access condition constrains the access to friends, and three tags are provided without an AEC; (ii) $ATR_{group} = \langle Cond_7, \{''common'','' group'','' close''\}, \emptyset\rangle$ is the same for the belonging to the group of the provider; (iii) $ATR_{hiking} = \langle Cond_4, \emptyset, \langle ''?tag'','' hiking''\rangle\rangle$ where the user can access the data if he is tagged with tag "hiking" in the graph created by the provider; (iv) $ATR_{fun} = \langle DACS_{colleagues-or-friends}, \{''fun'','' funny'','' :-)''\}, \emptyset\rangle$ where the user can access the data if the disjunctive ACS above is satisfied on the named graphs tagged with these three tags.

3 Access control for the Social Semantic Web

3.1 The ISICIL use case

The challenge of the ISICIL[11] project is to reconcile new web applications with formal representations and processes to integrate them into corporate practices for technological, and scientific monitoring. More specifically, ISICIL proposes to study and to experiment with the usage of new tools for assisting corporate intelligence tasks. These tools rely on Web 2.0 advanced interfaces, e.g., blog, wiki, social bookmarking, for interactions, and on semantic web technologies for interoperability and information processing.

[11] http://isicil.inria.fr/v2/index.php

In this context, the users can create webmarks, resources, e.g., wiki pages, and personal information, e.g., social relationships represented through an activity stream. All these data cannot be fully accessible by any other user on the Web. The idea is that the users should be allowed to define their own policies in order to grant the access to their data only to those users who have the features they require. In particular, the access control model has to consider the social dimension in which it is inserted. This leads to the need of defining a model where the users can easily define their access policies, e.g., by using tags, and the relation among them. The access control model has to rely on a vocabulary like S4AC able to define the fine-grained properties the user must satisfy to access the data. For instance, the WAC vocabulary[12] allows the user to specify access control lists (ACL). The ACL are of the form `[acl:accessTo <card.rdf>; acl:mode acl:Read, acl:Write; acl:agentClass <groups/fam#group>]`, which means that anyone in the group `<http://example.net/groups/fam#group>` may read and write `card.rdf`, but a drawback of this vocabulary is that it grants the access to a whole RDF document, e.g., `card.rdf`.

3.2 The Access Control Manager

The Access Control Manager (ACM), visualized in Figure 4, is the core module which allows the user to define, and check the access conditions.

First, the ACM provides a mean to the user (userA in Figure 4) to define its own access policies. The user accesses the ACM through the user interface which considers two kinds of users: the expert users, and the non expert ones. The expert users we consider are those users who are able to define their own access conditions directly writing the SPARQL 1.1 ASK queries. Non expert users, instead, are those users who need to be guided through the interface during the policies definition, as shown in Figure 5, and they can reuse and edit the policies defined by other users clarified by using the *hasParameter* property to explain the variables. The definition of the access policies includes in particular the definition of the Access Tagging Rules, such as (i) the set of access conditions, and the way they have to be evaluated, i.e., conjunctively or disjunctively, (ii) the set of tags the resources have to be associated with in order to apply these access conditions, and (iii) the binding to constrain the variables of the access conditions. After the definition of the policies, the user is allowed to see through the interface a preview of the result of the restrictions resulting from the application of the policies. In this way, the user can verify whether the result is the expected one, or not, and he can decide eventually to reformulate the policies.

Second, consider another user, (userB in Figure 4), who wants to access the data of userA. The request to access the data is first filtered by the ACM which will allow userB to access only the data he is granted access to. The ACM receives the query of the userB. Once the request of the userB is received, the ACM selects, by means of the module called Access Control Policies Selector

[12] http://www.w3.org/wiki/WebAccessControl

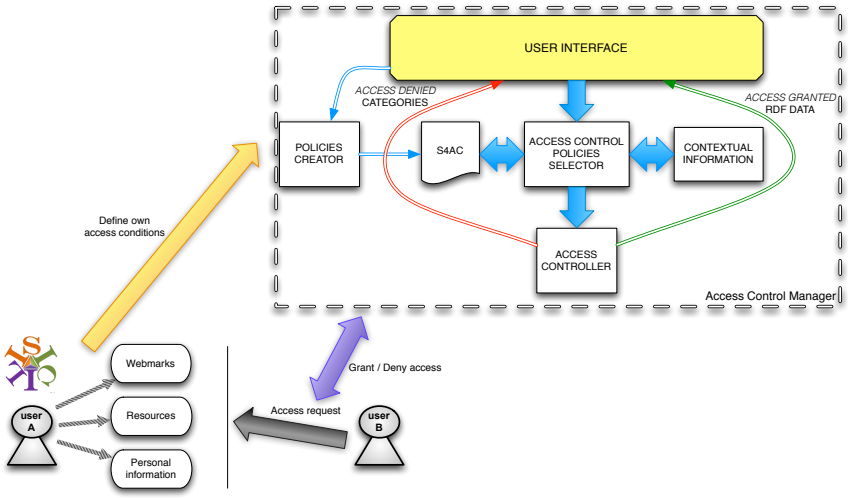


Fig. 4. The Access Control Manager.

(ACPS), which policy applies, depending on the requested operation. For instance, if the user uses a `SELECT` query, then the ACPS identifies all the policies which apply, and concern a `Read` access privilege. The ACPS performs two kinds of operations: (i) it checks the S4AC module which contains all the access conditions provided by userA to protect his data, in order to identify which access conditions apply, and (ii) it checks whether the contextual information, e.g., the temporal or spatial validity of the selected policies is satisfied. Note that we check whether the contextual constraints hold before checking the reminder of the policy. If the contextual constraints are not satisfied, then we already know that the access will not be granted. After the identification of the policies, and a positive checking of the contextual constraints, the Access Controller module matches the policies according to the userB's profile to identify what he can access. The Access Controller addresses a SPARQL `ASK` query which returns *true* if the access to the data is granted to userB. Note that userB will receive only the data he can access, and he does not know that there may be other data to which his query was addressed and that he cannot access. If the answer is *false*, then the Access Controller returns a failure, coupled with the categories causing the failure. The categories are natural language labels that are used to explain to the the user the reasons behind the failure of his query. These categories are provided to the Access Controller by the ACPS when it checks the ontology. An example of access policy composed by two access conditions that have to be conjunctively evaluated, a Bindings clause, and a Tag Set is visualized in Figure 3. The two ACs constrain the access to all the users who are colleagues of the data provider, and to all the resources tagged with ?*tag*, respectively. Moreover, the ACs are applied only to those named graphs tagged with "science",

and to the resource identified by the URI `<http://MyExample.net#wiki>`. If the conjunction is positively evaluated, then the access is granted with the privilege `s4ac:Update`. Otherwise, the access is denied, and the label "colleagues" is returned.

The developed prototype provides a user interface implemented in HTML 5, as visualized in Figure 5. It relies on the SPARQL query engine KGRAM/CORESE[13]. Briefly, the system uses the Binding SPARQL 1.1 to substitute the variable *?resource* with the URI of the resource to be accessed. The query is executed to obtain all the *ATR*s associated with the resource, and the data provider. CORESE returns these *ATR*s which contain the ACS. The `ASK` queries inside the single AC are executed on CORESE, and the returned booleans are conjunctively or disjunctively evaluated to grant or deny the access.

Fig. 5. The non-expert user interface for creating the access policies.

4 Related work

Sacco and Passant [8] present a Privacy Preference Ontology (PPO), built on top of WAC, in order to express fine-grained access control policies to an RDF file. They also specify the access queries with a SPARQL `ASK`, but their vocabulary does not consider the temporal and spatial validity of the privacy preferences, and the maximum number of accesses allowed. They rely entirely on the WAC vocabulary without distinguishing the `Write` actions. Their model does not allow to specify set of tags to limit the application of the policies to the resources

[13] http://www-sop.inria.fr/edelweiss/software/corese/

marked with those tags, and to specify conjunctive and disjunctive sets of privacy preferences.

Giunchiglia et al. [6] propose a Relation Based Access Control model (*RelBAC*), providing a formal model of permissions based on description logics. They require to specify who can access the data, while in our model and in [8] the provider can specify the attributes the user must satisfy.

The Access Management Ontology (AMO) [3] defines a role-based access control model. The AMO ontology consists of a set of classes and properties dedicated to the annotation of the resources, and a base of inference rules modeling the access strategy to carry out. This model again needs to specify who can access the data.

Abel et al. [1] present a model of context-dependent access control at triple level, where also contextual predicates are allowed, e.g., related to time, location, credentials. The policies are not expressed using Web languages, but they introduce an high level syntax then mapped to existing policy languages.

Hollenbach and Presbrey [7] present a system where the users can define access control on RDF documents, and these access controls are expressed using the WAC. Our model extends WAC for allowing the construction of more fine-grained access control policies.

Carminati et al. [4] propose a fine-grained on-line social network access control model based on semantic web technologies. The access control policies are encoded as SWRL[14] rules. This approach is also based on the specification of who can access the resources, i.e., the access request is a triple (u, p, URI), where the user u requests to execute privilege p on the resource located at URI.

Stroka et al. [9] present a preliminary proposal about securing the collaborative content on the platform KiWi. They consider global permissions, individual content item permissions, and RDF type based permission management. They do not specify the kind of access polices they can define.

Finin at al. [5] study how to represent RBAC using the OWL language. The authors show also the representation of policies based on general attributes of an action, similarly to what we present in this paper. The difference is that we specify the policies using SPARQL 1.1 ASK queries, where the Bindings clause is used to specify the values of the variables, and temporal and spatial constraints may be expressed too.

5 Conclusions

In this paper, we have introduced a fine-grained access control model for the social semantic web. This model is grounded on the S4AC ontology which allows the users of the social networks to define the access conditions for their data. In particular, these access conditions have the form of SPARQL 1.1 ASK queries, and they can be either conjunctively or disjunctively evaluated. Moreover, the access policies can be constrained w.r.t. the set of tags the resources are tagged

[14] http://www.w3.org/Submission/SWRL/

with, and an access evaluation context providing the bindings can be specified too. We have presented our Access Control Manager, in the context of the ISICIL platform. The manager has the aim to grant or deny the access to the users. Through a user interface which allows also non-expert users to interact with the system, the users can specify the access policies to protect their data. The manager looks for the policies which apply to the resource, and after checking the contextual constraints and the features of the user trying to access, it states whether the access is granted or not.

There are several lines to follow for future work. First of all, in this paper we assume that the user's information are trustworthy. Since this assumption is not always verified, we will investigate the adoption of methodologies able to assess the trustworthiness of the users. Second, a prototype of the Manager has been developed in the ISICIL platform. We aim at providing a more efficient implementation of the Manager, in order to fully integrate it into the platform.

References

1. Fabian Abel, Juri Luca De Coi, Nicola Henze, Arne Wolf Koesling, Daniel Krause, and Daniel Olmedilla. Enabling advanced and context-dependent access control in rdf stores. In *Proceedings of the 6th International Semantic Web Conference (ISWC-2007), LNCS 4825*, pages 1–14, 2007.
2. John Breslin, Alexandre Passant, and Stefan Decker. *The Social Semantic Web*. Springer-Verlag, Heidelberg, 2009.
3. Michel Buffa, Catherine Faron-Zucker, and Anna Kolomoyskaya. Gestion sémantique des droits d'accès au contenu : l'ontologie AMO. In Sadok Ben Yahia and Jean-Marc Petit, editors, *EGC*, volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, pages 471–482. Cépaduès-Éditions, 2010.
4. Barbara Carminati, Elena Ferrari, Raymond Heatherly, Murat Kantarcioglu, and Bhavani M. Thuraisingham. Semantic web-based social network access control. *Computers & Security*, 30(2-3):108–115, 2011.
5. Timothy W. Finin, Anupam Joshi, Lalana Kagal, Jianwei Niu, Ravi S. Sandhu, William H. Winsborough, and Bhavani M. Thuraisingham. ROWLBAC: representing role based access control in OWL. In Indrakshi Ray and Ninghui Li, editors, *SACMAT*, pages 73–82. ACM, 2008.
6. Fausto Giunchiglia, Rui Zhang, and Bruno Crispo. Ontology driven community access control. In *Proceedings of the 1st Workshop on Trust and Privacy on the Social and Semantic Web (SPOT-2009)*, 2009.
7. James Hollenbach, Joe Presbrey, and Tim Berners-Lee. Using RDF Metadata To Enable Access Control on the Social Semantic Web. In *Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK-2009)*, 2009.
8. Owen Sacco and Alexandre Passant. A Privacy Preference Ontology (PPO) for Linked Data. In *Proceedings of the 4th Workshop about Linked Data on the Web (LDOW-2011)*, 2011.
9. Stephanie Stroka, Sebastian Schaffert, and Tobias Burger. Access Control in the Social Semantic Web - Extending the idea of FOAF+SSL in KiWi. In *Proceedings of the 2nd Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2010)*, 2010.

LexiTags: An Interlingua for the Social Semantic Web

Csaba Veres
University in Bergen, Fosswinckelsgt. 6, 5020 Bergen,
Norway. Csaba.Veres@infomedia.uib.no

Abstract. The paper describes *lexitags*, a new approach to social semantic tagging whose goal is to allow users to easily enrich resources with semantic metadata from WordNet. This is a paradigm example of the Social Web and the Semantic Web working together: ordinary users help create the metadata so needed by the Semantic Web and in turn, Semantic Web technologies help those users get a richer experience from the Social Web. A family of simple user interfaces for lexitagging is described, as are some methods for the subsequent, automatic generation of lightweight ontologies. These ontologies are presented as an ideal *interlingua* for the Social Semantic Web.

Keywords: Social Web, Semantic Web, ontology, metadata, WordNet, linked data, rdf, folksonomy

1. Introduction

Two of the most exciting innovations for transforming the World Wide Web are "Web2.0" [1] and the "Semantic Web". Each has a separate vision for moving a relatively static Internet driven by focused content providers, to a dynamic and largely self managing entity enabled by large volumes of metadata. But while the general vision is shared, the details of the two approaches appear to be opposites. While Web2.0 is focused on free-form, user generated ad hoc metadata and opportunistic social organization, the Semantic Web is a vision containing strict and enforced data structures suitable for automated machine processing. Web2.0 has proven advantages in the ease of data creation and a correspondingly lower threshold for user adoption, but the lack of predefined structure may inhibit effective retrieval as the amount of unstructured metadata grows in volume. An obvious idea is to combine the two sets of technologies so that the users can have systems which behave as Web2.0 at the point of insertion, yet as Semantic Web at the point of retrieval. Following papers such as [2], it is now widely agreed in the community that the Semantic Web and the Social Web can benefit from each other.

In particular, the Information Architecture community has embraced *folksonomy*[1] as a way to enhance information management practices. An analogy is often made with the term *desire lines,* which comes from landscape architecture. The basic idea originates in the observation that, in spite of the careful planning undertaken by architects to lay out walking tracks in their meticulously designed spaces, one will often find emergent paths that have been forged by people who deviate off the planned tracks onto the grass or gravel of the spaces. The paths become entrenched when particular tracks are found useful by many people. It is similar in information spaces, where folskonomy describes the desire lines, representing informal tag based classification schemes that people find useful. The addition of formalized "desire lines" on the web would benefit emerging semantic platforms that rely on such metadata, especially with respect to querying and mining social semantic data.

This paper describes a set of tools and principles currently under development, which will help formalize folksonomy for the web. In the next section we describe some of the main problems with current tagging practice. Then we describe our approach to cleaning up tags and generating formal rdf based *semantic tags*. Following this, a method for automatically generating lightweight ontologies from semantic tags is described, and their use as a universal *interlingua* is

1 http://vanderwal.net/folksonomy.html

explained.

2. Folksonomy term problems

The basic problems with folksonomy terms from user tags are nicely summarized in [3]. The problems Mathes identifies are as follows.

- **Ambiguity** Since tags are mainly natural language terms, they are characterized by the inherent ambiguity of those terms. A special case of ambiguity can be seen in the proper identification of acronyms. As noted by Mathes: "Examining the front page on November 14, 2004 revealed one user tagging sites with ANT. After examining the other sites the user tagged with ANT, it was apparent this was an acronym for Actor Network Theory, in the domain of sociology. However, when examining the ANT tag across all users (Delicious apparently is not case sensitive in tags) most of the bookmarks were about Apache Ant, a project building tool in the Java programming language. Two completely separate domains and ideas are mixed together in the same tag."
- **Spaces**, multiple words Many services do not allow users to enter multiple word tags separated by spaces, so users improvise as in the example: "vertigovideostillsbbc". Perhaps more creatively, users concatenate words to express alternative names (design/css), or even hierarchical groupings (Devel/C++).
- **Synonyms**. Since there is no control on the vocabulary, one often finds multiple words or variants expressing the same concept, as in mac, macintosh, and apple (apple of course has the added problem of being ambiguous). Another manifestation of this problem is the indiscriminate use of the plural and singular of a term. The NISO guidelines for controlled vocabularies recommends the singular use.

[4] also summarize problems in tag use, and based on these observations they provide guidelines for creating "tidier tags". They conclude that the main problem with tags is *imprecision*. They flesh out this remark to include the already mentioned problems with ambiguity, synonymy and number, but add a few additional observations: "… the tags are often ambiguous, overly personalised and inexact. .. Plural and singular forms, conjugated words and compound words may be used, as well as specialized tags and 'nonsense' tags designed as unique markers that are shared between a group of friends or co-workers. The result is an uncontrolled and chaotic set of tagging terms that do not support searching as effectively as more controlled vocabularies do."

[4] performed some quantitative analyses on a set of randomly selected tags from delicious as well as the photo sharing site, Flickr. They made the following observations about the prevalence of various errors:

- **Misspellings, incorrect encodings, and compound words**: "By testing against multilingual dictionary software, we found that 40% of flickr tags and 28% of del.icio.us tags were either misspelt, from a language not available via the software used, encoded in a manner that was not understood by the dictionary software, or compound words consisting of more than two words or a mixture of languages."
- **Words that did not follow system conventions**: Almost 8% of the flickr tags and over 11% of the del.icio.us tags were plural forms of words.
- **Symbols used in tags**: "Symbols such as "# " were used at the beginning of tags, probably for an incidental effect such as forcing the del.icio.us interface to list the tags at the top of an alphabetical listing."

They also note after the quantitative evaluation that "However, we did find that single-use tags were less common than we had expected", suggesting a high degree of consensus in tagging behavior, and a correspondingly low degree of "personalised" tags. They additionally note that the high number of tags that were not words that can be found in a standard dictionary may be artificially high. In many cases the tags were misspelled or creative variants of dictionary words. Many examples of misspelling consisted of the transcription of characters across languages. For example, the Norwegian æ can be written as "ae". Sometimes the reason was the compounding of

words and letters as in “17thjuly”. Another prominent practice was the inclusion of geotagging information (latitude and longitude) in the tag. This was particularly popular in Flickr (perhaps unsurprisingly).

Based on the previous observations, it is fair to say that the predominance of tags are dictionary words, or compounds formed from dictionary words (or numbers). In support of this conclusion, [5] report that 82% of the top 100 tags on delicious.com appear in WordNet, and that this drops to a still respectable 79% for the top 1000, and 61% for the top 10000 tags. Apparently, all the mysticism surrounding tags notwithstanding, in the vast majority of cases tags are simply dictionary words or word compounds. It is in this vein that [4] recommend a number of simple guidelines to improve tagging practice. They propose a number of practices like standardized spelling and hyphenation practices, and a handful of useful heuristics for tag selection.

The problem with recommendations is that they can be difficult to enforce, or even convince people that they should try to follow them. For example, one could stipulate that tags should follow NISO recommendations [6] that count nouns appear in the plural form (e.g. dogs, toys) and mass nouns in the singular (e.g. water, furniture). But it is difficult to imagine that people will accept that they should always use the tags *movies, toys, knives,* rather than *movie, toy, knife* for example.

The solution which is suggested in this paper is at the outset a simple way to gently enforce these best practices through the tagging interface, by allowing people to simply tag with dictionary words, otherwise known as *lexical items*. It is for this reason that the tags themselves are called *lexitags*. Lexitags guarantee that every tag can be unambiguously connected with a known lexical item, while still allowing some flexibility in user behavior. For example, both *cat* and *cats* are allowed in the user interface, but both are linked to the lexical item {cat}. By keeping information about the surface form and lexical item separate, no information about user behavior is lost: the underlying semantics of the tag is captured, as well as the potentially significant choice of plural or singular. On the other hand, only acceptable spellings can be used, so the problem of misspellings, idiosyncratic spelling variations and so on, disappears.

3. Creating RDF-based knowledge using social media services

The primary lexical database in this project is WordNet. This is supplemented by DBPedia which provides terms missing in WordNet, such as names for emerging technologies and people. WordNet is perhaps the most well established electronic lexical database, whose development at Princeton University dates back to 1985. WordNet represents disambiguated word senses with synonym sets (*synsets*), which are equivalent terms enclosed in braces. For example some of the unique senses of the word *cat* are: {cat, true cat}, {guy, cat, hombre, bozo}, {cat, gossip}, {kat, khat, qat, quat, cat, Arabian tea, African tea}, {cat-o'-nine-tails, cat}, and so on. WordNet is a very large database, containing in total 206941 word-sense pairs including nouns, verbs, adjectives and adverbs. In addition, each synset contains lexical pointers to related synsets, where the relations are specific to grammatical category. For example nouns are included in (amongst other things) *hyponymy* and *meronymy* relations, but adjectives in *antonymy*. In summary, WordNet is an extensive database of English words, together with a rich set of lexical and semantic relations defined over the lexical items.

The simple idea, then, is to use WordNet as a source for disambiguating tags that are applied to resources by users, and to provide a simple interface where this can be achieved. The disambiguated tags are referred to as *lexitags* to honor their origins in the lexicon, or *semantic tags* to indicate that their interpretation is fixed relative to a semantic resource. In order to realize the interface design in a reference implementation, we designed a platform for social bookmarking, which we will refer to as LexiTags, and which was developed through a commercial startup company called LexiTags D.A. The company was jointly established by the author and his colleague Andreas Opdahl at the University in Bergen, and supported by a seed grant from Innovation Norway. It should be noted that there are at least two existing commercial ventures that are advertised as a “semantic bookmarking service”, making them similar to

LexiTags in this regard. These are Faviki and Zigtag. Zigtag, which has been running since early 2009[2] is the most similar in that it uses its own dictionary as tag definitions. Faviki uses WikiPedia concepts instead. However, there are fundamental differences in the motivation for these services and LexiTags. Zigtag and Faviki are bookmarking services, pure and simple. Their interest in "semantic tags" is to enhance findability on their site by providing equivalences between differently spelled tags (NYC, New York City, Big Apple, …), returning results for only one sense of an ambiguous tag, and so on. On the other hand LexiTags is simply a reference implementation of the lexitagging interface, with the focus being on the generation and exploitation of the metadata itself, rather than the underlying purpose of the reference implementation (which in this case happens to be bookmarking). The principles and algorithms are meant to be portable to any application, using the semantics in the generated metadata to bridge the divide in content across the services and applications. Metadata in the form of lexitags becomes an interlingua between applications.

To demonstrate the idea of a simple portable tagging interface, we will discuss here an iPhone interface which is currently in development. The iPhone interface communicates with the LexiTags service over http and can upload html bookmarks, but also photographs taken with the iPhone camera. The application can therefore serve as a tagging interface for bookmarking as well as a photo upload service.

The design principle is that lexitagging must be no more difficult than ordinary tagging otherwise people will not be inclined to use it. One key research problem is how to rank the possible senses so that the sense intended by the user is immediately available in the interface. The current iPhone tagging interface is shown in figure 1. On the left are two ambiguous tags, and on the right *cat* has been disambiguated (which is evident from the text below the tag). Notice that both URL entry field and photo choser are both available, and the user must chose which they use. URLs have to be manually entered at the moment, but ideally this will be linked to the web browser. In figure 1 we see that the user has chosen (or snapped) an image of a cat, and has assigned two tags "cat" and "cute". The tags are simply typed in the "Tags" field, and automatically marked as "undefined". This allows people to initially add tags freely. Once they have typed a few tags, users must tap each one to define it, which brings up the selection interface in figure 2.

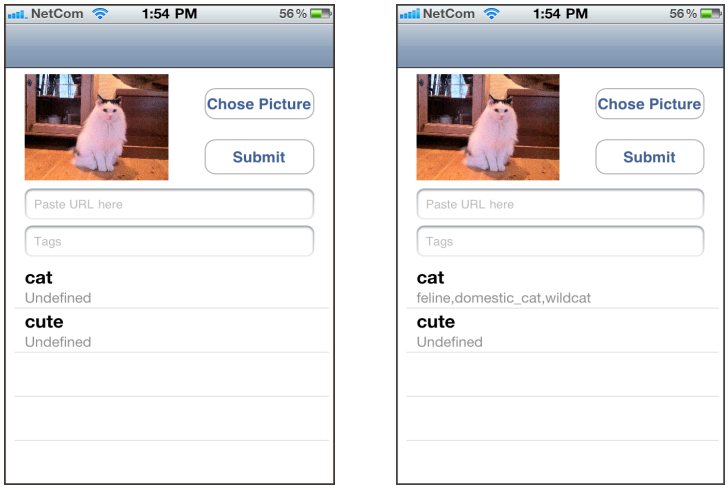


Fig. 1. Picture with two undefined tags and one defined tag on the right

On the left of figure 2 is the initial display, showing 5 possible choices. There are actually 10 senses of the word "cat", so one must scroll to see the others. It is here where the ranking becomes important, since ideally the desired sense will always appear in the top 5 choices. The iPhone choser currently uses a series of words related to each sense as a

2 http://www.readwriteweb.com/archives/semantic_tagging_service_zigtag_finally_launches.php

disambiguator, rather than the actual synset (in this example {cat, true cat}). This is because not all senses have near equivalent synonyms, so the synset in these cases is simply the word itself. For example the synset for the third sense of *cat* is simply {cat}, which would not be a useful disambiguator. We are still experimenting with the usefulness of this method for selecting the correct sense. We are also experimenting with an alternative display in the larger, browser based desktop client. The display in figure 3 shows for each sense the synsets from WordNet, as well as the full explanatory gloss. Usually one or other at least is available. In addition, each sense is preceded by a determiner in brackets. This is meant to help people with selecting the grammatical category: "(a, an, the)" are nouns, "(to)" are verbs and "(is)" are adjectives. Our feeling is that the simple iPhone interface is adequate, but we will need to experiment extensively before making any conclusive claims regarding usability, since the selection of the appropriate word sense is the most difficult and important role for the lexitags interface.

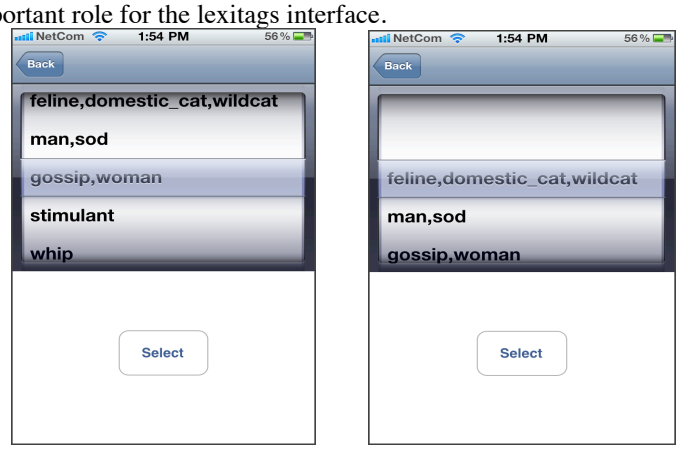


Fig. 2. Two stages of sense selection

Once the process is complete, the entry can be submitted. Any tag that has not been disambiguated by the user is simply discarded at this stage.

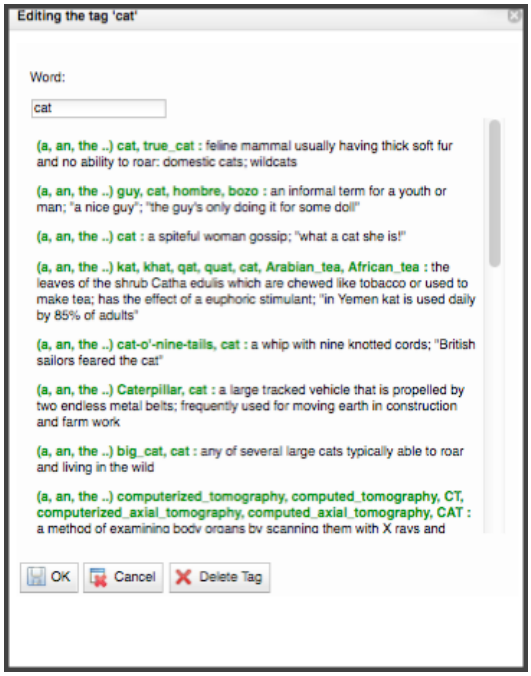


Fig. 3. An alternative tagging interface

A key problem is in ranking the possible senses such that the desired alternative is very near the top of the list for every tagging episode. Clearly this involves an estimate of the likelihood that a particular sense matches the content of the to-be-tagged item. If one can obtain

disambiguated key terms from the resource itself, then there are a number of useful algorithms for computing the similarity between those terms and each candidate sense of the tag [7]. When the resource is an html page, there are a number of obvious possibilities to obtain such contextual information. The simplest is to extract the title, or any metadata that is available, and use any of a number of open solutions which are available from the text processing community to disambiguate these terms. Of course the more sparse the retrieved text, the more difficult the disambiguation. Another option is to scrape the entire text of the html page and extract key summary terms. However, while this method could give the most accurate candidate ranking, it can become computationally expensive and may not return results sufficiently quickly for use in the tagging interface.

Currently we are using a much simpler approach, which is to use the tags themselves for disambiguating other tags. That is, once the user has sense selected the appropriate lexitag, then that can be used to rank any successive tags. The more tags that have been selected, the more accurate the algorithm can become. The biggest problem with this simplistic approach is that there is no disambiguating evidence for the first tag. However, in these cases we simply use the relative frequency of use, arguing that people are less likely to use infrequent senses of words as tags. There is no reason why these various techniques could not be used in complementary ways, combining ranking estimates based on the different sources. For example the initial disambiguating context could be a fast analysis of the title and some metadata, which would be replaced as the analysis of the text becomes available. In turn, this could be combined with the disambiguated lexitags as the user works his way through the tagging session. It is of course an empirical question to see which combination of these methods results in the best user experience.

The results of a tagging session are recorded in RDF, using a number of common standards including Dublin Core[3], FOAF[4] and Common Tag.[5] Figure 4 shows the format we have adopted from the Common Tags specification. The representation is straightforward, so we only point out the two relations ctag:label and ctag:means. The former is the word string used by the tagger, and the latter is a "dereferencable Resource that identifies the concept expressed by the Tag".[6] Of course in LexiTags this is a WordNet synset. This allows some separation between the word string and its meaning, accommodating the case where *NYC* and *NewYork* can both be used as tags, yet refer to the same concept. Because the application is based on open standards, all web sites which expose their data in the Common Tags format will automatically inter operate. Lexitags give extra value in that they can add semantically rich, disambiguated metadata to a URL that may be recorded on another site without rich metadata.

```
@prefix  ctag: <http://commontag.org/ns#> .
@prefix  wn:   <http://www.w3.org/2006/03/wn/wn20/instances/> .
@prefix  rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix  foaf: <http://xmlns.com/foaf/0.1/> .
@prefix  xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix  dc:   <http://purl.org/dc/terms/> .
[] rdfs:type ctag:TaggedContent ;
  ctag:isAbout "http://commontag.org/QuickStartGuide"^^xsd:anyURI ;
  ctag:tagged [
  ctag:means wn:dog-n-1;
  ctag:label "dog"@en;
  foaf:maker u1234;
  dc:created "20.01.2200"^^xsd:Date;
  ctag:taggingDate "22.22.2200"^^xsd:Date  ] .
```

Fig. 4. RDF representation of a tagging session

---

3 http://dublincore.org/
4 http://www.foaf-project.org/
5 http://commontag.org/Home
6 http://commontag.org/Specification#means

4. Ontologies for the Social Web

We have already mentioned the most straightforward advantages of using semantic tags for finding content on a bookmarking site. But the use of WordNet as the reference semantics provides far greater benefits. Lexitagging provides us with collections that are marked up with semantically disambiguated lexical items, which have rich associations to other lexical items in WordNet. We have taken advantage of this in developing a method for creating lightweight ontologies for social media sites. [8] reports an algorithm to extract general terms from a set of resources annotated with WordNet synsets. Basically, the algorithm infers maximally informative hypernyms (SuperTags) for user generated tags with the simple algorithm shown in figure 5. Nodes are only retained with this algorithm if they have two or more children, and are more than six nodes from the root nodes. These parameters are variable.

```
algorithm enrich Bookmark collection {
    forall Bookmarks in the collection {
        find all hypernyms and store them in chains
    }

    forall hypernym chains {
        find every hypernym that {
            either only has one unique child in the
                set of chains it appears in
            or appears as the sixth or higher element
                in any chain
        }
        mark these hypernyms as irrelevant
    }

    for all unmarked hypernyms {
        convert the hypernyms to a SuperTag of the
            Bookmark to which the tag chain belongs
    }
}
```

Fig. 5. Algorithm for maximally informative SuperTags.

We have implemented this algorithm in a web service which generates visualizations for a set of lexitagged resources. Figure 6 shows a typical visualization for a small set of tags.

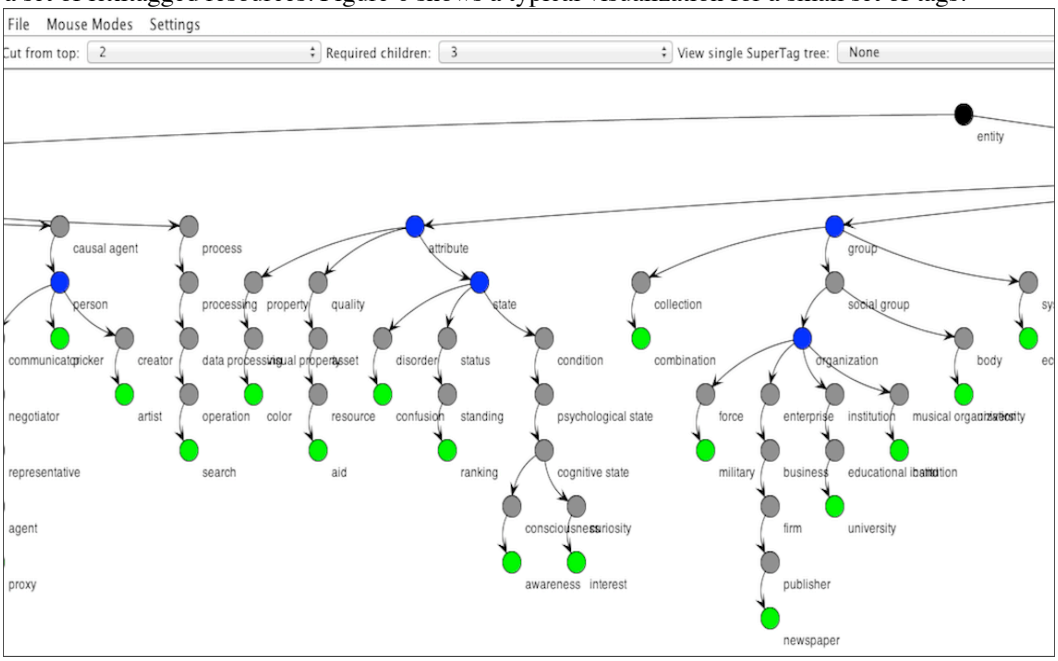


Fig. 6. Inferred SuperTags.

The figure shows the user assigned lexitags in green (light grey leaf nodes), and their

respective hypernym chains. The hypernyms eventually intersect at common nodes. The nodes colored blue (dark grey) are retained as SuperTags because they have three or more children. Light grey nodes (except the leaf nodes) are discarded because they have less than three, and black ones because they are too close to the root nodes. All of these parameters are adjustable in the interface, so it is possible to adjust the generality and inclusiveness of the nodes which are finally retained. Already in this small example we see some useful nodes emerge: *group, organization, attribute, state, person*. The emerging nodes can tell us about the nature of this particular collection. For example, in the sub tree originating from *group* we see *organization* but not *social group* emerge as an important node. The reason is that most resources in this part of the collection deal with either *university, newspaper,* or *military*. It is easy to imagine other collections where the predominant node would be *social node*.

The real power of using lexitags as a basis for lightweight ontologies becomes apparent when relations from WordNet are added to the inferred SuperNodes. For example, figure 7. shows the meronyms of *organization*. These can be added to the emerging ontology as properties. Once embellished with properties, the ontology becomes a rich representation of the key concepts in a social site, and can be used for various inference tasks. For example, if someone uses the tag *Apple inc.,* then this will be an example of an *organization*. Since the ontology tells us that organizations have a *quorum*, this could prompt an application to automatically fill in the names of the Apple board of directors and even suggest them as contacts in the social site. The result is a rich, dynamically emerging ontology which reflects the users attitudes to the underlying domain, and which can change if the concepts or the tagging behavior of users change.

organization, organisation -- (a group of people who work together)
HAS MEMBER: quorum -- (a gathering of the minimal number of members of an organization to conduct business)
HAS MEMBER: membership, rank -- (the body of members of an organization or group; "they polled their membership"; "they found dissension in their own ranks"; "he joined the ranks of the unemployed")

Fig. 7. Meronyms of *organization*

The results presented in this paper are preliminary, but the way forward is clear. The implementation of portable tagging interfaces will result in a growing number of resources tagged with lexitags. The resources could include traditional http bookmarks, geo tagged photographs, wiki and blog entries, and even local file systems. The automatically generated lightweight ontologies will add unique metadata to each site. However, because each site is marked up with the same lexitags, this will facilitate comparisons and sharing between the sites. In fact, we make the bold claim that lexitags (WordNet synsets) are an ideal *interlingua* for the social semantic web because it has the expressive power to align concepts between any arbitrary ontologies, yet is intuitive in the most basic sense of the word.

Notice that we are not advocating WordNet as a universal ontology. In fact, we are sympathetic to [9], who details a number of reasons for why the lexicon ought not be construed as an ontology at all. Ontologies attempt to model domains of interest with strict, mutually exclusive classes, while lexicons often use overlapping words to cover the semantics of the world. For example, consider the English words *error* and *mistake* and some of their hyponyms, which by definition denote kinds of mistakes or errors: *blunder* (an embarrassing mistake), *slip* (a minor inadvertent mistake), *lapse* (a mistake resulting from inattention), *faux pas* (a socially awkward or tactless act). But notice that a *slip* can also be a *blunder* and that a *faux pas*, which is itself a kind of of *blunder*, could also be just a *slip*. What licenses the use of the different words in natural language conversation is that they emphasize different dimensions of the concept being communicated: a *slip* is distinct from *mistake* because it does not (presumably) result from an *error* in judgment (i.e. it is inadvertent), whereas a *blunder* is distinguished by the fact that it causes embarrassment. But there is no reason that a *blunder* could not be inadvertent, and therefore also a *slip*. Words at a given level in the hyponym tree sometimes shift attention from

one distinguishing feature to another, rather than being non overlapping sub types of their hypernym.

WordNet may not be a universal ontology, but is powerful as an interlingua precisely for the same reasons that make language so powerful at communicating concepts. Flexibility allows one to finesse levels of detail but still communicate, and also allows one to reach arbitrary levels of precision when needed. When using lexitags as an interlingua, designers of individual ontologies can map their terms to specific interpretations in WordNet as the requirements demand. They can chose the mappings that reflect their particular world view: for example domains that require attribution of blame can map their terms to *slip* or *mistake* while everyone else can map to *error*. If it is important that people who use *cinema* are kept away from people who use *movie* [10] then this is possible, but they can still become acquainted when the distinction no longer matters.

Another interesting possibility is that the lexitags interface may help solve another problem with using WordNet as an ontology: *lexical gaps*. [9] points out the problem where an easily demonstrable *covert category* exists, but there is no word for it. For example, *things that can be worn on the body*. Since the lexitagging interface allows multi word tags, someone could use a general tag *body wear* with the two words appropriately disambiguated. This would then establish a new link between *body* and *wear*, as the lexical representation of the covert category.

Lexitags can also serve as an interlingua between formal ontologies and the social web. For example, SUMO [11] has an extensive set of links to WordNet which can be explored with the SIGMA knowledge engineering environment.[7] The links include equivalent as well as subsuming mappings. Any ontology that is mapped to SUMO is therefore automatically aligned with lexitags ontologies. Perhaps equally importantly, the EuroWordNet project oversees the creation of wordnets for many European languages,[8] and there are attempts at Chinese wordnets.[9] These projects constitute a major step towards making lexitags a universal interlingua for formal and semi formal metadata.

5. Related work

There is a large body of work whose aim is to exploit folksonomies for more effective information management. Most of the existing literature concerns the exploitation of statistical regularities in the way tags are assigned to resources by users. [12] suggests that the efforts can broadly be classified as (a) extracting semantics of folksonomies, including measuring relatedness, clustering, and inferring subsumption relations or (b) semantically enriching folksonomies, including collaborative structuring, and linking tags with professional vocabularies and ontologies.

One of the earliest demonstrations in the first vein was *clustering* on Flickr, where polysemous tags are displayed with co-occurring tags in different sets of images. For example, the tag *apple* has the following clusters: <mac, macbook, macintosh, computer, laptop, imac, keyboard, powerbook, osx, macbookpro>, <fruit, red, green, food, tree, macro, canon, orange, blossom, apples>, <ipod, iphone, music, nano, touch, shuffle, mp3, black, phone, ipodtouch>, and <nyc, newyork, manhattan, newyorkcity, ny>. The algorithm can identify photographs tagged with the different uses of *apple*: apple the fruit, apple the company, and the "Big Apple". However, this form of clustering is not simply lexical disambiguation since the company sense of *apple* is listed in two different clusters which reflect different distinguishing product lines for the company. An additional benefit is that different spelling variations of a tag are bundled into the same cluster as in *nyc, newyork, ny,* because these tags tend to co-occur with the same pictures. While the details of the Flickr algorithm are proprietary, various clustering algorithms were explored by [13].

In another interesting use of co-occurrence, [14] report a study in which their algorithm

---

7 http://sigma.ontologyportal.org:4010/sigma/WordNet.jsp

8 http://www.illc.uva.nl/EuroWordNet/

9 http://cwn.ling.sinica.edu.tw/

suggests new tags to users just in case they used an ambiguous tag for a resource. The tag is ambiguous because it also appears in a cluster of unrelated resources, as in the Flickr example. For example the spatially ambiguous tag *Cambridge* can co-occur either with *MA* or *UK*. In these situations one of these will be suggested as an additional disambiguator.

Clustering algorithms can identify different uses of tags, but they do not provide any semantics beyond this. [15] show that an analysis of the temporal and spatial distribution of tags can determine if a tag belongs to a place and/or an event. For example, they can identify that the tag *bay bridge* corresponds to a place, but *www2007* to an event.

Researchers have also investigated the possibility of inferring hierarchical relations between folksonomy terms. [16] also consider a probabilistic model of tag semantics in which ambiguity is directly observed through graphs of the distribution of concepts labeled by individual tags. For example *cooking* has a single very distinct distributional peak, whereas *XP* has several peaks corresponding to the various uses of the term. Because semantics is defined relative to the resources in the data set, the results are dynamic and depend on the current state of the concepts in the data set. But more interestingly, their probabilistic model can also infer hierarchical ordering among the tags by considering overlaps in the concepts covered. Another interesting attempt to infer hierarchies is to use conditional probabilities rather than distributional data. [17] inferred subsumption relations through conditional probabilities in tags. They say that X potentially subsumes Y if P(x|y >= t) and P(y|x < t), where t is a co-occurrence threshold. The algorithm can discover interesting subsumptions, like that between *san francisco* and *goldengatebridge, fishermanswharf, pier39*. On the other hand there are spurious probabilistic dependancies that lead to poor examples like *glass* subsuming *magnifying, blow, stained*. This highlights the problem with purely statistical procedures that are oblivious of syntactic or semantic constraints.

In terms of semantic enrichment there are several attempts to extend statistical approaches by extending folksonomies using resources such as Wikipedia, on line ontologies, and WordNet [18-20]. These resources are used in various ways, including to effectively cluster tags, for disambiguation, adding synonyms, and linking to annotated resources and ontology concepts. During this process the terms of the folksonomy are cleaned up and disambiguated, linked to formal definitions and given properties which make them more useful as ontologies. [21] also suggest a rich framework by which tags can acquire post hoc assignments to formal interpretations, including the categories of use suggested in [22].

There are also a few studies in which users are expected to contribute semantics at the time of tagging. [23] studies a corporate blogging platform which included a tagging interface. The tagging interface was linked to a domain ontology, and whenever someone typed a tag that had interpretations in the ontology the interface would present a choice of possible concepts to link the tag to. The ontology would also evolve as users typed new tags which were initially not in the ontology, but the scope of defined tags was limited by the ontology. [24] discuss a sophisticated Firefox plugin, *Semdrops*, which allows users to annotate web resources with a complex set of tags including *category, property,* and *attribute* tags. These are aggregated in a semantic wiki of the user's choosing. [25] reports on an open source bookmarking application (SemanticScuttle) that has been enhanced with *structurable tags* which are tags that users can enhance with inclusion and equivalence relations at the time of tagging. [26] describes *extreme tagging* in which users can tag other tags, to provide disambiguation and other relational information about tags.

Finally, the two previously mentioned commercial ventures Faviki and Zigtag should be mentioned as existing bookmarking services which make use of defined tags. Faviki uses Wikipedia concepts as common tags, and is able to aggregate tagged content according to Wikipedia categories. Since the defined tags are Wikipedia concepts, Faviki cannot semantically ground tags like *interesting, cool,* and *useful*. Zigtag uses dictionary entries, but also allows undefined tags, which make up a significant proportion of their tags.

This birds eye view of the literature shows that existing work is focused almost exclusively on the problem of extracting latent semantics from naive folksonomies composed of messy vocabularies rife with the problems of ambiguity and indeterminacy. In this respect the

work presented here represents a much less well explored effort in eliciting precise semantic tags at the time of tagging. The current work is distinguished from similar research along four major dimensions. First, Lexitags aims to provide a lightweight tagging tool that can be used to tag a wide range of content including html bookmarks, pictures, and local filesystem content. Second, we use WordNet as the primary semantic reference, exploiting the structure of WordNet to construct new relationships and lightweight ontologies. Third, no tags are allowed to be completely undefined, which makes for a more coherent tag collection. Fourth, Lexitag users are not expected to make any complex decisions when assigning semantic tags. They are not expected to contribute relational tags, and so on. They simply chose the sense of the word which they already had in mind when writing the tag.

6. Conclusion

The paper introduced the lexitags approach to social semantic tagging with simple lightweight tagging interfaces. Lexitags are tags whose semantics are grounded in disambiguated lexical items, and which stand in useful relations to other disambiguated lexical items. These form the basis of automatically generated lightweight ontologies which can take the role of universal interlingua between social applications in any domain, and in many non English languages.

Tags which have rich, unambiguous definitions make some aspects of previous work to make sense of tags, un necessary. There is no need to infer that spelling variations on a term have the same meaning, for example, because the distinction between word form and word meaning in lexitags already accommodates spelling variations. Similarly there is no need for disambiguation or clustering for the purpose of identifying different word senses. However, many of the current ideas can still be used in more refined ways. For example clustering is still useful but now at a more detailed level because we can focus on clusters within each sense. If we ignore the fruit sense of *apple*, for example, it may be possible to discover interesting clusters in the way the company name is used. Similarly, taxonomy inference for "tags-in-use" with any of the methods mentioned is still possible, but now it can be refined by taking into consideration the semantics of the tags. For example if subsumption can only occur between nouns, then *glass* will never subsume *magnifying*.

Semantic enrichment becomes much easier too, because lexitags are primarily WordNet synsets. As an example, WordNet already has a rich mapping to DBPedia, so embellishing the dynamically constructed ontology with Wikipedia facts is much simplified. This is the essence of the linked data movement, removing uncertainty and probability from data integration.

One of the most important claims is that WordNet is the ideal means by which to ground the semantics of common tags. This differentiates Lexitags from previous efforts such as Faviki, [21], and [24]. Faviki has chosen to use Wikipedia concepts instead, but we argue that WordNet is more useful as an interlingua because it is more flexible, has more general coverage of terms, and already has many mappings defined to resources such as DBPedia and SUMO.

In summary, this paper suggests that the tagging world be turned upside down. Rather than using clever algorithms for making sense of messy user generated tags, the clever algorithms should be used to help users generate tags that make sense in the first place.

References

1. Oreilly, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. (2007).
2. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web and Information Systems. 3, (2007).
3. Mathes, A.: Folksonomies-cooperative classification and communication through shared metadata. Computer Mediated Communication. (2004).
4. Tonkin, E., Guy, M.: Tidying up tags. D-Lib Magazine. 12, (2006).
5. Cattuto, C., Benz, D., Hotho, A.: Semantic grounding of tag relatedness in social bookmarking

systems. The Semantic Web-ISWC 2008. (2008).
6. National Information Standards Organization: NISO_vocabularies. (2005).
7. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics. 32, 13–47 (2006).
8. Veres, C., Johansen, K., Opdahl, A.: Browsing and Visualizing Semantically Enriched Information Resources. 2010 International Conference on Complex, Intelligent and Software Intensive Systems. 968–973 (2010).
9. Hirst, G.: Ontology and the Lexicon. Handbook on ontologies. (2009).
10. Shirky, C.: Ontology is Overrated--Categories, Links, and Tags. http://www.shirky.com/writings/ontology_overrated.html. (2007).
11. Niles, I., Pease, A.: Towards a Standard Upper Ontology. Proceedings of the international conference on Formal Ontology in Information Systems - FOIS '01. pp. 2–9. ACM Press, New York, New York, USA (2001).
12. Limpens, F., Gandon, F., Buffa, M.: Linking Folksonomies and Ontologies for Supporting Knowledge Sharing,
13. Begelman, G., Keller, P., Smadja, F.: Automated Tag Clustering:
Improving search and exploration in the tag space. Proc. of the Collaborative Web Tagging Workshop at WWW'06. (2006).
14. Weinberger, K., Slaney, M., van Zwoi, R.: Resolving tag ambiguity. Proceeding of the 16th ACM international conference on Multimedia, Vancouver, Canada. (2008).
15. Rattenbury, T., Good, N., Naaman, M.: Towards extracting flickr tag semantics. Proceedings of the 16th international conference on World Wide Web (2007).
16. Zhang, L., Wu, X., Yu, Y.: Emergent semantics from folksonomies: A quantitative study. Journal on Data Semantics VI. (2006).
17. Schmitz, P.: Inducing ontology from flickr tags. Collaborative Web Tagging Workshop at WWW2006. (2006).
18. Specia, L.: Integrating folksonomies with the semantic web. The semantic web: research and applications. (2007).
19. Angeletou, S., Sabou, M., al, E.: Bridging the gap between folksonomies and the semantic web: An experience report. In ESWC workshop. Bridging the Gap between Semantic Web and Web 2.0 (2007)
20. Van Damme, C., al, E.: Folksontology: An integrated approach for turning folksonomies into ontologies. In ESWC workshop. Bridging the Gap between Semantic Web and Web 2.0 (2007)
21. Limpens, F., Monnin, A., Laniado, D., Gandon, F.: NiceTag Ontology: tags as named graphs. In Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS '10). (2010).
22. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems, (2005).
23. Passant, A.: Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. Proceedings of International Conference on Weblogs … (2007).
24. Torres, D., Diaz, A., Skaf-Molli, H., Molli, P.: Semdrops: A Social Semantic Tagging Approach for Emerging Semantic Data. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011).
25. Huynh-Kim Bang, B., Dané, E., Grandbastein, M.: Merging semantic and participative approaches for organising teachers' documents. In J. Luca & E. Weippl (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008 (pp. 4959-4966). Chesapeake, VA: AACE., Vienna (2008).
26. Tanasescu, V., Streibel, O.: Extreme tagging: Emergent semantics through the tagging of tags. Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution, ESOE 2007, co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 12th, 2007. (2007).

Volatile Classification of Point of Interests based on Social Activity Streams

A. E. Cano, A. Varga and F. Ciravegna[1]

OAK Group, Dept. of Computer Science, The University of Sheffield, UK
{A.Cano, A.Varga, F.Ciravegna}@dcs.shef.ac.uk

Abstract. Location sharing services(LSS) like Foursquare, Gowalla and Facebook Places gather information from millions of users who leave trails in locations (i.e. chekins) in the form of micro-posts. These footprints provide a unique opportunity to explore the way in which users engage and perceive a point of interest (POI). A POI is as a human construct which describes information about locations (e.g restaurants, cities). In this work we investigate whether the collective perception of a POI can be used as a real-time dataset from which POI's transient features can be extracted. We introduce a graph-based model for profiling geographical areas based on social awareness streams. Based on this model we define a set of measures that can characterise a location-based social awareness stream as well as act as indicators of volatile events occurring at a POI. We applied the model and measures on a dataset consisting of a collection of tweets generated at the city of Sheffield and registered over three week-ends. The model and measures introduced in this paper are relevant for design of future location-based services, real-time emergency-response models, as well as traffic forecasting. Our empirical findings demonstrate that social awareness streams not only can act as an event-sensor but also can enrich the profile of a location-entity.

Keywords: Points of Interest, social awareness streams, social data mining, citizen sensing, emerging semantics

1 Introduction and Motivation

Recent studies in user profiling have proposed the use of social activity streams for modelling users' interest, activities and behaviour [11][1][3]. These studies explore a user's comments in windows of time for revealing hidden features; which can aid in profiling the user in real-time. Although people-entities have started to be modelled in real-time, little has been done in modelling other entities involved in the environment in which a user is immersed. One example of these entities is Location.

In terms of location-awareness, a Point of Interest (POI) has been so far modelled as a set of static data (e.g. name, address, geo-coordinates) and classified according to the type of services it provides. Nonetheless, there are diverse latent (or hidden) features which can describe volatile and temporal aspects of it. For example, in normal conditions London, UK can be classified as a city labelled as: Urban, Tourism, Fashion. However during the London riots(Aug 2011), the collective opinions gathered through social activity streams (i.e. Twitter) regarding this city, started profiling this place with

the following tags: looting,unrest,police. These tags clearly provide a temporal reclassification of this venue labelling it as for example: Political, Uprising, Violence.

In this paper, we investigate whether the supplement of situational knowledge extracted from social activity streams can be used to infer higher level contextual information, which can induce a transient representation of a venue. Given the real-time and volatile nature of events happening at a venue, providing an accurate classification of these events involve different challenges including the variation of the vocabulary and classes in which an event could be classified in time.

The contributions of this paper are as follows:

- *GeoLattice Awareness Streams*: We introduce a graph-based model for profiling geographical areas based on social awareness streams.
- *Approach to derive a transient semantic classification of a POI*: We present a novel approach for dynamically classifying POI based on location-based social footprints and DBPedia structured data. We define a set of measures that can characterise a location-based social awareness stream as well as act as indicators of volatile events occurring at a POI.
- *Empirical Study*: We applied this methodology in a dataset consisting of a collection of tweets generated at the city of Sheffield and registered over three week-ends.

The model and measures introduced in this paper are relevant for design of future location-based services, real-time emergency-response models, as well as traffic forecasting.

2 Related Work

Little work has been done in classifying POIs based on location-based social activity streams. However, there are several research directions closely related to POI classification. Analysing the contextual meanings of places has long attracted attention by researchers in fields like social interaction, environmental psychology, ubiquitous computing and spatial data mining. Researchers on social interaction and environmental psychology have documented the way in which mobile users tend to provide information about location when they are asked about their current activity [7][12]. Schegloff [10] noted that during a conversation, attention is exhibited to: 1) 'where-we-know-we-are'; 2) 'who-we-know-we-are'; 3) 'what-we-are-doing-at-this-point-in-conversation'; from which a '*this* situation' can be translated in some '*this* conversation, at *this* place, with *these* members, at *this* point in its course'. This contextual knowledge has been used to infer a users' situational features including a person's level of availability or interruptibility.

The role of geography and location in online social networks has recently attracted increasing attention. Experimental work done on location awareness has shown that location sharing services (LSS) (e.g. Foursquare) are used to express not only users' whereabouts but also their moods, lifestyle and events [2]. In their work, Barkhuus et al. allowed users to tag areas and build a repartee in a group. They pointed out four different types of location labels that participants used in their study, including: 1) geographic references, 2) personal meaningful place, 3) activity-related labels, and 4) hybrid labels.

Cheng et al.[4] modelled the spatial distribution of words in Twitter's user-generated content for predicting user's location. Following a top-down approach they propose a probabilistic framework for estimating a Twitter user's city-level location based on the content of the user's tweets even on the abscence of any geospatial cues. Although their approach is content-based and can automatically indetify words in tweets with a strong geo-scope, they don't provide a topical categorisation of a given geo-scope.

Further work from Cheng et al [13] study mobility patterns of users in location sharing services (LSS), they correlate social status, geographic and economic factors with mobility and perform a sentiment-based analysis of post for deriving unboserved context between people and locations.

Lin et al [8] derive a taxonomy of different place naming methods, showing that a person's perceived familiarity with a place and the entropy of that place (i.e. the variety of people who visit it) strongly influence the way people refer to it when interacting with others. Based on this taxonomy, they present a machine learning model for predicting the place naming method people choose. Ireson and Ciravegna [6] study toponym resolution (i.e. the allocation of specific geolocation to target location terms) using Flickr data. They construct an SVM classifier for predicting location labels associated to a location term. Their model makes use of information context features including geo-tag media, users' contacts'related tags.

Regarding place descriptions based on location sharing services (LSS), Hightower [5] redefines a place as an evolving set of both communal and personal labels for potentially overlapping geometric volumes. He highlights that a meaningful place can capture the venue's demographic, environmental, historic, personal or commercial significance.

Our work is in line with Hightower's definition of a place, however rather than study location-sharing practices we aim to study how location-based generated content can be modelled for discoverying topics or categories that classify a place on time.

3 GeoLattice Awareness Stream

Following the Tweetonomy model suggested by Wagner and Strohmaier[11], we introduce a formalisation for describing the comments related to a geographical region in time; we refer to it as GeoLattice Awareness Streams.

The W3C POI Working Group [1] defines a POI as a human construct which describes information about locations. According to their definition, a POI is not limited to a set of coordinates and an identifier but also can include a more complex structure like for example a three dimensional model o a building, opening and closing hours etc.

As mentioned in the previous section, location sharing services provide a classification of their points of interest according to the type of service they provide (e.g. Food, Nightlife Spots), however these categories are static and do not reveal any information about the type of events occurring in a given venue. The key idea of our approach is to enrich a POI by associating transient categories emerging from social activity streams regarding this POI.

Definition 1. *A GeoLattice Awareness Stream can be defined as a sequence of tuples* $\mathtt{S} := (Poi_{q1}, C_{q2}, R_{q3}, Y, ft)$ *where*

[1] W3C POI Working Group, http://www.w3.org/2010/POI/

- $\mathtt{Poi}, \mathtt{M}, \mathtt{R}$ *are finite sets whose elements are called Points of Interest, Messages and Resources;*
- *Each of these sets is qualified by* $\mathtt{q1}$, $\mathtt{q2}$ *and* $\mathtt{q3}$ *respectively (explained below);*
  - *The qualifier* $\mathtt{q1}$ *for a Point of Interest (poi) includes for example name, geographical-bounding area, and geo-coordinates.*
  - *The qualifier* $\mathtt{q2}$ *for a message* m *considers for example the message's source (e.g Facebook, Twitter) and it's geo-coordinates.*
  - *The qualifier* $\mathtt{q3}$ *for a resource* r *considers:* $\mathtt{R}_{cat}$ *(category),*$\mathtt{R}_k$ *(keywords),* $\mathtt{R}_h$ *(hashtags).*
- $\mathtt{Y}$ *is the ternary relation* $\mathtt{Y} \subseteq \mathtt{Poi} \times \mathtt{M} \times \mathtt{R}$ *representing a hypergraph with ternary edges. The hypergraph of a GeoLatice Awareness Stream* $\mathtt{Y}$ *is defined as a tripartite graph* $\mathtt{H}(\mathtt{Y}) = \langle \mathtt{V}, \mathtt{E} \rangle$ *where the vertices are* $\mathtt{V} = \mathtt{Poi} \cup \mathtt{M} \cup \mathtt{R}$, *and the edges are:* $\mathtt{E} = \{\{poi, m, r\} \mid (poi, m, r) \in Y\}$.
- f_t *is a function that assigns a temporal marker to each Y;* $f_t : Y \rightarrow T$.

Given a GeoLattice awareness stream S, a POI awareness stream can be defined as the sequence of tuples from S where:

$\mathtt{S}(\mathtt{Poi}') = (\mathtt{Poi}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', ft)$, and $\mathtt{Y}' = \{(\mathtt{poi}, \mathtt{m}, \mathtt{r}) \mid \mathtt{poi} \in \mathtt{Poi}' \vee \exists \mathtt{poi}' \in \mathtt{Poi}', \tilde{\mathtt{m}} \in \mathtt{M}, \mathtt{r} \in \mathtt{R} : (\mathtt{poi}', \tilde{\mathtt{m}}, \mathtt{r}) \in \mathtt{Y}\}$ i.e., a POI Awareness Stream is the aggregation of all messages which are related to a certain set of points of interest $\mathtt{poi} \in \mathtt{Poi}'$ and all resources and further points of interest related with these messages.

4 Transient Semantic Classification of a POI

4.1 Problem Statement

Comments extracted from social activity streams can be described as semi-public, natural-language messages produced by different users and characterised by their brevity. Given these characteristics and the variation in the vocabulary appearing on a POI awareness stream comments, finding relevant categories that can accurately qualify a comment is a challenging task.

Definition 2. *We define a temporal classification of a Point of Interest as the aggregation of* $\mathtt{R}_{cat}$ *category resources qualifying messages contained in a specific window of time denoted by* $[t_s, t_e]$. *An* $\mathtt{S}(\mathtt{Poi}')[t_s, t_e]$ *is defined as* $\mathtt{S}(\mathtt{Poi}')$ *where* $ft : Y \rightarrow T, t_s \leq ft \leq t_e$.

Given the above definition, our task consists on obtaining category resources $\mathtt{R}_{cat}$ which can classify a poi within a window of time $[t_s, t_e]$. In this section, we introduce a strategy for categorising points of interest.

The POI categorisation within a window of time could enable reactive services (e.g. targeting advertisements to users based on a users location and the POI categorisation, emergency-response).

4.2 Entity-Based Discovery of Transient Categories

Our intuition is to use the categorisation of the messages' resources generated from a Point of Interest awareness stream ($\mathtt{S}(\mathtt{Poi}')$) taken in windows of time ($[t_s, t_e]$), to induce a categorisation function. Figure 1 presents an overview of our approach.

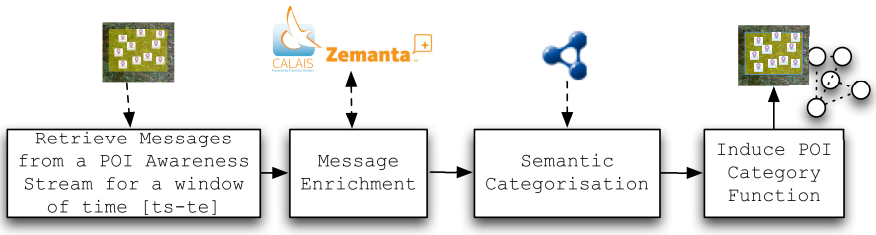


Fig. 1. Category Induction Pipeline: Messages are retrieved from a POI awareness stream. DBPedia categories are derived for each enriched message. These set of categories are used to induce a transient categorisation of a Point of Interest.

Message Enrichment Given a message from a POI awareness stream $\mathtt{S}(\mathtt{Poi}')$, we perform a lightweight *message enrichment* by using Zemanta [2], and OpenCalais [3]. These services perform entity-extraction on the input message identifying resources which can be qualified as: $\mathtt{R}_o$ (organisations – entities recognised as an organisation), $\mathtt{R}_p$ (people –entities recognised as a person), $\mathtt{R}_l$ (location – entities recognised as a location) and $\mathtt{R}_{li}$ (links resources). These services also provide DBPedia concepts relevant to the message. Consider the example in Figure 2, where the extracted entities and DBPedia concepts for a Twitter message are shown.

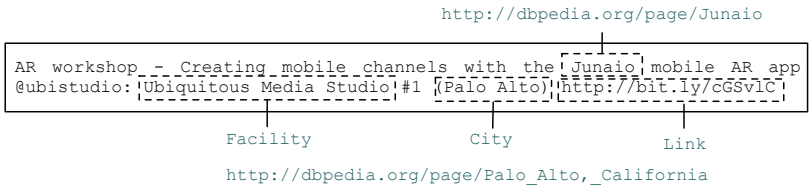


Fig. 2. Message Enriched with Zemanta and OpenCalais services. These service return entity labels as well as DBPedia concepts related to the message

Semantic Categorisation In order to semantically categorise a POI stream's message (m), we search for DBPedia concepts which are relevant to the extracted entity-based resources, and aggregate these concepts to those already suggested by the message enrichment services. Given a resource (r) we extract DBPedia *categories* and *broader categories* from the DBPedia Linked Data Graph (D) using the following construct:

[2] Zemanta, http://www.zemanta.com/
[3] OpenCalais, http://www.opencalais.com/

$$\mathtt{R}_{cat}(r) = \{x_{cat} \cup x_{broaderCat} | \\ < r, \text{dcterms:subject}, x_{cat} > \\ \wedge < x_{cat}, \text{skos:broader}, x_{broaderCat} > \in D \} \tag{1}$$

For each resource (r) we SPARQL query DBPedia retrieving the collection of categories (dcterms:subject) and parent categories (skos:broader) of r. Using the previous construct, we derive the categories presented in Table 4.2 for the resource Palo_Alto contained in the example of Figure 2. These categories become a resource category R_{cat} of the POI awereness stream ($S(Poi')$).

Entity	**Category**
(of type City) `Palo_Alto`	`dcterms:subject` *Palo_Alto,_California*
	`skos:broader` *Populated_places_in_Santa_Clara*
	`skos:broader` *University_towns_in_the_United_States*
(of type Thing) `Junaio`	`dcterms:subject` *Augmented_reality*
	`skos:broader` *Mixed_reality*

Table 1. Categories and broader categories derived for the entities extracted from the comment in Fig 2

Induce Category Function After applying the *semantic categorisation* technique to all messages belonging to a POI stream taken from a window of time $[t_s, t_e]$, we need to weight them in order to identify the relevant categories.

In order to do so, we utilise the resource category stream ($S(R'_{cat})$) of a POI stream ($S(Poi')$), which is the collection of all category resources classifying the POI stream's messages. For characterising the POI stream ($S(Poi')$) based on the category resources we propose two metrics:

1. *Category Entropy of a Stream*, which indicates the topical diversity of the stream. We defined the category entropy in terms of the POI stream's vocabulary as :

$$\mathtt{CE(c)} = -\sum_{w \in R_k} P(w|c) * log(P(w|c)) \tag{2}$$

where w is a word in the POI stream's vocabulary ($S(R'_k)$), and c is a category in the POI stream's categories ($S(R'_{cat})$). Low category entropy levels reveal that a stream is dominated by few categories, while a high category balance reveals a higher topical diversity. In normal conditions (i.e. no special events happening), we would expect for example to obtain a low category entropy levels for a POI stream referring to a Restaurant, since the messages would be classified within a limited set of categories related to Food. While for a POI stream referring to a city

in normal conditions (no particular events happening), we would expect to observe higher category entropy levels since the topical diversity would be higher.
However if normal conditions are broken, and unexpected (or volatile) events start to happen, we would expect to observe an increment in the category entropy levels of Restaurant POI stream, and a decrement in the category entropy levels of a City POI stream. The category entropy acts in this way as an indicator of volatile events.

2. *Mutual Information (MI)*, measures the information that two discrete random variables share. In this work we consider the following:
   - *Categories-Hashtags (MI)*

$$\mathtt{I(C;H)} = \sum_{c \in R_{cat}} \sum_{h \in R_h} p(c,h) * log \frac{p(c,h)}{p(c)p(h)} \tag{3}$$

   where c is a category in the POI stream's categories ($S(R'_{cat})$) and h is a hashtag in the POI stream's hashtags ($S(R'_h)$) and p(c,h) is the joint probability distribution function of C and H, with marginals p(c) and p(h).
   - *Categories-Keywords (MI)*

$$\mathtt{I(C;K)} = \sum_{c \in R_{cat}} \sum_{w \in R_k} p(c,w) * log \frac{p(c,w)}{p(c)p(w)} \tag{4}$$

   where c is a category in the POI stream's categories ($S(R'_{cat})$) and w is a word in the POI stream's keywords ($S(R'_k)$).
   - *Hashtags-Keywords (MI)*

$$\mathtt{I(H;K)} = \sum_{h \in R_h} \sum_{w \in R_w} p(h,w) * log \frac{p(h,w)}{p(h)p(w)} \tag{5}$$

   where h is a hashtag in the POI stream's hashtags ($S(R'_h)$).

The higher the mutual information, the more one random variable is relevant to the other.

5 Experiments

In this section we discuss our approach for evaluating the accuracy of the strategies proposed in Section 4 by using the formalisation introduced in Section 3. In order to identify a transient categorisation of a point of interest we decided to investigate a POI stream $S(Poi')$ in windows of time of one week-end.

5.1 Dataset

The corpus used for our study consists of Twitter messages taken over three week-ends in the city of Sheffield. Since we aim to study patterns emerging from volatile events we registered a week-end in normal conditions (i.e. no events happening) from 2011-06-10 to 2011-06-13 as control and two more week-ends in which especial events occurred.

The especial events were the Sheffield Food Festival (from 2011-07-08 to 2011-07-11) and the Sheffield Tramlines Music Festival (from 2011-07-22 to 2011-07-25). The data was collected using the Twitter Streaming API[4] with the public firehose and filtering by geographical area (using Sheffield's bounding geo-coordinates).

For each week-end dataset we removed stop words and applied the approach presented in Section 4.2, extracting hasthags, keywords and entity resources as well as DBPedia categories for these resources. The statistics for each stream is summarised in Table 2.

Week-End	Tweets	Users	Hashtags	Links	GeoTagged	RT	Reply
Common	5853	649	9%	5%	27.11%	2.8%	40.6%
Food Festival	11203	726	18%	4.2%	40.7%	4.2%	40.7%
Tramlines	13381	899	9%	24%	14.8%	9%	39.3%

Table 2. General Statistics, percentages of messages containing hashtags, links, geotagged, RT (retweeted) and Reply (tagged as a reply-tweet)

Week-End	Hashtags	Resources[a]	Categories[b]
Common	9%	1475	9495
Food Festival	18%	2681	830
Tramlines	9%	1912	9770

[a] DBPedia resources derived from the messages
[b] DBPedia categories derived from the resources

Table 3. Streams hashtags, and categories.

5.2 Results and Discussion

First we analyse the most frequent hashtags in the three datasets presented in Table 4. Although trends in hashtags are useful for detecting changes in a stream, hashtags tend to present high ambiguity, and a frequent use of abbreviations. These are some of the reasons why hashtags are not enough to provide a categorisation by themselves.

We calculated the categories' entropies for each of the three datasets' categories. The categories entropy distributions are shown in Figure 3. We can observe that the stream taken from Sheffield in normal conditions (labelled as "Week End" in the graph) presents denser regions in higher entropy levels.

[4] https://dev.twitter.com/

Order	Common	Food Festival	Tramlines
1	ff	ff	tramlines
2	sheffdocfest	foofighters	ff
3	blogsmoda	sheffield	buskersbus
4	ofs	notw	replacewordinamoviewithgrind
5	bbcf	totb	sheffield
6	blkstg	bbcf	amywinehouse
7	nosleeptilleadmill	titp	swfc
8	underwearshongs	swfc	allabouttonight
9	articmonkeys	sonishphere	hallamfm
10	beards	believe	forgetramlines

Table 4. Top 10 Most Frequent hashtags

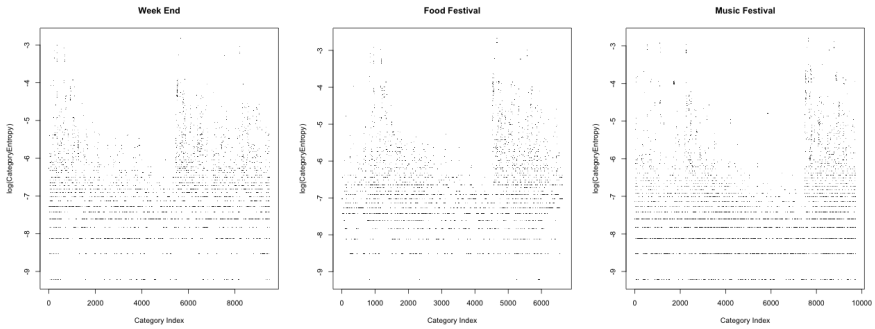


Fig. 3. Category Entropies vs. Category Index

Since lower category entropy levels provide a better information gain, we pick a category entropy threshold from which to pick categories. For these data sets and following Figure 3 we picked -9 as a threshold obtaining: 255 categories for the common week-end, 28 categories for food festival, and 562 categories for Tramlines. Table 5 shows the top 21 categories for each stream.

It is important to notice that we are not biasing the results by picking a priori hashtags relevant to the week-end events, but rather the categories emerge from category entropy analysis. From Table 5, very disparate categories appeared for the week-end in normal conditions ("common"), while for the Food Festival week-end we find categories which appear to be related either to external events or future events (Music_Festivals), as well as categories related to a current event (Food_companies_of_the_United_Kingdom). Incidentally for the food festival week-end we found two sets of semantically coherent categories, the first (categories from 13-17) matches an external event related to the 2012 Olympic tickets sales, while the second (categories 18-23) appears to be closely

Order	Common	Food Festival	Tramlines
1	History_of_the_Middle_East	Music_festivals_by_country	Arts_occupations
2	Mediterranean	American_Roman_Catholics	Music_industry
3	Near_East	American_people_by_ethnic_or_national_origin	Disco
4	Western_Asia	Food_companies_of_the_United_Kingdom	Dance_music_by_subgenre
5	Geography_of_Iraq	Public_opinion	DJing
6	Geography_by_country	Youth	Electronic_music
7	Cultural_history	Students	New_York_culture
8	Argentine_culture	Education	New_York_City
9	Argentine_society	Adolescence	Rock_music_genres
10	Nicaraguan_culture	Sport_and_politics	Rock_music
11	Languages_of_Colombia	Athletic_culture_based_on_Greek_antiquity	Underground_culture
12	Zambian_culture	Athletics_in_ancient_Greece	Postmodernism
13	Ike_&_Tina_Turner	Olympic_culture	Types_of_subcultures
14	Sun	Olympics	Youth_culture_in_the_United_Kingdom
15	Social_groups	Sport_and_politics	British_culture
16	Corporate_groups	Olympic_competitors	Youth_culture
17	Cognition	Sports_competitors_by_competition	Pejorative_terms_for_people
18	Prejudice	La_Liga	Slang
19	Critical_thinking	People_associated_with_Glasgow	Stereotypes
20	Social_class_subcultures	Football_in_Spain	European_Union_member_states
21	Romani_loan_words	Footballers_in_Spain_by_club	European_Union

Table 5. Top 21 Categories (sorted by category entropy (decreasing order))

relevant to an event involving Spanish football. We can observe that the categories obtained for the Tramlines Music Festival are more semantically coherent compared to the other two week-ends. This could be due to a higher relevance of the tramlines event compared to other events occurring at the same time in the city or externally.

Although some of the categories emerging from the category entropy analysis give an insight of endemic events, there are also other categories which provide information of events occurring externally. Hence, a Point of Interest considered as a Location-Entity presents the "meformer" and "informer" patterns observed by Naaman et al. [9] in Person-Entity activity streams. In this case the "Meformer" pattern refers to a self focus of a Location-Entity, presenting information about endemic events, while the "Informer" pattern refers to an information sharing of external events, not necessarily related to this Location-Entity.

In order to provide a context in which the category is being used, we use the mutual information between categories and hashtags (see Equation 3), from which we obtain a set of hashtags that can be used to further derived related keywords (see Equation 5)

Category	Hashtag	Keywords
heightSlang	#jobs, #jheeze, #rihanna, #neversayneverdvd	earth, swag, concert
Music_Industry	dance_music	party,music,record

Table 6. Hashtags and Keywords derived for two category using mutual information (see Equation 3)

6 Conclusions and Future Work

The identification of category resources $\mathtt{R}_{cat}$ from a POI awareness stream $\mathtt{G_a}(\mathtt{P}')$ can be considered as a multi-class, multi-label classification task. This becomes challenging when no assumptions can be made a priori on the type of classes that will classify future events. Our approach semantically enriches the information of the social stream by providing a DBPedia based categorisation.

We have presented a formalisation for describing geographically bounded social awareness streams, we have also provided an approach for deriving transient categorisations of points of interest. We have applied our methodology on a data set and we have presented an empirical analysis of our results.

Future work includes a quantitative evaluation of this methodology by using larger datasets in which events have been identified a priori, and against which we can evaluate the emerging categories resulting from our approach.

Questions still remain on how we could determine a semantic coherence metric, which could induce broader category clusters. A semantic cluster of these categories can provide a better insight to the kind of events to which they refer to. Take for example the categories found for the Tramlines event, although we know these categories are related to music, we still haven't inferred the broader category "Music_Festival".

Acknowledgements A.E. Cano is funded by CONACyT, grant 175203.Andrea Varga is funded by the SAMULET project, co-funded by TSB and Rolls-Royce plc/

References

1. F. Abel, Q. Gao, G. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *In proceedings of Extended Semantic Web Conference 2011*, May 2011.
2. L. Barkhuus, B. Brown, M. Bell, S. Sherwood, M. Hall, and M. Chalmers. From awareness to repartee: sharing location within social groups. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 497–506, New York, NY, USA, 2008. ACM.
3. A. Cano, S. Tucker, and F. Ciravegna. Capturing entity-based semantics emerging from personal awareness streams. In *Proceedings of the Workshop on Making Sense of Microposts (MSM2011)*, May 2011.
4. Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.
5. J. Hightower. From position to place. pages 10–12, 2003.
6. N. Ireson and F. Ciravegna. Toponym resolution in social media. In *Proc., 9th International Semantic Web Conference*. ISWC 2010, 2010.
7. E. Laurier. Why People Say Where They Are During Mobile Phone Calls. *Environment and Planning D: Society and Space*, 2000.
8. J. Lin, G. Xiang, J. I. Hong, and N. Sadeh. Modeling people's place naming preferences in location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 75–84, New York, NY, USA, 2010. ACM.

9. M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW '08: Proc., 2010 ACM conference on Computer supported cooperative work*, pages 189–192, 2010.
10. E. Schegloff. Notes on a conversational practice: formulating place. *in Studies in Social Interaction Ed D Sudnow (Free Press)*, 1972.
11. C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.
12. A. Weilenmann. "i can't talk now, i'm in a fitting room": formulating availability and location in mobile-phone conversations. *Environment and Planning A*, 35(9):1589–1605, 2003.
13. K. L. D. Z. S. Zhiyuan Cheng, James Caverlee. Exploring millions of footprints in location sharing services. In *5th International Conference on Weblogs and Social Media (ICWSM)*, ICWSM '11. ACM, 2011.

Semantic Technologies to Support the User-Centric Analysis of Activity Data

Mathieu d'Aquin, Salman Elahi and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
{m.daquin, s.elahi, e.motta}@open.ac.uk

Abstract. There is currently a trend in giving access to users of online services to their own data. In this paper, we consider in particular the data which is generated from the interaction between a user and an organisation online: activity data as held in websites and Web applications logs. We show how we use semantic technologies including RDF integration of log data, SPARQL and lightweight ontology reasoning to aggregate, integrate and analyse activity data from a user-centric point of view.

1 Introduction

Social interactions on the Web, especially between individual users and organisations, rely on the exchange of personal data. As discussed in the article "Show Us the Data! (It is ours after all)" in the New York Times by Richard H. Thaler[1], while being heavily exploited by online organisations, these data are rarely made accessible to the users themselves. However, many initiatives have emerged recently arguing that obtaining and being able to exploit such data could be very beneficial to individual users. The *mydata* project in the UK[2] for example focuses on consumer data. At Google, the *Data Liberation Front*[3] has been formed to push the deployment of mechanisms allowing users to extract their data from Google services. In relation to this, there is currently a wide expansion of the idea of self-tracking, with new forms of applications being created based on social and personal data on the Web (see e.g., [1, 2]).

There are however specific challenges that appear when trying to apply such a user-centric perspective on activity data. Activity data typically sits in the logs of websites and Web applications, and are exploited by online organisations, in an aggregated form, to provide overviews of the interactions between the organisation's online presence and its users (most commonly in the form of website analytics). UCIAD[4] is a short project with the aim to experiment with the technological challenges that are faced when trying to invert the perspective

[1] http://www.nytimes.com/2011/04/24/business/24view.html
[2] http://www.bis.gov.uk/news/topstories/2011/Apr/better-choices-better-deals
[3] http://www.dataliberation.org/
[4] http://uciad.info

on activity data: provide individual users with an overview of their interactions with the online organisation.

This raises a number of challenges for which the use of semantic technologies seem to provide adequate solutions:

Fragmentation and heterogeneity: Activity data is typically held in log files that have different formats, and might not be easily integratable from one system (website, application) to another.

User identification: Recognising and identifying a user within the data is typically a problem faced by any activity data analysis. However, when taking a user-centric perspective, a user needs to be identified over several systems and the consequences of inaccurately recognising a user can be more critical.

Data analysis: Activity data is generally available through raw, uninterpreted logs from which meaningful information is hard to obtain.

Scale: Tracking user activities through logs can generate immense amounts of data. Typical systems cope with such scale through aggregating data based on clusters of users. Here, we need to keep the whole set of data for each individual user available to provide meaningful analysis of their interaction with the organisation in a user-centric way.

In this paper, we show how we investigated and handled these challenges through relying on semantic technologies, especially RDF for the low level integration and management of data, ontologies for the aggregation of heterogeneous data and their interpretation, and lightweight ontological reasoning to support customisable analysis of user-centric activity data. We also discuss how these components have been put together in a demonstrator platform, the UCIAD platform, providing user-centric views on activity data related to several websites of the Open University.

2 Activity Data Integration - Base Architecture

There are two reasons why we believe semantic technologies can benefit the analysis of activity data in general, and from a user-centric perspective in particular. First, ontology related technologies (including OWL, RDF and SPARQL) provide the necessary flexibility to enable the "lightweight" integration of data from different systems. Not only we can use ontologies as "pivot" models for data coming from different systems, but such models are also easily extensible to take account of the particularities of the systems available, but also to allow for custom extensions to deal with particular users, making personalised analysis of activity data feasible.

The overall architecture of the activity data infrastructure set up for the UCIAD project is shown in Figure 1. Its goal is to support the extraction from a variety of logs, of homogeneous representations of the traces of activity data present in these logs and store them in a common semantic store so that they can be accessed and queried by the user. We use RDF as a common data model, and a triple store providing SPARQL querying facilities for storing and accessing the

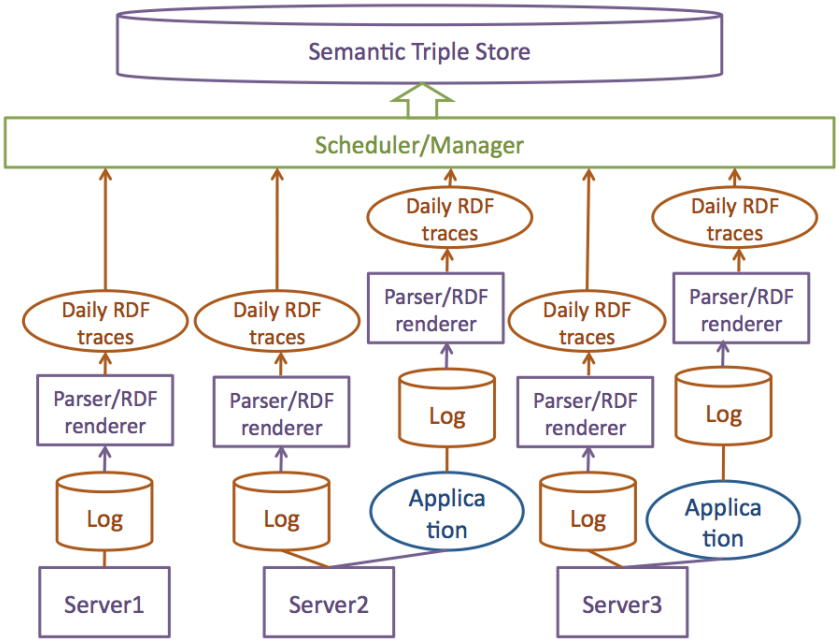


Fig. 1. Overview of the architecture of the UCIAD platform.

data.[5] Information from logs is extracted on a daily basis and represented using the ontologies described in the next section, which together with the semantic store represent the basis of the platform to provide user-centric views on activity data.

3 Aggregating Heterogeneous Activity Data - The UCIAD Ontologies

Compared to other domains, the advantage of user activities is that there is a lot of data to look at. This might be seen as an issue (from a technical and conceptual point of view), but in reality, this allows us to apply a bottom-up approach to building the ontologies necessary to achieve our goal: modelling through characterising the data, rather than through conceptualising the domain from established expert knowledge. It also gives us an insight into the scale of the task, and the need for adapted tools to support both the ontological definition of specific situations, and the ontology-based analysis of large amounts of traces of activity data.

[5] We use OWLIM (http://www.ontotext.com/owlim) which provides scalable storage and lightweight reasoning facilities.

3.1 Identifying Concepts and their Relations

The first step in building our ontologies is to identify the key concepts, i.e., the key notions, that we need to tackle, bearing in mind that our ultimate goal is to understand activities. We rely extensively on website logs as sources of activity data. In these cases, we can investigate requests both from human users and from robots automatically retrieving and crawling information from the websites. The server logs in question represent collections that can be seen as traces of activities that these users/robots are realising on websites. We therefore need to model these other aspects, which correspond to actions that are realised by actors on particular resources. We propose three ontologies to be used as the basis of the work in UCIAD:

The Actor Ontology is an ontology representing different types of actors (human users vs. robots), as well as the technical settings through which they realise online activities (computer and user agent).

The Sitemap Ontology is an ontology to represent the organisation of webpages in collections and websites, and which is extensible to represent different types of webpages and websites.

The Trace Ontology is an ontology to represent traces of activities, realised by particular agents on particular resources (here, mostly webpages). As we currently focus on HTTP server logs, this ontology contains specific sections related to traces as HTTP requests (e.g., HTTP methods are represented as instances of "Action" and HTTP response codes are included within "Response" type objects). It is however extensible to other types of traces, such as specific logs for particular applications.

3.2 Reusing Existing Ontology

When dealing with data and ontologies, reuse is generally seen as a good practice. Apart from saving time from not having to remodel things that have already been described elsewhere, it also helps anticipating on future needs for interoperability by choosing well established ontologies that are likely to have been employed elsewhere. We therefore investigated existing ontologies that could help us define the notions mentioned above. Here are the ontology we reused:

The FOAF ontology (http://xmlns.com/foaf/spec/) is commonly used to describe people, their connections with other people, but also their connections with documents. We use FOAF in the Actor Ontology for human users, and in the Sitemap Ontology for webpages (as documents).

The Time Ontology (http://www.w3.org/TR/owl-time/) is a common ontology for representing time and temporal intervals. We use it in the Trace Ontology.

The Action ontology (http://ontology.ihmc.us/Action.owl) defines different types of actions in a broad sense, and can be used as a basis for representing elements of the requests in the UCIAD Trace Ontology, but also as a base typology for actions. It itself relies on a number of other ontologies, including its own notion of actors.

While not currently used in our base ontologies, other ontologies can be considered at a later stage, for example to model specific types of activities. These include the Online Presence Ontology (OPO[6]), as well as the Semantically-Interlinked Online Communities ontology (SIOC[7]).

The current version of the ontologies developed as part of this work are available at https://github.com/uciad/UCIAD-Ontologies.

4 Identifying and Extracting User Activity Data

Once activity data have been extracted and represented according to the ontologies briefly described above, the next step consists in identifying and aggregating, in this data the traces of activities realised by a particular user, in order to create a user-centric view of his or her interactions with the considered systems (websites, applications).

4.1 Overview

The information the UCIAD platform collects regarding users can be seen as similar to the one basic analytics systems have. The user is rarely seen directly, as the interaction is mediated through a "user agent": a software programme running on a particular computer. Each HTTP request is associated with the ID of the user agent realising it, and the IP address of the corresponding computer. Several analytics systems use the combination of these two parameters to recognise a user with a reasonable level of accuracy. The disadvantage however is that the same user can be using different agents (e.g., different browsers) and different computers (or even mobile phones) to access the Web.

In UCIAD, we have the advantage that it is very likely that the user will connect to the UCIAD platform using the same agents and computers they usually use to access the Web, and especially the considered websites. The "settings" the user is using can therefore be detected at the time of logging in, and be attached to the user account. These settings will then be used to aggregate all the activity data that have been realised using the same computer and user-agent, and be added to the set of activity data for the particular user.

In addition, this provides a convenient mechanism to aggregate information realised on different computers and different settings. The user can log again in the UCIAD platform with a different browser, or a different device. When that happens, as described in Figure 2, the current setting will simply be added to the list of known settings for this user, and contribute another set of activity data around this particular user.

A setting, in our ontology, corresponds to a computer (generally identified by its IP address) and an agent (generally a browser), identified by a complex string such as

[6] http://online-presence.net/opo/spec/
[7] http://sioc-project.org/ontology

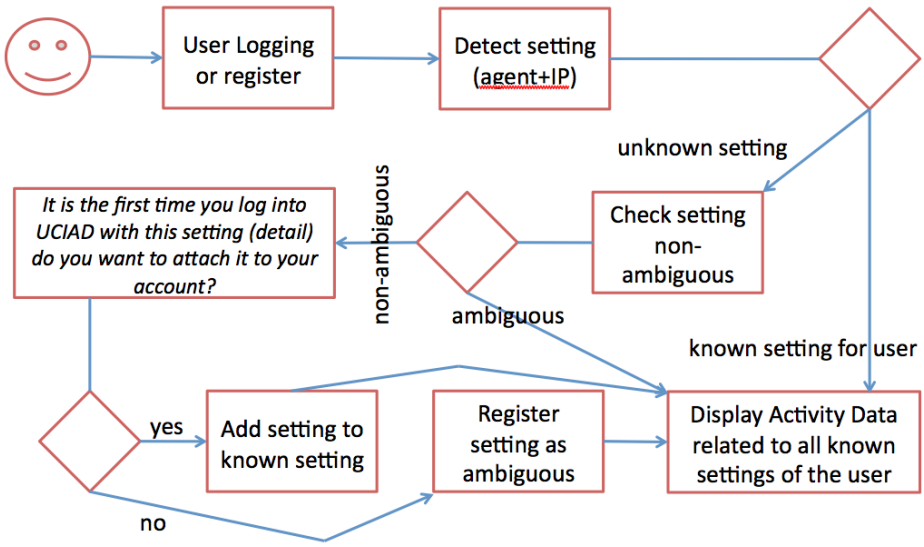


Fig. 2. Associating user accounts to their settings.

```
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_6) AppleWebKit/534.24
(KHTML, like Gecko) Chrome/11.0.696.68 Safari/534.24)
```

Such a setting can be associated to a user based on a representation following our ontologies described above, such as in the example below:

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:actor="http://uciad.info/ontology/actor/">
<rdf:Description rdf:about="http://uciad.info/actor/mathieu">
 <actor:knownSetting
  rdf:resource="http://uciad.info/actorsetting/4eafb6e074f46857b1c0b4b2ad0aa8e4"/>
 <actor:knownSetting
  rdf:resource="http://uciad.info/actorsetting/c97fc7faeadaf5cac0a28e86f4d723c9"/>
 <actor:knownSetting
  rdf:resource="http://uciad.info/actorsetting/eec3eed71319f9d0480ff065334a5f3a"/>
</rdf:Description>
<rdf:Description
 rdf:about="http://uciad.info/actorsetting/4eafb6e074f46857b1c0b4b2ad0aa8e4">
 <actor:hasComputer rdf:resource="http://uciad.info/computer/4eafb6e074f46857b1" />
 <actor:hasUserAgent rdf:resource="http://uciad.info/useragent/c0b4b2ad0aa8e4"/>
 </rdf:Description>
<rdf:Description rdf:about="http://uciad.info/computer/4eafb6e074f46857b1">
 <actor:hasIPAddress>187.108.24.45</actor:hasIPAdress>
```

```
</rdf:Description>
<rdf:Description rdf:about="http://uciad.info/useragent/c0b4b2ad0aa8e4">
 <actor:hasAgentID>Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_6)
   AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.68 Safari/534.24)
 </actor:hasAgentID>
</rdf:Description>
</rdf:RDF>
```

This indicates that the user http://uciad.info/actor/mathieu has three settings. These settings are all on the same computer and correspond to the Safari and Chrome browsers, as well as the Apple PubSub agent (used in retrieving RSS feeds amongst other things).

4.2 Extracting User-Related Data

Managing the activity data regarding a particular user corresponds to creating a sub-graph of the complete graph of raw activity data we collect from logs, based on the information about the known settings of the user. To identify a user, we rely here on the settings used to realise the activity. Each trace of activity is realised through a setting (linked to the trace by the *hasSetting* ontology property). Knowing the settings of a user therefore allows us to list the traces that correspond to this particular user through a simple query. Using a SPARQL Construct query, we can create a model, i.e. an RDF graph, that contains all the information related to the user's activity on the considered websites:

```
PREFIX tr:<http://uciad.info/ontology/trace/>
PREFIX actor:<http://uciad.info/ontology/actor/>
construct {
  ?trace ?p ?x.
  ?x ?p2 ?x2.
  ?x2 ?p3 ?x3.
  ?x3 ?p4 ?x4
} where{
  <http://uciad.info/actor/mathieu> actor:knownSetting ?set.
  ?trace tr:hasSetting ?set.
  ?trace ?p ?x.
  OPTIONAL {{?x ?p2 ?x2}.
  OPTIONAL {{?x2 ?p3 ?x3}.
  OPTIONAL {?x3 ?p4 ?x4}}}
}
```

The results of this query correspond to all the traces of activities in the collected data that have been realised through known settings of the user http://uciad.info/actor/mathieu, as well as the surrounding information. These data, materialised as an RDF graph, can therefore be considered on its own, as a user-centric view on the activity data available through integrated logs.

4.3 Managing Access Right over Semantic Data

We store, manipulate and reason over activity data using Semantic Web technologies, namely RDF, a triple store with inference capabilities and SPARQL for querying. As part of the UCIAD platform, we needed a mechanism to restrict the queries being sent to only the part of the data that the current user has access to: his/her own subgraph of activity data.

Unfortunately, most current triple stores, and especially the one we are employing, do not provide sufficiently fine-grained access control mechanisms, allowing to associate sub-graphs to particular users. We therefore implemented our own mechanism, which can be seen as a generic recipe for access control over activity data.

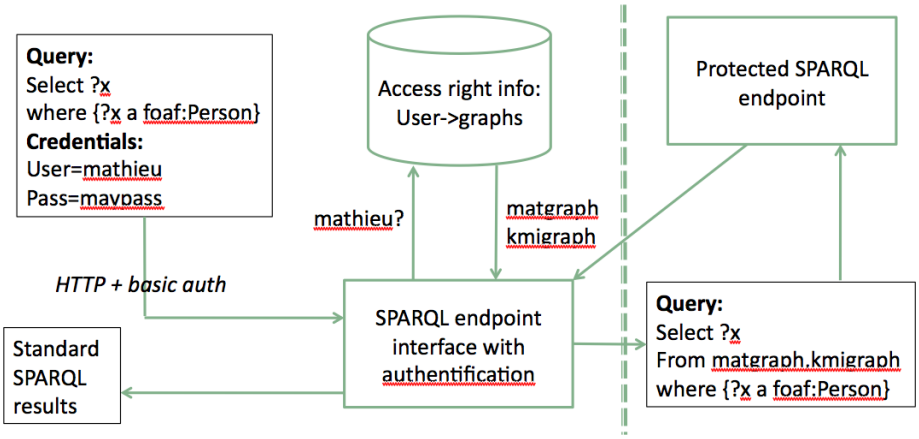


Fig. 3. Overview of the mechanism for access right to data in a SPARQL endpoint.

The idea, as depicted in Figure 3, is that the actual SPARQL endpoint giving access to all the data for all the users is being hidden using standard security measures so that it can only be accessed by our own system. We then implement a "proxy SPARQL endpoint" that can handle basic HTTP authentication. When receiving a query, this proxy endpoint will check the credential of the user and see what sub-graphs the user has access to, so that it can modify the query to restrict it to these sub-graphs only (using the FROM clause in SPARQL). It can then send the query to the real, hidden SPARQL endpoint and forward the results back to the user.

While this mechanism is relatively simple, it offers an appropriate level of flexibility, allowing to define arbitrary subgraphs and user definitions as a model for access control.

5 Interpreting and Analysing Activity Data through Lightweight Ontology Reasoning

Here, we want to use the ontologies we have created, and extend them, so that they can support the interpretation and analysis of the extracted activity data. What we want to achieve is, through providing ontological definitions of different types of activities and resources, to be able to characterise different types of traces and classify them as evidences of particular activities happening.

The first step in realising such inferences is to characterise the resources over which activities are realised – in our case, websites and webpages. Our ontologies define a webpage as a document that can be part of a webpage collection, and a website as a particular type of webpage collection. As part of setting up the UCIAD platform, we declare in the RDF model the different collections and websites that are present on the considered server, as well as the URL patterns that make it possible to recognise webpages as parts of these websites and collections. These URL patterns are expressed as regular expressions and an automatic process is applied to declare triples of the form $page_1\ isPartOf\ website_1$ or $page_2\ isPartOf\ collection_1$ when the URLs of $page_1$ and $page_2$ match the patterns of $website_1$ and $collection_1$ respectively.

The base ontologies we have defined can then be extended to represent particular categories of resources, depending on their properties. We for example declare a particular website as a $Wiki$. We can also declare a webpage collection that corresponds to RSS feeds, using a particular URL pattern, and use an ontology expression to declare the class of $WikiUpdate$ as the set of webpages which are both part of a $Wiki$ and part of the $RSSFeed$ collection, i.e., in the OWL abstract syntax

```
Class(WikiUpdateFeed complete
    intersectionOf(Webpage
      restriction(isPartOf someValuesFrom(RSSFeed))
      restriction(isPartOf someValuesFrom(Wiki))))
```

We can similarly define the activity of checking and federating updates from a wiki by creating the class of traces of activities (requests) realised on a $WikiUpdateFeed$ using an $RSSClient$ as user agent. Another example would be defining the class $ExecutingASPARQLQuery$ as the one of sending a request to a page of the type $SPARQLEndpoint$ using a *query* parameter.

Such definitions can be added to the repository, which, using its inference capability, will derive that certain pages are $WikiUpdateFeed$s, and certain activities correspond to $ExecutingASPARQLQuery$ without this information being directly provided in the data, or the rule to derive it being hard-coded in the system. We can therefore engage in an incremental construction of an ontology characterising websites and activities generally, in the context of a particular system, or in the context of a particular user.

6 Implementation: the UCIAD Platform

We realised the UCIAD platform as a demonstrator, where a user can register to the platform with some setting details and browse his or her activity data as they appear on several Open University websites (mostly, an internal wiki system and the Open University's linked data platform – data.open.ac.uk).[8]

The current "in development" version of the platform implements and demonstrates the following components described above:

User management: As the user registers into the UCIAD platform, his current setting is automatically detected, and other settings (other browsers) that are likely to be associated to him or her are also included. As the user registers, the settings are associated to his account and the activity data realised through these settings are extracted.

Extracting user-centric activity data: As described in Section 4.2, the settings associated with the user are used to extract the activity data around this particular user, creating a sub-graph corresponding to his or her activity.

Ontologies to make sense of activity data: The ontologies are used in structuring the data according to a common schema and to provide a base to homogeneously query data coming from different systems. As discussed above, they can also be extended (specified) so that different categories of activities and resources can be represented, and reasoned upon.

Ontological reasoning for analysis: Activity data is organised according to different categories (traces, webpages, websites, settings, etc.) coming from the base ontologies, but also according to classes of activities, resources, etc. that have been specially added to cover the websites and the particular user in this case (see Section 5). Here, we extended the ontologies in order to include definitions of activities relevant to the use of a wiki and a data platform. For example, we define "Executing a SPARQL Query" as an activity that takes place on a SPARQL endpoint with a "query" parameter, or "Checking Wiki Updates" as an activity on a Wiki page that is realised through an RSS client.

Browsing data according to ontologies: We rely on an homemade "browser" that we use in a number of projects and that can inspect ontology classes and members of these classes, generating graphs and simple statistics for these classes and members.

7 Discussion and Future Work

While the UCIAD platform provides an interesting first attempt at demonstrating the feasibility of user-centric activity data based on semantic technologies, a number of challenges are left to be considered before such technologies could be deployed in realistic settings to provide Web users with an appropriate view on

[8] see http://uciad.info/ub/2011/08/final-post-putting-things-together-with-a-demo/ for a description and a video of this demonstrator.

their own activity data.

The first, technical challenge is scalability. Indeed, triple stores such as OWLIM can now handle very large amounts of data (see the benchmark tests in [3, 4]). However, activity data in the form of traces from logs are enormous. Indeed, an average Web server from the Open University would serve a few million requests per month. Each request (summarised in one line in the logs) is associated with a number of different pieces of information that re-factor in terms of our ontologies, concerning the actor (IP, agent), the resource (URL, website it is attached to, server), the response (code, size) and other elements (time, referrer). We can obtain between 20 and 50 triples per request. This leads us to amounts of data in the order of 100 million triples per month per server (each server can host many websites). In theory, OWLIM should cope with such a scale, even if we consider several servers over several months. However, the data we are uploading to OWLIM is complex, and has a refined structure. Some objects (user settings, URLs) would appear very connected, while others would only appear in one request, and share only a few connections. From our experience, it is not only the number of triples that should be considered, but also the number of objects. A graph where each object is only associated with 1 other object through 1 triple might be a lot more difficult to process than one with as many triples, but shared amongst significantly less nodes (see [5]).

This scale issue is amplified when inference mechanisms are applied. OWLIM handles inferences at loading times. This means that not only the number of triples uploaded onto the store are multiplied through inferences, but also that immensely more resources are required at the time of loading these triples, depending not only on the size of what is uploaded, but also on its complexity (and, as mentioned above, our data is complex) and on the size of what is already stored. Originally, our approach was to have one store holding everything with inferences, and to extract from this store data for each user. We changed this approach to one were the store that keeps the entire dataset extracted from logs does not make use of inference mechanisms. Data extracted for each user are then transferred into another (necessarily smaller) store for which inferences apply.

A less technical challenge for approaches to activity data relying on a user-centric perspective is the identification of user-related data and their distribution. Indeed, as we explained in Section 4, we identify users based on a number of indicators detected at the time the user is registering and logging in. These indicators are far from being 100% accurate. Other types of systems can cope with inaccuracy as they are generally eliminated or reduced when the data is being aggregated. However, here, providing activity data to the wrong user could create critical privacy issues that need to be considered. More robust security mechanisms, as well as more accurate user identification mechanisms (using for example the cookies employed by Web tracking systems) would need to be deployed.

Another crucial element concerns the distribution of the data. One of the important aspects of user-centric data is that the user should be able to export his or her own data, in order to exploit them for their own benefit. The ownership of the data is not however very clear in this case. It is data collected and delivered by our systems, but that are produced out of the activities of the user. We believe that in this case, a particular type of license is needed, which would give control to the user on the distribution of their own data, but without opening it completely.

References

1. d'Aquin, M., Rowe, M., Motta, E.: Self-tracking on the web: Why and how. In: W3C Workshop on Web Tracking and User Privacy. (2011)
2. d'Aquin, M., Elahi, S., Motta, E.: Personal monitoring of web information exchange: Towards web lifelogging. In: Web Science 2010, WebSci10: Extending the Frontiers of Society On-Line, poster. (2010)
3. Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. Journal of Web Semantics **3**(2) (2005)
4. Bizer, C., Schultz, A.: The berlin sparql benchmark. International Journal on Semantic Web and Information Systems **5**(2) (2009)
5. Duan, S., Kementsietsidis, A., Srinivas, K., Udrea, O.: Apples and oranges: A comparison of RDF benchmarks and real RDF datasets. In: SIGMOD. (2011)

Social Network Aggregation Using Face-Recognition

Patrick Minder and Abraham Bernstein

University of Zurich, Dynamic and Distributed Informations Systems Group
{minder,bernstein}@ifi.uzh.ch

Abstract. With the rapid growth of the social web an increasing number of people started to replicate their off-line preferences and lives in an on-line environment. Consequently, the social web provides an enormous source for social network data, which can be used in both commercial and research applications. However, people often take part in multiple social network sites and, generally, they share only a selected amount of data to the audience of a specific platform. Consequently, the interlinkage of social graphs from different sources getting increasingly important for applications such as social network analysis, personalization, or recommender systems. This paper proposes a novel method to enhance available user re-identification systems for social network data aggregation based on face-recognition algorithms. Furthermore, the method is combined with traditional text-based approaches in order to attempt a counter-balancing of the weaknesses of both methods. Using two samples of real-world social networks (with 1610 and 1690 identities each) we show that even though a pure face-recognition based method gets outperformed by the traditional text-based method (area under the ROC curve 0.986 vs. 0.938) the combined method significantly outperforms both of these (0.998, $p = 0.0001$) suggesting that the face-based method indeed carries complimentary information to raw text attributes.

1 Introduction

With the rapid growth of the social web an increasing number of people started to replicate their off-line preferences and lives in an on-line environment. Indeed, the usage of social network sites (SNS) such as Facebook, Google+, or LinkedIN the use of messaging services (e.g., Twitter), tagging systems (e.g., del.ico.us), sharing and recommendation services (e.g., Last.fm) has not only increased immensely, but the activities on these site become an integral element in the daily lives of millions of people. Hence, the social web provides an enormous source for social network data collection.

Often people take part in multiple of these SNSs. In some cases this multi-participation arises from necessity, as some features may only be provided by some sites and not by others. However, in most cases, it is also the result of free choice. The many services allow people to "partition" their lives (e.g, they may

use facebook for the private- and LinkedIN for the professional network). In fact, the construction of site-specific identities enables the possibility to gain multiple personalities as identifying features, such as the email address can be changed easily—an effect that has been called "multiplicity" by Internet researchers [21]. Hence, users will continue to maintain multiple identities even if one service will cater to all their needs.

At the same time, the identification of users for interlinking data from different and distributed systems is getting increasingly important for different kind of applications. In personalization, the use of cross-site profiles is essential as the incorporation of multi-source user profile data significantly increases the quality of preference recommendations [4]; In social network analysis, the merging of multiple networks provides a more complete picture of the overall social graph and helps to minimize the data selection bias on which most single-site studies suffer [1]; and trust networks can be created by aggregating relationships among network participants [17]. Even if the semantic web were to become immensely popular the increased usage of a global identifier may not simplify universal identification of a person, as some sites may not use the same identifiers or even totally ignore the identification scheme and the users may choose—to ensure their multiplicity—to maintain multiple identifiers. In fact, Mika et al. [16] argue that the key problem in the area of extraction of social network data—the disambiguation of identities and relationships—still remains, as different social web applications refer to relationship types, attributes, or tastes in profiles in different ways and do not share any common key for the identification of users. As a consequence, both researchers and practitioners (such as marketers) are placed in front of a complicated research question: *how can we combine the multitude of information available about a person in the multiple SNSs to develop a holistic, combined (and as complete as possible) user model when the identity of the user in different sites is difficult to combine?*

Current proposals for interlinking social network profiles based on comparing text-based attributes of user profiles [4] or using the network structure [13] have the drawback that these methods scale poorly or they need to contain some overlap in the relationship structure and result in a large computational expenditure respectively. In this paper we propose to enhance current text-based methods—in absence of semantic metadata — by combining it with face recognition algorithms. Specifically, we propose to use face-recognition software to compare the images uploaded by users on different SNSs as an additional feature for identity merging. As we show, this statistical entity resolution procedure significantly enhances the merging precision of two SNSs. Consequently, *the contribution of this paper are: (1)The presentation of an enhanced identity merging framework to incorporate images; (2) The presentation of an algorithm that merges identities based on face recognition software. (3) The combination of traditional text-based and the introduced image-based merge-approach to counter-balance the respective weaknesses of each of the approaches.*

To this end, we first ground our idea by giving an overview of related work and introducing the fundamental concepts of entity resolution (i.e. re-identification)

and face-recognition. Then we present our novel re-identification technique and discuss our prototype. Finally, we evaluate our procedure empirically on three real-world datasets and close with a discussion of the limitations, future work and some general conclusions.

2 Related Work

Winkler [26], showed that with a minimal set of attributes a large portion of the US population can be re-identified based on US Census data. Furthermore, Gross et al. [10] showed that about 80% of social network sites user provide enough public data for a direct re-identification and that at least 61% of the published profile images on Facebook.com allow a direct identification by a human.

Carmagnola et al.[4] and Bekkermann et al. [2] provide a cross-system identity discovery system, which is based on text-based identification probability calculations, whereby public available textual attributes of social network sites are analyzed by their positive, respectively negative, influence on identification. Further, [3] suggest the use of key phrase extraction for the name disambiguation process, which is also used in POLYPHONET [14] for interlinking web pages

[13] and [22] provide re-identification algorithms based on network similarity. These system provide high accuracy, but lack on computational complexity and time expenditure.

A lot of research concerns shared approaches [12]: Especially, the application of common semantic languages , such as the FOAF ontology[1], the SIOC (Semantically-Interlinked Online Communities) ontology[2] for online communities or the SCOT (Social Semantic Cloud Of Tags) ontology[3] for tagging systems. Such systems are desirable, but not widely spread in reality. The most well-known system based on such data is FLINK [15].

3 Theoretical Foundations

In this section, we present the theoretical foundations for our approach. First, we present a formal model for entity resolution and then succinctly explain the basics of face-recognition. Both foundations are used in our framework.

3.1 Entity Resolution and the Fellegi-Sunter Model

Entity resolution can be defined as *the methodology of merging corresponding records from two or more sources* [26]. Consider for example a profile about "Peter J. Miller" and another one about "Peter Jonathan Miller" on two different SNS. Entity Resolution tries to decide if these two profiles belong to the same

[1] http://www.foaf-project.org / http://xmlns.com/foaf/spec/20100101.html
[2] http://sioc-project.org/
[3] http://scot-project.org/

person or not. Therefore, entity resolution assumes that an individual shares similar features in different environments which can be used to identify an entity, even though no common key is defined. Generally, to complicate the resolution process, there are different entities that share similar attribute values.

Most current re-identification approaches are variants of the Fellegi-Sunter model—a distance- and rule-based technique. The Fellegi-Sunter Model determines a match between two entities by computing the similarity of their attribute (or feature) vectors [9]. Specifically, given entities $a \in \mathbb{A}$ and $b \in \mathbb{B}$, where both $\mathbb{A}$ and $\mathbb{B}$ are the set of entities in SNS A and B, it tries to assign each pair (a, b) of the space $\mathbb{A} \times \mathbb{B}$ to a set $\mathbb{M}$ or $\mathbb{U}$ whereby:

$$\mathbb{M} := \text{is the set of true matches} = \{(a, b); a \in \mathbb{A} \wedge b \in \mathbb{B} \wedge a = b\}$$
$$\mathbb{U} := \text{is the set of non-matches} = \{(a, b); a \in \mathbb{A} \wedge b \in \mathbb{B} \wedge a \neq b\}$$

It does so using a comparison function γ that computes the similarity measures for each of the n comparable attributes of the entities and arranges these in a vector:

$$\gamma(a, b) = \{\gamma^1(a, b), ..., \gamma^n(a, b)\}$$

Based on the comparison vector $\gamma(a, b)$ a decision rule L now assigns each pair (a, b) to either to the set M or U as follows:

$$(a, b) \in \begin{cases} M & \text{if p(M } |\gamma) \geq \text{p(U } |\gamma) \\ U & \text{otherwise} \end{cases}$$

whereby $p(M|\gamma)$ is the probability that the comparison vector γ belongs to the match class and $p(U|\gamma)$ that γ belongs to U. In other words, the Fellegi-Sunter model treats all pairs of possible matches as independent. Recently several authors argued that this independence offers the opportunity for enhancements. Singla et al [18], e.g., proposes such an enhancement based on Markov logic.

3.2 Face-Recognition and the Eigenface Algorithm

The face provides an enormous set of characteristics that the human perception system uses to identify other individuals. The problem of face-recognition can be formulated as follows *"Given still or video images of a scene, identify or verify one or more person in the scene using a stored database of faces. Available collateral information [...] may be used in narrowing the search (enhancing recognition)"* [25, p. 4]. Accordingly, face-recognition includes [25]: *(1)* The detection and location of an unknown number of faces in an image [11]; *(2)* The extraction of key facial-features; and *(3)* The identification [25, p. 12] which includes a comparison and matching of invariant biometric face signatures [25, p. 14 - 16]. The identification can either be done by using *holistic matching*, *feature-based matching*, or *hybrid matching methods* which concern the whole face, local features— e.g. the location or geometry of the nose —or both as an input vector for classification respectively [25, p. 14].

Our re-identification framework uses the holistic face-recognition algorithm *Eigenface* [20] based on Principal Component Analysis (PCA) and covering all relevant local and global features [20]. The Eigenface approach tries to code all the relevant extracted information of a face image, such that the encoding can be done efficiently, allowing for a comparison of the information to a database of encoded models [25, p. 67].The Eigenface algorithm can be split up into two parts:

(1) Representation of the Image Database in Principal Component Vectors Based on PCA, the principal components of a face-image are extracted, by (1) acquiring an initial set of face images; (2) Defining the face space by calculating the eigenvectors (Eigenfaces) from the set and eliminating all but k best eigenvectors with the highest eigenvalues, by using PCA; and (3) Presenting each known individual by projecting their face image onto the face space.

Therefore, an image $I(x, y)$ can be interpreted as a vector in a N-dimensional space, where $N = rc$ and r are the rows and c columns of the image [20] . Every coordinate in the N-dimensional vector $I(x, y)$— the *image space* —corresponds to a pixel of the image. This representation of an image obfuscates any relationship between neighboured pixels as long as all images are rearranged in the same manner. Thus the average face of the initially acquired training set $\Gamma := \{\gamma_1, \gamma_2, ..., \gamma_m\}$ can be calculated by

$$\bar{\gamma} = \frac{1}{m} \sum_{n=1}^{m} \gamma_n.$$

and the distance between an image and the average image is measured by $\phi_i = \gamma_i - \bar{\gamma}$. Whereby, the orthonormal vectors define an Eigenface with the eigenvectors:

$$u_l = \sum_{k=1}^{M} e_{lk} \phi_k \forall i \in [1, M]$$

whereby the eigenvectors e_l are calculated from the covariance matrix $L = AA^\top$, where $L_{mn} = {\phi_m}^\top \phi_n$ and $A = [\phi_1, \phi_2, ..., \phi_M]$. The derivation of the best eigenvectors out of the covariance matrix is presented in [19]. The k significant eigenvectors of L span an k-dimensional face space—a subspace of the $N \times N$ dimensional image space—where every face is represented as a linear combination of the Eigenfaces [20] [25, p. 67 - 72].

(2) The Identification Process The identification respectively verification of an image is processed by: (1)Subtracting the mean image from the new face images and projecting the result onto each of the eigenvectors (Eigenfaces); (2) Determining if the image is a face by calculating the distance to the face space and comparing it to a defined threshold; and (3)If it is a face, classifying the weight pattern as a known or unknown individual by using a distance metric, such as the Euclidian distance.

Thus, a new face image $I(x, y)$ will be projected into the face space by $\omega_k = u_k^\top (\gamma - \bar{\gamma})$ for $\forall k = [1, ..., M']$. The weight matrix $\Omega^\top = [\omega_1, ..., \omega_{M'}]$ represents the influence of each eigenvector on the input image. Hence, given a threshold θ_ε, if the face class k, which minimizes the Euclidian distance is

$$\varepsilon_k = \| (\Omega - \Omega_k) \| \text{ and } \theta_\varepsilon > \varepsilon_k \quad (1)$$

then the image will belong to the same individual. Else the face is classified as unknown. Furthermore, the distance between an image and the face space can be characterised by the squared distance between the mean-adjusted input image:

$$\varepsilon^2 = \| (\phi - \phi_f) \| \text{ , where } \phi = \gamma_k - \bar{\gamma} \text{ and } \phi_f = \sum_{i=1}^{M'} \omega_i u_i \quad (2)$$

Therefore, a new face image $I(x, y)$ will be calculated as a non-face image if $\varepsilon > \theta_\varepsilon$, as known face image if $\varepsilon < \theta_\varepsilon \wedge \varepsilon_k < \theta_\varepsilon$ and as an unknown face image if $\varepsilon < \theta_\varepsilon \wedge \varepsilon_k > \theta_\varepsilon$.

4 Re-Identification Framework

Our theoretical re-identification framework for user disambiguation in a social network aggregation and cross-system personalization process. is based on the Fellegi-Sunter-Model. The presented algorithms calculate the probability that two user profiles belong to the same entity, and incorporates the ability to incorporate images as an additional feature based on the Eigenface method. Therefore, the framework provides three kind of methods: a pure face-recognition based, a text-attribute based, and joined re-identification method.

The methods follow a simple re-identification algorithm. Assume, two sets $\mathbb{A} = \{a_1, a_2, ..., a_m\}$ and $\mathbb{B} = \{b_1, b_2, ..., b_n\}$ of user profiles from two different SNSs. Each profile is characterized by a set of text attributes and a single profile image. We can now define $\mathbb{E} = \{e_1, e_2, ..., e_z\}$ as the set of different individuals, who have a profile in one or both social networks. Consequently, the re-identification algorithm is based on the following three subtasks:

1. *Attribute Comparison:* The attributes of two social network profiles are compared pairwise. The result is a comparison vector $\gamma(a_i, b_j) = \{d_1, d_2, ..., d_n\}$, where n is the number of compared attributes and $d_k \in [0, 1]$ indicates the distance between the values of the k^{th}-attribute of the profiles a_i and b_j. Therefore, a distance d_k of 0 indicates, that the two attribute instances are completely equal, and a value of 1 indicates the opposite.
2. *Matching Probability Calculation:* Then, based on the comparison vector $\gamma(a_i, b_j)$, the probability $\rho(a_i, b_j)$, that a pair (a_i, b_j) belongs to the same entity, is calculated.
3. *Merging Task:* Finally, if probability $\rho(a_i, b_j)$ is greater or equal to a threshold value $\theta \in [0, 1]$ (*i.e.*, $\theta \geq \rho(a_i, b_j)$) then the profiles a_i and b_j are assumed to belong to the same person.

4.1 Attribute Comparison and Matching Probability Calculation

The following three generic methods allow the comparison of n different attributes and the calculation of a matching probability. The methods cover the

first two subtasks of the above introduced re-identification algorithm.

(1) Pure Face-Recognition Based Method The method re-identifies user profiles only by the application of the face-recognition algorithm *Eigenface* on profile images. Hence, $\forall a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$, the probability $\rho(a_i, b_j)$, that two profiles a_i and b_j belong to the same entity $e_l \in E$, is defined as:

$$\rho(a_i, b_j) = \varepsilon_{ij}(a_i, b_j) = \| \ (\Omega_{a_i} - \Omega_{b_j}) \ \| \in [0, 1]$$

Whereas, it is assumed that the profile images are projected into the face space by $\omega_{a_i} = u_k^\top (a_i - \bar{\gamma})$ and $\omega_{b_j} = u_k^\top (b_j - \bar{\gamma})$. Additionally, the set $\mathbb{B}$ is used as training set for the initialization task, thus $\Gamma = \mathbb{B}$.

(2) Text-Attribute Based Method The algorithm re-identifies user profiles by the application of text-attribute comparison. The attributes are compared with the token-based *QGRAM* algorithm [7]. Note that spelling errors minimally affects the similarity when using *QGRAM*, as it uses q-grams instead of words are used as tokens. For the k^{th}-attribute the algorithm computes a normalized distance $d(a_{ik}, b_{jk}) \in [0, 1]$, where the distance is zero, if the value of the k^{th}-attribute of a_i and b_j are syntactically equivalent. As we discuss in Section 6, we considered *name, email address, birthday, city* as a minimal set of text attributes in the experiments as they where shown to be strong indicators for identification [5] [26] [10] and other attributes such as address or phone number are often not accessible. As a result, the matching probability is calculated by a logistic function [8]:

$$\rho(a_i, b_j) = \frac{\exp(Y_T(a_i, b_j))}{1 + \exp(Y_T(a_i, b_j))} \in [0, 1]$$

where

$$Y_T(a_i, b_j) = \alpha_0 + \sum_{k=1}^{n} \alpha_k d(a_{ik}, b_{jk})$$

The intercept α_0 and regression coefficients $\{\alpha_1, ..., \alpha_n\}$ for the linear regression model $Y_T(a_i, b_j)$ are learned by a logistic regression on a specific training set.

(3) Joined Method Finally, the two previously described methods are joined to a method that uses both face-image-based and text-attribute-based identification. Thus, for all pairs of profiles $a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$, it is assumed that the matching probability is equal to:

$$\rho(a_i, b_j) = \frac{\exp(Y_J(a_i, b_j))}{1 + \exp(Y_J(a_i, b_j))} \in [0, 1]$$

where

$$Y_J(a_i, b_j) = \alpha_0 + \sum_{k=1}^{n} \alpha_k d(a_{ik}, b_{jk}) + \beta \varepsilon_{ij}(a_i, b_j)$$

Again, the intercept α_0 and regression coefficients $\{\alpha_1, ..., \alpha_n, \beta\}$ for the linear regression model $Y_J(a_i, b_j)$ are learned by a logistic regression on a specific training set.

4.2 Merging Task

Finally, based on one of the above introduced matching probabilities, a pair (a_i, b_j) is called to belong to the same entity (*i.e.*, $(a_i, b_j) \in \mathbb{M}$), if:

$$\forall a_i \in A \wedge b_j \in B : \theta \geq \rho(a_i, b_j) \longrightarrow \mathbb{M} \tag{3}$$

5 Prototype

Our re-identification framework consists of four major components. Currently, the *Data Gathering and Acquisition* module enables the acquisition of network data from the social network sites Facebook, LinkedIn, Twitter and Flickr, whereby only concerns public available data. The *Data Preprocessing* module preprocesses the crawled data by transforming profile attributes into an internal schema and establish connections between profiles for each relationship in the source network. The implementation provides functionality for both the integration of text attributes and profile images. For the integration of profile images, we use an implementation of the face detection algorithm *OpenCV*[4] *HaarClasifier*[23] provided by the Faint[5] open source project. The algorithm returns the coordinates of every face region on an input image, whereby one region of the n returned regions is randomly selected and resized to a 50×50-pixel image. The *Matching* module performs a pairwise comparison of all possible profiles pairs (a_i, b_j), where $a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$. The goal of the matching task is to calculate the comparison vector $\gamma(a_i, b_j)$ and matching probability $\rho(a_i, b_j)$ for each of the methods introduced in Section 4.1. The module uses text-based algorithm QGRAM provided by the open-source project SimMetrics[6], and our own implementation of the Eigenface algorithm. Finally, The *Merging* module merges the data sources to an aggregated social graph based on rule introduced in Section 4.2.

6 Experiments

We evaluated the accuracy of the framework based on two experiments. In the first experiment we determined various input parameters, the intercept and the coefficients for the two regression models. The second experiment benchmarked the suitability of profile images for user disambiguation in the pure face-recognition and joined method against the text-based matching algorithm.

6.1 Experiment 1: Determining the Parameters

In the first experiment two social networks with a size of 47 and 45 were generated from data crawled on Facebook. 36 of these users had a profile in both

[4] http://sourceforge.net/projects/opencvlibrary/
[5] http://faint.sourceforge.net/
[6] http://www.sourceforge.net/projects/simmetrics/

networks. The profile image was randomly selected from all public available published images in the specific Facebook profile. We performed a pairwise comparison of the two sets, whereas for each pair the attribute similarities were stored as a quintuple $[name, emailaddress, birthday, city, image_{similarity}]$ whilst varying the number of Eigenfaces in the $image_{similarity}$ computation. Finally, the optimal number of Eigenfaces and parameters for the two linear models were calculated using a logistic regression model in SPSS[7].

Performance metric The profile image similarity measurements based on Eigenfaces were compared using Receiver Operating Characteristics (ROC) curves. The ROC-curve graphs the true positive rate (y-axis) respectively sensitivity against the false positive rate (x-axis) respectively 1 - Specificity, where an ideal curve would go from the origin (0,0) to the top left (0,1) corner, before proceeding to the top right (1,1) one [24, p. 244 - 225]. The area under the ROC-curve (AUC, also called c-statistic in medicine) can be used as a single number performance metric for the merge accuracy. In contrast to the traditional precision, recall, or f-measure it has the advantage that both the ROC-curve and the AUC are independent of the prior data-distribution and, hence, serve as a more robust metric to compare the performance of two approaches.

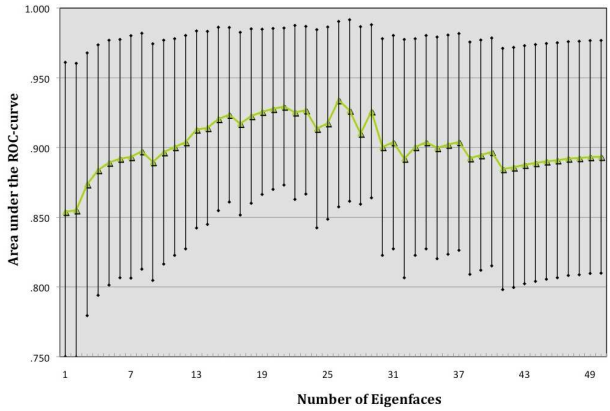


Fig. 1: Showing the influence of the number of Eigenfaces on the area under the ROC-Curve based on data of the first experiment and a confidence interval of 95%

Results As illustrated in Figure 1, the number of Eigenfaces influences on the accuracy of match. The accuracy of the algorithm increases when increasing the number of Eigenfaces until a specific barrier, where any increase in its numbers is not beneficial or even detrimental to the overall performance. Thus, the Eigenface algorithm should use between 50 to 60% of the top-most Eigenfaces—a result similar to [24]. The resulting input parameters for the linear models are shown in Table 1.

Computational Costs The computational costs for the face-image comparison is higher than for single text-based comparison. On our test-machine (an

[7] http://www.spss.com/

Apple iMac computer with a 3.06 GHz Intel Core 2 Duo processor and 4 GB of RAM) the comparison of the four concerned text-attributes takes between 10 to 20ms per pair without data preprocessing; the image-based comparison alone takes 25 to 35ms/pair. Additionally, once per image, the face preprocessing, including face-detection and image resizing, takes between five and six seconds.

Attribute	α_0	α_{Name}	α_{Email}	$\alpha_{Birthday}$	α_{City}	β
Text-Based Method Y_T	-0.319	25.655	-1.763	9.750	25.334	-
Joined Method Y_J	-6.659	26.656	0.234	11.536	18.272	8.788

Table 1: Input parameter for the regression based text-based and joined method models learned on the dataset of the first experiment and used in the second experiment as input.

6.2 Experiment 2

For the second experiment we collected a subgraph of both Facebook and LinkedIn. Departing from the first author's profile we collected 1610 (Facebook) respectively 1690 (LinkedIn) profiles and manually determined that 166 users where present in both samples. We compared all these profiles with the three approaches using the input parameters determined in Experiment 1. **Results** Figure 2 graphs the ROC curves for the three methods. Note that whilst the text method (AUC=0.986) outperforms the pure image-based method (AUC=0.938), the combined method (AUC=0.998) significantly outperforms either methods ($p = 0.001$, $p = 0.0001$ compared with a non-parametric method described by DeLong [6]).

6.3 Discussion, Limitations and Future Work

As the above results show the combined method clearly outperforms each of others. It is interesting to observe that the ROC-Curve of both text-based and the image-based method both shoot almost straight up until about (0,0.9). Then the text-based method flattens out whilst the combined one continues to rise. This suggests that the element of the method's accuracy is contributed mostly by the image-based method. Only then does the image-based method contribute additional predictive power. When looking at the regression parameters this suggestion receives some additional support as the parameters for the *Email* and *City* lose in their contribution whilst the algorithm relies more on the $Name$, $Image$, and interestingly the $Birthday$.

Obviously, all these results are limited by the usage of only one, albeit real-world, dataset and will have to be validated with a others. Also, our experiment assumed that we knew the semantic alignment of the text-attributes. When merging only two SNS this assumption seems reasonable, when more are involved the this alignment may introduce additional error. Consequently, we probably overestimated the accuracy of the textual method.

Last but not least, a real-world system would probably not perform a full pairwise comparison to limit the computational expenditure but use some optimization approach.

We intend to investigate all these limitations in our future work.

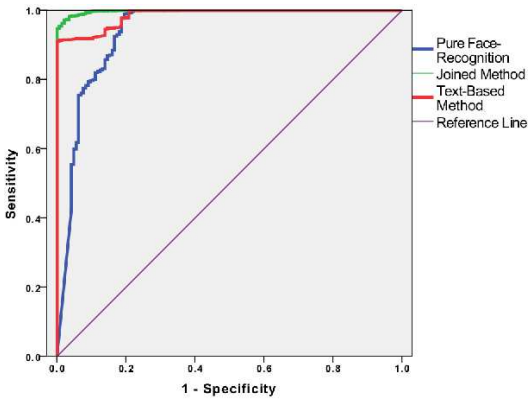


Fig. 2: Results of the second experiment merging two subnetworks of Facebook and LinkedIn

7 Discussion and Conclusion

In this paper we proposed an extension of the traditional text-attribute-based method for re-identification in social networks using the images of profiles. The experimental results show that the pure face-recognition based re-identification method does not compete the traditional text-based methods in accuracy and computational performance. A combined method, however, significantly outperforms the pure text-based method in accuracy suggesting that it contains complementary information. As we showed this combined method significantly improves the accuracy of a social network system merge. Consequently, we believe that it provides a more solid basis for both researchers and practitioners interested in investigating multiple SNSs and facing the problems of multiplicity.

References

1. Bachmann, A., Bird, C., Rahman, F., Devanbu, P., Bernstein, A.: The missing links: Bugs and bug-fix commits. In: ACM SIGSOFT / FSE '10: Proceedings of the 18th International Symposium on the Foundations of Software Engineering (2010)
2. Bekkerman, R., McCallum, A.: Disambiguating web appearnaces of people in a social network. In: Procceding of the WWW 2005 (2005)
3. Bollegara, D., Matsuo, Y., Ishizuka, M.: Extracting key phrases to disambiguate personal names on the web. In: Proceeding to CICling 2006 (2006)
4. Carmagnola, F., Cena, F.: User identification for cross-system personalisation. Information Sciences: an International Journal 1-2(179), 16–32 (2009)
5. Carmagnola, F., Osborne, F., Torre, I.: User data distributed on the social web: how to identify users on different social systems and collecting data about them. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (2010)
6. DeLong, E., DeLong, D., Clarke-Pearson, D.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44(44), 837 – 845 (1988)

7. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1) (2007)
8. Fahrmeir, L., Pigeot, I., Tutz, G.: In: Statistik - Der Weg zur Datenanalyse. Springer-Verlag Berlin Heidelberg New York (2003)
9. Fellegi, I., Sunter, A.: A theory for record linkage. Journal American Statistic Association (64), 1183 – 1210 (1969)
10. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Workshop On Privacy In The Electronic Society, Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society. pp. 71 – 80 (2005)
11. H Demirel, TJ Clarke, P.C.: Adaptive automatic facial feature segmentation. Proc. of 2nd International Conference on Automatic Face and Gesture Recognition pp. 277 – 282 (196)
12. Leonard, E., Houben, G.J., van der Sluijs, K., Hidders, J., Herder, E., Abel, F., Krause, D., Heckmann, D.: User profile elicitation and conversion in a mashup environment. In: Int. Workshop on Lightweight Integration on the Web, in conjunction with ICWE 2009 (2009)
13. Malin, B.: Unsupervised Name Disambiguation via Social Network Similarity (2006)
14. Matsuo, Y., Mori, J., Hamasaki, M., Ishizuka, M.: Polyphonet: An advanced social network extraction system. In: Procceding of 15th International World Wide Web Conference (2006)
15. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics 3(211 - 223) (Jan 2005)
16. Mika, P., Gangemi, A.: Descriptions of social relations. In: Proceedings of the First Workshop on Friend of a Friend, Social Network and the Semantic Web (2004)
17. Rowe, M.: Interlinking distributed social graphs. In: Proc. Linked Data on the Web Workshop, 18th Int. World Wide Web Conference (2009)
18. Singla, P., Domingos, P.: Entity resolution with markov logic. In: ICDM Sixth International Conference on Data Mining 2006 (2006)
19. Sirovich, L., Kirby, M.: Application of the karhunen-loève procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence . . . 12, 103 – 108 (1990)
20. Turk, M., Pentland, A.: Face recognition using eigenfaces. Conference on Computer Vision and Pattern Recognition (Jan 1991)
21. Turkle, S.: Cyberspace and identity. Contemporary Sociology 28(6), 643 – 648 (1999)
22. Veldman, I.: Matching Profiles from Social Network Sites. Master's thesis, University Twente (2009)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recogntiion (2001)
24. Wechsler, H.: Reliable Face Recognition Methods - System, Design, Implementation and Evaluation. Springer Media LLC (2"6)
25. Wenyi Zhao, R.C.: Face Processing - Advanced Modeling and Methods. Academic Press (2006)
26. Winkler, W.: The state of record linkage and current research problems. Tech. rep., Statistical Research Division, U.S. Census Bureau. (1999)

Towards Mining Semantic Maturity in Social Bookmarking Systems

Martin Atzmueller[1], Dominik Benz[1], Andreas Hotho[2], and Gerd Stumme[1]

[1] Knowledge & Data Engineering Group, University of Kassel,
34121 Kassel, Germany
{atzmueller,benz,stumme}@cs.uni-kassel.de
[2] Data Mining & Information Retrieval Group, University of Würzburg,
97074 Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

Abstract. The existence of emergent semantics within social metadata (such as tags in bookmarking systems) has been proven by a large number of successful approaches making the implicit semantic structures explicit. However, much less attention has been given to the *factors* which influence the "maturing" process of these structures over time. A natural hypothesis is that tags become semantically more and more mature whenever many users use them in the same contexts. This would allow to describe a tag by a specific and informative "semantic fingerprint" in the context of tagged resoures. However, the question of assessing the quality of such fingerprints has been seldomly addressed.
In this paper, we provide a systematic approach of mining semantic maturity profiles within folksonomy-based tag properties. Our ultimate goal is to provide a characterization of "mature tags". Additionally, we consider semantic information about the tags as a gold-standard source for the characterization of the collected results. Our initial results suggest that a suitable composition of tag properties allows the identification of more mature tag subsets. The presented work has implications for a number of problems related to social tagging systems, including tag ranking, tag recommendation, and the capturing of light-weight ontologies from tagging data.

1 Introduction

Social metadata, especially collaboratively created keywords or tags, form an integral part of many social applications such as BibSonomy[3], Delicious[4], or Flickr[5]. In such social systems, many studies of the development of the tagging structure have shown the presence of *emergent semantics* (e.g., [3]) in the set of human-annotated resources. That is, the semantics of tags develop gradually depending on their usage.

[3] http://www.bibsonomy.org
[4] http://www.delicious.com
[5] http://www.flickr.com

Due to this important observation, one can regard this development as a process of "semantic maturing". The basic idea is that knowledge about a set of cooccurring tags is sufficient for determining synonyms with a certain reliability. The underlying assumption is that tags become "mature" after a certain amount of usage. This maturity will then be reflected in a stable semantic profile. Thus, tags that have arrived at this stage can be regarded as high-quality tags, concerning their encoded amount of emergent semantics.

In this paper, we utilize folksonomy-based tag properties for mining profiles indicating "matured tags", i.e., high-quality tags that can be considered to convey more precise semantics according to their usage contexts. The proposed properties consist of various structural properties of the tagging data. e.g., centrality, or frequency properties. For a semantic grounding, we analyze the applied tagging data with respect to tag-tag relations in Wordnet, for assessing the "true" semantic quality. Our contribution is thus three-fold: We provide and discuss different tag properties that are useful in determining semantic maturity profiles of tags. These are all obtained considering the network structure of folksonomies. Additionally, we obtain a detailed statistical characterization of semantic tag maturity profiles in a folksonomy dataset. Finally, we provide a list of useful indicators for identifying "mature tags" as well as synonyms in this context.

Applications of the obtained knowledge concern the construction of lightweight ontologies using tagging knowledge [18], tag recommendation [14,19], or tag ranking [16]. All of these utilize selection options and/or ranking information about sets of tags, for initial setup and refinement. Tag ranking approaches, for example, can benefit from a "maturity ranking" for filtering purposes.

The rest of the paper is structured as follows: Section 2 discusses related work. After that, Section 3 introduces basic notions of the presented approach, including folksonomy-based tag properties, and the applied pattern mining method. Then, we describe the mining methodology in detail, discuss our evaluation setting and present the obtained results. Finally, Section 5 concludes the paper with a summary and interesting directions for future research.

2 Related Work

While the phenomenon of collaborative tagging was discussed in its early stages mainly in newsgroups or mailing lists (e.g. [17]), a first systematic analysis was performed by [10]. One core finding was that the openness and uncontrolledness of these systems did not give rise to a "tag chaos", but led on the contrary to the development of stable patterns in tag proportions assigned to a given resource. [5] reported similar results and denoted the emerging patterns as *"semantic fingerprints"* of resources. [18] presented an approach to capture emergent semantics from a folksonomy by deriving lightweight ontologies. In the sequel, several methods of capturing emergent semantics in the form of (i) tag taxonomies [12], (ii) measures of semantic tag relatedness [6], (iii) tag clusterings [22] and (iv) mapping tags to concepts in existing ontologies [1] were proposed.

Most of the above works provided evidence for the *existence* of emergent tag semantics by making certain aspects of it explicit; however, the question

which *factors* influcence its development were seldomly discussed. Despite that, a common perception seemed to be that a certain amount of data is necessary for getting a "signal". Golder and Hubermann [10] gave a rough estimate that *"after the first 100 or so bookmarks"*, the proportions of tags assigned to a resource tended to stabilize. This suggested the rule "the more data, the better semantics". This assumption was partially confirmed by Körner et al. [15], who analyzed the amount of emergent semantics contained in different folksonomy partitions. More data had a beneficial effect, but the *user composition* within the partitions turned out to be crucial as well: Sub-folksonomies induced by so-called "describers", which exhibit a certain kind of tag usage pattern, proved to contain semantic structures of higher quality. Halpin [11] showed that the tag distribution at resources tends to stabilize quickly into a power-law, as a kind of "maturing" of resources. In contrast, our work targets the maturing of tags themselves.

Hovever, to the best of our knowledge none of the aforementioned works has systematically addressed the question if there exists a connection between *structural* properties of tags and the quality of semantics they encode (i.e. their "semantic maturity"). In this work, we aim to fill this gap.

3 Preliminaries

In the following sections, we first briefly present a formal folksonomy model and a folksonomy-based measure of tag relatedness. Then, we detail on the structural and statistical tag properties serving as a basis for mining maturity profiles. After that, we briefly summarize the basics of the applied pattern mining technique.

3.1 Folksonomies and Semantic Tag Relatedness

The underlying data structure of collaborative tagging systems is called *folksonomy*; according to [13], a folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, respectively. Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$. An element $y \in Y$ is called a *tag assignment* or TAS. A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$.

Folksonomies introduce various kinds of relations among their contained lexical items. A typical example are cooccurrence networks, which constitute an aggregation indicating which tags occur together. Given a folksonomy (U, T, R, Y), one can define the post-based *tag-tag cooccurrence graph* as $\mathbb{G}_{cooc} = (T, E, w)$, whose set of vertices corresponds to the set T of tags. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of posts that contain both t_1 and t_2, i.e. $w(t_1, t_2) := \mathrm{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$

For assessing the semantic relatedness between tags we apply the *resource context similarity* (cf. [6]) computed in the vector space $\mathbb{R}^R$. For a tag t, the vector $v_t \in \mathbb{R}^R$ counts how often the tag t is used for annotating a certain resource $r \in R$:

$$v_{tr} = \mathrm{card}\{u \in U \mid (u, t, r) \in Y\}\,.$$

Based on this representation, we measure vector similarity by using the cosine measure, as is customary in Information Retrieval: If two tags t_1 and t_2 are represented by $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^X$, their cosine similarity is defined as: $\text{cossim}(t_1, t_2) := \cos \measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}$. In prior work, we showed that this measure comes close to what humans perceive as semantically related [6].

3.2 Folksonomy-Based Tag Properties

For folksonomy-based tag properties, we can utilize aggregated information such as frequency, but also properties based on the network structure of the tag-tag co-occurrence graph. The properties below are based on prior work in related areas. They are abstract in that sense, that none of them considers the textual content of a tag. Therefore, all properties are language independent since the only operate on the folksonomy structure, on aggregated information, or on derived networks. Below, we describe the different folksonomy-based properties, and also discuss their intuitive role regarding the assessment of tag maturity.

Centrality Properties In network theory the centrality of a node $v \in V$ in a network G is usually an indication of how important the vertex is [20]. Because important nodes are usually well-connected within the network, one can hypothesize that this connectedness corresponds to a well-established semantic fingerprint. On the other hand, high centrality might correspond to a relatively "broad" meaning – in the context of our study, we avoid the latter by restricting ourselves to single-sense tags (see Section 4). Applied to our problem at hand, we interpret centrality as a measure of maturity, following the intuition that more mature terms are also more "important". We adopted three standard centralities (degree, closeness, betweenness). All of them can be applied to a term graph $\mathbb{G}$:

- According to *betweenness centrality* a vertex has a high centrality if it can be found on many shortest paths between other vertex pairs:

$$bet(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

  Hereby, σ_{st} denotes the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through v. As its computation is obviously very expensive, it is often approximated [4] by calculating the shortest paths only between a fraction of points.
  It seems intuitive, that tags with a high betweenness centrality are closer to important (semantic) hubs, and therefore more mature themselves. In essence, higher values should indicate semantic maturity.
- A vertex ranks higher according to *closeness centrality* the shorter its shortest path length to all other reachable nodes is:

$$clos(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)} \tag{2}$$

$d_G(v,t)$ denotes hereby the geodesic distance (shortest path) between the vertices v and t. A tag with a high closeness value is therefore more close to the core of the Folksonomy. Therefore, it seems intuitive to assume, that more central tags according to this measure should have a higher probability of being more mature.

- The *degree centrality* simply counts the number of direct neighbors $d(v)$ of a vertex v in a graph $G = (V, E)$:

$$deg(v) = \frac{d(v)}{|V| - 1} \tag{3}$$

Comparred to the other metrics, degree centrality is a local measure since it only takes into account the direct neighbourhood of a tag within the network. According to the degree, a tag could be linked to both semantically mature and non-mature tags. In this sense, it seems intuitive to assume that other factors need to be taken into account; then an estimation of the effect of the degree centrality can be considered.

Frequency Properties One first idea about tag maturity considers the fact that tags that are used more often *can* get more mature, since they can exhibit a more specific fingerprint. However, this does not guarantee maturity of tags. Therefore, we consider the frequency of a tag as a candidate for the analysis.

- We capture the *resource frequency* property *rfreq* which counts the number of resources tagged by a given tag t according to

$$rfreq(t) = \text{card}\{r : \exists(u, t', r) \in Y, t = t'\} \tag{4}$$

For the semantic assessment of tag, an intuitive hypothesis could be that the semantic profile of a tag gets more concise when more and more resource are tagged with it. However, this is not necessarily a criterion for mature tags since the development of the semantic profile could still be relatively fuzzy.

- The *user frequency* property *ufreq* counts the number of users that applied the tag t:

$$ufreq(t) = \text{card}\{u : \exists(u, t', r) \in Y, t = t'\} \tag{5}$$

Similar to the resource frequency, more users should help to focus the semantic profile of a tag due to the refinement of its usage patterns.

3.3 Pattern Mining using Subgroup Discovery

Subgroup discovery [21,2] aims at identifying interesting patterns with respect to a given target property of interest according to a specific interesting measure. In our context, the target property is given by a quality indicator for tags. The top patterns are then ranked according to the given interesting measure. Subgroup discovery is especially suited for identifying local patterns in the data, that is, *nuggets* that hold for specific subsets: It can uncover hidden relations captured in small subgroups, for which variables are only significantly correlated in these subgroups.

Formally, a database $D = (I, A)$ is given by a set of individuals I (tags) and a set of attributes A (i.e., tag properties). A *selector* or *basic pattern* $sel_{a=a_j}$ is a boolean function $I \to \{0, 1\}$ that is true, iff the value of attribute a is a_j for this individual. For a numeric attribute a_{num} selectors $sel_{a \in [min_j; max_j]}$ can be defined analogously for each interval $[min_j; max_j]$ in the domain of a_{num}. In this case, the respective boolean function is set to true, iff the value of attribute a_{num} is in the respective range.

A *subgroup description* or (complex) *pattern* $p = \{sel_1, \ldots, sel_d\}$ is then given by a set of basic patterns, which is interpreted as a conjunction, i.e., $p(I) = sel_1 \wedge \ldots \wedge sel_d$. A subgroup (extension) sg_p is now given by the set of individuals $sg_p = \{i \in I | p(i) = true\} := ext(p)$ which are covered by the subgroup description p. A subgroup discovery task can now be specified by a 5-tuple (D, C, S, Q, k). The target concept $C \colon I \to \Re$ specifies the property of interest. It is a function, that maps each instance in the dataset to a target value c. It can be binary (e.g., the quality of the tag is high or low), but can use arbitrary target values (e.g, the continuous quality of a given tag according to a certain measure). The search space 2^S is defined by set of basic patterns S. Given the dataset D and target concept c, the quality function $Q \colon 2^S \to \mathbb{R}$ maps every pattern in the search space to a real number that reflects the interestingness of a pattern. Finally, the integer k gives the number of returned patterns of this task. Thus, the result of a subgroup discovery task is the set of k subgroup descriptions $res_1, \ldots, res_k$ with the highest interestingness according to the quality function. Each of these descriptions could be reformulated as a rule $res_i \to c$.

While a huge amount of quality functions has been proposed in literature, cf. [9], many interesting measures trade-off the size $|ext(p)|$ of a subgroup and the deviation $c - c_0$, where c is the average value of the target concept in the subgroup and c_0 the average value of the target concept in the general population.

We consider the quality function *lift*, which measures just the increase of the average value of c in the subgroup compared to the general population:

$$lift(p) = \frac{c}{c_0}\,, \text{if } |ext(p)| \geq \mathcal{T}_{Supp}\,, \text{ and 0 otherwise}\,.$$

with an adequate minimal support threshold $\mathcal{T}_{Supp}$ considering the size of the subgroup. Usually, the analysis is performed using different minimal size thresholds in an explorative way. It is easy to see, that both types of quality measures are applicable for binary and continuous target concepts.

4 Mining Semantic Tag Maturity

For a given Folksonomy and its tagging dataset, we apply the following steps: Using the dataset, we construct the tag properties discussed in Section 3.2. As we will see below, the "raw" properties do not correlate sufficiently with semantic maturity. Therefore, we consider the dataset at the level of high-quality subgroups of semantically matured tags, and apply pattern mining using the *lift* quality function for this task. As an evaluation, we apply a gold-standard measure of semantic relatedness derived from WordNet [8].

4.1 Methodology

For the purpose of assessing the degree of semantic maturity of a given tag, a crucial question is how to measure this degree in a reliable and semantically grounded manner. In prior work [6] we identified folksonomy-based measures of semantic relatedness, which are among others able to detect potential synonym tags for a given tag. The most precise measure we found was the *resource context relatedness*, which is computed in the vector space $\mathbb{R}^R$. For a tag t, the vector $\boldsymbol{v}_t \in \mathbb{R}^R$ is constructed by counting how often a tag t is used to annotate a certain resource $r \in R$: $v_{tr} := \text{card}\{u \in U \mid (u, t, r) \in Y\}$. This vector representation can be interpreted as a "semantic fingerprint" of a given tag, based on its distribution over all resources. Our intuition for capturing the degree of maturity is based on the following argumentation chain:

1. The better the semantic fingerprint of a tag t reflects the meaning of t, the higher is the probability that the resource context relatedness yields "true" synomyms or semantically closely related tags $t_{sim1}, t_{sim2}, ...$ for t
2. If the most related potential synonym tag t_{sim1} is a "true" synonym of t (as grounded against the WordNet synset hierarchy), then the semantic fingerprint of t is regarded as semantically mature.
3. Otherwise, we consider the similarity in WordNet between t and t_{sim1} as an indicator for the maturity of the tag.

Please note, that we are using purely folksonomy-based measures (i.e., resource context relatedness) as a proxy for semantic similarity, because WordNet is not available for all tags. Simply spoken, this approach regards a tag as semantically mature if the information encoded in its resource context vector suffices to identify other tags with the same meaning. Naturally, this requires the existence of a sufficiently similar tag, which cannot be guaranteed. Therefore, this is not a sufficient but a necessary criterion. However, we think that the approach is justified, because the process of maturing is not restricted to isolated tags, but takes place similar to a "co-evolution" among several tags belonging to a certain domain of interest. As an example, if the topic of *semantic web* is very popular, then a relatively broad vocabulary to describe this concept will emerge, e.g. `semantic_web`, `semanticweb`, `semweb`, `sw`, ... In such a case, the maturity of a single tag would "correlate" with the existence of semantically similar tags within the same domain of interest. In general, it is important to notice that our methodology is also applicable to narrow folksonomies when replacing the resource context relatedness with the tag context relatedness (see [6]).

4.2 Semantic Considerations

For assessing the semantic similarity between tags we apply WordNet [8], a semantic lexicon of the English language. WordNet groups words into *synsets*, i.e., sets of synonyms that represent one concept. These synsets are nodes in a network; links between these represent semantic relations. WordNet provides a distinct network structure for each syntactic category (nouns, verbs, adjectives and

adverbs). For nouns and verbs, it is possible to restrict the links in the network to (directed) *is-a* relationships only, therefore a subsumption hierarchy can be defined. The *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. Since the *is-a* WordNet network for nouns and verbs consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies; the graph is then fully connected so that several graph-based similarity metrics between pairs of nouns and pairs of verbs can be defined. In WordNet, we measure the semantic similarity using the taxonomic shortest-path length *dist*; the WordNet similarity $wns = 1 - \frac{dist}{max_{dist}}$ is then normalized using the maximum distance max_{dist}.

In addition to the WordNet similarity, we consider two additional indicators:

- The *Maturity Indicator* (mat) is a binary feature and measures if a tag has reached a certain maturity according to the WordNet information, i.e., the indicator is true, if we observe a WordNet similarity $wns \geq 0.5$.
- The *Synonym-Indicator* (syn) is a binary feature that specifies, if a tag-pair is in a synonym relation, i.e., the WordNet similarity $wns = 1$.

Since we consider the semantic fingerprint of tags using folksonomy information, we restrict the analysis to WordNet terms with only one sense; otherwise advanced word-sense disambiguation would be necessary in order to compare the correct senses in the WordNet synsets.

4.3 Dataset

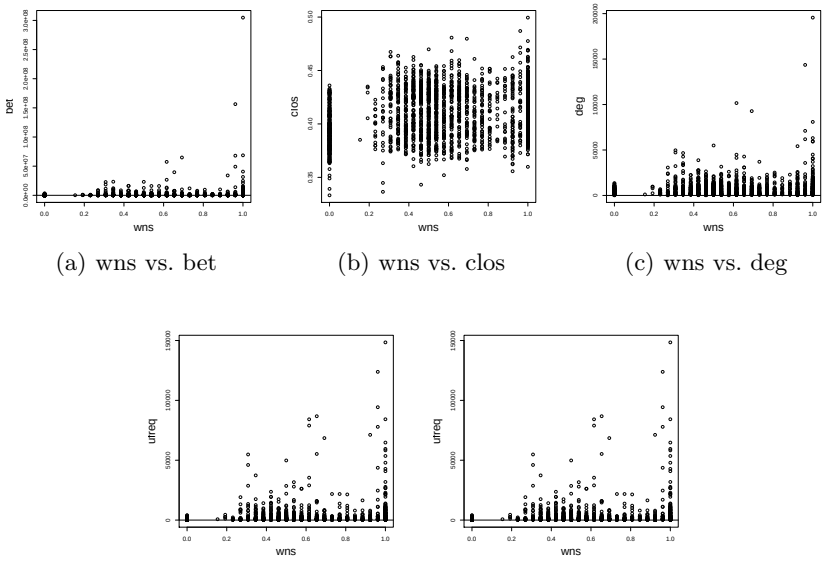
For our experiments we used data from the social bookmarking system del.icio.us, collected in November 2006. In total, data from 667,128 users of the del.icio.us community were collected, comprising 2,454,546 tags, 18,782,132 resources, and 140,333,714 tag assignments. For the specific purpose of our papers, some preprocessing and filtering was necessary: For the purpose of "grounding" the true semantic content of a tag t, we are applying vector-based measures to compute similar tags t_{sim}. Hence, we must assure that (i) the vector representation is dense enough to yield meaningful similarity judgements and (ii) there exist sufficiently similar tags t_{sim}. For these reasons, we first restrict our dataset to the 10.000 most frequent tags of delicious (and to the resources/users that have been associated with at least one of those tags). The restricted folksonomy consists of $|U| = 476,378$ users, $|T| = 10,000$ tags, $|R| = 12,660,470$ resources, and $|Y| = 101,491,722$ tag assignments. In order to assure the existence of sufficient "similarity partners" for each tag, we filter all tags whose cosine similarity to their most similar tag is lower than 0.05. As a last step, we only considered tags with exactly a single sense in WordNet in order to eliminate the influence of ambiguity. After all filtering steps, we considered a total of 1944 tags. We are aware that this is a strong limitation regardint the number of considered tags – however, because the problem at hand as well as our experimental methodology is sensitive towards a number of factors (like ambiguity or folksonomy-based similarity judgements), our focus is to start with a very "clean" subset. As a followup, it would of course be interesting to include more tags given the results on the clean subset are promising.

Table 1. Correlation between WordNet Similarity (wns), Maturity Indicator (mat), Synonym-Indicator (syn) and the different tag properties.

	bet	*clos*	*deg*	*rfreq*	*ufreq*
wns	0.15	0.20	0.20	0.21	0.18
mat	0.09	0.14	0.12	0.15	0.12
syn	0.12	0.14	0.13	0.15	0.15

We calculated all tag properties given the described co-ocurrence network, and discretized these using the standard MDL method of Fayyad & Irani [7] considering the WordNet similarity as a target class.

Statistical Characterization Figure 1 and Table 1 provide a first glance on the applied data. Each circle in Figure 1 represents one of the 1944 tags. Concerning the WordNet similarity (wns), we observe, that there is little correlation with the tag properties; Furthermore, we observe even lower correlations considering the two indicators *mat* and *syn*. Therefore, pattern mining using subgroup discovery is very suited for mining semantic tag profiles, since it also considers correlations in rather small subgroups described by combinations of different influence factors.



(a) wns vs. bet (b) wns vs. clos (c) wns vs. deg

(d) wns vs. rfreq (e) wns vs. ufreq

Fig. 1. Scatterplots for the WordNet Similarity (wns) vs. different tag properties.

4.4 Results

We applied pattern mining for the presented dataset using the tag properties as attributes, and the target concepts (wns, mat, syn) discussed above. Concerning the Wordnet Similarity (wns) and the *lift* quality function with a minimal subgroup size $n = 40$, we obtained the top patterns shown in Table 2. Lines 1-10

Table 2. Top patterns for target concept wns, split according to the different lengths of the patterns (mean in dataset: 0.54).

#	lift	mean	size	pattern
1	1.43	0.78	46	$ufreq > 13.0\%$
2	1.28	0.70	77	$rfreq > 3.0\%$
3	1.26	0.69	162	$clos > 64.7\%$
4	1.24	0.68	275	$deg > 6.0\%$
5	1.23	0.67	140	$bet > 1.0\%$
6	1.19	0.65	523	$ufreq > 1.0\%$
7	1.15	0.63	761	$rfreq > 0.1\%$
8	1.15	0.63	627	$bet > 0.2\%$
9	1.13	0.62	871	$clos > 47.0\%$
10	1.05	0.57	1519	$deg > 1.0\%$
11	1.32	0.72	51	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$
12	1.28	0.70	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$
13	1.28	0.70	246	$deg > 6.0\%$ AND $rfreq > 0.1\%$
14	1.33	0.73	74	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
15	1.30	0.71	119	$bet \in [0.2\%, 1.0\%]$ AND $deg > 6.0\%$ AND $rfreq > 0.1\%$

Table 3. Top patterns for the target concept "Maturity Indicator" (mean: 0.59)

#	lift	p	size	pattern
1	1.52	0.91	44	$ufreq > 13.0\%$ AND $clos > 64.7\%$
2	1.49	0.89	46	$ufreq > 13.0\%$
3	1.33	0.80	73	$rfreq > 3.0\%$ AND $clos > 64.7\%$
4	1.31	0.78	77	$rfreq > 3.0\%$
5	1.25	0.75	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$
6	1.24	0.74	246	$deg > 6.0\%$ AND $rfreq > 0.1\%$
7	1.21	0.72	115	$bet \in [0.03\%, 1.0\%]$ AND $ufreq > 1.0\%$
8	1.21	0.72	275	$deg > 6.0\%$
9	1.20	0.72	162	$clos > 64.7\%$
10	1.18	0.70	588	$clos > 47.0\%$ AND $rfreq > 0.1\%$
11	1.36	0.81	74	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
12	1.33	0.80	86	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $rfreq > 0.1\%$
13	1.30	0.77	105	$bet \in [0.03\%, 1.0\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
14	1.26	0.75	108	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 554$

show only basic patterns (one selector), while the lines 11-15 indicate more complex patterns. These results show that high betweenness and high closeness as intuitively expected. The influence of the degree centrality is not as prounounced as the other centralities, while higher degree also improves semantic maturity. Furthermore, a relatively high user frequency seems like the best indicator for high quality tags. Additionally, relatively high resource frequency is also a top indicator for semantic maturity.

If we consider the "maturity indicator" as the binary target concept, we obtain the patterns shown in Table 3. We observe similar influential properties as discussed above, however, the user and resource frequency combined with a medium or high closeness show the best performances.

Table 4. Top 5 patterns for the target concept "Synonym Indicator" (mean: 0.13)

#	lift	p	size	pattern
1	3.61	0.50	46	$ufreq > 13.0\%$
2	2.61	0.36	47	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$ AND $ufreq > 1.0\%$
3	2.53	0.35	77	$rfreq > 3.0\%$
4	2.40	0.33	51	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$
5	2.28	0.32	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$

Looking at the "synonym indicator" results shown in Table 4, we observe, that the tag properties identified above have an even more pronounced influence, since the increase in the target concept (the lift) is between 2 and 3, indicating an increase in the mean target share of the synonym indicator in the subgroups by 100% to 200%. An example for a small subgroup containing only synonyms is described by the pattern: $bet \in [1326142, 1.0\%]$ AND $ufreq > 13.0\%$ consists of the tags "wallpaper", "templates" and "bookmarks".

5 Conclusion

In this paper, we have presented an approach for mining semantic maturity of tags in social bookmarking systems. We applied pattern mining for identifying subgroups of tags with mature semantic fingerprints according to different tag properties. These were based on structural and statistical folksonomy properties and computed using the tag co-occurrence information and tag/user frequency information. We provided a detailed analysis of the different properties, and presented a case study using data from del.icio.us. The results indicate the influence of several properties with interesting orders of magnitude for the del.icio.us dataset. For example, the number of users plays a crucial role for the process of semantic maturing; however, the addditional consideration of centrality properties can help to identify subsets of tags with a higher degree of maturity.

For future work, we plan to extend our proposed methodology to larger tag sets, including less frequently used tags and especially the notion of semantic "immaturity". Furthermore we plan to include further tag properties, also including temporal aspects like the amount of time a tag is present in the system. Additionally, we aim to evaluate the method on more datasets from diverse social systems.

Acknowledgements

This work has partially been supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University, and by the EU project EveryAware.

References

1. Angeletou, S.: Semantic Enrichment of Folksonomy Tagspaces. In: Int'l Semantic Web Conference. LNCS, vol. 5318, pp. 889–894. Springer (2008)

2. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05). pp. 647–652. Edinburgh, Scotland (2005)
3. Benz, D., Hotho, A., Stumme, G.: Semantics Made by You and Me: Self-emerging Ontologies can Capture the Diversity of Shared Knowledge. In: Proceedings of the 2nd Web Science Conference (WebSci10). Raleigh, NC, USA (2010)
4. Brandes, U., Pich, C.: Centrality Estimation in Large Networks. I. J. Bifurcation and Chaos 17(7), 2303–2318 (2007)
5. Cattuto, C.: Semiotic dynamics in online social communities. The European Physical Journal C - Particles and Fields 46, 33–37 (August 2006)
6. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: The Semantic Web, Proc.Intl. Semantic Web Conference 2008. vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
7. Fayyad, U.M., Irani, K.B.: Multi-interval Discretization of continuousvalued Attributes for Classification Learning. In: Thirteenth International Joint Conference on Articial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
9. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3) (2006)
10. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems. Journal of Information Sciences 32(2), 198–208 (April 2006)
11. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: Proc. of WWW2007. pp. 211–220. ACM, New York, NY, USA (2007)
12. Heymann, P., Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Tech. rep., Computer Science Department, Standford University (April 2006)
13. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: The Semantic Web: Research and Applications. LNAI, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
14. Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag Recommendations in Folksonomies. In: Proc. PKDD 2007. Lecture Notes in Computer Science, vol. 4702, pp. 506–514. Berlin, Heidelberg (2007)
15. Körner, C., Benz, D., Strohmaier, M., Hotho, A., Stumme, G.: Stop Thinking, start Tagging - Tag Semantics emerge from Collaborative Verbosity. In: Proc. of WWW2010. ACM, Raleigh, NC, USA (apr 2010)
16. Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J.: Tag Ranking. In: Proc. of WWW2009. pp. 351–360. WWW '09, ACM, New York, NY, USA (2009)
17. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata (December 2004)
18. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Proc. Intl. Semantic Web Conf. LNCS, vol. 3729, pp. 522–536. Springer (2005)
19. Sigurbjörnsson, B., van Zwol, R.: Flickr Tag Recommendation Based on Collective Knowledge. In: Proc. of WWW2008. WWW '08, ACM (2008)
20. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge Univ Pr (1994)
21. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97). pp. 78–87. Springer Verlag, Berlin (1997)
22. Zhou, M., Bao, S., Wu, X., Yu, Y.: An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. pp. 680–693 (2008)

Proceedings of the 4th International Workshop

Social Data on the Web

Workshop at the 10th International Semantic Web Conference

Alexandre Passant, DERI NUI Galway, Ireland
Sergio Fernández, Fundación CTIC, Spain
John Breslin, DERI NUI Galway, Ireland
Uldis Bojārs, University of Latvia, Latvia

Social Data on the Web (SDoW) workshop series

http://sdow.semanticweb.org/