USER MODELING, ADAPTATION and PERSONALIZATION
BIG ISLAND OF HAWAI'I, JUNE 20-24, 2010
UMAP 2010

2010 SASWeb

International Workshop on

Adaptation in Social and Semantic web

SASweb 2010

http://ailab.dimi.uniud.it/en/events/2010/sasweb/

This volume is published by Federica Cena, Antonina Dattolo, Styliani Kleanthous, Carlo Tasso, David Bueno Vallejo, and Julita Vassileva.

ISSN 1613-0073

Proceedings of the International Workshop on Adaptation in Social and Semantic Web (SASweb – 2010), Hawaii, EEUU, June 21, 2010.
Edited by: Federica Cena, Antonina Dattolo, Styliani Kleanthous, David Bueno Vallejo, Carlo Tasso, and Julita Vassileva
Submitted by: Antonina Dattolo

Published on CEUR-WS:
ONLINE: http://CEUR-WS.org/Vol-590/
ARCHIVE:
ftp://SunSITE.Informatik.RWTH-Aachen.DE/pub/publications/CEUR-WS/Vol-590.zip

SASweb 2010 site:
http://ailab.dimi.uniud.it/en/events/2010/sasweb/

Table of Contents

Papers

Adaptive Retrieval and Composition of Socio-Semantic Content for Personalised Customer Care

Ben Steichen, Vincent Wade

Knowledge and Data Engineering Group, School of Computer Science and Statistics
Trinity College, Dublin, Ireland

{Ben.Steichen, Vincent.Wade}@cs.tcd.ie

Abstract. The parallel rise of the Semantic and Social Web provides unprecedented possibilities for the development of novel search system architectures. However, many traditional search systems have so far followed a simple one-size-fits-all paradigm by ignoring the different user information needs, preferences and intentions. In the last number of years, we have begun to see initial evidence that personalisation may be applied within web search engines, however little detail has been published other than adaptation based on user histories. Moreover, current implementations often fail to combine the mutual benefits of both structured and unstructured information resources. This paper presents techniques and architectures for leveraging socio-semantic content and adaptively retrieving and composing such content in order to provide personalised result presentations. The system is presented in a customer care scenario, which provides an application area for personalisation in terms of available heterogeneous resources as well as user preferences, context and characteristics. The presented architectures combine techniques from the fields of Information Retrieval, Semantic Search as well as Adaptive Hypermedia in order to enable efficient adaptive retrieval as well as personalised compositions.

Keywords: Adaptive Information Retrieval, Adaptive Result Composition, Socio-Semantic Search, Personalised Search

1 Motivation

The vast growth of the World Wide Web has resulted in search engines playing an integral part in people’s daily pursuit for information. In particular, with the rise of the *Social Web*, or Web 2.0, a significant part of the growing number of resources constitute user-generated content such as forum posts, tags, media uploads, etc. Although web search engines have become very efficient at indexing, retrieving and ranking unstructured documents (including such Web 2.0 resources), traditionally they have often followed a one-size-fits-all paradigm: the same results are returned in the same form and order for each user with the same query. More recently the notion of *Personalised Information Retrieval (PIR)* has emerged in research projects in order to retrieve more relevant results for users’ personal information needs [1]. However, the conceived solutions have mainly focussed on improving ranked list scores by boosting documents depending on their similarity to a mined user profile. They do not take into account the different search expertise, preferences or knowledge levels of users, nor do they make use of search strategies in order to assist more complex informational queries. The rise of the Semantic Web has provided new possibilities for representing information using semantic data formats such as ontologies, allowing the development of *Semantic Search (SS)* systems. However, the current state of the art of such systems has largely followed the IR approach of ranking relevant documents and presenting them in ranked lists. They have so far failed to use semantic knowledge to provide an improved guidance for querying users. The field of *Adaptive Hypermedia (AH)* has traditionally focussed on providing such guidance using personalised result compositions and presentations through multi-dimensional adaptivity. However, their reliance on heavily marked up content has often hampered the inclusion of open-corpus documents such as user-generated content.

This paper proposes to combine techniques and architectures from PIR, SS and AH in order to provide Adaptive Information Retrieval and Composition. The proposed system consolidates both social and semantic data sources and provides a single query interface that supports personalised query responses. Customer Care is used as an example field where such a personalised system can be applied, since in addition to providing traditional technical documentation, many organisations now provide their own versions of community resources where users increasingly engage in forums in order to solve technical problems. By applying our search system across these different data sources, we are able to provide users with result compositions that are (i) personalised to their own needs with respect to the product, (ii) semantically structured according to organisational knowledge and (iii) combined from closed (semantic) as well as open (social) content.

2 Related Work

A variety of techniques and technologies have been developed in several research fields in order to i) search across increasingly large volumes of data and ii) tailor the content retrieval towards users' personal interests and preferences. A broad characterisation of such techniques reveals three distinct research areas: Personalised Information Retrieval, Semantic Search and Adaptive Hypermedia.

The field of Information Retrieval [2] has typically focused on improving ranked result lists using one-size-fits-all algorithms. More recently, *Personalised Information Retrieval* systems make use of personal information (e.g. gathered from previous search interactions [3]) in order to either expand the original user query with personalised keywords [4] and logical operators (e.g. AND, OR, NOT) [5], or to bias traditional ranking algorithms towards more personally relevant information [6]. Alternative composition and presentation attempts such as result clustering [7] have most often been confined to keyword frequency calculations, largely lacking a more fine-grained representation of i) the knowledge space that is being queried and ii) the user's personal knowledge state and preferences.

In order to overcome this lack of structured representations of both domain and user models, *Semantic Search engines* draw from the expressive power of ontologies, which can be used for modelling and reasoning across the knowledge space as well as user interests [8]. Although early Semantic Search systems often made use of manual one-to-one mappings between documents and ontological concepts, more "lightweight" systems [9][10] now (semi-) automatically annotate documents using multiple concepts drawn from ontologies. These annotations can then be used in order to rank open corpus documents not only by their statistical similarity to a user's keywords, but also by ranking them according to the importance of their particular annotations [10]. The usage of semantic user models such as in [8] has advanced the field to more personalised rankings of documents, however the sole dimension of adaptation has again been that of user interests. Moreover, user guidance has so far been largely neglected, as documents have mostly been composed and presented in a flat ranked list format, failing to guide the user through the result space.

Adaptive Hypermedia (AH) [11] is a field that has inherently focussed on providing multi-dimensional adaptation by creating personalised information compositions and presentations. Since the earliest systems such as AHA! [12] and APeLS [13], their focus has been on providing information compositions, which contain documents that are not only adaptively *selected* for the particular users, but also *sequenced* according to current user knowledge states as well as to a variety of user preferences. Moreover, presentational cues such as link hiding [12] or link annotation [14] provide additional navigation guidance across the document space. This increased adaptivity is facilitated by a new type of model called the Adaptation Model [12] or Narrative Model [13]. This model describes the strategy by which concepts can be traversed to support particular objectives. For example, a "how to" query of an inexperienced user might have a narrative that would first choose content containing a general introduction of the topic and its concepts, followed by examples on how to carry out the queried task. However, AH systems have inherently been hampered by their reliance on fine-grained concept-to-content indexing of the document space, making it hard to incorporate "unknown" open corpus data.

An additional search paradigm that has emerged over the last years is the notion of social search or collaborative recommendation. In these systems, users are presented with documents or items that are either globally popular [15] or recommended by users with similar interests (e.g AMAZON[1] recommendations). With the growth of online communities, these techniques might become increasingly powerful for future adaptive and personalised search. However, such collaborative techniques are out of the scope of the research presented in this paper.

In conclusion, the major gap in current search systems lies in the failure to augment Personalised Information Retrieval with Semantic Search and Adaptive Hypermedia techniques in order to create Personalised Result Compositions and Presentations. In order to overcome this gap, search systems need to integrate the notion of query adaptation based on a wider variety of user characteristics in order to enable more personalised retrieval. Moreover, the expressive power of ontologies that drives Semantic Search systems needs to be integrated in order to model both the knowledge domain as well as the system users. Finally, the Adaptive Hypermedia notion of a Narrative Model needs to be incorporated in order to i) retrieve documents that most closely correspond to the current domain and user model states and ii) adaptively compose and present the results to improve the guidance of users.

[1] http://www.amazon.com

3 Methodology

In order to study and address the identified gaps in current adaptive search techniques, a vertical application area is needed, which provides i) the necessary heterogeneous content and ii) an authentic evaluation scenario. For the research presented in this paper, a case-based study of customer care has been chosen, which represents an application area where users are currently already searching across both structured (closed corpus) and unstructured (open corpus) content. Additionally, this case study provides the necessary context for addressing different user information needs, skills and preferences.

4 Personalised Customer Care

Customer care is a crucial area for companies wishing to establish long-term relationships with their user base. Despite offering a strong product or service, it is often the post-purchase assistance that influences a user's decision to consider purchasing more products or services from this particular company [16]. However, it is surprising that the type of support in this massive area has been confined to the simple one-size-fits-all paradigm. Users are left having to either consult complete user manuals in order to find the relevant section for the problem in question, or perform a keyword query and search through traditional ranked result lists regardless of their personal background in terms of product knowledge, skills and preferences.

From a technical perspective, there are three types of help files that are available for supported products. First of all, a company internally produces *technical documentation* that is often sliced to a fine granularity in order to assure their reuse in the case of software updates. These smaller units are then compiled into manuals in order to be shipped as complete user guides. By composing these knowledge items into manuals, chapters and sections, companies provide a great array of implicit metadata information that can potentially be used for adaptive and personalised retrieval. In addition to these highly structured data sources, companies often produce a second type of documents, which contain *knowledge resources* that have been generated by support staff following a direct interaction with customers. These types of documents are generally less structured than technical support documents, containing limited metadata such as topic categorisation. Nevertheless, these articles contain valuable information for an end-user who might be facing a similar issue. Finally, a third type of documents is emerging increasingly with the rise of the social web, or Web 2.0. Users increasingly engage in *community forums*, asking questions to the general user community in the hope that either a similar problem might have been solved previously or that a user in the community has the technical knowledge to identify the problem area. In terms of technical markup, these documents contain the least structure for several reasons. First of all, users inherently use different terminologies depending on their linguistic and technical background. Secondly, when users categorise or tag forum posts, they might have differing intentions and perceptions of what might be relevant for future use. Finally, even if users agree on the type of tags, categorisations and language, the problems of synonymy and polysemy increase the mismatch between user-generated terms and the organisational terminology.

It becomes apparent that current customer care is not lacking in terms of support document quantity, but rather in terms of aggregating and structuring existing content in order to make it i) consistent, ii) reusable and iii) suitable for adaptation and personalisation. Hence it is necessary to develop new techniques and architectures for structuring and aggregating the different document types. Additionally, new search architectures are required that leverage such improved data models in order to make full usage of the complete document space.

5 Structuring Heterogeneous Content

The heterogeneous support content that is available for software products needs to be transformed to a semantically richer form in order to allow reasoning, adaptation and personalisation across it. As mentioned earlier, Semantic Web technologies such as ontologies represent an opportunity to base such structuring and markup on. The different types of content can be broadly categorised by their amount of existing metadata and structure. Consequently, different types of usage can be drawn from each: whereas *highly structured* content (such as technical documentation) can be used to derive an ontology of the knowledge domain, *unstructured* content (such as forum posts) can be marked up in order to provide querying users with a larger range of problem solutions. Key challenges in using marked up content and ontologies lie in identifying (i) how high quality markup needs to be, (ii) how extensive the vocabulary can be and (iii) how extensive the ontology needs to be.

5.1 Structuring organisational content

Organisational structured content is often of a fine granularity in order to ensure its reusability for future product updates. Transforming both the *individual knowledge items* as well as their *compositions* (e.g. from product manuals) to a domain ontology allows the content to be more reusable and suitable for adaptation and personalisation.

First of all, for each individual knowledge item, there exist a number of content fields such as title, paragraph, procedure, etc., as well as metadata fields such as index terms (i.e. keywords) or media type (e.g. text, image). By modelling the different fields as ontological classes, each knowledge item and its constituent parts can be populated as instances of these classes. This is particularly useful in the case of content and metadata fields that can be used for reasoning and adaptation (e.g. a metadata field indicating a procedure). For example, if a particular user has only just installed the product, explanatory items should introduce the user to a particular feature first, before showing a detailed procedure on how to configure this feature. The difficulty of an item can also be inferred from a variety of structural features contained such as the number of procedural steps, the content length, the number of paragraphs, etc. Corporate product documentation is often extensively marked up to a deep structural level, allowing such a detailed content analysis.

Secondly, the composition of knowledge items is transformed to ontological form by creating classes for the hierarchical components of the document (see Figure 1).

Moreover, components such as chapters, sections and subsections often contain additional data (e.g. overview titles), providing valuable information about the overall subject of its constituent knowledge items. The individual content items (e.g. chapters, sections, subsections) are used to populate the different ontological classes as instances, with instance relationships ensuring the ability to reason across connected items. For example, if a section explains a particular product feature, its subsections typically provide more detailed information. Again in the case of a less experienced user, it is important to not only show the detailed information about how to configure a particular feature, but also to introduce the feature with the explanation that is contained in a higher level section.

By transforming the complete technical documentation into classes and instances, a domain ontology can be created, which accurately describes the subject area from the point of view of the product provider. In particular, implicit knowledge from the existing item compositions in product manuals is effectively transformed into a form that allows making this knowledge explicit using ontological reasoning. Since the technical documentation is marked up consistently according to predefined schemas, most of the transformations can be applied automatically. However, in order to extract additional, more high-level concepts, a certain amount of manual effort is involved. For example, in the case of several product manual chapters referring to the same product features (one chapter explaining its installation, another one its configuration), the domain ontology should capture these cross-chapter relationships. Unless such references to higher level concepts (e.g. particular product features) are mentioned explicitly in the document markup, a domain expert needs to manually add these ontology classes and relationships.

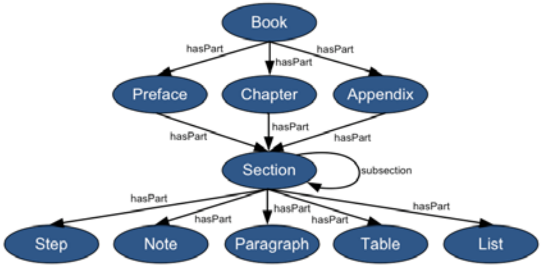


Fig. 1. Document Structure modelled as Ontology Classes

5.2 Annotating user-generated content

After a domain ontology has been generated, it is possible to link new "unknown" documents with the existing ontological instances. Two separate components are needed in order to generate i) the right granularity from the open corpus content and ii) conceptual indexing according to the ontological structure (see Figure 2).

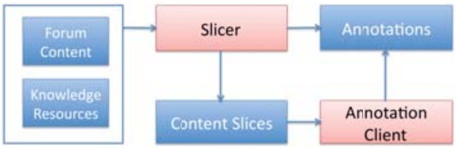


Fig. 2. Annotation Architecture

First of all, a content slicer described in [17] is responsible for transforming the original documents into fine-grained “slices”. Such slices are viewed as stand-alone pieces of information, containing their own semantic properties and metadata. During the slicing of the original open-corpus data (i.e. forum content and knowledge resources), structural as well as semantic analysis techniques are applied in order to generate fine-grained knowledge items as well as an initial set of metadata fields.

In a second step, the Web 2.0 concept of “crowd sourcing” is used to generate additional and more accurate annotations by presenting the content slices and their initial associated metadata to voluntary annotators (similar to [18]). Ideally, this socio-semantic annotation client is embedded within the actual community forum, allowing the initial content generators to tag their own posts. The domain ontology is also available to annotators as a preferred vocabulary in order to correspond their conceptual understanding of the slice to the terminology of the underlying semantic knowledge representation. The ontology is presented in hierarchical form, allowing annotators to easily browse and select concepts for the displayed slice. Furthermore, the annotation user interface includes several drop-down lists, which offer an annotation vocabulary for additional properties, such as the difficulty or interactivity level of the content. Finally, the selected annotations are stored in a triple store.

As a result of this two-stage approach, the original user-generated forum content as well as the knowledge resources have been annotated and consequently integrated with the semantic knowledge representation of the domain ontology. Even if the annotations are not as complete or accurate as the fine-grained technical documentation, they nevertheless enable partial reasoning, adaptation and personalisation during the content retrieval and composition stage.

6 Knowing and adapting to the user

Knowing the different characteristics, context and preferences of users is crucial for the development of any adaptive and personalised system. In the particular case of Personalised Customer Care, there are a number of user characteristics that product providers can adapt on.

First of all, a customer is using one or more *particular products or services* out of a potentially large portfolio from the company in question. Instead of leaving users sorting through search results in order to find the information that is related to their particular product, a system can automatically adapt the information retrieval and result composition accordingly. Secondly, upon each interaction with a search system, the user has a particular *product state*. For example, a user might have just purchased the product, consequently finding him-/herself at the “product installation and activation state”. Other examples would be the state of “configuring” after installation or the execution of “pro-active actions” (e.g. the user simply wants to find out more about a certain feature) or “re-active actions” (e.g. an error message has occurred in the product and the user wants to solve the problem). Another characteristic of a customer is one’s personal *knowledge state,* which depends on previous interactions with the product and the search system. Users could range from being complete novices to being considerable experts regarding particular parts of the product. Additionally, users can have different *content preferences*, for example some users might prefer looking at the content that contains the procedures for solving a problem, whereas other users might prefer to consult explanations or overviews first. Also, *language preferences* of users can be used during the adaptation phase, given that most of the content produced by a company as well as the community forums are available in different languages. In particular, consider a user who types in a keyword query in his/her native language other than English. If this particular user also speaks English, the system can adaptively retrieve additional resources in the case of poor coverage in the user's native language.

In addition to these user characteristics and preferences, there are additional axes of adaptation that arise at query time. A particular query can have a *question type*, which represents the type of intent of the user's question. For example, a user can have a query that is a "what"-type question, which requires an explanation as an answer. On the other hand, a "how" question requires the result to be a tutorial or procedure that the user has to follow in order to solve a particular problem. Additionally, the *preferred answer structure* might vary from query to query. For example, some queries are preferably answered with a "highly structured" result composition (including overviews, explanations, tutorials, related items, etc.), whereas a "quick" answer would simply provide a tutorial or reference resources (e.g. registry entry values, etc.).

The different user characteristics and preferences are stored using a hybrid user model, consisting of simple key-value pairs (e.g. for language preferences), semantic structures that mirror the domain ontology (i.e. overlay user model), as well as keyword vectors that represent users' historical interactions with the system (i.e. based on resources a user has looked at/clicked on).

7 Retrieval and Composition System architecture

In order to provide multi-dimensional adaptation, the domain and user models need to be consolidated with the Adaptive Hypermedia concept of a Narrative/Adaptation Model (as mentioned in section 2). This model contains the particular rules on i) what should be adapted on and ii) how the adaptation should occur. In this section, a Retrieval and Composition system architecture will be explained, which incorporates these three models in order to deliver Personalised Customer Care. The retrieval and composition process is broken down in several stages (see Figure 3) and incorporates influences from the areas of Adaptive Hypermedia, Semantic Search and Information Retrieval. In particular, this work extends an initial prototype presented in [18], which has already proven the benefits of personalised retrieval and composition of open-corpus content in an educational scenario.

In the first stage, a user is requested to input a standard keyword query, along with a drop-down selection of query types (i.e. what/how). Additionally, users indicate their current activity or intent regarding the product, i.e. getting started, reacting to a problem, etc. Ideally, this property would already be stored in a user model (e.g. from previous interactions with the product or search engine), thus not requiring a user to manually select this information. The keyword query is executed on an indexed version of the domain ontology, yielding a collection of instance results. From these results, several statistics can be generated. First of all, it is possible to determine which "conceptual area" of the domain ontology has yielded the most results, i.e., which are the high level concepts that have the most results. For example, by ordering the results by their corresponding chapter or subject, one can infer the particular part of the domain ontology that contains many of the keywords. Additionally, by analysing the search results, it is also possible to generate statistics about the type of content that is retrieved, such as the activity-level (i.e. amount of procedures and tutorials), the compositional properties (number of detailed subsections results), etc.

These initial statistics are used in a second stage to group results and to extend the subject space in order to personalise the results shown to the user. By consolidating the initial results with the domain and user model, a strategy is then applied to provide a "storyline" across the conceptual space. Particular ontological relationships of the initial results are followed depending on user model preferences. For example, in the case of a user who has just purchased the product, knowledge items (i.e. instances) that focus on installing and activating the product are added to the results. Another example would be to add related instances that fill a particular user's current knowledge gap (e.g. overview resources about a product feature, related features, etc.). Also, the activity level and difficulty level of instances influence their inclusion in the result space based on the user model preferences. At the end of this second step, a complete personalised result space has been selected from the domain ontology, which is not only more personally relevant than the initial results, but also more diversified and complete, containing additional relevant instances that would not have been found using conventional keyword search. The different results are composed according to their ontological relationships (provided by the domain ontology), their subject coverage, as well as their relevance to the querying user.

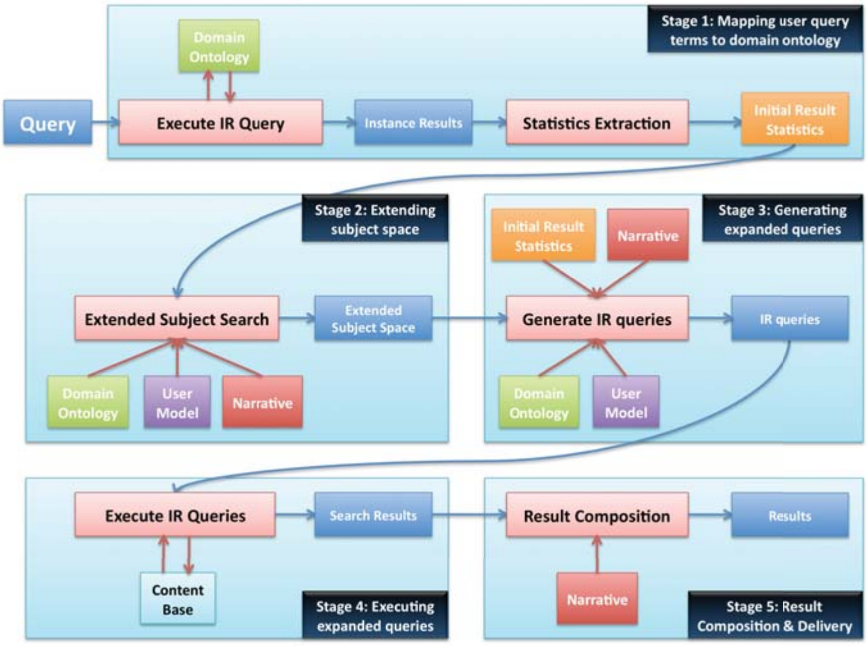


Fig. 3. Retrieval and Composition System Architecture

In stages 4 and 5, additional resources are retrieved by generating and executing expanded information retrieval queries across the user-generated content base. For each instance result in the extended subject space, an adapted query is generated, which contains the various aspects of resources that should be retrieved (in terms of keywords and metadata attributes). By indexing the content as well as the user-generated annotations, structured queries can be used to retrieve topically as well as personally relevant data. Additionally, logical operators and query term weights are used in order to also minimise an overlap between the different result sets.

In the final step, the different results are composed together with the instance results from the domain ontology in order to provide a complete result space. The combined sets are grouped, sequenced and linked according to the particular structure of the personalised subject space that was generated in step 2. This additional notion of sequence or narrative corresponds to a typical Adaptive Hypermedia presentation that guides users through the result space rather than presenting a flat list [13]. For example, for a novice user, advanced features are preceded by simpler (overview-type) resources, and followed by additionally relevant/related results. Also, due to these highly structured and personalised characteristics of the result space, additional Adaptive Hypermedia techniques can be applied. For example, on the result overview page, visual cues and link annotations guide a user to the currently most appropriate items to look at. Lastly, the composition of both organisational content as well as user-generated content ensures structure while still maintaining great topic coverage.

8 Ongoing Work

The system implementation is currently being completed using a variety of technologies. The organisational content has been transformed into the Web Ontology Language (OWL)[2] using customised scripts, whereas the annotation store consists of a standard installation of the ARC triple store[3]. To ensure both efficiency as well as reasoning capabilities, the domain ontology is stored in both eXist[4] (which allows efficient indexing using the built-in Lucene[5] functionality), as well as its ontological form (for reasoning during the extended subject search stage). The retrieval and composition system builds on work presented in an educational scenario [18] (see Figure 4) and uses an Adaptive Engine to consolidate the User, Domain and Narrative Models. Ontological reasoning is performed within the Adaptive Engine using the Jena Framework[6]. Similarly, the extended queries are generated by the rules encoded in the narrative, which can either be scripted (JavaScript), or rule-based (Drools[7]). The adapted queries are executed on an indexed version of the annotated content slices and the results are presented in a web-based interface using JSP and JavaScript.

The system evaluation will consist of authentic users performing activities over the domain content, with assessment measures focussing on retrieval accuracy and

[2] http://www.w3.org/2004/OWL/

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

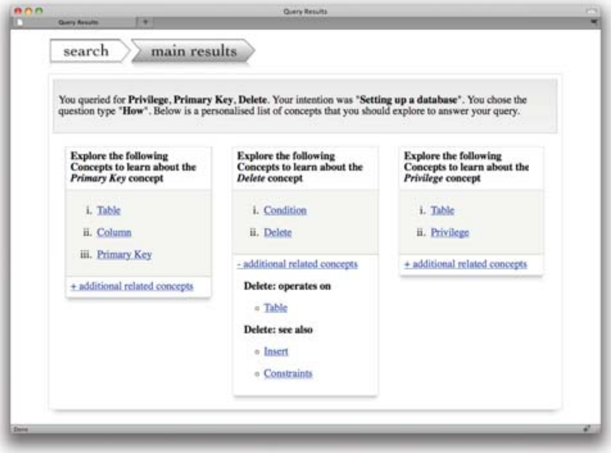
[4] http://exist.sourceforge.net/

[5] http://www.exist-db.org/lucene.html

[6] http://jena.sourceforge.net/

[7] http://www.jboss.org/drools/

appropriateness, as well as the general task assistance in terms of task completion time and user effort. A second evaluation will capture typical user queries, which will be used as test evaluations of system response accuracy by product experts.



You queried for **Privilege**, **Primary Key**, **Delete**. Your intention was "**Setting up a database**". You chose the question type "**How**". Below is a personalised list of concepts that you should explore to answer your query.

Explore the following Concepts to learn about the *Primary Key* concept

i. Table
ii. Column
iii. Primary Key

+ additional related concepts

Explore the following Concepts to learn about the *Delete* concept

i. Condition
ii. Delete

- additional related concepts

Delete: operates on

- Table

Delete: see also

- Insert
- Constraints

Explore the following Concepts to learn about the *Privilege* concept

i. Table
ii. Privilege

+ additional related concepts

Fig. 4. Result presentation in educational scenario

9 Conclusions

This paper has presented a novel approach to providing personalised information retrieval and composition from a variety of heterogeneous data sources. The presented architectures for structuring and retrieving both structured and user-generated content combine the latest advances in Personalisation, Semantic Search, Information Retrieval as well as the Social Web. Firstly, existing content resources are leveraged and structured in order to make them reusable, as well as suitable for adaptation and personalisation. Secondly, large sets of user-generated content are annotated using a socio-semantic annotation tool. Finally, an adaptive retrieval and composition architecture is responsible for aggregating the different data sources into personalised result presentations, which guide users towards relevant and appropriate resources.

The system is presented in a Customer Care scenario, which provides both the necessary heterogeneous data sources, as well as the context for different user information needs and preferences. It makes full usage of existing organisational structured knowledge and applies this across the user-generated content. The resulting user experience is a vastly improved customer care service, which provides an automated personalised assistance without the need of technical support staff intervention. Existing socio-semantic resources are hence leveraged and combined not only to improve customer satisfaction, but also to save costs for the product provider.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Trinity College Dublin. We would like to acknowledge the contributions of the Localisation Department at Symantec that have provided us with a variety of customer care content, especially Fred Hollowood, Johann Roturier and Jason Rickard.

References

1. Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. Personalized Search on the World Wide Web. In: The Adaptive Web, LNCS, vol. 4321, pp. 195-230. (2007)
2. Baeza-Yates, R. A. and Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc. (1999)
3. Speretta, M., and Gauch, S.: Personalized Search Based on User Search Histories. In: Web Intelligence, WI2005, pp. 622–628 (2005)
4. Teevan, J., Dumais, S. T., and Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, SIGIR '05. (2005)
5. Koutrika, G., Ioannidis, Y.: A Unified User Profile Framework for Query Disambiguation and Personalization. In: Proceedings of the Workshop on New Technologies for Personalized Information Access, PIA2005, pp. 44–53, Edinburgh, Scotland, UK (2005)
6. Micarelli, A., and Sciarrone F.: Anatomy and empirical evaluation of an adaptive web-based information filtering system. In: User Modeling and User-Adapted Interaction, 14, 2-3, pp. 159-200 (2004)
7. Xu, S., Jin, T., and Lau, F. C.: A new visual search interface for web browsing. In: Second ACM international Conference on Web Search and Data Mining, Barcelona, Spain (2009)
8. Cantador, I., Fernández, M., Vallet, D., Castells, P., Picault, J., and Ribière, M.: A Multi-Purpose Ontology-Based Approach for Personalised Content Filtering and Retrieval. In: Advances in Semantic Media Adaptation and Personalization. Studies in Computational Intelligence, vol. 93, pp. 25-51 (2008)
9. Baruzzo, A., Dattolo, A., Pudota, N., and Tasso, C.: Recommending New Tags Using Domain-Ontologies. In: Proceedings of the 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology - Volume 03 (2009)
10. Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., and Castells, P.: Semantic Search Meets the Web. In: Proceedings of the 2008 IEEE international Conference on Semantic Computing, pp. 253-260, Washington, DC, USA (2008)
11. Brusilovsky, P., and Maybury, M.T.: From adaptive hypermedia to the adaptive web. In: Communications of the ACM, vol. 45, pp. 30-33 (2002)
12. De Bra, P., Aerts, A., Berden, B., de Lange, B., Rousseau, B., Santic, T., Smits, D., and Stash, N.: AHA! The adaptive hypermedia architecture. In: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT '03. (2003)
13. Conlan, O., Wade, V., Bruen, C. and Gargan, M.: Multi-model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning. In: Adaptive Hypermedia and Adaptive Web-Based Systems, AH2002, LNCS, vol. 2347, pp. 100-111. (2002)
14. Sosnovsky, S., Brusilovsky, P., Lee, D. H., Zadorozhny, V., and Zhou, X.: Re-assessing the Value of Adaptive Navigation Support in E-Learning Context. In: Proceedings of the 5th international Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2008, LNCS, vol. 5149, pp. 193-203. (2008)
15. Brusilovsky, P., Chavan, G., and Farzan, R.: Social adaptive navigation support for open corpus electronic textbooks. In: Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2004, LNCS, vol. 3137, pp. 24–33 (2004)
16. Reichheld, F.: The One Number You Need. In: Harvard Business Review, December (2003)
17. Levacher, K., Hynes, E., and Wade V.: A Framework for Content Preparation to Support Open-Corpus Adaptive Hypermedia. In: International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy (2009)
18. Steichen, B., Lawless, S., O'Connor, A., and Wade, V.: Dynamic hypertext generation for reusing open corpus content. In: Proceedings of the 20th ACM conference on Hypertext and hypermedia, Torino, Italy (2009)

A Cognitive Perspective on Emergent Semantics in Collaborative Tagging: The Basic Level Effect

Tobias Ley[1,2], Paul Seitlinger[2]

[1] Know-Center[1], Inffeldgasse 21a, 8010 Graz, Austria,
tley@know-center.at
[2] Cognitive Science Section, University of Graz, Universitätsplatz 2, 8010 Graz, Austria
paulchristian.seitlinger@edu.uni-graz.at

Abstract. Researching the emergence of semantics in social systems needs to take into account how users process information in their cognitive system. We report results of an experimental study in which we examined the interaction between individual expertise and the *basic level advantage* in collaborative tagging. The basic level advantage describes availability in memory of certain preferred levels of taxonomic abstraction when categorizing objects and has been shown to vary with level of expertise. In the study, groups of students tagged internet resources for a 10-week period. We measured the availability of tags in memory with an association test and a relevance rating and found a basic level advantage for tags from more general as opposed to specific levels of the taxonomy. An interaction with expertise also emerged. Contrary to our expectations, groups that spent less time to develop a shared understanding shifted to more specific levels as compared to groups that spent more time on a topic. We attribute this to impaired collaboration in the groups. We discuss implications for personalized tag and resource recommendations.

Keywords: Tagging, Categorization, Personalized Recommendation

1 The Basic Level Effect in Collaborative Tagging

Emerging semantics in social systems is a topic that has sparked significant interest in the research community. In collaborative tagging systems [5], for example, it has been suggested that a community of users negotiates meaning in a collaborative sensemaking process [4] that would lead to a stabilization of the used vocabulary over time [5]. Some have suggested that this process could be an alternative to the usually top down driven engineering of ontologies [3] [8] [11]. An example for this is the Software SOBOLEO (Social Bookmarking and Lightweight Engineering of Ontologies [1])

[1] The Know-Center is funded within the Austrian COMET Program (Competence Centers for Excellent Technologies) under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. Work on this paper has been partially funded within the MATURE project (www.mature-ip.eu) by the European Community.

that was used in our study. Here, users collaboratively tag bookmarks and then use the tags to build a shared vocabulary and a taxonomic structure.

Our conjecture is that besides an understanding of the social (e.g. [2]) and pragmatic processes (e.g. 6], it is equally important to understand the underlying cognitive processes in collaborative tagging for offering effective recommendations. For example, it has been shown that human categorization processes are highly variable and adaptive. Categorization does, for instance, vary on the level of specificity depending on a number of factors. Therefore, our intention with the study reported here is to look at the temporal dynamics in the collaborative tagging environment both in terms of the tagging activities and the associated cognitive processes over time. By doing so, we would like to gain a better understanding of the variability in human categorization as it can be observed in such an environment, and thereby enhance current personalized tag recommendation mechanisms provided both in the process of tagging and in the process of browsing tag clouds and resource collections. This should enhance the emergence of stable patterns in these environments.

The term *basic level advantage* [10] has been introduced to describe a preferred level of taxonomic abstraction when classifying objects of the real world (e.g. a preference for the term "dog" as opposed to "mammal" or "poodle"). In human communication, the basic level has an important role as it contains categories that are most easily retrieved from memory and have a high degree of information value in describing objects. Among many others, an advantage for the basic level has been shown when people verify the categories of pictures of objects [10], or in a free naming paradigm [12]. While the role of the basic level advantage in collaborative tagging is often acknowledged [1] [5] [6], surprisingly little empirical research exists to inform design decisions. In their study of *delicious*, Golder & Huberman [5] suggest that popular tags which are introduced very early for a certain bookmark correspond to categories of the basic level. The authors also find that the tag distribution for a certain bookmark quickly stabilizes over time suggesting an emerging consensus.

The authors also point to a potential problem with the basic level advantage that arises with differing levels of expertise. They hypothesize that there should be systematic variations across individuals of "what constitutes a basic level". In collaborative tagging, this basic level variation is a potential drawback. When resources are described on varying levels of specificity, it makes retrieval of information more difficult both for experts and for novices. While for the former, the information value of a basic level category is too low, for the latter the specific categories are not sufficiently well represented in memory, and, hence, their labels difficult to comprehend.

The hypothesized basic level variation is in line with cognitive research which has found a basic level shift in various categorization paradigms, such as generating attributes of category objects, free naming of category labels or verifying category membership [9] [10] [12]. Basic level shift for more experienced persons leads to better availability in memory of category members and their attributes on more specific levels of the taxonomy. Following sensemaking research, we expect that in a collaborative tagging environment a growing expertise in the domain can be observed over time. Therefore, we hypothesize that users will use more specific categories, will show better availability of these in memory and will ascribe more importance to more specific categories, when they collaboratively tag for a longer as compared to a shorter duration of time.

2 An Experimental Study

To test this hypothesis, we asked four groups of students to collaboratively collect bookmarks related to their course subject and describe them with tags. Two of the groups had to work on a topic for the whole duration of the semester (10 weeks), the other two groups switched their topic at half time. Our hypothesis was that the *long duration (ld)* groups would form a stronger representation in memory of the more specific tags and that they would rate their relevance higher than the *short duration (sd)* groups. Collaborative tagging among the students was realized through the social bookmarking system SOBOLEO. In SOBOLEO, the tags and the tag taxonomy that is collaboratively created are shared by all users of the system.

2.1 Participants and Procedure

The study took place in the context of a university course on cognitive models in technology enhanced learning at the University of Graz. Subjects (N=25, mean age M=23.3, SD=1.2) were psychology students participating for course credit. After an introduction to SOBOLEO, a computer literacy questionnaire and a word association test eliciting participants' knowledge about central concepts of the given topics were administered to the participants. Subjects were then assigned to four groups of 6 or 7 participants which were equivalent according to their scores on the word association test and computer literacy questionnaire. Each group was provided with their own SOBOLEO instantiation only accessible by personal usernames and passwords.

E-mails were then sent out to inform the participants of the topic they had to work on together with access details for their SOBOLEO environment. Two groups were asked to research the topic "the use of Wikis in enterprises", the other two groups "the use of Weblogs in universities". They were asked to prepare these topics as if they were collaboratively working on a report of presentation. Both topics were chosen because they were related to the course subject and because we expected the participants to have only little prior knowledge about them.

During the whole duration of the study (ten weeks) each student was expected to post two relevant bookmarks per week to the SOBOLEO environment and describe them with meaningful tags. The students were also required to collaboratively organize their tag collection with the help of the SOBOLEO taxonomy editor. To facilitate the emergence of consensus, the students were also encouraged to utilize the SOBOLEO chat and an external discussion forum.

After five weeks (at halftime), the SOBOLEO environments of two of the four groups were cleared. They had to start from scratch and to work on the other topic for another five weeks, making them the short duration (*sd*) groups. The other two groups continued with their prior topic, making them the long duration (*ld*) groups. Right before this topic switch, we also controlled for the fact that there still were no differences between the two conditions in the word association test. At the end of the semester, the association test and the relevance rating were administered to the 25 students in a group setting using a sample of tags they had created so far.

2.3 Tag Samples, Tag Specificity and Dependent Measures

By the end of the 10-week period, the four groups had created N=213 tags from which n=76 tags were drawn as a sample. To yield the independent variable *tag specificity,* tags were drawn from three different levels of the SOBOLEO taxonomies the students had created: *General* tags were drawn from the taxonomy levels 1, *medium* tags from level 2, and all tags below level 2 were allocated to the *specific* tags. From each of the four SOBOLEO environments, 19 tags were randomly drawn: three general (e.g. "weblogs", "e-learning by collaborating"), eight medium (e.g. "kinds of weblogs", "psychology of weblogs") and eight specific tags (e.g. "videoblogs", "microblogging"). Hence, the entire sample consisted of 76 tags: 12 general, 32 medium and 32 specific tags.

As a dependent measure, a relevance rating was collected at the end of the semester asking subjects to rate each tag sampled from their own SOBOLEO environment on a five-point Likert scale ranging from strongly relevant to strongly irrelevant for describing and organizing resources. By averaging the ratings of all group members a mean relevance rating for each tag was obtained. An association test was also conducted at the end of the semester. This test elicits implicit knowledge about concepts underlying verbal representations. Subjects were confronted with tags as stimulus words and asked to write down all associations coming to their mind. Response time was confined to 30 seconds. By counting the number of associations, the test informs about the strength of representation of concepts in memory. Stimulus words were the same tags used for the relevance rating. Again we averaged the number of associations of all group members to obtain a mean number of associations for each tag.

3 Results

Figure 1 displays the mean *number of associations* (left) and the mean *relevance rating* (right) as a function of tag specificity and duration obtained at the end of the study. These results indicate a basic level advantage, i.e., a strong representation of categories represented by general tags. Independent of duration, general tags at level 1 seem to evoke more associations (M=4.43, SD=0.21) than medium tags at level 2 (M=2.95, SD=0.13) and specific tags at level 3 (M=2.99, SD=0.13).

Secondly, a level - group interaction is emerging, but it is in the opposite direction than we had expected. Contrary to our expectations, *sd* groups achieved more associations and higher relevance ratings than *ld* groups for medium and specific tags. This was confirmed by a duration (*ld* and *sd*) × tag specificity (*medium* and *specific*) multivariate analyses (MANOVA) on the variables *number of associations* and *relevance rating*. The main effect for *duration* proved highly significant ($F_{2,58}$=9.82, p<.01) explaining 25% of variance in the dependent variables and indicating a strong effect ($\hat{f} > .40$). Neither the main effect *tag specificity* nor the interaction between *duration* and *tag specificity* were significant. To further determine the nature of the significant effect, two univariate analyses (ANOVAs) for each of the dependent variables were conducted. Both results match our descriptive pattern. Averaging over *medium* and *specific* tags, the *sd* groups achieve more associations (M=3.29, SD=0.61) than the *ld* groups (M=2.65, SD=0.81; $F_{1,59}$=13.01, p<.01). The same applies to the relevance rating ($F_{1,59}$=9.12, p<.001): the judged relevance of *medium* and *specific* tags is higher in *sd* groups (M=2.56, SD=0.60) than in *ld* groups (M=2.22, SD=0.69).

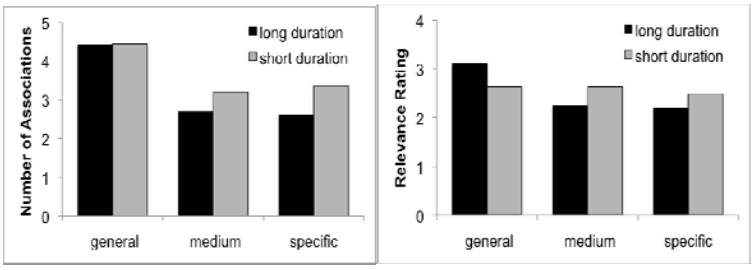


Fig. 1. Number of associations (left) and relevance rating (right) for general, medium and specific tags in long duration (10 weeks) and short duration (five weeks) groups

Results of a post-hoc questionnaire that had been administered to the students at the end of the semester give insight into these counterintuitive findings. First, all groups indicated they had been dissatisfied with the communication mechanisms (the SOBOLEO Chat and discussion forum). Albeit having worked on their topic for a longer time, groups of the *ld* condition gave significantly lower ratings when asked for the understanding of the topic (M=1.67 on a 5-point Likert scale, SD=1.23) than *sd* groups (M=2.69, SD=0.75; $F_{1,23}$=6.44, p<.05). Additionally, *ld* groups (M=1.92, SD=1.00) perceived a lower quality of their taxonomy than *sd* groups (M=2.92, SD=0.86; $F_{1,23}$=7.33, p<.05). Free text answers indicate that especially students in *ld* groups found it more difficult to collaboratively work on the shared taxonomy in SOBOLEO and they felt that the exercise had resulted in a chaotic collection of bookmarks and tags where it was rather difficult to keep an overview.

4 Discussion and Outlook

We conclude from the study that a strong basic level effect could be observed for an implicit memory measure (number of associations) as well as an explicit measure (relevance rating), where for the latter this only showed for one of the groups. However, our manipulation (duration of engagement with a topic) was obviously not effective in producing a stronger representation in memory. Quite to the contrary, the fact that environments of students in *sd* groups were cleared after half time actually helped them to build a more effective and shared external knowledge representation. The negative effect for *ld* groups was exacerbated by missing effective communication mechanisms in the SOBOLEO system. Similarly, we assume that it was students from *sd* groups that developed a more shared and stronger internal representation. If this was the case, then there is clear evidence for a shift in the basic level. This already showed after a comparatively little time (5 weeks), and produced a strong and also practically significant effect (an increase of 0.64 associations on average).

Results of this study have practical significance for tag and resource recommendation in collaborative environments (e.g. [2]) as they suggest that effective tag recommendations need to take tag specificity into account. Experts in a domain would benefit from more specific tag recommendations or from recommendations of resources with more specific tag assignments. The study also suggests that temporal dynamics need to be taken into account where shifts in basic level already take place after a few weeks of collaboration. Finally, in case tag specificity could be captured, this would also have implications for user modelling as the level of expertise pertaining to a certain topic could be derived for any user from his or her tag assignments.

A limitation of our results relates to the manual creation of the taxonomy by students which extends the (normally flat) folksonomy by a taxonomic relation. For our future work, we plan to draw on statistical approaches, such as [3] who found different tag similarity measures (tag co-occurrence vs. distributional measures) to correspond to different taxonomic relationships between tags. Moreover, these results seem to be moderated by particular behavioural tendencies of users using the tagging system [6].

References

1. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. In: 16th int'l WWW conference, pp. 1-10 (2007).
2. Carmagnola, F., Vernero, F., Grillo, P.: SoNARS: A Social Networks-Based Algorithm for Social Recommender Systems. In: 17th int'l UMAP conference, pp. 223-234 (2009).
3. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: 7th ISWC Conference, pp. 615-631 (2008).
4. Fu, Wai-Tat.: The microstructures of social tagging: a rational model. In: Proceedings of the ACM 2008 conference on Computer supported cooperative work, pp.229-238 (2008).
5. Golder, S., Huberman, B.A. The Structure of Collaborative Tagging Systems. J. Information Sciences 32, 198-208 (2006).
6. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop Thinking, Start Tagging: Tag Semantics Emerge From Collaborative Verbosity. In: 19th Int'l WWW Conference, ACM, New York, 2010.
7. Marlow, C., Naaman, M., Boyd, D., Davis, M: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proc. HYPERTEXT '06, ACM Press, 31-40 (2006).
8. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. J. Web Semantics: Science, Services and Agents on the World Wide Web 5, 5-15 (2007).
9. Rogers, T.T., Patterson, K.: Object categorization: reversals and explanations of the basic-level advantage J. Experimental Psychology: General 136, 451-469 (2007).
10. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. J. Cognitive Psychology 8, 382-439 (1976).
11. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: 4th ESWC2007 European Semantic Web Conference, pp. 624--639. Springer, Heidelberg (2007).
12. Tanaka, J.W., Taylor, M.: Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder? J. Cognitive Psychology 23, 457--482 (1991).

Personalized Recommendation of Integrated Social Data across Social Networking Sites

Yuan Wang[1], Jie Zhang[2], and Julita Vassileva[1]

[1]Department of Computer Science
University of Saskatchewan, Canada
[2]School of Computer Engineering
Nanyang Technological University, Singapore
[1]{yuw193, jiv}@cs.usask.ca [2]{zhangj}@ntu.edu.sg

Abstract. We have developed a dashboard application called "SoC-Connect" for integrating social data from different social networking sites (e.g. Facebook, Twitter), which allows users to create personalized social and semantic contexts for their social data. Users can blend their friends across different social networking sites and group them in different ways. They can also rate friends and/or their activities as favourite, neutral or disliked. We compare the results of applying five different machine learning techniques on previously rated activities and friends to generate personalized recommendations for activities that may be interesting to each user. The results show that machine learning can be usefully applied in predicting the interest level of users in their social network activities, thus helping them deal with cognitive overload. A visualization technique that has been shown to work well in previous work is applied to display personalized recommendations.

1 Introduction

Social Networking Sites (SNSs) have changed how people communicate: nowadays, people spend more time on SNSs than ever, and prefer communication via SNSs over emails [1]. Despite the diversity of SNSs and the fact that social media enriches people's lives, current SNSs have the limitation of poor user data interoperability [2]. User-generated contents, users' online activities, and their friendships are scattered over different places. It becomes increasingly inconvenient for users to manage their social data and constantly check many sites to keep track of all recent updates. People may also keep different accounts on the same SNS in order to protect their privacy or other purposes. In addition, users are often overwhelmed by the huge amount of social data, especially friends' activities (status updates).

In this paper, we present an approach for recommending social activities in a dashboard application called "SoCConnect", described in [3], for integrating social data from different SNSs (e.g. Facebook, Twitter), which allows users to create personalized social and semantic contexts for their social data. More specifically, through SoCConnect, users can blend their friends across different

SNSs to become an “integrated” friend account in SocConnect. Users can create groups for their friends who may share some common features and do some activities together. In the current work, we add the functionality that allows users to rate friends and/or their activities as favourite, neutral or disliked. To relieve users’ cognitive overload, we also apply different machine learning techniques to learn their preferences on activities based on their interactions with SCcConnect and to provide personalized recommendations of activities that are interesting to them. Evaluation results show the good performance of these techniques and especially good for some of them. A visualization technique developed in our previous work [4] is also used to display the personalized recommendations.

Section 2 presents the functionalities of SoCConnect. The approach for personalized recommendation of social networking activities is described in Section 3, followed by an experimentation in Section 4 to evaluate the performance. Related work on social data integration and recommendation is presented in Section 5. Finally, Section 6 summarizes the contributions of our work and discusses future research directions.

2 SoCConnect Dashboard

In this section, we provide a brief description about the functionalities of our dashboard application SoCConnect, the results of user studies supporting our design decisions for the functionalities, and the implementation of the system.

2.1 Functionalities

SoCConnect retrieves users’ friends information and their activities on different SNSs. It provides three functional categories, “managing friends”, “rating friends and activities”, and “personalized recommendation of activities”.

The first functional category, “managing friends” contains two functions: blending friends and grouping friends. In most cases, there is some level of overlap between the sets of a user’s friends on different social networking sites. Our system allows the user to merge the different accounts of a friend across SNSs, to create a single “integrated”(or “blended”) friend account for this friend in the user’s SoCConnent dashboard. The second function is to group friends. Users can put their friends, both individual SNS accounts and “integrated” accounts, into groups. This function allows users to express the context and semantics of friendships, which could be the shared characteristics, interests or activities between friends.

The second functional category, “rating friends and activities” allows users to rate friends or friends’ activities as favourite or disliked. The favourite activities are bookmarked, which can be revisited more easily. By rating, users are able to specify a semantic characteristic (currently limited to postive/negative) of their relationships with their friends and express their preferences on activities that they find more or less interesting and valuable.

The third functional category, "personalized recommendation of activities" recommends activities that may be interesting to users, making use of the previous ratings and the information about friend groups.

2.2 Motivation for these Functionalities

We conducted a user study to evalute our design decisions for the functionalities of SoCConnect. [1]A total number of 16 subjects (all students) were involved in this study, distributed over both gender and major (Computer Science or Non-CS). They were asked questions related to the functionalities during interviews. We provide here only the most relevant results.

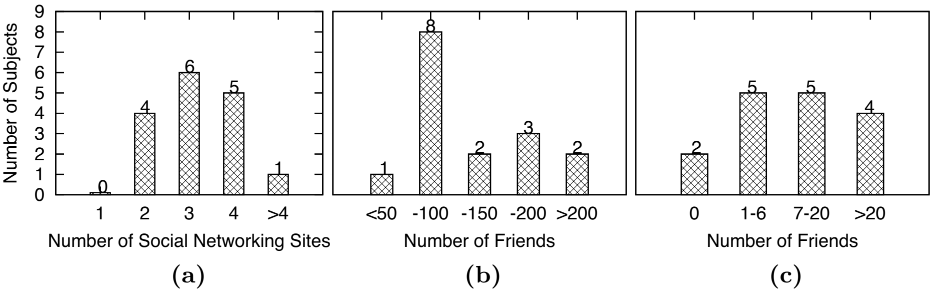


Fig. 1. (a) Number of Frequently Used SNSs; (b) Total Number of Friends on All SNSs; (c) Number of Friends Who have Accounts on More Than Two SNSs

All subjects have frequently used more than one SNSs (see Figure 1(a)). Most of them have frequently used more than two SNSs. Most of them also have more than 50 friends in total (see Figure 1(b)). Almost a half of the subjects have at least 100 friends. While it can be argued that this mini-study involved only students, this group presents the majority of users on most SNSs. For example, users of age 18-35 represented collectively 90% of the users on Facebook in 2008.[2] An e-business report from 2009 shows that 75% of the adults aged 18-25 have accounts on a SNS.[3]

The subjects were asked about the number of their friends who have user accounts on more than two SNSs. Only two subjects do not have such friends (see Figure 1(c)). More than a half of the subjects have at least 7 such friends. Several subjects (25% of all subjects) have more than 20 such friends. 81.25% of subjects answered that these friends were active on different sites and most of the friends have identical activities on these sites. 75% of subjects want to view these friends' activities in one place. These results confirm a strong need for the

[1] The study was approved by the Behavioural Ethics Research board of the University of Saskatchewan

[2] http://social-media-optimization.com/2008/05/social-network-user-demographics/

[3] http://emarketer.com/Article.aspx?R=1006882

function of blending friends. A significant majority (about 90%) of subjects have some friends who share similar interests, preferences or demographic information, or do some activities together. They want to create a group for these friends and include in groups some friends on different sites. These results support strongly our function of grouping friends.

In order to check if users would be willing to describe semantically their relationship and the content on their SNSs, we asked the subjects whether they want to tag friends and activities. Only half of them (54.25%) provided a positive answer. Tagging requires cognitive effort. The subjects were not sure whether they want to spend much time on tagging. Some subjects also feel that not many updates (friends' activities) are important. They prefer to tag only important activities or friends to revisit later. Instead of tagging with a word or a phrase, it requires less effort to mark an activity or friend as "favourite" or "disliked". Moreover, users of Twitter are familiar with this way of marking updates that they may wish to revisit later. This is why we provide the function of allowing users to add friends and activities as favourite or disliked, instead of a tag of any possible phrases.

The majority (68.75%) of the subjects said that they feel overwhelmed by the number of their friends' updates in one SNS. The number of updates will increase significantly when the friends' accounts across different SNSs are integrated by an application like SoCConnect. Thus, it is necessary to provide recommendations to help users navigate through their long list of friends' updates.

2.3 Semantics of SNS Data

To represent the semantics of social for generating recommendations, we design a generic ontology consisting of four main classes: SNS account (SNSAccount), integrated account (person), activity, and group. "SNSAccount" represents a user account on a SNS. "Person" represents a person who holds one or more SNS accounts. "Activity" represents generic information about activities appearing on SNSs. Each activity has a type. It can be a user update, e.g. a new friend added by the user, or an update by a third party application, e.g. a game such as FarmVille (farmville.com), and MafiaWars (mafia-wars.com) or other applications for Facebook, or specific clients (e.g. Tweetie, Twitdroid) or applications (e.g. Bit.ly) for Twitter. The activity may contain text and different types of media, such as pictures, videos and links. It may also have a target identifying the targeted user. "Group" represents a user-defined group for keeping friends together. A member of a group can be a SNSAccount or a Person.

3 Personalized Recommendations in SocConnect

One common problem of social networking site is information overload[4]. As indicated in our user studies, most of the activities from friends are not very

[4] http://www.stormdawg.com/2009/10/12/social-networking-and-information-overload/

important or interesting. Christian Kreutz in his blog calls this problem "network overload"[5].

everything in the list because as indicated in our user studies, most of the activities from friends are not very important or interesting. SocConnect aims to provide a personalized recommendation on activities to individual users according to a prediction generated using their preferences on previous social data. In this section, we will present a comparison of several machine learning techniques that can be used to predict users' preferences on activities and the approach selected for visualization of the personalized recommendations.

3.1 Learning User Preferences on Activities

Users directly express their preferences on activities by using the function of rating activities as favourite or disliked activities. Based on the ratings, SocConnect can learn users' preferences and predict whether they will be interested in new similar activities from friends. Machine learning techniques are often used for learning and prediction. SocConnect applies the classic techniques of Decision Trees, Support Vector Machine [5], Naive Bayes, Bayesian Networks, and Radial Basis Functions [6]. In brief, decision tree learning is one of the most widely used techniques to produce discrete prediction about whether a user will find an activity interesting. It classifies an instance into multiple categories. Naive Bayes Classifier and Bayesian Belief Networks are the two commonly used Bayesian learning techniques. The method of Radial Basis Functions belongs to the category of instance-based learning to predict a real-valued function. Support Vector Machines have been shown promising performance in classification problems. The implementation of these techniques bases Weka 3.7.0. The performance of these techniques on learning users' preferences on their social network activities will be presented and compared in Section 4. The one providing the best performance will be used by our system.

Table 1. Features of Activities for Used Learning

Features	**A Set of Possible Values**
Actor	actor's SNS account ID
Actor Type	favourite; neutral; disliked
Activity Type	upload album; share link; upload a photo; status upload; use application; upload video; reply; twitter retweet; etc
Source	Facebook; Twitter; etc
Application	foursquare; FarmVille; etc
Rating	favourite, neutral, disliked

[5] http://www.crisscrossed.net/2009/10/15/network-overload-the-burden-to-deal-with-too-many-social-network-sites/

To work with the above learning techniques, an activity needs to be represented by a set of features. Table 1 summarizes a list of relevant features and some of their possible values. Each activity has an actor (creator). SocConnect allows a user to add friends into a favourite or disliked list. Using these two features, we will be able to learn whether a user tends to be always interested in some particular friends' activities or activities from a particular type of friends. As discussed in Section 2.3, each activity has a type. We also take into account the sources which activities come from, such as Facebook and Twitter, since often users have a particular purpose for which they predominantly use a given SNS, e.g. Facebook for fun, Twitter for work-related updates. From this feature, we can find out whether a user is only interested in activities from particular social networking sites source. Different applications used to generate those activities are also useful to consider. For example, if a user's friend plays "MafiaWars" on Facebook but this user does not, the status updates generated from the "MafiaWars" application may be annoying to the user. We leave out the textual content of activities. One reason is that many activities, such as video uploads, do not have any textual content. Another reason is that activities may contain non-Latin language characters and the current meta-data of activities cannot reflect which language the actor is using, which makes text analysis difficult and expensive.

After learning from a user-annotated list of activities from his or her friends, each of which is represented by a set of the feature values, our system is able to predict whether a new activity from a friend will be considered as "favourite", "neutral" or "disliked" by the user. We assign an approximate weight to the new activity as follows:

$$w = \begin{cases} 0.5 & \text{if predicted as favourite;} \\ 0 & \text{if predicted as neutral;} \\ -0.5 & \text{if predicted as disliked.} \end{cases} \quad (1)$$

These predictions are based on the features of each activity. The next section presents how the social context, expressed by the user by grouping friends in SocConnect, influences the recommendations.

3.2 Heuristic to Supplement Learning

As described earlier, SocConnect allows users to create groups and add friends into the groups. A group implies the existence of some commonalities among the members of the group or some activities that group members have been doing together. The group information provides an indirect indication about users' preferences on activities. For example, if many activities of members in a given groups are considered as favourite by a user, the activities of the other friends classified by the user in this group will also be likely interesting to the user. Based on this heuristic, we extend the results of machine learning, by adjusting the weight of an activity. More specifically, for a friend in a group, if the number of favourite activities of other group members is larger than that of disliked activities, the weight of each activity from this friend will be increased. Otherwise, the weight will be decreased. Formally, suppose that the number of liked (marked as "favourite") activities of other group members in the group is F, and the number of disliked activities from them is D, then the weight of an activity from the friend will be updated as follows:

$$w = w + 0.5 \times \frac{F - D}{F + D} \tag{2}$$

Note that in extreme cases where every activity of the other group members is considered favourite, the weight of the friend's activity will be increased by 0.5. On another hand, if every activity of the other group members is considered disliked, the weight of the friend's activity will be decreased by 0.5. Also note that w stays the same if every activity of the other group members is considered neutral by the user ($F + D = 0$). For a friend who belongs to several groups, the effect of the heuristic on the weight of the friend's activity will be averaged over these groups.

This extension brings two extra levels of user interests in activities, namely "very favourite" and "very disliked". Together, we have a range of five levels of distinction for user interests, which has been commonly used in many popular rating systems, such as Amazon (amazon.com) and TripAdvisor (tripadvisor.com). The mapping between the interest levels of users in activities and the numerical weight for the activities is summarized in Table 2.

Table 2. Interest Level, Activity Weight and Colour Presentation

Interest Level	Activity Weight	Colour
Very Favourite	$0.6 \leq w \leq 1$	Persimmon
Favourite	$0.2 \leq w < 0.6$	Tawny
Neutral	$-0.2 \leq w < 0.2$	Maroon
Disliked	$-0.6 \leq w < -0.2$	Burgundy
Very Disliked	$-1 \leq w < -0.6$	Thyrian purple

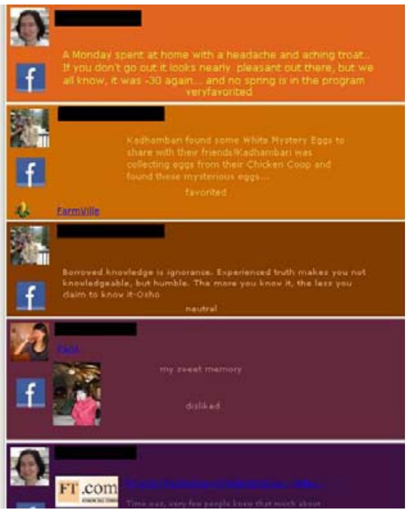
3.3 Adaptive Presentation of Recommendations in Visualization

The recommendations for the activities that the user may find interesting are integrated in the display of the activities in the activity stream that the user views in the interface of SocConnect. Colour in a spectrum that allows people with the most common type of colour-blindness (red-green) to distinguish,[6] is used to represent if an activity is recommended or unrecommended according to the predicted interest level calculated for the activity (Table 2). In this way the recommendation is unobtrusive, and can be easily ignored, but in the same time, it is intuitively clear for the user since it uses the metaphor "hot" item (displayed

[6] Images can be tested for appearance with simulated colour blindness at: http://www.colblindor.com/coblis-color-blindness-simulator/

in bright orange background, yellow text and large font) and "cold" item (dark purple background, blue text and small font). The metaphor allows representing a spectrum of recommendations with a larger number of values than 5, but we have picked 5 colours to represent transitions from hot through neutral (earth colour) to cold.

We have tested a visualization of items with different levels of interestingness using this metaphor with users in a study in previous work [4] and it was shown to work very well in quickly focussing user attention to the recommended items, while still allowing them to explore all items. This kind of visualization has been successfully deployed in the Comtella-D system in four classes with over hundred students for 2 years. That is why we decided to use it in SocConnect.



[redacted]

A Monday spent at home with a headache and aching troat.. If you don't go out it looks nearly pleasant out there, but we all know, it was -30 again.... and no spring is in the program
veryfavorited

[redacted]

Kadhambari found some White Mystery Eggs to share with their friends!Kadhambari was collecting eggs from their Chicken Coop and found these mysterious eggs...
favorited

FarmVille

[redacted]

Borrowed knowledge is ignorance. Experienced truth makes you not knowledgeable, but humble. The more you know it, the less you claim to know it-Osho
neutral

[redacted]

my sweet memory
disliked

[redacted]

Time out, very few people know that much about

Fig. 2. An Example of Visualization

4 Evaluation

We carried out another study to evaluate the performance of the five machine learning techniques on predict user preferences on social activities. Twelve subjects were involved in our evaluation. Five of them are from Saskatoon, Canada, and the other seven are from New Jersey, USA. A half of them are students and the other half are workers. We collected from the subjects the recent Facebook and Twitter activities from their friends. Ten of the subjects are experienced users of Facebook and Twitter. For each of these subjects, we collected 100 recent activities of friends. The other two subjects are relatively new users of Facebook and Twitter. For each of them, we collected around 50 recent activities of friends. We asked all subjects to rate their friends and activities. On average, they rated 38% of their friends as favourite or disliked friends and 45% of the activities as favourite or disliked, thus representing quite a diverse data sample.

A 10-fold cross validation was performed on the collected data from each subject, and the average performance of the machine learning techniques over the activities of all subjects are reported in Figure 3. Although the performance difference among these techniques is not very significant, support vector machine (SVM) provides the best performance, and it correctly classifies 74.1% of instances in the testing data. RBF performs the worst (70%). The performance of Naive Bayes and that of Bayesian Belief Network are about the same (around 72.6%). Decision Tree performs a little better (71.4%) than RBF.

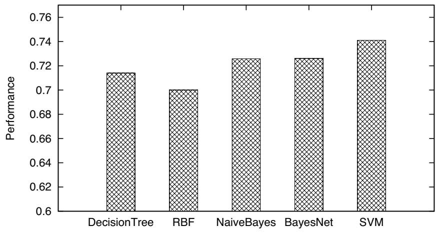


Fig. 3. Performance Comparison among Machine Learning Techniques

By looking closely into the predicted results, we found that many instances were misclassified by only one interest level, i.e. from “favourite” to “neutral” or from “disliked” to “neutral” and vice versa. We consider these as smaller mistakes. We summarize in Figure 4 the percentage of more serious misclassification from “favourite” to ‘disliked” and vice versa. We can see that only a very few (less than 3%) activities have been misclassified from “favourite” to “disliked” and vice versa. SVM consistently shows its best performance in this case. Overall, the experimental results confirm the good performance of machine learning techniques in learning social networking users’ preferences on their friends’ activities. SVM is particularly recommended in this context.

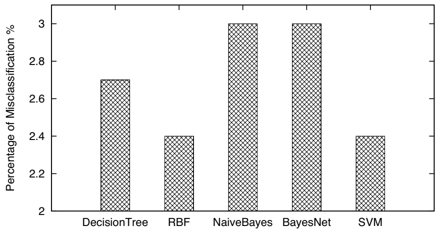


Fig. 4. Percentage of More Serious Misclassification

We also performed the validation on only 50% of collected data. More specifically, for each subject, we randomly selected 50% of collected instances. For each half of the data, we performed the same 10-fold cross validation to test the performance of the machine learning techniques. We repeated this process for 10 times to get the average performance when using only 50% of collected data.

Results shown in Figure 5 indicate that the performance when using 50% of data is consistently lower than that when using all data for the five machine learning techniques. This implies that the performance of personalized recommendation on social activities can be much improved when more data is collected from users, as users continuously use our system.

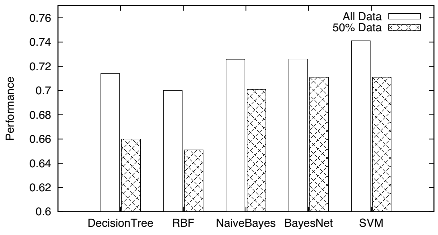


Fig. 5. Performance When Using All Data verses Performance When Using 50% Data

Using Weka's feature selection function, we can see which features are more important for individual users. We summarize in Table 3 the number of subjects for whom each feature was the most important one in the prediction. For all users, the feature "Actor" is the most important. "Actor Type", "Activity Type" or "Application" are more important for different users. The source of activities (i.e. whether they come from Twitter or Facebook) turns out to be not important. This interesting difference represents the diversity of social networking users' criteria in judging whether an activity is interesting to them, reflected in their ratings. Some users mainly care about their close friends' activities. Some users care more about the applications that generate the activities, which are usually the games they are playing. The implication is that learning the user type may be useful in personalized recommendation of activities. We leave this for future work.

Table 3. The Most Important Features

Features	Actor	Actor Type	Activity Type	Application	Source
Number of Subjects	12	4	3	3	0

5 Related Work

There have been some attempts to create personal portals that aggregate a user's accounts on different social networking sites, for example, the Seesmic Desktop (seesmic.com), power.com and the social web browser Flock (flock.com). They allow the user to view her pages on different social networking sites in one place. In this way, the users do not have to login to many different sites to view the updates of their friends. However, these applications do not allow users to blend or group their friends from different places.[7] They provide just a single-login

[7] http://news.cnet.com/8301-17939_109-10109878-2.html

interface in which users can switch between different tabs, one for each social networking site.

Bojars et al. [7] have been working on the SIOC project (Semantically-Interlinked Online Communities). This project shares similar focus with our work: social network portability and semantic web technologies. They proposed the SIOC ontology, which mainly focuses on users, implicit friendship, and social contents (primarily photos and discussions) in online communities such as online forums and Weblogs where contexts of social data are not so different.

In contrast, we focus mainly on developing a user-centric approach for integrating users' social data (including explicit friendship) on different SNSs, and that allows users to organize their social data and to create their personal contexts for the social data. We also provide personalized recommendation of friends' activities from different SNSs that are interesting to users.

Most recommender systems use collaborative filtering [8–10] based on the sharing of user ratings. While many SNSs deploy algorithms based on the analysis of social network structure to recommend new friends to the user, there haven't been many approaches to recommend contents on SNSs. One such approach is SoNARS. It takes a hybrid approach, combining results from collaborative filtering and content-based algorithms [11]. Dave Briccetti developed a Twitter desktop client application called TalkingPuffin (talkingpuffin.org). It allows users to remove "noise" (uninteresting updates) by manually muting users, retweets from specific users or certain applications. Currently, SocConnect focuses on automatically providing recommendations of social networking activities mainly based on the features of the activities.

6 Conclusions and Future Work

Our work has four contributions: 1) integration of social data from different SNSs; 2) allowing users to define their personal contexts of social data, including their integrated friends who may have SNS accounts on different SNSs, groups of their friends who share commonalities and activities from the users' own perspective, as well as their interest level (favourite, neutral or disliked) for friends and activities; 3) personalized recommendation of activities that may be interesting to individual users; 4) suggestion of a particular machine learning method for user preferences that has the best performance among five compared methods (SVM). A fifth potential contribution is the visualization of personalized recommendations integrated in the interface for viewing the activities, once its benefits are evaluated with users. Together, the personal dashboard application SocConnect provides users with a tool of integrating social data across different SNSs and with the convenience to selectively view friends' activities that are interesting to them.

For future work, next step will be to conduct user studies on the user interface to evaluate the usability of the visualization of recommendations and the appropriateness of the proposed heuristic to supplement machine learning. We are interested in exploring more deeply the relative importance of different features of social networking activities, to further improve the performance of personalized recommendation of activities. Other features that may be worth looking at include textual content of activities and the targeted friends of friends in activities.

7 Acknowledgement

This work has been supported by the Canada NSERC Discovery Grant of the third author.

References

1. Chisari, M.: The future of social networking. In: Proceedings of the W3C Workshop on the Future of Social Networking. (2009)
2. Erétéo, G., Buffa, M., Gandon, F., Leitzelman, M., Limpens, F.: Leveraging social data with semantics. In: Proceedings of the W3C Workshop on the Future of Social Networking. (2009)
3. Wang, Y., Zhang, J., Vassileva, J.: SocConnect: A user-centric approach for social networking sites integration. In: Proceedings of the International Conference on Intelligent User Interface (IUI) Workshop on User Data Interoperability in the Social Web. (2010)
4. Webster, A., Vassileva, J.: Personal relations in online communities. In: Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland, Springer LNCS (June 2006) 223–233
5. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods: Support Vector Learning. MIT Press (1999)
6. Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)
7. Bojars, U., Passant, A., Breslin, J., Decker, S.: Social network and data portability using semantic web technologies. In: Proceddings of the Workshop on Social Aspects of the Web. (2008)
8. Resnick, P., Lacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceddings of the ACM conference on Computer supported cooperative work. (1994)
9. Kautz, H., Selman, B., Shah, M.: Referral web: combining social networks and collaborative filtering. Communications of the ACM **40**(3) (1997) 63–65
10. Goo, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., , Riedl, J.: Combining collaborative filtering with personal agents for better recommendations. In: Proceddings of the sixteenth national conference on Artificial intelligence. (1999)
11. Carmagnola, F., Vernero, F., Grillo, P.: Sonars: A social networks-based algorithm for social recommender systems. In: Proceddings of the 17th International Conference on User Modeling, Adaptation, and Personalization. (2009)

Semantic Enrichment of Social Media Resources for Adaptation

Oliver Schimratzki, Fedor Bakalov, Adrian Knoth, and Birgitta König-Ries

Friedrich Schiller University of Jena, Germany
oliver.schimratzki@gmx.de,
{fedor.bakalov | adrian.knoth | birgitta.koenig-ries}@uni-jena.de

Abstract. With more and more dynamic content available on the web, we need systems that aggregate and filter information from different sources to provide us with only the information we are really interested in. In this paper, we present one such system, the CompleXys portal, aimed at users interested in complexity or subtopics thereof. It accesses a large variety of different information sources, among them calendars, news sites and blogs, semantically annotates and categorizes the retrieved content and displays only relevant content to the user.

1 Introduction

The amount of dynamic content available on the web is rapidly growing. It becomes more and more difficult for users to keep track of all relevant information - in particular since it becomes more and more overwhelming to manually separate relevant from irrelevant content. Even if a user has identified a variety of news sites and blogs that often contain information she is interested in, those sites will also contain lots of information the user is *not* interested in. Purely syntactic filtering based, e.g., on keywords, as offered by today's tools, offers only a partial solution. What is really needed is semantic filtering, i.e., filtering based on some "understanding" of the content. This will allow for higher precision, i.e., fewer irrelevant articles displayed, and higher recall, i.e. less relevant articles deleted, and will thus increase user confidence in using the tools.

In this paper, we present the CompleXys portal, an information site that will provide users with personalized access to information related to the topic of complexity. CompleXys harvests information from a large variety of sites, ranging from event calendars to blogs and news sites. It semantically annotates the retrieved content. These annotations are then used to categorize the retrieved items and to decide whether they are sufficiently related to complexity or should be discarded. In the future, CompleXys will use the categorization for a more fine-grained personalization, displaying the most relevant items most prominently and providing recommendations to the user.

In the remainder of this paper, after a brief discussion of related work in Section 2 we take a closer look at CompleXys and the underlying technologies: Section 3 provides an overview of the CompleXys architecture. We will then focus on the most interesting part of this architecture, namely the semantic content annotator which will be presented in Section 4. Finally, Section 5 contains a summary and an outlook on our future work.

2 Related Work

In this paper we describe an architectural solution and an approach to providing a personalized access to the variety of resources residing on the Web and in intranets. To achieve this, we combine the approaches and technology from three areas of research, namely content aggregation, semantic content annotation, and content-based recommender systems.

Content aggregation, though a relatively new field, has already achieved the state of maturity. Apart from the multitude of research proposals [16, 10, 11], there exists a number of industry standards and commercial applications of content aggregators. Really Simple Syndication (RSS)[1] and Atom[2] formats have been successfully used by a large number of Web and desktop application for aggregating various types of content, including but not limited to calendar information, news, blog entries, and podcasts. The iCalendar[3] format is used by many applications for aggregating appointments and events from multiple calendar systems. Personal Web portals like *i*Google[4] and My Yahoo![5] allow their users to place different types of content harvested through RSS and Atom feeds on their personal pages. Portals like Technorati[6] aggregate information on more or less specific topics. RSS filtering tools like Feed Rinse[7] allow the user to define keyword based filters on RSS feeds to get rid of irrelevant items. These tools work, however, on a purely syntactic level.

The field of **semantic content annotation** mainly deals with the challenges related to availability of well-formed metadata for the unstructured text resources, which is essential for achieving high recall and precision of information retrieval. A number of approaches to semantic content annotation have been reported in the literature [5, 9, 7]. GATE [4] has become one of the most widely used open source frameworks for implementing natural language processing (NLP) tasks. The framework empowers developers to implement such components as tokenizers, sentence splitters, part-of-speech taggers, gazetteers, semantic taggers, and the components for identifying relationships among the entities in the text. A number of NLP systems leverage GATE and its components for semantic tagging of content; these include but not limited to the KIM platform [14], MUSE [12], and Ont-O-Mat [8].

[1] http://web.resource.org/rss/1.0/

[2] http://tools.ietf.org/html/rfc4287

[3] http://tools.ietf.org/html/rfc5545

[4] http://www.google.com/ig

[5] http://my.yahoo.com

[6] http://technorati.com/

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

Availability of machine-processable metadata of content is one of the most essential requirements for the **content-based recommender systems** [13]. These systems recommend relevant content to the user based on the semantic description of available resources and the user's personal preferences. The relevant content is selected by analyzing the content metadata and the user's profile and identifying the items that match the user's individual interests. A number of systems leveraging this approach have been proposed. CHIP [17], for instance, is capable of recommending the user artworks from multiple museum collections. For recommendation, the system leverages the semantic description of artworks and the user's personal interests in the domain of cultural heritage, which the system identifies based on the user's explicit ratings of artworks and semantic relations among the art topics. Other examples of the systems leveraging similar recommendation approach are the Personal Reader Framework [3] and Personal Learning Assistant [6].

3 Overall Architecture

The CompleXys portal aggregates a multitude of different sources from the Internet, categorises the retrieved content, applies semantic annotation and finally presents the filtered and personalised results to the user. Table 1 shows a schematic overview of CompleXys' architecture and its data flow in a left to right manner, which basically implements the Input-Processing-Output model.

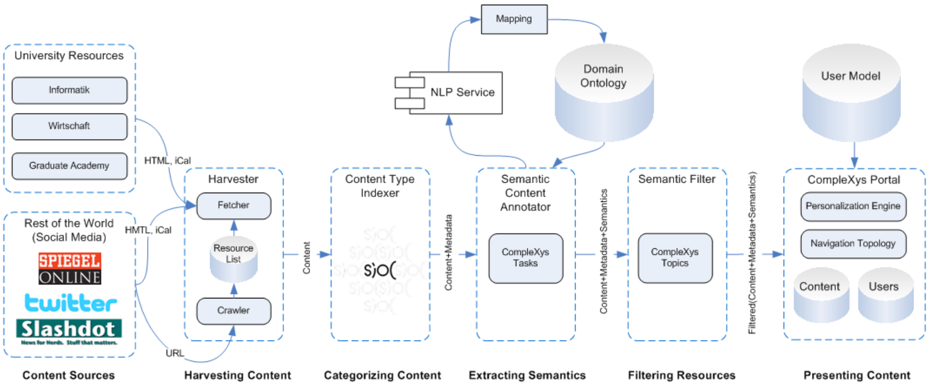


Fig. 1. Overview of the CompleXys portal. Resources are fetched and stored in crawler database, then semantically annotated and finally presented to the user, if they match his personal preferences and interests.

On the input site, the harvester retrieves arbitrary content and stores a mangled version in the crawler database. Since the particular source of each entry is known, this step also provides content type indexing for free.

The crawler database is fully generic and hence supports any kind of input source. Figure 1 shows the underlying schema. The DBMS will increment the unique key *id* for every newly retrieved entry. All later processing steps make use of this key: querying the crawler database for new content simply means querying for all *id*s higher than the last known or processed *id*.[8]

Column	**Type**	**Modifiers**
id	bigint	UNIQUE
source	character varying(255)	
content	text	
internal_id	character varying(255)	UNIQUE not null

Table 1. Database layout of the crawler database. This database contains already fetched items from a potentially large variety of sources. *id* is supposed to be monotonically increasing, while *internal_id* holds a suitable hashsum (e.g. MD5) of the cached resource. The *source* field specifies the origin of the item stored in *content*.

The content itself is blindly stored as text (BLOB), semantic parsing is delayed to subsequent stages in the processing pipeline. The *source* column contains the SIOC content type, it serves as a type indicator to the processing modules.

The crawler is idempotent, that is, it can be run several times without storing already known content again. This property is achieved by *internal_id*, another column set to be a unique key. For each retrieved entry, the crawler calculates a suitable MD5 hash and stores both, content and its derived hash into the database. If this entry has been already fetched, the DBMS will prevent inserting a duplicate MD5 hash into *internal_id* and consequently avoid storing known content again. Obviously, finding an appropriate way for calculating the MD5 hash is crucial. CompleXys currently has built-in support for two different source types: it can directly retrieve calendar events from SQL databases and arbitrary HTML input from the web via RSS, but more resource types can be added via crawler plugins. Generating a suitable hash representing the content usually differs among sources and is hence individually implemented in each such crawler plugin.

The SQL crawler connects to specified source databases in the University's network and harvests information about upcoming events, potentially of interest to the user. Since CompleXys strictly adhere to UTF-8 character encoding throughout the whole processing, the crawler is responsible to convert any source specific encoding, e.g., from Latin1 to UTF-8. This way, subsequent processing modules do not have to take care for different character encodings.

The retrieved SQL calendar events are normalised into a standardised template as shown in Figure 2.

[8] SELECT * FROM crawler WHERE id > already_seen

```
def fill_template(params)
    "<sioc:Item rdf:about=\"#{params[:source]}-#{params[:their_id]}\">\n" +
    "\t<vevent:dtstart>#{params[:date]}</vevent:dtstart>\n" +
    "\t<dcterms:creator>#{params[:speaker]}</dcterms:creator>\n" +
    "\t<vevent:location>#{params[:affi]}</vevent:location>\n" +
    "\t<dcterms:title>#{params[:title]}</dcterms:title>\n" +
    "\t<dcterms:abstract><![CDATA[#{params[:abstract]}]]>\n\t</dcterms:abstract>" +
    "\t<vevent:url>#{params[:url]}</vevent:url>" +
    "\t<vevent:dtend>#{params[:endtime]}</vevent:dtend >\n" +
    "\t<dcterms:modified>#{params[:lastupdate]}</dcterms:modified>\n" +
    "</sioc:Item>"
end
```

Fig. 2. Standardised ruby template for calendar events. All occurrences of *params* are substituted by values retrieved from a SQL based event management system.

The crawler finally calculates the appropriate MD5 hash for this event by concatenating the source prefix (a constant arbitrary string), the event's primary key in the foreign database and the provided last-update timestamp. This way, the MD5 hash of the concatenation is different for each event from every source database. Even more, updates to already retrieved events have a different timestamp, and consequently, a new MD5 hash together with this updated content will end up in the crawler database. Whenever the CompleXys portal encounters multiple entries for the same source URI in its crawler database, younger rows are updates to already known events.

In addition to SQL calendar events, the harvester has a HTTP crawler for arbitrary HTML content. URLs are extracted from RSS feeds specified in a static configuration file (see Figure 3).

Whenever possible, the crawler tries to use the print version of a document to remove navigation menus, advertisements and other unrelated noise. If the source already provides a more structured representation, e.g., iCal format, it will be used instead. Likewise the SQL crawler, the HTTP crawler wraps retrieved content into a SIOC[9] schema as depicted in Figure 4, generates a suitable MD5 hash and tries to store the result in the crawler database. Again, this insert will fail if the content is already known.

At this stage, all entries in the crawler database are simply unstructured raw text. Unless already provided by the source, there is no semantic information available, yet. However, semantic annotation is required to decide if a given content item is of interest to the user. The next section will explain in detail how this is done.

Once semantic annotation has been provided, relevant items are displayed to the user of the CompleXys portal categorized in appropriate domains. We are currently working on integrating our approach to personalization into CompleXys. This will allow to adapt the information provided to individual user

[9] http://sioc-project.org/ontology

```
<?xml version="1.0" encoding="UTF-8"?>
<rss-channels>
   <channel>
      <url>http://scienceblogs.com/sample/technology.xml</url>
      <source>scienceblogs.com</source>
      <type>blogs</type>
      <name>ScienceBlogs - Technology</name>
      <category>Technology</category>
   </channel>
   <channel>
      <url>http://scienceblogs.com/sample/medicine.xml</url>
      <source>scienceblogs.com</source>
      <type>blogs</type>
      <name>ScienceBlogs - Medicine</name>
      <category>Medicine</category>
   </channel>
       <channel>
       <url>http://www.wired.com/wiredscience/feed/</url>
       <source>wired.com</source>
       <type>blogs</type>
       <name>Wired Science</name>
       <category>Science</category>
   </channel>
</rss-channels>
```

Fig. 3. Example settings file for CompleXys' HTTP crawler.

```
public String wrapNewsItem(NewsItem newsItem){
  String wrappedNewsItem =
    "<sioc:Post rdf:about=\"" + newsItem.link + "\">\n" +
       "\t<dcterms:title>" + newsItem.title + "</dcterms:title>\n" +
       "\t<dcterms:created>" + newsItem.pubDate + "</dcterms:created>\n" +
       "\t<sioc:topic rdfs:label=\"" + newsItem.category + "\"/>\n" +
       "\t<sioc:content>\n" + "<![CDATA[" + newsItem.content +
       "]]>\n\t</content>\n" +
    "</sioc:Post>";
  return wrappedNewsItem;
  }
```

Fig. 4. Wrapper code for encapsulating newsfeed items into SIOC and DublinCore.

needs: Only information relevant to a specific user (and not to complexity in general) will be provided, the most important information will be displayed most prominently, related information (and possibly related users) will be recommended etc. Underlying this adaptation is a user interest model realized as an overlay over the domain model, that collects user interests based on the interactions of the user with the system and also allows the user manual adaptations [1, 2].

4 Semantic Content Annotators

The Semantic Content Annotators pursue the purpose of extracting semantic data from incoming text documents and of annotating this data back to the resources. Furthermore, they are meant to decide whether a given resource is relevant for the topic of complexity and to categorize it by means of corresponding topical concepts.

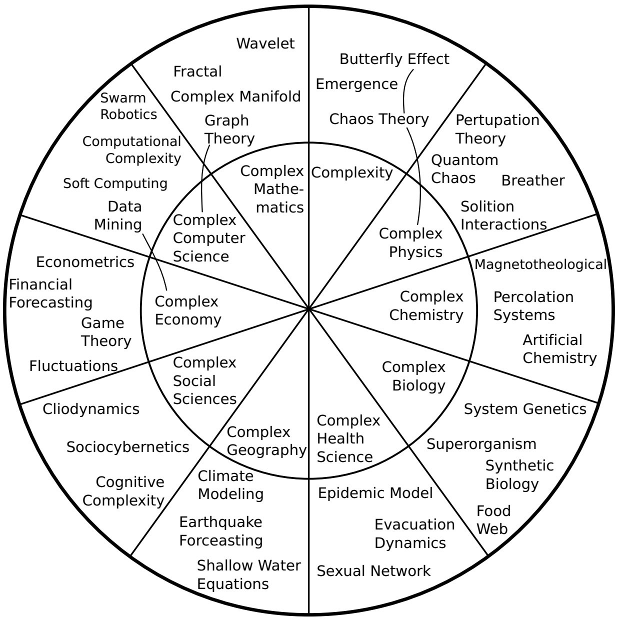


Fig. 5. The CompleXys taxonomy

Both, the annotation and the categorization tasks rely on an ontology, that represents the domain knowledge space of complexity. It is implemented as a SKOS[10] taxonomy and shallowly organized within two hierarchical levels - ten

[10] http://www.w3.org/TR/skos-reference/

main categories and 297 appendant terms. Furthermore some terms are interconnected by the relation type *related*, to express either topical closeness between two terms or an ambiguity of belonging, when a term could be assigned to more than one main category. Figure 5 shows an excerpt of the model as taxonomy circle[11]. The main categories are displayed in the inner circle, while the outer circle contains examples of their appendant terms. The connections between some of the terms are exemplarily for the use of the *related* relationships.

Figure 6 visualizes the architectural composition of the Semantic Content Annotator. It is structured as a parallel working pipeline, utilizing the standard java concurrency package[12] for its implementation. The current pipeline consists of five components, which are called CompleXys Tasks. The Crawled Content Reader and the Content Writer take care of an internally valid data structure and of persistency tasks. In contrast, the inner Complexys Tasks are the actual processing units. They analyze the resources, extract the semantic data and finally annotate it back to the text. The analysis is based on existing NLP services from various contexts. These services are called by using intermediate GATE modules, so that the Complexys Tasks need not care about the technical details of the annotation, but just have to adjust the modules according to their needs and evaluate the solutions.

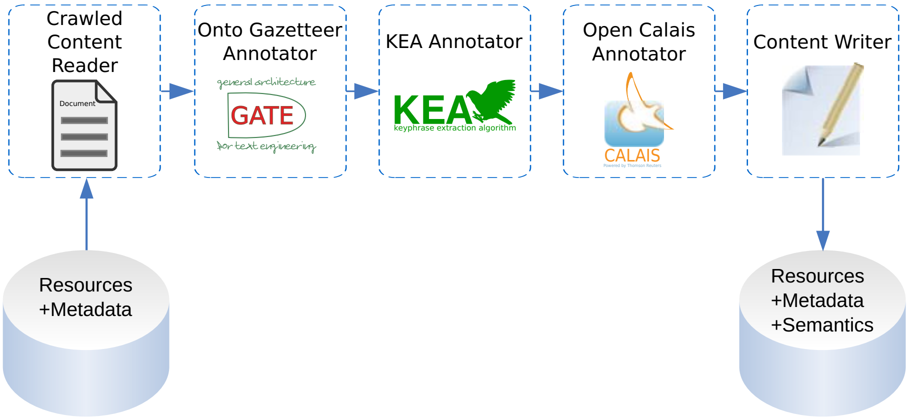


Fig. 6. The Semantic Content Annotators

The Crawled Content Reader is the first component of the pipeline and its main purpose is to gather the documents from the input data store and to prepare them for the succeeding tasks. It wraps the new resources into the internally used

[11] The complete ontology can be found at http://www.minerva-portals.de/o/complexys.rdf.

[12] http://java.sun.com/j2se/1.5.0/docs/api/java/util/concurrent/package-summary.html

GATE data format, embeds them into the corresponding persistency layer and sends them into an output queue for further processing in the pipeline.

The Onto Gazetteer Annotator searches the text for keywords, that are listed in the gazetteer files and annotates found terms with the corresponding taxonomy concepts. The frequency of occurring annotations can then be used as a simple indicator for the categorization. The central element of this component is the OntoGazetteer or semantic tagger, that is included in the information extraction system ANNIE[13]. It is not directly applicable to the SKOS CompleXys taxonomy, but can make use of a derived, rule-based version. Therefore, every main category of the domain model gets its own *.lst* gazetteer file, wherein all subordinate terms are listed one per line. A file *mappings.def* defines the mapping rules from the *.lst* files to SKOS concepts. However, the expressiveness of the gazetteer data is very limited, so the relationships can not be transformed.

The KEA Annotator also categorizes a document into the concepts of the CompleXys domain model. It is based on the Keyphrase Extraction Algorithm KEA[14], that analyzes texts in order to identify the most important words or word groups for each one. The idea of leveraging this behavior for the task of semantic data extraction is, that KEA is implicitly capable of scoring terms according to their text importance. While the OntoGazetteer is capable of answering the question "Do taxonomy terms occur in the text and how often?", KEA goes one step further and additionally tries to answer "Are these terms relevant for the text?". In order to do so, it utilizes additional factors like the relative term occurrence in a single text, compared to the occurrence in all processed texts or the SKOS *related* relationships as weight boosting functions. To ensure that the keyphrases are matchable to the domain model anyway, it simply uses the CompleXys taxonomy as a controlled vocabulary for the extraction process. To use this functionality the older KEA GATE plugin was manually adapted to the new KEA version 5.0, which allows the controlled indexing. As categorization model CompleXys is trained with the CiteULike-180 data set[15]. First evaluations indicate, that a well adjusted KEA Annotator is capable of outperforming the competing OntoGazetteer solution by means of precision.

The Open Calais Annotator utilizes the OpenCalais[16] metatagging web service to semantically annotate named entities, events and facts in the text. The so obtained data is not yet used for the domain categorization, but links the data to the wide external set of Calais' stored semantical knowledge base. To exploit these relations has great potential in further improving the categorization, but also for other features like enriching the displayed resources in the front end with additional information.

Finally the Content Writer ensures, that every document is correctly stored in the *Semantic DB*, before the pipeline terminates. However, it also checks if a document has actually exceeded the critical threshold of Onto Gazetteer

[13] http://gate.ac.uk/sale/tao/splitch6.html

[14] http://www.nzdl.org/Kea/index.html

[15] 11.01.2010: http://maui-indexer.googlecode.com/files/citeulike180.tar.gz

[16] http://www.opencalais.com/

annotations or Kea annotations, that marks the relevancy for the domain of complexity. If a document fails to pass this test, it is deleted. Furthermore, the annotations of a document are counted and mapped to their corresponding main categories. The document is ultimately regarded as being a member of the most frequently occurring categories.

5 Conclusion and Future Work

In this paper, we have described the CompleXys portal, an information system about complexity, as an example for a system that allows to automatically aggregate, semantically annotate and filter content stemming from a wide variety of sources. We believe that in times of a rapidly growing amount of content that is being dynamically created in ever increasing rates, such systems are an absolute necessity to ensure that users do not “drown in information” and at the same time do not miss relevant information. Only with such intelligent support will we be able to take advantage of this information revolution.

Up to now, the parts of CompleXys dealing with information harvesting, annotating and filtering have been implemented. A first evaluation shows that CompleXys reaches, indeed reasonable precision and recall with acceptable runtime. For more details please refer to [15]. Right now, we are working on integrating our approach to personalization into CompleXys. Once this has been done, the portal will be launched as an information site for the members of the research focus area “Analysis and Management of Complex Systems” at our university and for the general public.

References

1. F. Bakalov, B. König-Ries, A. Nauerz, and M. Welsch. A Hybrid Approach to Identifying User Interests in Web Portals. In *Proc. of the 9th Int. Conf. on Innovative Internet Community Systems*, 2009.
2. F. Bakalov, B. König-Ries, A. Nauerz, and M. Welsch. IntrospectiveViews: An interface for scrutinizing semantic user models. In *Proc. of the 18th Int. Conf. on User Modeling, Adaptation, and Personalization*, 2010.
3. R. Baumgartner, N. Henze, and M. Herzog. The Personal Publication Reader: Illustrating web data extraction, personalization and reasoning for the semantic web. In *Proc. of the 2nd European Semantic Web Conference*, 2005.
4. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust nlp tools and applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.
5. E. Desmontils and C. Jaquin. Indexing a web site with a terminology oriented ontology. In *The Emerging Semantic Web*, pages 181–197. IOS Press, 2002.
6. P. Dolog, N. Henze, W. Nejdl, and M. Sintek. Personalization in distributed e-learning environments. In *Proc. of the 13th Int. World Wide Web Conf.*, 2004.
7. R. Guha, R. McCool, and E. Miller. Semantic search. In *Proc. of the 12th Int. Conf. on World Wide Web*, 2003.

8. S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM - Semi-automatic CREAtion of Metadata. In *Proc. of the 13th Int. Conf. on Knowledge Engineering and Knowledge Management*, 2002.
9. A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2(1):49–79, 2004.
10. M. Kowalkiewicz, T. Kaczmarek, and W. Abramowicz. MyPortal: robust extraction and aggregation of web content. In *Proc. of the 32nd Int. Conf. on Very Large Data Bases*, 2006.
11. M. Lalmas and V. Murdock, editors. *Proc. of the Workshop on Aggregated Search held in conj. with the 31st Int. ACM SIGIR Conf.*, 2008.
12. D. Maynard, V. Tablan, K. Bontcheva, H. Cunningham, and Y. Wilks. MUSE: a MUlti-Source Entity recognition system. *Computers and the Humanities*, 2003.
13. M.J. Pazzani and D. Billsus. Content-based recommendation systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter 10, pages 325–341. Springer, 2007.
14. B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. KIM a semantic platform for information extraction and retrieval. *Natural Language Engineering*, 10(3-4):375 – 392, 2004.
15. O. Schimratzki. An approach for semantic enrichment of social media resources for context dependent processing, Diploma Thesis, University of Jena, 2010. Fulltext: http://www.minerva-portals.de/publications/theses/an-approach-for-semantic-enrichment-of-social.
16. M. Shilman. Aggregate documents: making sense of a patchwork of topical documents. In *Proc. of the 8th ACM Symp. on Document Engineering*, 2008.
17. Y. Wanga, N. Stash, L. Aroyoa, P. Gorgels, L. Rutledge, and G. Schreiber. Recommendations based on semantically enriched museum collections. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):283–290, 2008.

Trust and Reputation in Social Internetworking Systems

Lora Aroyo[1], Pasquale De Meo[2], Domenico Ursino[2]

[1] Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[2] DIMET, Università "Mediterranea" di Reggio Calabria, Via Graziella, Località Feo di Vito, 89060 Reggio Calabria, Italy
l.m.aroyo@cs.vu.nl,demeo@unirc.it,ursino@unirc.it

Abstract A Social Internetworking System (*SIS*) is the scenario arising when Web users decide to affiliate to multiple social networks. Recent studies show an increasing user tendency of creating multiple identities on different social systems and exposing, in each of them, different traits of their personalities and tastes. This information provides a better picture of user needs and enhances the quality of services they can use. In the next years a large growth of *SIS* phenomenon is foreseeable. In order to boost the level of user participation in a *SIS*, suitable mechanisms capable of discerning reliable users must be designed. We propose a model to represent a *SIS*, a software architecture to gather real data and analyze the structural properties of a *SIS*. In concrete use cases with different contexts and different levels of protection of data, we introduce an ontology-based model to compute trust and reputation in a *SIS*. This research is collaborative effort between the Vrije Universiteit Amsterdam and the University of Reggio Calabria in the context of a Marie Curie Fellowship.

1 Introduction

Social media applications, such as blogs, multimedia and link sharing sites, question and answering systems, wikis and online forums, are growing at an unprecedented rate and are estimated to generate a significant amount of the contents currently available on the Web [11]. Social media applications are a significant part of a more meaningful kind of applications, named *Web 2.0 applications*, which aim to provide a platform for information sharing and collaboration among users on the Web.

In social media applications, users form communities, typically modelled as *social networks*. Users are driven to get in touch and become friends of other users, create and publish their own contents (like videos or photos), share these contents with others, rate and comment contents posted by others. Examples of popular Web-based social networks are Facebook, MySpace and LinkedIn.

The value of social networks expresses in multiple ways. For instance, users may take an advantage of their interactions with other users to find information relevant to them or they can explore connections existing in a social network to get in touch with user with whom they may profitably interact: many Web users, as an example, indicate that they were able to get a job through their contacts in LinkedIn[1].

A further advantage is that social networks allow to disseminate new knowledge in a widespread fashion, to diffuse innovations, to spread opinions (e.g., social or political messages) among members, to advertise new products [13].

The power of social networks has been fully recognized by institutional actors like museums, TV broadcaster, academic and government institutions. For instance, the Rijksmuseum Amsterdam is exploring the added-value of providing artworks online, allowing users to express their opinions on them or contribute to describing artwork's. Furthermore, major European broadcasters, such as BBC and RAI are experimenting with Web 2.0 technologies to improve interactivity and participation of TV consumers.

Users often decide to affiliate to multiple social networks: for instance, in a recent survey, Ofcom found that 39% of UK adults with at least one social networking profile has indeed two or more profiles[2]. We call *Social Internetworking System* (hereafter SIS) the scenario arising when many users decide to affiliate with multiple social networks. Companies are discovering the potential of social internetworking and are promoting systems capable of supporting social internetworking tasks. Popular examples of commercial social internetworking systems are *FriendFeed* and *Gathera*.

The main goal of these systems is to offer a technological platform to ensure data portability among different social networks. The major bottleneck for the success of a SIS is the absence of mechanism that helps users in finding other "reliable" users with whom they can profitably interact and discloses the presence of malicious users or spammers.

In the past, significant research efforts have been done to define and handle trust and reputation, as a large body of literature highlights [1,9,10,12,16,19].

However, in our opinion, several reasons suggest a further investigation. A first research question is to provide a model capable of representing a SIS, its components and their relationships. In addition, it is necessary to gather real data about a SIS in order to understand its structural properties and clarify to what extent a SIS differs from traditional social networks.

A second issue depends on the fact that the concepts of trust and reputation may assume different meanings according to the scenario in which a user operates in. For instance, in communities like Question & Answering systems (in which users are allowed to pose questions, to answer questions raised by other users and, finally to rate received answers), the reputation of a user coincides with his level of expertise on a particular topic; in a Web community like YouTube, the reputation of a user coincides with the quality of contents he generated. This

[1] http://www.mainstreet.com/article/career/employment/social-media-job-seeker-s-best-new-tool

[2] http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/socialnetworking/annex3.pdf

requires to define a procedure to compute, in an abstract and general fashion, the reputation of a user and to specialize it in concrete domains.

As a final research challenge, it would be necessary to define a model to represent trust and reputation in different contexts. In addition, it is useful to observe that, in different contexts, different policies for accessing, publishing and re-distributing data may exist. For instance, in the case of a TV broadcaster which delivers online part of its archive of resources, users are allowed for instance to use some resources (e.g., for educational purposes) but are forbidden to re-use those protected by copyright. To address these issues, it would be beneficial to design an ontology to model the key concepts of trust and reputation in different environments characterized by different levels of protection of data.

In this paper we describe our plans to define and handle trust and reputation in the context of a SIS. Our research activities will be carried out in the context of a *Marie Curie Intra-European Fellowships for Career Development (IEF)* project, a funding opportunity provided by EU Commission. The paper is structured as follows: in Section 2 we review existing trust and reputation models and illustrate the challenges arising in a SIS. In Section 3 we illustrate a model to represent the features of a SIS along with a software architecture we are currently implementing to gather real data from a SIS and analyze its structural properties. In Section 4 we provide a general model to compute trust and reputation in a SIS and illustrate the steps we are planning to specialize it in real contexts; in particular, we plan to adapt our notion of trust on data gathered within two research projects, namely *NoTube* [15] and *Agora* [2]. In Section 5 we discuss a possible ontology-based model to represent trust and reputation in environments characterized by different levels of protection of data. Finally, in Section 6 we draw our conclusions.

2 Background and challenging issues

In virtual communities the term *trust* is generally exploited to indicate the reliance that a community member associates with another one. Trust values are "local parameters" in the sense that specifying the trust of a user A toward a user B is equivalent to indicate how much A perceives B as reliable.

The opinion of the whole community of users toward a member of the community itself is known as *reputation.* In the past, the issue of computing and handling trust and reputation in virtual communities has been deeply investigated and several models and approaches to facing it have been proposed. his research them is gaining more an more relevance in the context of Web 2.0 and social recommender systems: for instance, [7] suggests to cluster users on the basis of their trust relationship. Such a methodology, coupled wit a memory-based recommender system is able to yield high quality recommendations.

Here we discuss some of these approaches and outline the challenges we encounter in the context of a SIS. Existing approaches can be classified into two categories:

Graph-Based Approaches. A first category of approaches model a user community as a graph G in which nodes represent users [1,9,10,19]. An edge linking two nodes v and u indicates that the user v explicitly trusts the user u. The graph G is usually sparse because a user typically evaluates a handful of other users; as a consequence, various techniques have been proposed to *infer* implicit trust relationships. In detail, the approach of [1] applies a maximum network flow algorithm on G to compute trust between any pair of users. In [9] the authors apply a modified version of the Breadth First Search algorithm on G to infer multiple values of reputation for each user; these values are then aggregated by applying a voting algorithm to produce a final (and unique) value of reputation for each user. The approach of [10] considers paths up to a fixed length k in G and propagates the explicit trust values on them to obtain the implicit ones. In [19] trust values are computed by applying a spreading activation algorithm.

Graph-based approaches leverage on explicit trust relationships declared between pairs of users. As a consequence, they neglect to consider a broad range of activities that, in a SIS (e.g., the activity of rating resources) are a precious and reliable indicator of trust.

Link-Based approaches. A second category of approaches use ranking algorithms such as PageRank [5] or HITS [14], which have been successfully applied in the context of Web Search, to find trust values. For instance, [12] proposes an approach based on PageRank to measure peer reputation in a peer-to-peer network. The approach of [16] defines a probabilistic model of trust which strongly resembles that described in [12]; however, differently from this last, the approach of [16] computes and handles trust values and not reputation values. In [6] the authors present an algorithm which computes global reputation values in a peer-to-peer network; the proposed algorithm uses a personalized version of PageRank along with information about the past experiences of peers.

Experimental tests indicated that link-based methods can obtain precise results and are often *attack-resistant*, i.e., they can resist to attempts conceived to manipulate reputation scores.

We observe that in some approaches trust is conceived as a *measure of performance*. For instance, in [12], the trust of a peer depends on the success of downloading a file from it and, then, trust depends on parameters like the number of corrupted files stored in the peer or the number of connections with the peer that have been lost. By contrast, in our case, trust should quantitatively encode the confidence of a user in the opinions formulated by other ones.

We can observe that both graph-based and link-based approaches try to model trust and reputation in a "force-mass-acceleration" style. In other words, these approaches try to capture all factors influencing trust and reputation and combine them in a set of equations. The resulting model is too complicated to be handled because a lot of parameters capable of influencing trust should be considered. In our opinion, the assessment of trust and reputation critically depends on the concrete domain in which we are operating in and we believe that an universal model of trust is not possible.

To better clarify this concept, we report some results emerging from the *PrestoPrime* project [4], an EU financed project devoted to study and develop practical solutions for the long-term preservation of digital media objects, programmes and collections.

In the context of PrestoPrime, two pilot demonstrators were developed. In the first one, in a game-like environment, users were asked to label videos by applying simple keywords (*tags*). Experiments with users showed that a satisfactory measure of trust between a pair of users who do not know each other can be obtained by considering the tags they apply to label a video and computing the degree of match of the set of tags they inserted.

In the second demonstrator, users were provided with a small annotation environment allowing them to label museum objects with four main entry fields (i.e., "who", "where", "what", and "when"). This allowed us to create links between museum objects on the basis of the key dimensions "who", "where", "what", and "when"; as an example, objects coming from different museum collections can be tied if they refer to the same artistic and historical context and this produces a more complete description of cultural movements.

The notion of trust developed in the context of the first demonstrator is not applicable for the second one, and other factors influencing trust and reputation need to be studied.

A further challenge we are in charge of studying depends on the fact that, in some cases, real organizations often decide to make available on the Web their own resources and often allow end users to enrich their descriptions through metadata like tags. For instance, think of the case of public TV broadcasters like BBC which offers online a large number of contents referring to its TV programmes. Each organization may use different policies for accessing, distributing and labelling the contents they produce and disseminate. For instance, a digital content may published online only in some specific cases (e.g., if the material must be used in education) while its usage is forbidden for commercial purposes. This proves that, in the process of defining trust and reputation, it is necessary to consider not only the application context but also te level of data protection about available resources.

3 Defining a basic model of social internetworking

The first goal of our research is to find a suitable model to represent a Social Internetworking System (*SIS*) and interactions between humans that can take place in it.

To this purpose, our model must fit two requirements:

- *Requirement 1.* The model should be rich enough to represent a wide range of *heterogeneous entities* (i.e,. users, resources, posts, comments, ratings, and so on) and their *interactions* (e.g., users may declare to be friends or they may rate resources).
- *Requirement 2.* The model should be easy to manipulate and intuitive.

Clearly, Requirements 1 and 2 are conflicting each other and a suitable trade-off is compulsory. Traditional approaches to modelling social networks are usually based on *graphs*. Nodes in graphs represent social network actors (e.g., users) while edges identify relationships between them.

We believe that graph-based models are not satisfactory in the context of a SIS for several reasons.

A first weakness relies on the role the nodes would have if we would decide to represent a SIS through a graph. Generally, a social network consists of *homogeneous* nodes, i.e., all nodes represent objects sharing the same nature. As claimed by Requirement 1, in a SIS *heterogeneous* entities co-exist and these heterogeneities must be properly modelled.

A further limitation is that graph-based models are able to represent *one-dimensional networks*, i.e., edges of a graph specify that only *one* particular kind of relationship may exist between nodes. On the contrary, we expect that a SIS should be represented through a *multi-dimensional* network because various type of interactions may involve entities of the same type or of different nature: for instance, an edge should link a user u and a resource r to indicate that u has posted r or an edge should tie two users to indicate that they declared to be friends.

Finally, edges in graphs highlight *binary relationships* between nodes they link. In a SIS, it could be useful to consider n-ary relationships (e.g., an edge may glue together a user u, a resource r and a tag t under the hypothesis that u applied t to label r).

We are currently studying a more sophisticated model in which a SIS is represented through an *hypergraph* such that: *(i)* nodes are labelled and the label of a node reflects the nature of the object represented by the node itself; *(ii)* multiple hyperedges may run between two nodes to indicate that multiple interactions may take place between two arbitrary entities; *(iii)* hyperedges denote relationships involving two or more entities.

In addition to defining a model to represent a SIS, we are also interested in gathering data from real social networks in order to understand the properties showed by a real SIS. For instance, it would be interesting to check whether properties typical of real social networks (e.g., the small world phenomenon) still emerge in a SIS.

Such a task is quite complex because, in different networks, a user may have different identities so it would be extremely hard to join information scattered across multiple networks.

To address this issue, we used the *Google Social Graph API* [3]. Social Graph API allows human users or software applications to access public connections between people on the Web. In particular, Social Graph API can be queried through an HTTP request and is able to return two kind of results:

- *A list of public URLs that are associated with a person.* For instance, given a user u, Social Graph API reveals the URLs of the blog of u and his Twitter page.

– *A list of public declared connections between people.* For instance, it returns the list of persons who, in at least one social network, have a link to the blog of u or any other page referable to u.

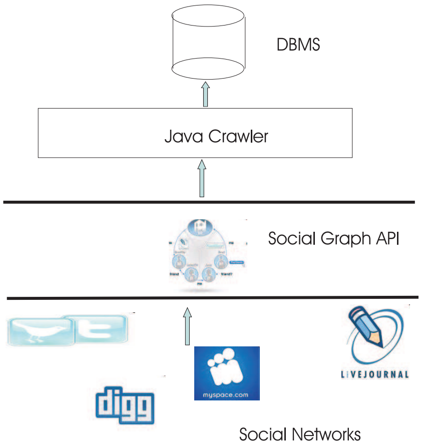


Figure 1. The architecture of our crawler

At the moment of writing, we have implemented a simple crawler to explore and gather data from a SIS. The architecture of our system is shown in Figure 1. In particular, a Java crawler invokes Social Graph API by sending a *seed URL* associated with a user u; the API sends back the list of people who are somewhat related with u. The Java crawler gets these data and launches a Breadth-First-Search like procedure to find new URLs and connections. Retrieved results are permanently stored in a relational DBMS implemented in MySQL.

4 Defining reputation in a Social Internetworking System

As a further step, we are interested in studying a model of reputation for a SIS. The notion of trust/reputation in the context of a SIS (and, in general, for Web 2.0 applications) is hard to define because, as shown in Section 2, in different contexts, they may assume different meanings.

We propose a methodology to compute reputation in a SIS which operate in *two* stages: in the former stage we review and analyze the factors capable of influencing the value of reputation in a SIS. In the latter stage we consider some concrete domains and specialize our methodology to them.

Intuitively, we assume that the reputation of a user depends on the following facts:

The reputation of a user depends on the relationship he created in the SIS. We suggest to use the hypergraph model introduced in Section 3 to represent users and their relationships in a SIS. Past user interactions are analyzed to determine the level of trust a user u confers to a user v and this information is used to weight edges in the hypergraph representing the SIS. Since a user generally interact with an handful of other users, the hypergraph we obtain is *sparse* and a suitable algorithm to propagate trust values is necessary. Currently, we are planning to use a link-based algorithm like PageRank. At the end of this step, we are able to generate a vector $\mathbf{r}'$ such that the i-th component of $\mathbf{r}'$ equals to the reputation of the i-th user.

Users with high level of reputation are also those who produce high quality resources. The quality of a resource could be computed by consider the average rating it got and, then, resources with a high average rating are also high quality resources. To avoid biases, we can pose a further requirement: the number of ratings received by a resource must be *statistically significant*, i.e., we can consider only resources which received at least N_{min} ratings, being N_{min} a suitable threshold. Such a requirement would avoid that resources evaluated by a small number of users are deemed better than resources rated by a large mass of human users.

The procedure described above resembles that applied in many social systems like YouTube or Digg to evaluate the quality of a resource. We believe that such a procedure is affected by several fallacies and it may incur in harsh inaccuracies. In fact, spam or malicious users may tend to provide generous evaluations to artificially inflate the evaluation of a resource. As a consequence, we need a more complicated framework capable of putting together the reputation of users, the quality of resources they post and the evaluations associated with resources. At the current stage of the project we are considering, as a possible solution, the following criterium:

> *A user has a high reputation if he authors high quality resources. A resource, in its turn, is of high quality if it gets a high average rating and it has been posted by users with high reputation.*

The intuition provided above relies on a *mutual reinforcement principle* that is similar, to some extent, the approach underlying HITS [14] algorithm. The principle outlined above easily turns into a set of linear equations. In fact, let n be the number of users composing a SIS and let m be the number of resources they authored. Let $\mathbf{r}''$ be an n-th dimensional array such that the i-th entry of $\mathbf{r}''$ equals to the reputation (to compute) of the i-th user and let $\mathbf{q}$ be an m-th dimensional array such that the j-th entry of $\mathbf{q}$ equals to the quality (to compute) of the j-th resource. Finally, let $\mathbf{e}$ be an m-th dimensional array such that the j-th entry of $\mathbf{e}$ equals to the average rating of the j-th resource and let $\mathbf{A}$ be an n-by-m matrix such that $\mathbf{A}_{ij}$ equals 1 if the i-th user posted the j-th resource and 0 otherwise.

According to this notation, we can write the following equations:

$$\mathbf{r}'' \propto \mathbf{A}\mathbf{q} \qquad (1)$$

$$\mathbf{q} \propto \mathbf{A}^T \mathbf{r}^{''} + \mathbf{e} \tag{2}$$

In both Equations 1 and 2, the symbol $\propto$ means "is proportional to". As for Equation 1, the i-th row of the the product $\mathbf{Aq}$ specifies the sum of the qualities of the resources authored by the i-th user. This immediately follows from the definition of product between a matrix and a vector. Interestingly enough, the $\mathbf{A}$ matrix can be interpreted as the adjacency matrix of a bipartite graph whose nodes represent users and resources and edges link a user to the resources he authored. In Equation 2, the symbol $\mathbf{A}^T$ is the *transpose* of $\mathbf{A}$. As in the previous case, $\mathbf{A}^T$ matrix can be viewed as the adjacency matrix of a bipartite graph whose nodes represent resources and users and edges link a resource to the user who authored it. Observe that the same model holds if we assume that a resource has been posted by one user or it has been posted by multiple users. The product $\mathbf{A}^T\mathbf{r}^{''}$ is an m-th dimensional vector whose j-th entry specifies the reputation (or the sum of the reputations) of the user (users) who posted the j-th resource.

By plugging Equation 2 into Equation 1 we obtain:

$$\mathbf{r}^{''} \propto \mathbf{A}\left[\mathbf{A}^T\mathbf{r}^{''} + \mathbf{e}\right] \Rightarrow \mathbf{r}^{''} - \mathbf{A}\mathbf{A}^T\mathbf{r}^{''} \propto \mathbf{Ae} \Rightarrow$$

$$\Rightarrow \mathbf{r}^{''}\left[\mathbf{I} - \mathbf{A}\mathbf{A}^T\right] \propto \mathbf{Ae} \Rightarrow \mathbf{r}^{''} \propto \left[\mathbf{I} - \mathbf{A}\mathbf{A}^T\right]^{-1}\mathbf{Ae}$$

Since $\mathbf{AA}^T$ is *symmetric*, its eigenvalues are real [17]. In particular, $\left[\mathbf{I} - \mathbf{AA}^T\right]^{-1}$ can be easily and effectively approximated by computing the *dominant eigenvector* of $\mathbf{AA}^T$. Such a result is of great practical impact because there exist efficient numerical methods to compute dominant eigenvector of a symmetric matrix (think of Lanczos method [17]) and, then, our methodology is suitable also if the size of $\mathbf{A}$ gets very large; such a case if quite common in real cases because, in traditional social sites the number of users and resources they generate (which correspond to the number of rows and columns of $\mathbf{A}$) is huge.

Finally, we merge the arrays $\mathbf{r}'$ and $\mathbf{r}^{''}$ into a single reputation value $\mathbf{r}$ as follows:

$$\mathbf{r} = \alpha\mathbf{r}' + (1-\alpha)\,\mathbf{r}^{''} \tag{3}$$

The coefficient $\alpha \in [0, 1]$ is instrumental in weighting the contributions coming from link analysis and the analysis of resources generated by a user. We plan to tune α by applying a linear regression technique.

Once a theoretical model of reputation in a SIS has been defined, our intention is to specialize it in concrete domains. In particular, we are interested in monitoring and analyzing the behaviour of users in long-term experiments associated with different domains; the notion of reputation, from abstract concept turns into a concrete tool to aid user in better taking advantage of potentialities offered by the SIS. Experiments on real users allow us to get an iterative assessment of the strengths and weaknesses of our notion of reputation as well as indications for improvement.

To this purpose, we will use data gathered in the context of two research projects, namely: *(i) NoTube* (an EU financed project on interactive television) [15], and *(ii) Agora* [2] (a Dutch funded project on digital museums and audio-visual archives). In the context of interactive television, trust/reputation values represent the level of expertise of a user. This information could be exploited to select in a personalized fashion contents to propose to the user. As for digital museums we can study what parameters in the user behaviour are relevant for producing authoritative annotations and what are the motivations for users to participate in this labelling process. This information could be instrumental in better using human mass potential in annotating artworks.

5 Building an ontology-based model of trust and reputation in a Social Internetworking System

Once we have defined the concept of trust and reputation in concrete domains it is advantageous to create a model capable of representing them in different domains. To this purpose we plan to design an *ontology* capable of specifying how reputation and trust specialize in different application contexts.

To the better of our knowledge, there are few approaches to designing ontologies to model trust. For instance, in [8], *TrustOntology* is presented. This is an OWL ontology allowing each user to indicate the people he trusts. Trust information is automatically composed to infer new values of trust for newcomer users. In [18], the authors suggest a trust protocol in which the decision about the trustworthiness of a message depends on many factors like the creator (*who*) of the message (*what*), time (*when*), location (*where*), and intent (*why*). An ontology to capture factors influencing trust and a set of functions to evaluate trust is presented.

Our goal is different from that of [8] and [18] we want to model how reputation specializes in different contexts. In addition our ontology can be used to represent a scenario in which different organizations decide to make available on the Web their own resources. An organization (e.g., a cultural institution) may let its users to freely use, copy and re-distribute available resources. Another organization may apply different policies to protect data because some resources can be freely disseminated, other resources are not accessible because protected by copyright and, finally, some resources can be published online and re-used for some purposes (e.g., as educational material) but their access is forbidden in other cases.

The availability of such an ontology would offer us, on the long run, the possibility of designing complex software applications running across multiple social networks. For instance, we can think of a content-based recommender system operating as follows:

1. A user issues a query.
2. The query is forwarded to multiple social system and a list of resources matching the query is retrieved by each social system.

3. Retrieved resources are ranked on the basis of the reputation of the users who created, on the application context and on the rights for its distribution.
4. A global list is produced by merging the previous ones.

Such an application is, in our opinion, capable of introducing relevant novelties in the research field of Recommender Systems. In fact, the proposed application is able to sift through different social sites (while traditional Recommender Systems usually operate on a single resource repository) and is able to rank resources on the basis of multiple and criteria.

6 Conclusions

In this paper we introduce the concept of Social Internetworking System, i.e., the scenario arising when Web users decide to affiliate to multiple social networks. We propose a model to represent a SIS and describe the main components of a software architecture we are implementing to gather real data from a SIS and analyze its structural properties. In concrete use cases with different contexts and different levels of protection of data, we introduced an ontology-based model to compute trust and reputation in a SIS. This research is collaborative effort between the Vrije Universiteit Amsterdam and the University of Reggio Calabria in the context of a Marie Curie Fellowship.

In the future we plan to gather a large amount of data about a SIS and carry out an empirical study on them. The goal is to understand whether some properties of real social networks (like small world phenomenon, power law distribution of in-degree and out-degree distributions, and so on) if they are still confirmed in a SIS or if significant deviations emerge.

A further research line is to carry out a detailed review of existing literature on the meaning of trust and reputation in different social site. Finally, we plan to test the effectiveness of our ontology-based model with an experiment involving real users. In particular, the validation phase will be strictly tied to the activity of designing our ontology; in fact, we shall use feedbacks provided by users to revise the structure of our ontology.

References

1. Advogato's trust metric. `http://www.advogato.org/trust-metric.html`, 2000.
2. Agora: Creating the historic fabric for and providing web-enabled access to objects in dynamic historical sequences. `http://agora.cs.vu.nl/`, 2010.
3. Google Social Graph API. `http://code.google.com/intl/it-IT/apis/socialgraph/`, 2010.
4. Prestoprime: Keeping audiovisual contents alive. `http://www.prestoprime.org/`, 2010.
5. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

6. P. Chirita, W. Nejdl, M. T. Schlosser, and O. Scurtu. Personalized Reputation Management in P2P Networks. In *Proc. of the ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*, CEUR Workshop Proceedings, Hiroshima, Japan, 2004. CEUR-WS.org.
7. T. DuBois, J. Golbeck., J. Kleint, and A. Srinivasan. Improving recommendation accuracy by clustering social networks with trust. In *proc. of the ACM RecSys'09 Workshop on Recommender Systems & the Social Web*, 2009.
8. J. Golbeck. Trust ontology. `http://www.schemaweb.info/schema/SchemaDetails.aspx?id=171`, 2010.
9. J. Golbeck and J.A. Hendler. Inferring binary trust relationships in web-based social networks. *ACM Transactions on Internet Technology*, 6(4):497–529, 2006.
10. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proc. of the International Conference on World Wide Web (WWW '04)*, pages 403–412, New York, NY, USA, 2004. ACM.
11. P. Heymann, G. Koutrika, and H. Garcia-Molina. Can Social Bookmarks Improve Web Search? In *Proc. of International Conference on Web Search and Data Mining (WSDM 2008)*, pages 195–206, Stanford, California, USA, 2008. ACM Press.
12. S.D. Kamvar, M.T. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. In *Proc. of the International Conference on World Wide Web (WWW 2003)*, pages 640–651, Budapest, Hungary, 2003. ACM Press.
13. J. Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51(11):66–72, 2008.
14. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
15. L. Nixon and L. Aroyo. NoTube – Making TV a Medium for Personalized Interaction. In *Proc. the European Interactive TV Conference (EuroITV 2009)*, pages 22–25, Leuven, Belgium, 2009. University of Leuven.
16. M. Richardson, R. Agrawal, and P. Domingos. Trust Management for the Semantic Web. In *Proc. of International Conference on Semantic Web (ISWC 2003)*, pages 351–368, Sanibel Island, FL, USA, 2003. Lecture Notes in Computer Science, Springer.
17. G.W. Stewart. *Matrix Algorithms: Basic Decompositions (Volume 1)*. Society for Industrial Mathematics, 1998.
18. S. Toivonen and G. Denker. The Impact of Context on the Trustworthiness of Communication: An Ontological Approach. In *Proc. of the ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*, volume 127 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2004.
19. C. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, 2005.

SAS WEB 2010

http://ailab.dimi.uniud.it/en/events/2010/sasweb