Journal of Universal Computer Science, vol. 16, no. 16 (2010), 2196-2213
submitted: 31/3/10, accepted: 31/7/10, appeared: 28/8/10 © *J.UCS*

Extraction of Contextualized User Interest Profiles in Social Sharing Platforms

Rafael Schirru
(German Research Center for Artificial Intelligence, Germany and
University of Kaiserslautern, Germany
schirru@dfki.de)

Stephan Baumann
(German Research Center for Artificial Intelligence, Germany
baumann@dfki.de)

Martin Memmel
(German Research Center for Artificial Intelligence, Germany and
University of Kaiserslautern, Germany
memmel@dfki.de)

Andreas Dengel
(German Research Center for Artificial Intelligence, Germany and
University of Kaiserslautern, Germany
dengel@dfki.de)

Abstract: Along with the emergence of the Web 2.0, E-learning more often takes place in open environments such as wikis, blogs, and resource sharing platforms. Nowadays, many companies deploy social media technologies to foster the knowledge transfer in the enterprise. They offer Enterprise 2.0 platforms where knowledge workers can share contents according to their different topics of interest.

In this article we present an approach extracting contextualized user profiles in an enterprise resource sharing platform according to the users' different topics of interest. The system analyses the social annotations of each user's preferred resources and identifies thematic groups. For every group a weighted term vector is derived that represents the respective topic of interest. Each user profile consists of several such vectors that way enabling recommendation lists with a high degree of inter-topic diversity as well as targeted context-sensitive recommendations.

The proposed approach has been tested in our Enterprise 2.0 platform ALOE. A first evaluation has shown that the method is likely to identify reasonable user interest topics and that resource recommendations for these topics are widely appreciated by the users.

Key Words: user modeling, topic detection, Web 2.0 resource sharing, E-Learning 2.0, Enterprise 2.0

Category: H.3.1 - Content Analysis and Indexing

1 Introduction

With the emergence of the Web 2.0 a shift in user behavior has been observed. An increasing amount of users are not only consumers but also producers of content. They access and share bookmarks, photos, videos, and the like in different kinds of Web 2.0 platforms. The changing user attitudes also affect the way people learn. Two trends in E-Learning can be observed: Firstly, E-Learning today often takes place in open environments. E. g., when facing a task for which information has to be acquired, a knowledge worker might seek help from blogs and use the blog infrastructure to directly discuss solutions with her colleagues. The second trend builds on the idea of staff members making their knowledge explicit about dealing with particular tasks and problems. More and more companies offer resource sharing platforms, wikis, and the like where employees can share information.

1.1 Influences of the Web 2.0 on E-Learning

The participative character of the Web 2.0 ([O'Reilly, 2005]) has influenced modern E-Learning. Subsequently we'll describe the concept of E-Learning 2.0 as learning taking place in open environments where learners are not only consumers but also producers of learning content. Afterwards the idea of the Enterprise 2.0 is presented as an aggregation of Web 2.0 technologies helping to foster the knowledge transfer in the enterprise. Enterprise 2.0 platforms are the main field of application for our presented approach as they allow the knowledge workers to share resources according to their different topics of interest.

In 2005 Stephen Downes coined the term *E-Learning 2.0* to account for learning taking place in open platforms ([Downes, 2005]). In that time E-Learning content mostly consisted of learning objects (little building blocks) that could be put together or organized. Standards bodies provided specifications on how to sequence and organize the learning objects into courses and how they should be packaged for delivery. The dominant learning technology employed was the learning management system (LMS) which took learning content and organized it as courses divided into modules and lessons. Such systems do no longer fulfill the needs of today's internet users who approach work, learning, and play in a different manner. The digital natives capture information from text, images, audio, and video from different sources in parallel. According to Downes they prefer on-demand access to media, constantly communicate with their friends and are as likely to create their own content as to purchase a book or a CD. Student-centered designs manifest these trends. They are characterized by greater autonomy for the learner and emphasize active learning (including creation, communication, and participation). The technologies being used are well-known from the Web 2.0 and incorporate blogs, podcasting, wikis, as well as resource sharing platforms.

The concept of the *Enterprise 2.0* has been introduced by McAfee as a collection of Web 2.0 technologies for generating, sharing, and refining information ([McAfee, 2006]). Companies can buy or build these technologies in order to uncover the practices and outputs of their knowledge workers. His proposed SLATES framework consists of the following six components:

- **Search:** McAfee cites a Forrester study ([Morris et al., 2005]) which revealed that less than 50% of the intranet users reported to find the content they were looking for. Searches on the internet however are more likely to lead to successful search experiences (87%). This indicates that besides good intranet page layouts and navigation aids, there is a demand for improved keyword search on many platforms.

- **Links:** Google showed that the exploitation of the link structure between web pages can significantly improve search results ranking. Intranets could also profit from this approach however it requires that many people can add links, not only the small group of people that develop the portal.

- **Authoring:** The example of Wikipedia has shown that group authorship can have convergent, high-quality content as output. In enterprises blogs and wikis should enable every staff member to share knowledge, insights, experiences, and the like.

- **Tags:** Besides improved keyword search, the study found that staff members would appreciate an improved categorization of content. Web 2.0 resource sharing platforms usually collect a large amount of resources and outsource the process of categorization (tagging) to their users. In enterprise platforms this could reveal patterns and processes in knowledge work by means of social navigation (see which tags the colleagues used, which pages they visited, ...).

- **Extensions:** Often tagging is extended by automating categorization and pattern matching. Recommender systems serve as a well-known example. Based on the preferences a user expressed in the past, they recommend resources with similar content, resources that are preferred by the user's peers and the like.

- **Signals:** Checking the intranet for new content of a certain topic regularly is a tedious task. Feed technologies such as RSS and Atom can be used to inform the users of new content matching their topics of interest automatically.

1.2 Outline

Our work focuses on the automatic identification of a user's different topics of interest in Enterprise 2.0 resource sharing platforms. We assume that knowledge workers share resources according to their interest topics in such systems. Our approach analyzes the metadata of the knowledge workers' preferred resources and identifies thematic groups, from which contextualized user interest profiles are built. In the short-term we can use these profiles to assemble recommendation lists with a high degree of inter-topic diversity by recommending items from a user's different topics of interest. In [Ziegler et al., 2005] it has been shown that a high level of topic diversification in recommendation lists leads to an improved user satisfaction with the recommender system. In the long-term contextualized user profiles should be used to provide context-sensitive recommendations that meet the user's current needs and preferences.

We tested our method in the ALOE system ([Memmel and Schirru, 2007]), a resource sharing platform which is currently developed in the Knowledge Management group of DFKI. A first evaluation has shown that our approach is likely to identify reasonable user interest topics. Also the users widely agreed that they would appreciate resource recommendations according to these topics.

The remainder of this paper is structured as follows: Section 2 describes related work in the field of topic-based recommender systems and topic detection and tracking (TDT). It further reports about our previous work on a recommender system based on user interests in the C-LINK system. In Section 3 our proposed approach is presented. We depict our idea of contextualized user interest profiles and describe how they can be obtained by applying algorithms from the domain of TDT. Next in Section 4 we present the use case in which our method has been tested and report on preliminary evaluation results in Section 5. In Section 6 we summarize our findings and present ideas for future work.

2 Background

The higher-level goal of our work is the provision of resources recommendations matching a knowledge worker's different topics of interest. For that purpose we extract contextualized user profiles by applying algorithms from the domain of topic detection and tracking (TDT) in our first step. The current Section describes related work in the field of topic-based recommender systems and TDT [see Section 2.1]. Further we report on our previous work on recommendations based on user interests in the C-LINK project [see Section 2.2].

2.1 Related Work

A user modeling approach that takes a user's different topics of interest into account is presented in [Middleton et al., 2001]. Middleton et al. describe the

Quickstep recommender system which unobtrusively monitors the browsing behavior of its users. The target users of the systems are scientists that need to be informed about new papers in their field of interest as well as older papers relating to their work. The system applies supervised machine-learning coupled with an ontological representation of topics to elicit user preferences. It uses a multi-class behavioral model with classes representing paper topics that way allowing domain knowledge to be used when the user profile is constructed. The system works as follows: User browsing behavior is monitored unobtrusively via a proxy server that logs every URL browsed during the user's working activity. Overnight, a machine-learning algorithm classifies browsed URLs and saves the classified papers in a paper store. The interest profile is derived from explicit feedback and browsed topics. Recommendations are computed based on the user's current topics of interest and the classified paper topics. The generated recommendation lists contain items from the user's three most current topics of interest.

[Ziegler et al., 2005] aim at improving topic diversification by balancing top-N recommendation lists according to the users' full ranges of interests. In their recommender system each item is associated with features from a domain taxonomy like, e. g., author, genre, and audience in the domain of books. The proposed algorithm takes a top-N recommendation list and selects a (much) smaller subset of items with a low degree of intra-list similarity. The final recommendation list is built by gradually adding items that keep intra-list similarity low and are recommendable according to traditional collaborative filtering algorithms. The approach presented by Ziegler et al. assumes that features from a domain taxonomy are annotated for each item. In Enterprise 2.0 platforms such features are not always available as resources are contributed by the community of users and many systems do not want to place the burden of annotating content with concepts from a formal taxonomy on the users. Instead they rely on lightweight approaches such as tagging in order to classify content.

To detect the topics in the metadata profiles of the users' preferred resources our system uses algorithms from the domain of topic detection and tracking. TDT is concerned with finding and following new events in a stream of documents. In [Allan et al., 1998] the following TDT tasks have been identified: First is the segmentation task, i. e., segmenting a continuous stream of text into its several stories. Second, there is the detection task which comprises the retrospective analysis of a corpus to identify the discussed events and the identification of new events based on on-line streams of stories. Third is the tracking task where incoming stories are associated with events known in the system. In this work we focus on the detection of topics in the profiles of the users' preferred resources.

[Schult and Spiliopoulou, 2006] consider the problem of finding emerging and persistent themes in accumulating document collections which are organized in rigid categorization schemes such as taxonomies. They propose ThemeFinder, an algorithm for monitoring evolving themes from accumulating document collections. The algorithm works as follows: In the first period, it clusters all documents in the collection. In the following periods, it clusters the new documents with the old feature space and compares the new clusters to the ones found in the previous period. If the clusters of two adjacent periods are similar with regard to their themes and if the quality of the clustering is not declining significantly, then the original feature space is kept. Otherwise a new feature space is build for the documents of the latest period and the next comparison. Thematic clusters are represented by a label, consisting of a set of terms that have a minimal support in the associated cluster. Thematic clusters that survived over several periods, despite re-clustering and changes of the feature space, will become part of the classification scheme.

2.2 Our Previous Work

The idea to automatically extract a user's different topics of interest has risen from the experiences we made in the C-LINK project in 2008. The C-LINK system is a Web 2.0 conference organization system that has been built on top of the ALOE platform [see Section 4]. It is a social sharing tool allowing conference participants to exchange, for instance, material related to their talks. C-LINK also provides social networking facilities such as finding users, e. g., according to their affiliation, exchanging messages, a chat room, and a whiteboard. A content-based recommender system has been integrated into the platform allowing event recommendations as well as recommendations of potentially interesting users based on a user's research topics. Figure 1 shows the welcome page of the C-LINK system.

2.2.1 Recommendations in C-LINK

Content-based recommendations in C-LINK have been realized by integrating three different tools developed at DFKI:

- The *ALOE* platform [see Section 4] is used as the underlying system for resource sharing and collection of social metadata.
- *DynaQ* [Agne et al., 2006] is a desktop search engine for document based personal information spaces. It has a Lucene backend (`http://lucene.apache.org`) thus enabling high-performance, full-featured text search. In C-LINK, DynaQ is used for matching metadata profiles of users and events.

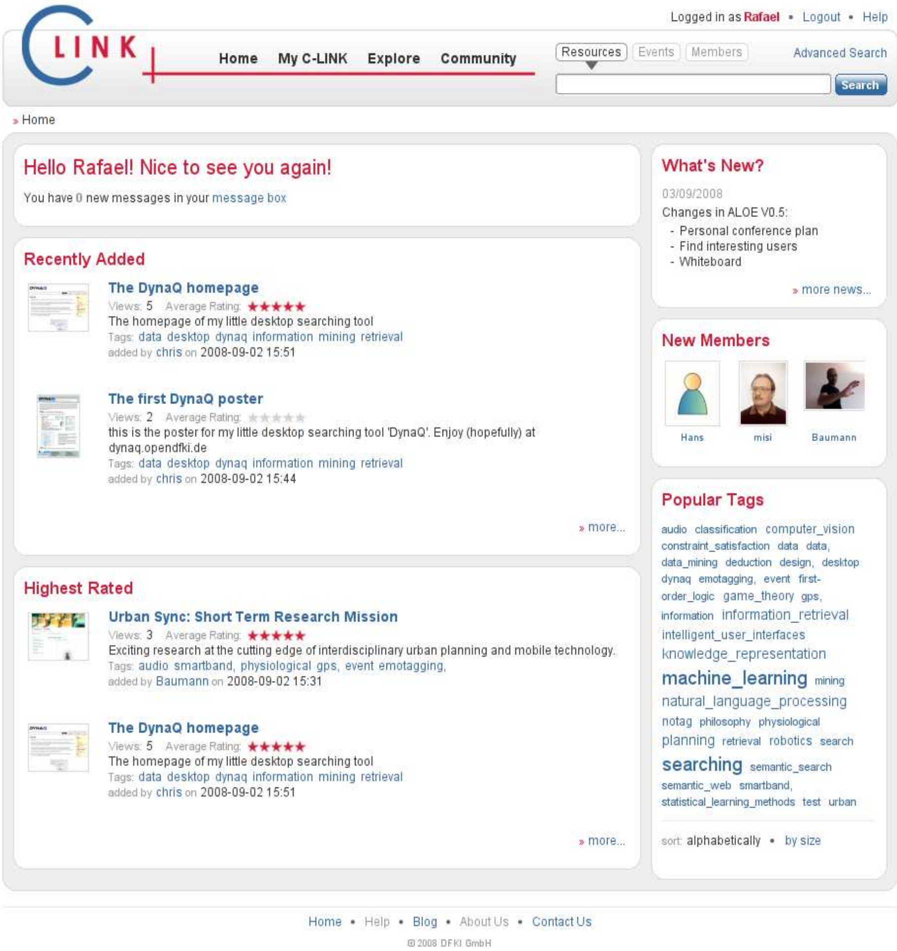

Figure 1: Welcome page of the C-LINK system.

– *MyCBR* [Stahl and Roth-Berghofer, 2008] is an integrated Case-Based Reasoning tool that extends the Protégé ontology editor. In the C-LINK system, MyCBR is used to model the similarities between different research topics.

There are three different kinds of items in the C-LINK system that are relevant for recommendations: resources (i. e., user-contributed contents), users, and events. For each of these items metadata profiles are composed which consist of user-contributed metadata, the full texts of the associated resources (where available) as well as manually annotated research topics. The detailed constitution of the metadata profiles is shown in Table 1.

Resources:	creator, description, title, full text
Events:	research topics (manually annotated), resource metadata profile of associated conference paper
User:	research topic (from user profile), annotated tags, resource metadata profiles of portfolio resources

Table 1: Composition of metadata profiles of resources, events, and users in the C-LINK system.

Whenever a user requests event recommendations, her current metadata profile is determined in the DynaQ backend. The user's research interests are extended by similar research topics as defined in MyCBR. The resulting query is matched against the profiles of the conference events. Finding similar users is performed analogously by extending the current user's metadata profile with related research topics and than matching it against the profiles of the other users in the system.

2.2.2 Review of the C-LINK Approach

Using manually annotated interest topics leads to good recommendation results. However there are two drawbacks of such an approach: First, a domain taxonomy of topics might not be available for every resource sharing platform. In our ALOE system, the users share resources according to their research interests, about software development but also about topics in which they are interested privately. Setting up a domain taxonomy for such an open world scenario might not always be feasible. Second, it is widely recognized that the success of research sharing platforms is among others based on their ease of use. Requiring the users to annotate resources with concepts from a taxonomy aggravates the contribution process and might hinder the usage of the system. For these reasons we aim at an approach that captures the interest topics of the users unobtrusively as a side effect of the normal usage of the system.

3 Proposed approach

Subsequently we'll describe how contextualized user interest profiles can be obtained by applying algorithms for the domain of TDT on the metadata profiles of a user's preferred resources. Section 3.1 describes the requirements that have to be met in order to make our method applicable. Next, in Section 3.2 our idea of contextualized user profiles is refined. Finally, Section 3.3 describes how the profiles can be obtained.

3.1 Requirements

There are two requirements that have to be met in order to make our approach applicable:

First, we need an Enterprise 2.0 *resource sharing platform* in which knowledge workers share and annotate contents according to their different topics of interest. For this purpose we have chosen the ALOE system [see Section 4].

Second, a *textual representation* of the resources in the platform has to be available. The ALOE system supports resource contributions for a variety of content types. Users can either contribute bookmarks to the system or upload files such as office documents, images, audio, and video. Even though automatic approaches to analyze the content of images and videos are currently investigated (e. g., [Ulges et al., 2009] for videos and [Duan et al., 2009] for images), it is still difficult to extract a textual representation of these resources today. In [Li et al., 2008] it has been shown that social metadata is likely to describe the content of resources appropriately. So we decided to exploit the users' annotations to capture the content of the resources in the system.

3.2 Modeling Contextualized User Profiles

As stated by Schwarz ([Schwarz, 2006]), the term *context* is used in different disciplines (e. g., linguistics and psychology) and understood in many different ways. Therefore when talking about context it is necessary to talk about its application as well as the scenario in which it is used. In our system, we assume that a knowledge worker has different topics of interest. E. g., a software engineer might be interested in the Java programming language, the Linux operating system, and in punk rock music. When talking about the user's current context, we refer to the topic of interest that is currently relevant for her.

Our system applies textual data mining techniques on the metadata profiles of the users' preferred resources [see Section 3.3] thus finding thematic groups that represent a users different topics of interest. For every identified topic a weighted term vector consisting of at most ten terms is calculated. The weights are in accordance with the relevance of the associated term for the respective topic. A schematic representation of a contextualized user interest profile is depicted in Figure 2.

A representation of user interests as weighted term vectors meets the requirements of our short-term and long-term goals:

In the *short-term* we aim at providing resource recommendation lists with a high degree of inter-topic diversity. For that purpose we can formulate data base queries where each query consists of the terms of one topic vector. When storing the metadata profiles of the resources, e. g., in a Lucene index, also the term weights can easily be exploited. The final recommendation list can be composed by selecting resources matching different interest topics of a user.

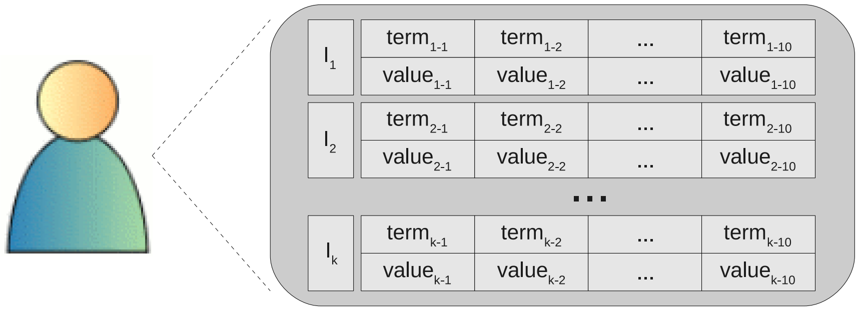

Figure 2: Schematic representation of a user interest profile.

We plan to achieve our *long-term* goal by exploiting the topic vectors to determine a user's current context. This might be achieved by matching the metadata profiles of currently visited resources against the available interest topic vectors of the active user. Whenever a topic is identified as being relevant at the current point in time, recommendations should be generated that meet the user's current needs and preferences.

3.3 Topic Extraction

We identify the knowledge workers' topics of interest by applying textual data mining techniques on the metadata profiles of her preferred resources in ALOE. A resource is considered as preferred when a user has contributed it to the system, added it to her portfolio or has given it a positive rating (i. e., a rating value bigger than three on a five point rating scale).

For these resources, metadata profiles are composed which are worked up and then fed to a clustering algorithm. The process steps of our topic extraction algorithm are depicted in Figure 3 and will be described in detail subsequently:

3.3.1 Data Access

For every preferred resource we compose a profile that consists of the title and the tags that have been annotated by the current user, as these metadata fields are considered to reflect the content of the associated resources appropriately in most cases. Experiments have been conducted that also included the description of the resources. However the best clustering results were achieved when only the titles and the tags of the resources were used.

3.3.2 Preprocessing

We convert the terms contained in the metadata profiles to lower case characters, remove punctuation characters, and stop words. Further stemming is

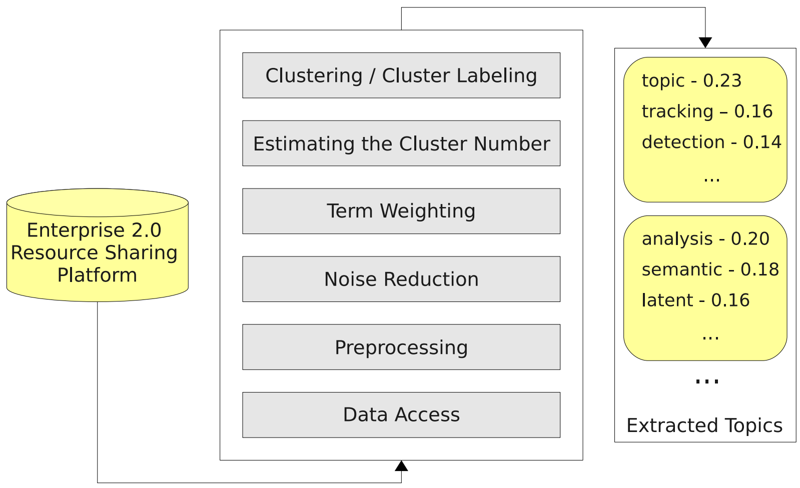

Figure 3: Topic extraction process steps.

applied to bring the terms to a normalized form. We use the Snowball stemmer (`http://snowball.tartarus.org/`) for this purpose. The normalized profiles of the resources are represented according to the "bag-of-words" model, i. e., they are represented as vectors where the features correspond to the terms in the corpus (i. e., the set of the current user's preferred resources) and the feature values are the counts of the words in the respective metadata profiles.

3.3.3 Noise Reduction

Very rare and very frequent terms are not considered helpful to characterize resources. As a consequence dimensions representing these terms are removed. To reduce the noise that is inherent in social metadata we experimented with dimensionality reduction based on Latent Semantic Analysis ([Deerwester et al., 1990]). However the positive impact of the application of this technique still has to be examined in greater depth.

3.3.4 Term Weighting

Terms that appear frequently in the metadata profile of one resource but rarely in the whole corpus are likely to be good discriminators and should therefore obtain a higher weight. We use the TF-IDF measure ([Jones, 1972]) which is widely applied in information retrieval systems in order to achieve this goal.

3.3.5 Clustering and Cluster Labeling

To be able to cluster the set of a user's preferred resources we need to find a reasonable number of clusters in our data first. For this purpose we follow an approach which is based on the residual sum of squares (RSS) in a clustering result. For document clustering and cluster label extraction we apply non-negative matrix factorization.

We *estimate the number of clusters* in the data set as described in [Manning et al., 2009], page 365. First we define a range in which we expect to find the number of topics. We chose a range between 2 and 20 for our experiments, however the boarders are configurable in our algorithm. For each potential cluster size k ($2 \leq k \leq 20$) we run K-Means i-times (we chose $i = 10$), each time with a different initialization. We compute for each clustering the residual sum of squares (RSS) and the minimum RSS over all i clusterings (denoted by $\widehat{RSS}_{min}(k)$). Then we take a look at the values $\widehat{RSS}_{min}(k)$ and search for the points where successive decreases in $\widehat{RSS}_{min}$ become significantly smaller (please note that $RSS_{min}(k)$ is a monotonically decreasing function in k with minimum 0 for $k = N$ with N being the number of documents). The first five such values $k-1$ are stored as reasonable cluster sizes. We store five values in order to enable clusterings according to different granularities. If broad clustering granularity is desired we take the first reasonable number of clusters, for finer granularity the second, and so on.

Using non-negative matrix factorization (NMF) for *document clustering* has firstly been introduced by [Xu et al., 2003]. The authors show that NMF-based document clustering is able to surpass latent semantic indexing and spectral clustering based approaches.

NMF finds the positive factorization of a given positive matrix. It is applied on the term-document matrix representation of the document corpus. In the latent semantic space which is derived by applying NMF, each axis represents the base topic of a document cluster. Every document is represented as an additive combination of these base topics. Associating a document with a cluster is done by choosing the base topic (axis) that has the highest projection value with the document. Formally NMF is described as follows:

Let $W = \{f_1, f_2, ..., f_m\}$ be the set of terms in the document corpus after our preprocessing steps. The weighted term vector X_i of a document is defined as

$$X_i = [x_{1i}, x_{2i}, ..., x_{mi}]^T \tag{1}$$

with x_{ij} being the TF-IDF weights of the terms f_i as described before.

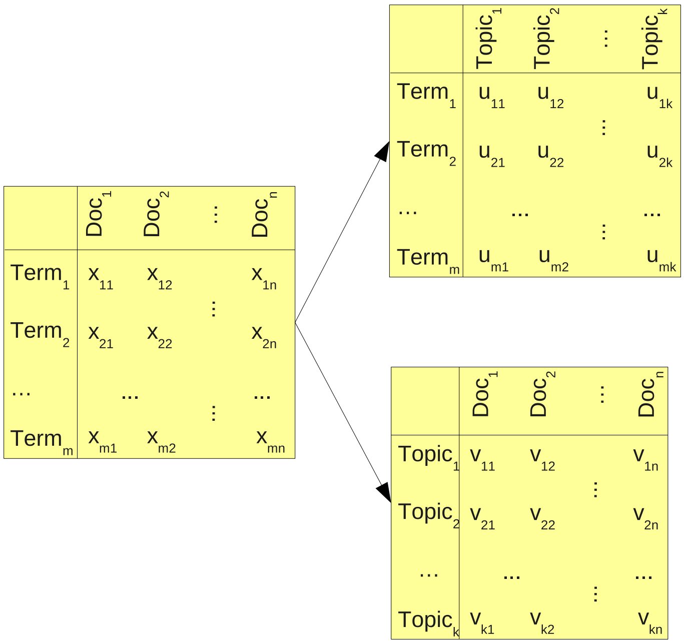

Figure 4: Factorization of the term-document matrix by the NMF algorithm.

We assume that our document corpus consists of k clusters. The goal of NMF is to factorize X into non-negative matrices U $(m \times k)$ and V^T $(k \times n)$ which minimize the following objective function:

$$J = \frac{1}{2} \parallel X - UV^T \parallel \tag{2}$$

$\parallel \cdot \parallel$ denotes the squared sum of all the elements in the matrix.

Each element u_{ij} of matrix U determines the degree to which the associated term f_i belongs to cluster j. For cluster labeling we simply choose for each cluster the ten terms with the highest degree of affiliation. Terms with a relevance value of less than 25% of that of the most relevant term in the cluster are discarded. Analogously each element v_{ij} of matrix V represents the degree to which document i is associated with cluster j. To cluster the documents, again we assign every document to the cluster with the highest degree of affiliation. If a document i clearly belongs to one cluster x then v_{ix} will have a high value compared to the rest of the values in the i'th row vector of V. The matrix factorization is depicted in Figure 4.

4 Use Case: The ALOE System

ALOE (`http://aloe-project.de/AloeView`) is a Web 2.0 resource sharing platform designed for learning content of arbitrary format ([Memmel and Schirru, 2007]). It supports sharing of bookmarks and all kinds of files (images, audio, video, office documents, etc.). ALOE provides tagging, commenting, and rating functionalities. It offers search facilities with ranking options that take the usage of resources into account (such as most viewed, highest rated, most commented). A group concept has been implemented that enables users to contact and exchange resources with other users that share similar topics of interest. ALOE has been deployed as Enterprise 2.0 platform at the Knowledge Management department of the German Research Center for Artificial Intelligence (`http://www.dfki.de`).

The ALOE system enables the knowledge worker to share content according to her topics of interest. In order to make resources easily retrievable they are annotated with metadata by the community of users. Whenever users add resources to their portfolio they have to annotate a title and tags. Optionally a description, author, and licensing information may also be added. That way ALOE meets our requirements as stated in Section 3.1. Figure 5 shows the ALOE details page for a resource. Selected metadata fields such as title (1), description (2), and tags (3) have been highlighted.

Our first evaluation experiment was performed with a participant group consisting of eight staff members of the Knowledge Management group of DFKI. Seven of the participants were researchers (junior to senior), one participant was a software engineer. Every participant had expressed preferences for at least twenty resources in the system. A questionnaire was sent to these users showing the terms which represented their identified topics of interest. For each of these topics the users had to answer three questions:

Q1: Has the topic of interest correctly been detected?

Q2: How would you describe the topic in your own words?

Q3: Would you like to get recommendations for the topic?

5 Preliminary Results

Evaluating the performance of our method is still difficult as we are missing a large amount of users contributing resources to the ALOE system according to their topics of interest. Table 5 shows for each participant of the evaluation experiment, how many topics were identified by the system, how many of them were classified as correctly identified by the user and for how many of these topics the user would appreciate recommendations.

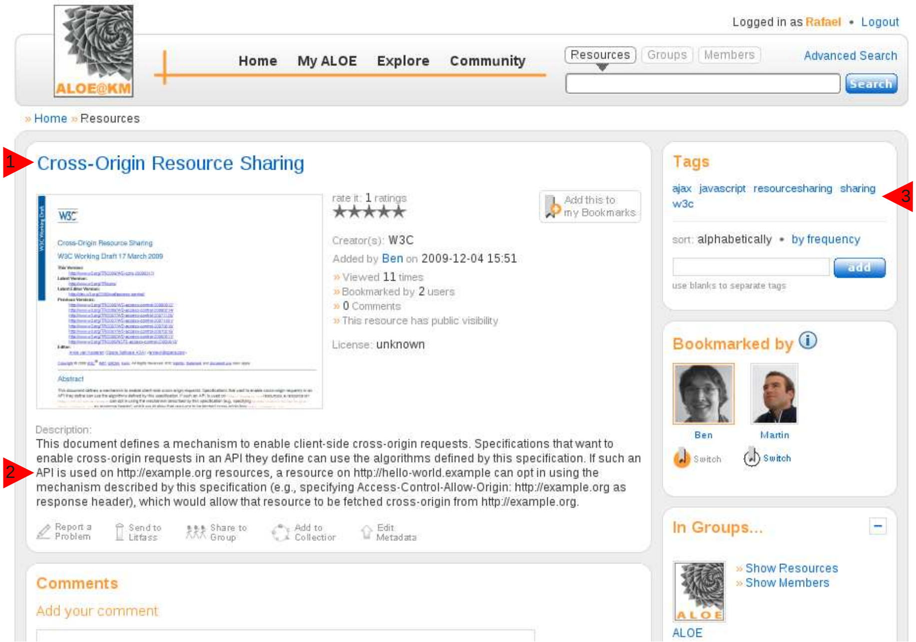

Figure 5: Detail view of an ALOE resource. Selected user-generated metadata is highlighted.

User	detected topics	correct topics	recommendations desired
A	3	3	3
B	2	2	2
C	3	1	1
D	10	9	7
E	3	3	3
F	9	6	6
G	6	5	2
H	3	3	3
Sum	39	32	27

Table 2: Results of the preliminary evaluation study. For each user the number of detected topics is juxtaposed to the number of correctly detected topics and the number of topics for which the users would appreciate recommendations.

Altogether 39 user interest topics have been identified, in average 4.875 per user. 32 of the topics were classified as correct by the users, i. e., in average four topics per user. For 27 topics the users said that they would appreciate resource recommendations, 3.375 in average per user. Each user in average classified 84.17% of her identified topics as correct. User C is an outlier, only one of three identified topics has been classified as correctly identified.

One problem of the metadata in the ALOE system is, that English and German descriptions are mixed. Even different languages within one metadata profile are possible (e. g., a title in German together with English tags). Currently our algorithm supports preprocessing steps such as stop word removal only for one language per metadata profile thus leading to the problem that stop words sometimes remain in the data set and introduce noise in the topic terms. In one extreme case we had a topic label consisting of two German articles ("der die") that were not recognized as stop words and thus were not filtered out in the preprocessing steps.

6 Conclusion and Future Work

In this paper we presented an approach on how to apply algorithms from the domain of topic detection and tracking for user modeling. The method identifies a user's different topics of interest in a Web 2.0 resource sharing platform unobtrusively. Each topic is represented as a weighted term vector. Our first evaluation experiment has shown that the users were able to associate the topic labels with their real topics of interest thus showing that the method is likely to capture reasonable user interest topics. Also the users widely agreed that they would appreciate resource recommendations for the identified topics.

The method can support E-Learning in at least two scenarios. In [Memmel and Schirru, 2007] the ALOE system has been introduced as a sharing platform for learning resources and metadata. In such a scenario our approach can identify the topics on which a learner is working and can provide resource recommendations matching these topics. Currently ALOE is deployed as Enterprise 2.0 platform in the Knowledge Management group of DFKI. Here our approach identifies the knowledge workers' topics of interest in research, software engineering, and also topics in which the users are interested privately. In such a scenario our approach can foster a targeted knowledge transfer among staff members according to their personal interests.

Our next step will be to integrate a content-based recommender into the ALOE system that exploits the contextualized user interests profiles. Whenever a user requests recommendations, the database should be queried according to resources matching her different interest topics and a recommendation list with a high degree of inter-topic diversity should be assembled. In the long-term we plan to exploit the usage tracking functionalities in the ALOE system to determine a user's current context. Whenever one of her interest topics is detected as relevant at the current point in time, targeted recommendations should be provided according to her needs and preferences.

Acknowledgments

This reasearch has been financed by the IBB Berlin in the project "Social Media Miner", and co-financed by the EFRE fonds of the European Union.

Special thanks to Dr. Thomas Roth-Berghofer for helpful suggestions and guidance.

References

[Agne et al., 2006] Agne, S., Reuschling, C., and Dengel, A. (2006). DynaQ - dynamic queries for electronic document management. In *Proceedings IEEE-EDM.*, pages 56–59. IEEE International Workshop on the Electronic Document Management in an Enterprise Computing Environment. Hong Kong, China.

[Allan et al., 1998] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

[Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

[Downes, 2005] Downes, S. (2005). E-learning 2.0. *eLearn*, 2005(10).

[Duan et al., 2009] Duan, M., Ulges, A., Breuel, T. M., and Wu, X.-q. (2009). Style modeling for tagging personal photo collections. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA. ACM.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

[Li et al., 2008] Li, X., Guo, L., and Zhao, Y. E. (2008). Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA. ACM.

[Manning et al., 2009] Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press, online edition.

[McAfee, 2006] McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review*, 47(3):21–28.

[Memmel and Schirru, 2007] Memmel, M. and Schirru, R. (2007). ALOE - a socially aware learning resource and metadata hub. In Wolpers, M., Klamma, R., and Duval, E., editors, *Proceedings of the EC-TEL 2007 Poster Session*. CEUR workshop proceedings. ISSN 1613-0073.

[Middleton et al., 2001] Middleton, S. E., De Roure, D. C., and Shadbolt, N. R. (2001). Capturing knowledge of user preferences: ontologies in recommender systems. In *K-CAP '01: Proceedings of the 1st international conference on Knowledge capture*, pages 100–107, New York, NY, USA. ACM.

[Morris et al., 2005] Morris, M., Pohlmann, T., and Young, G. O. (2005). How do users feel about technology? Forrester Research.

[O'Reilly, 2005] O'Reilly, T. (2005). What is web 2.0 design patterns and business models for the next generation of software. [Online; accessed 22-August-2008].

[Schult and Spiliopoulou, 2006] Schult, R. and Spiliopoulou, M. (2006). Discovering emerging topics in unlabelled text collections. In Manolopoulos, Y., Pokorný, J., and Sellis, T. K., editors, *ADBIS*, volume 4152 of *Lecture Notes in Computer Science*, pages 353–366. Springer.

[Schwarz, 2006] Schwarz, S. (2006). A context model for personal knowledge management applications. In Roth-Berghofer, T., Schulz, S., and Leake, D. B., editors, *Modeling and Retrieval of Context, Second International Workshop, MRC 2005, Edinburgh, UK, July 31 - August 1, 2005, Revised Selected Papers*, volume 3946 of *Lecture Notes in Computer Science*, pages 18–33. Springer.

[Stahl and Roth-Berghofer, 2008] Stahl, A. and Roth-Berghofer, T. R. (2008). Rapid prototyping of CBR applications with the open source tool myCBR. In Bergmann, R. and Althoff, K.-D., editors, *Advances in Case-Based Reasoning*. Springer Verlag.

[Ulges et al., 2009] Ulges, A., Koch, M., Borth, D., and Breuel, T. (2009). Tubetagger - youtube-based concept detection. In *Proc. Int. Workshop on Internet Multimedia Mining*. IEEE Computer Society.

[Xu et al., 2003] Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA. ACM.

[Ziegler et al., 2005] Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, New York, NY, USA. ACM.