Personalized Web - Science, Technologies and Engineering

11th Spring 2012 PeWe Ontožúr,
Modra-Piesok, Slovakia, April 2012
Proceedings

**Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko,
Jozef Tvarožek (Eds.)**

Proceedings in
Informatics and Information Technologies

Personalized Web – Science,
Technologies and Engineering
11th Spring 2012 PeWe Workshop

Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko,
Jozef Tvarožek (Eds.)

Personalized Web – Science, Technologies and Engineering

11th Spring 2012 PeWe Workshop
Modra – Piesok, Slovakia
April 1, 2012
Proceedings

Slovakia Chapter

PeWe Group

S T U

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
Faculty of Informatics and Information Technologies

Proceedings in
Informatics and Information Technologies

Personalized Web – Science, Technologies and Engineering
11th Spring 2012 PeWe Workshop

Editors

Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko, Jozef Tvarožek

Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia

Production of this publication was partially supported by
- the Gratex IT Institute, within the framework of GraFIIT Foundation (www.gratex.com)
- the Scientific Grant Agency of Slovak Republic, grants No. VG1/0675/11, VG1/0971/11
- the Slovak Research and Development Agency under the contract No. APVV-0208-10

Visit PeWe (Personalized Web Group) on the Web: pewe.fiit.stuba.sk

Executive Editor: Mária Bieliková
Cover Designer: Peter Kaminský

Published by:
Nakladateľstvo STU
Vazovova 5, Bratislava, Slovakia

ISBN 978-80-227-3693-0

Preface

The Web influences our lives for more than 20 years now. During these years, it has continuously been enjoying a growing popularity due to, among other things, its progressive change from passive data storage and presentation vehicle to the infrastructure for software applications and to the place for communication, interaction, discussions and generally collaboration. As the Web has an influence on our work, entertainment, friendships, it attracts more and more researchers who are interested in various aspects of the Web, seeing it from various perspectives – as a science, a place for inventing various technologies or engineering the whole process.

Research in the field of the Web has more than 10 years of tradition at the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava. Moreover, topics related to the Web attract many students. This volume is entirely devoted to students and their research. It contains extended abstracts of students' research projects presented at the 11th PeWe (Personalized Web Group) Workshop on the Personalized Web, held on April 1, 2012 in nice environment of Modra – Piesok, Pension Univerzitka near Bratislava. It was organized by the Slovak University of Technology (and, in particular, its Faculty of Informatics and Information Technologies, Institute of Informatics and Software Engineering) in Bratislava. Participants are students of all three levels of the study – bachelor (Bc.), master (Ing.) or doctoral (PhD.), and their supervisors.

The workshop covered several broader topics related to the Web, which served for structuring these proceedings:

- Search, Navigation and Recommendation,
- User Modelling, Virtual Communities and Social Networks,
- Domain Modelling, Semantics Discovery and Annotations.

The projects presented by students were at different levels according to the study level (bachelor, master or doctoral) and also according the progress stage achieved in each particular project. Moreover, we invited to take part also two of our bachelor study students who take an advantage of our research track offered within their study programme and who were just about to start their bachelor projects – *Samuel Molnár* and *Pavol Zbell.*

Bachelor projects:

- *Ľuboš Demovič, Martin Konôpka, Marek Láni, Matúš Tomlein:* Personalized User Interface for Web Surfing in Conditions of Slow and Intermittent Internet Connection
- *Peter Dulačka:* Validation of Music Metadata via Game with a Purpose

- *Eduard Fritscher:* Educational Content Recommendation Based on Collaborative Filtering
- *Tomáš Jendek:* Gathering Information on User Environment
- *Ondrej Kaššák:* Named Entity Recognition for Slovak Language
- *Marek Kišš:* Building Domain Model via Game With a Purpose
- *Martin Lipták:* Automated Public Data Refining
- *Andrea Šteňová:* Feedback Acquisition in Web-based Learning
- *Ján Trebuľa:* Group Recommendation Based on Voting
- *Juraj Višňovský:* Association Rules Mining from Context-enriched Server Logs

Master projects (started in the current academic year):

- *Pavol Bielik:* Integration and Adaptation of Motivational Factors into Software Systems
- *Roman Burger:* Personalized Reading Resources Organization
- *Máté Fejes:* Recognizing User's Emotion in Information System
- *Róbert Horváth:* Augmenting the Web for Facilitating Learning
- *Peter Krátky:* User Modeling Using Social and Game Principles
- *Peter Macko:* Unified Search of Linked Data on the Web
- *Štefan Mitrík:* Context-Aware Physical Activity Recommendation through Challenges
- *Balázs Nagy:* Metadata Collection for Effective Organization of Personal Multimedia Repositories Using Games With a Purpose
- *Jakub Ševcech:* Navigation Using Annotations in Web Documents
- *Michal Tomlein:* Method for Social Programming and Code Review

Master projects (started in the previous academic year)*:*

- *Anton Benčič:* Information Recommendation Using Context in a Specific Domain
- *Peter Kajan:* Discovering Keyword Relations
- *Marcel Kanta:* Trend-Aware User Modeling with Location Aware Trends
- *Peter Korenek:* Emotion Classification of Microblogs Based on Appraisal Theory
- *Milan Lučanský:* Acquiring Web Site Metadata by Heterogeneous Information Sources Processing
- *Róbert Móro:* Personalized Text Summarization
- *Ivan Srba:* Encouragement of Collaborative Learning Based on Dynamic Groups
- *Peter Svorada:* Modeling a Tutor for E-Learning Support
- *Márius Šajgalík:* Decentralized User Modeling and Personalization
- *Tomáš Uhercík:* Acquiring Metadata about Web Content Based on Microblog Analysis
- *Maroš Unčík:* User Modeling in Educational Domain

Doctoral projects

- *Michal Kompan:* Group Recommendations for Adaptive Social Web-based Applications
- *Tomáš Kramár:* Analyzing Temporal Dynamics in Search Intent

- *Eduard Kuric:* Search in Source Code based on Identifying Popular Fragments
- *Martin Labaj:* Explicit and Implicit Feedback in Recommendation
- *Dušan Zeleník:* Context Influencing our Behavior
- *Michal Holub:* Discovering Relationships between Entities in Web-based Digital Libraries
- *Karol Rástočný:* Knowledge Tags Repository
- *Jakub Šimko:* Games and Crowds: Authority Identification

Considerable part of our research meeting this year was devoted to a *hack-day-like activity* chaired by Tomáš Kramár and Dušan Zeleník. After the last year's successful PeWeProxy workshop, we decided to base the workshop on two events that happened in the life of our institution since the last year:

- we operate a Hadoop cluster, suitable for processing large amounts of data,
- we have started a cooperation with Azet, largest Slovak internet company, with the aim of improving the click through rate (CTR) of their advertisement. For research purposes, we have available a large dataset of advert impressions and their respective clicks.

The initial motivation was help our bachelor and master students to familiarize themselves with both Hadoop platform and Azet dataset and take advantage of innovative potential of PeWe group and unique connection of bright young researchers on various levels of their competences to practice processing large data sets, which can influence their projects towards real usability of the results. Even within constraints such limited time (4 hours plus instructions given few days before the workshop), teams, which were not used to work together and no internet connection (so the workshop chairs built small cluster of commodity PCs, just for the purposes of the workshop) we get really interesting results and exciting presentations on various dependencies in analyzed data.

Our workshop hosted for the sixth time recessive activity organized by the *SeBe (Semantic Beer) initiative* aimed at exploration of the Beer Driven Research phenomenon. SeBe workshop was chaired by Marián Šimko, Jakub Šimko and Michal Barla. It included various activities such as mutual beer tasting or newly established collaborative social game named *Tap and Reduce*.

More information on the PeWe workshop including presentations is available in the PeWe group web site at pewe.fiit.stuba.sk. Photo documentation is available at mariabielik.zenfolio.com/ontozur2012-04.

PeWe workshop was the result of considerable effort by our students. It is our pleasure to express our thanks to the *students* – authors of the abstracts, for contributing interesting and inspiring research ideas. Special thanks go to Katka Mršková and Saška Bieleková for their effective support of all activities and in making the workshop happen.

April 2012

Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko, Jozef Tvarožek

Table of Contents

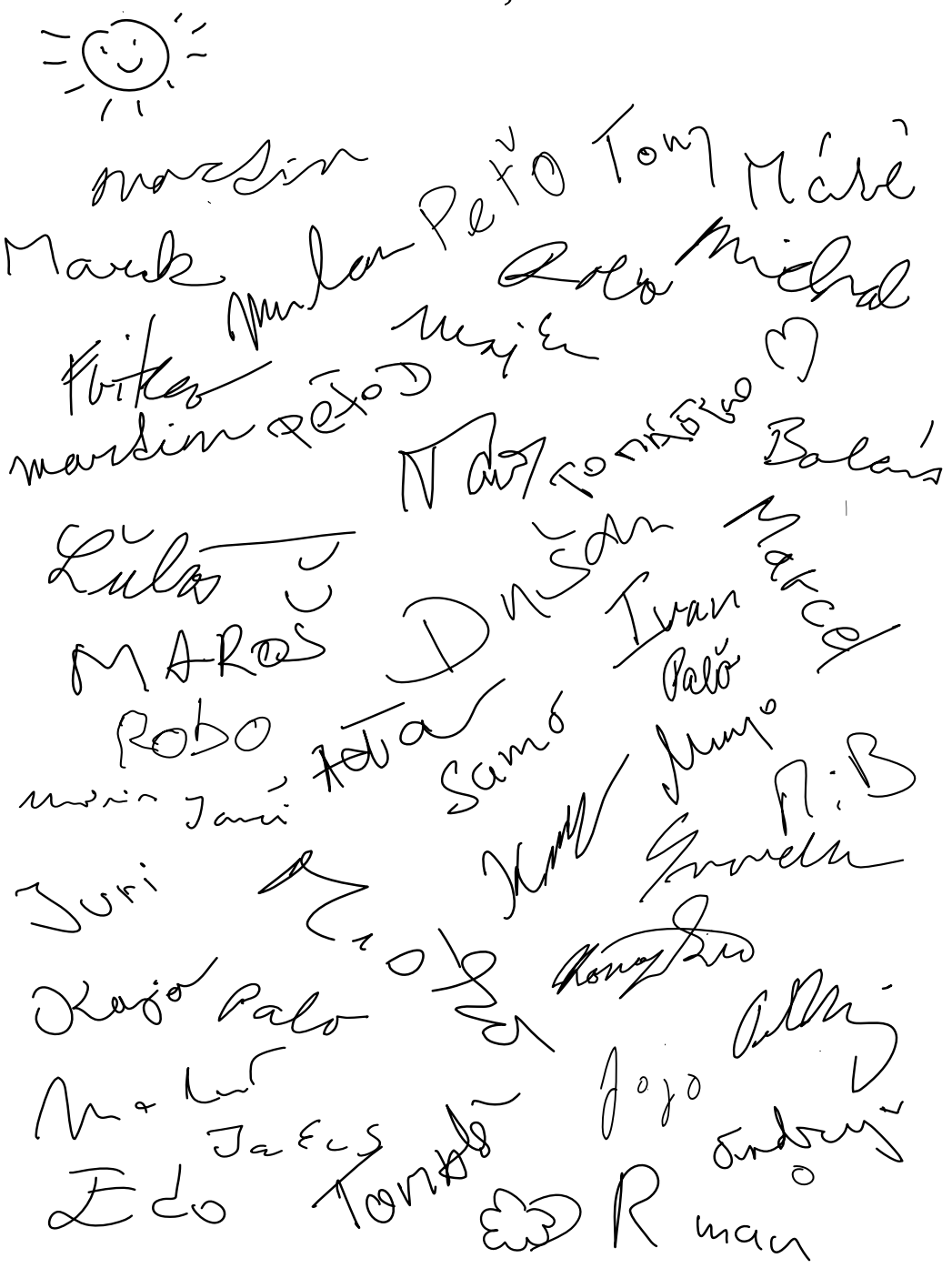
Students' Research Works

Search, Navigation and Recommendation

User Modeling, Virtual Communities and Social Networks

Workshop Events Reports

Workshop participants
Modra – Piesok, Univerzitka



PeWe Ontožúr Participants in Action

Search, Navigation and Recommendation

Information Recommendation Using Context in a Specific Domain

Anton BENČIČ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
bencican@live.sk

Adaptive and personalization methods all engage in recommendation process. A recommendation process consists of decisions that have to be made in order to deliver a resource to the user. More often than not adaptive and personalization methods engage in only a single decision, that is, *what* to deliver to the user. This is especially true for methods that work on demand, delivering the content when a query for it is made. These methods generally do not consider whether it is the right time to deliver the content, the extent of the content to be delivered and the way it will be presented, which are the three other important decisions in the recommendation process.

Apart from methods working on demand, there are also methods that work proactively. A proactive method decides on the action to take by considering various criteria. This can be as simple as setting the sound profile according to the user-defined time windows [1] or as complicated as recommending the right music for a given user and his friends around him considering their mood [2]. Our project is aimed at designing a method that is able to effectively and efficiently learn what actions should be performed in what situations, and then use this model to aid end-user applications in autonomous decision process by recommending actions for a given situation.

To accomplish this we devised a rule-based method that uses different classes of contextual information as antecedents for rules that are formed using feedback from the end-user application. Both rules and situations have a certainty assigned to them that are used further on in the recommendation process to compute the final score for the available actions. The strength of our method is in its domain independence. All situation classes, situations and possible actions are defined on-the-fly by the particular application which makes it possible to define correct models for different domains as each may require a different set of situation classes that may be important. The situation classes and situations that are fed to our method pose as abstract symbols with no required background meaning which makes it accessible for processing outside the client's device without facing any privacy or security concerns. This is a very

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 3–4.

important feature because even though the data have no meaning assigned they can still be used to support collaborative sharing of user models. Another important aspect is *time sensitivity principle* which we employed to accommodate preference changes that may happen over time. This can be for example a user that usually reads daily news during commute, but now decides to travel individually by car which naturally alters his preference toward reading during commute.

In order to evaluate our method's performance we have implemented a simulation framework and a mobile news recommendation application. Using the simulation framework we have performed a series of simulations for refinement of our method's parameters, so as to be able to reliably identify situations for recommending specific actions. After a series of these simulations and adjustments to our method we were able to reliably identify peeks that corresponded with the virtual user's preferences.

To evaluate our method in a real-life scenario we have also created a mobile news application that uses our recommendation framework for identifying appropriate situations for an autonomous news push. This application will be used in a real-life experiment. Since the target domain is news recommendation, the contextual information classes and recommendation actions are tailored to it as well. The specific situation classes that we identified in the news recommendation domain are:

- Time information: *time of day, day of week, week of month and month of year*
- Weather information: *weather, temperature, wind, pressure*
- Location information: *identifies whether a user is in motion or at some place*
- Calendar events: *identifies proximity of a calendar event*
- Dead time: *identifies parameters of a user session in our news application*

The situations from these classes are fed continuously into our recommendation framework together with action feedback which distinguished between good time for a news push and not a good time for a news push.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Ala-Siuru, P., Tapani, R.: Understanding and recognizing usage situations using context data available in mobile phones. In: *ubiPCMM'06: Proc. of the 2nd International Workshop on Personalized Context Modeling and Management for UbiComp Applications*, (2006).

[2] Shin, D., Lee, J.-w., Yeon, J., Lee, S.-g.: Context-Aware Recommendation by Aggregating User Context. In : *CEC*, Vienna, (2009), pp.423-430.

Personalized Reading Resources Organization

Roman BURGER*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
roman.arnold.burger@gmail.com

Considering the vast amount of accessible information and resources in the Web, one must carefully choose what to read and what to ignore. Regardless of our "web habits", chances are that our workspace eventually gets overwhelmed with considerable number of resources stored for later reading. The very common cause is that resources processing speed of a typical user is lower than resources retrieval speed of the user.

Keeping track of resources organization and structure consequently becomes a tedious task, possibly distracting us from reading of the individual resources. Users typically manage theirs workspaces manually which makes significant overhead while working with the saved resources. Occasionally, it is not uncommon that users eventually abandon the management of resources due to the load. In our project we explore ways for automatic organization management of resources.

Related work [1, 2] focuses mostly on content analysis of documents, and thus producing merely semantic relations between documents. While semantic relations are important, they usually cannot fully capture project relations. By project relations we mean relations defining collections of all resources used for particular projects. Using this representation, a project could be anything ranging from a large study assignment to a simple collection of blogs about an interesting topic.

In this project, we aim to propose a method for automatic organization of emerging resources. To actually capture the project relations we will take two sources of information into account. The first one will be documents' metadata, the second will be documents' context obtained via implicit user feedback. Metadata is chosen because it easily accessible and accurate. Part of metadata could be actually keywords extracted from content, but we will focus on other metadata such as creation time or last edit time as well. This metadata has higher chance in helping to discover the project relations, because they are describing the actual usage patterns of the user. Context of the documents will be used to aid the process that determines project relations of the resources, since resources within the same project should share a similar context.

Our proposed method (Figure 1) first extracts metadata from resources under consideration. After metadata extraction phase, clustering algorithm will be deployed

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 5–6.

to generate initial collections of resources. User will be provided with tools to personalize resource collections to his/hers needs and by logging these actions we will get implicit feedback on precision of generated resources structure. The next time user adds a resource to the system, we can use the context of the existing collections to cluster the new resource.

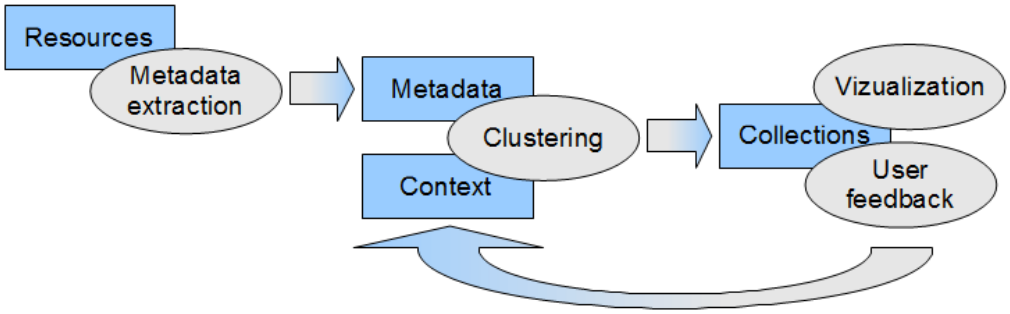


Figure 1. Proposed process of generating automated personalized collections.

On top of resources organization, recommendation of reading order could be employed. With automatic personalized organization and processing order recommendation, user would need little to none effort maintaining the workspace. User would only need to pop out first item from the recommended resources list while automatic organization and visualization takes care of providing the user with the right collection.

The proposed method will be evaluated in a user experiment. We will try to evaluate overall usefulness and focus mostly on organization method and visualization. Validation of the organization method will be done in a user experiment in which we will measure the amount of user corrections made in the automatically generated structure. We hypothesize that the number of user interventions in the resource organization provided by our method should decrease over time as users use our method for organizing resources.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Corrêa, R. F., Ludermir, T. B.: Semantic mapping and K-means applied to hybrid SOM-based document organization system construction. In: *Proc. of the 2008 ACM symposium on Applied computing*, ACM, (2008), pp. 1112-1116.

[2] Prinz, W., Zaman, B.: Proactive support for the organization of shared workspaces using activity patterns and content analysis. In: *Proc. of the 2005 international ACM SIGGROUP conference on Supporting group work*, ACM, (2005), pp. 246-255.

Enhancing Web Surfing Experience in Conditions of Slow and Intermittent Internet Connection

Ľuboš DEMOVIČ, Martin KONÔPKA, Marek LÁNI, Matúš TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
team.ownet@gmail.com

The Web became a mass media allowing its users to quickly and comfortably search for information. It provides a way to acquire knowledge from various study fields and allows people to stay in touch with the current situation in the world. Thanks to the Web, its users may express their thoughts and discuss their problems with people around the world without the need to know them personally or travel long distances.

Despite of the advancements in information and telecommunication technologies, slow and intermittent Internet connection is still a serious issue in many places of the World and is most visible in developing countries. Web surfers from these countries are extremely patient as there is not much they can do to deal with unexpected cut-offs, slow and incomplete downloads of webpages [1], all preventing the Web to become a really useful tool for dealing with everyday problems.

There are several problems in utilization of available connection. Even though the Internet connection is often shared among several users, it does not take into account repetitive accesses to the same resources. The same website or YouTube video is being downloaded again and again by different users even though it has not changed since last visit. At the same time, the browsing session is usually performed in subsequent peaks. The Internet link is idle most of the time when the user reads the content of a downloaded website or during the night and is overloaded when the user decides to access another website. There are almost no "background" jobs taking advantage of flat-rate link (even if slow and intermittent) [1].

Another problem is that users surf on their own. However, a group of users, e.g., students during computer class, share interests and information needs and could take advantage of collaborative surfing to achieve their goals more quickly and efficiently. If a valuable resource is identified by a group member, others can be notified about it.

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 7–8.

We propose OwNet solution to enhance Web surfing experience directly and indirectly, mainly in conditions of slow and intermittent Internet connection.

We took direct approach for saving connection by caching web objects to save bandwidth and reduce latency for fetching requested web objects [2]. We also implemented our own caching and cache invalidation algorithms. Being aware that our application will be used on older computers, we tried to optimize it. We also proposed a combination of existing prefetching approaches to get an algorithm tailored to our needs. Pro-active downloading and caching of not yet requested web objects of user's interest may result in user's perception of having a faster Internet connection.

We took indirect approach by applying collaborative tools in order to eliminate the problem with users of the same group browsing individually. By recommending and rating interesting websites they can help others find useful information faster.

OwNet solution consists of three modules which are individually independent but they complement one another:

- Local client proxy application responsible for handling requests from client applications, e.g., Internet browser. Users use their Internet browser to access the Web in the same way as they would without using it.
- Local server proxy application serves local clients applications, preferably within a same organization, e.g., in a school. This enables caching of Web content within an organization, prefetching and advantages of collaborative tools
- Central service as a means to find out which cached objects are outdated, reduces the load on local proxy servers and their Internet connection. It ensures that cached objects are updated only if they were changed on the Internet

We contacted few Slovak NGOs that operate in rural areas of Africa to discuss possibilities of OwNet deployment in these places. We are currently in progress of deploying OwNet to computer lab in Nanyuki High Shool in Kenya. In addition, we have worked to deploy OwNet to Slovak schools, which can also take advantage of its caching and collaborative features.

We believe that the Internet is an important source of information and a crucial part of modern education. OwNet helps people with slow or intermittent connections to the Internet make better use of the information that it provides.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Johnson, D.L., Pejovic, V., Belding, E.M., van Stam, G.: Traffic Characterization and Internet Usage in Rural Africa. In: Proceedings of the 20th international conference companion on World Wide Web. WWW '11, New York, NY, USA, ACM, 2011, pp. 493-502.

[2] Kroeger, T.M, Long, D.D.E., Mogul, J.C.: Exploring the Bounds of Web Latency Reduction from Caching and Prefetching. In: Proceedings of the USENIX Symposium on Internet Technologies and Systems, Berkeley, CA, USA, USENIX Association, 1997, pp. 17-20.

Educational Content Recommendation Based on Collaborative Filtering

Eduard FRITSCHER*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
eduard.fritscher@gmail.com

In the past students could only study from handwritten books and for achieving their goals (e.g., passing an exam or expanding their knowledge) they could only rely on the teacher to guide them through the vast amounts of information. But times have changed. Due to technologies such as the Web students have access to all information they could ever possibly need, without necessary guidance. Some people do not need any guidance, they intuitively know which path to choose; but for those who do, recommendation systems are the perfect tool to guide them through the lectures.

There are all sorts of recommendation systems which are based on different methods. Some of them recommend based on geographical data, some learn from the user, some recommend based on previous user actions; the list of techniques is quite long. But every method can be traced back to two main recommendation types: content-based or collaborative. In this project we create a recommender system that is based on a hybrid approach combining both collaborative and content-based aspects, and realize it within the ALEF (Adaptive LEarning Framework), an educational system which was created by Slovak University of Technology [1]. The main goal of our recommendation is to guide a student through the courses recommending studying materials that he or she will need to successfully pass the course.

The ALEF system consists of so-called *learning objects* of three types: *text-explanations*, *questions* and *exercises*. *Metadata* is used to conceptually describe learning objects. The learning objects are linked to *concepts*, which are represented as relevant domain terms. In our method, we leverage conceptual structure to derive student similarities, which are crucial for obtaining *learning objects* to be recommended. In our approach, the *learning objects* are displayed in a sorted tag cloud. The tag cloud is sorted by the degree of similarity between the users and the degree of knowledge relation between the users and the *concepts*.

During navigating in the system, a student can access learning objects using the main menu. By selecting a learning object in the menu, a student expresses his or her

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 9–10.

learning intent and the recommender recommends a set of learning objects appropriate for the student. Since our recommender uses a hybrid approach, we can receive a different output for every combination of user and learning object.

In our method we use relationships associated with learning objects and concepts, which are a part of the domain model and the user model within the educational system, e.g., *UserToConceptRelation*, *ConceptToLearningObjectRelation*, *ConceptToConceptRelation*.

The recommendation can be described by the following steps:

1. The input is the user and the accessed learning object.
2. All concepts, which are related to the learning object, are fetched.
3. All concepts, which are related to the previously fetched concepts, are fetched.
4. The similarity is calculated between users.
5. The concepts of the most similar users are mapped to our relevant concepts.
6. The most relevant concepts are transformed back to learning objects.
7. Learning objects that have the highest rating are recommended.

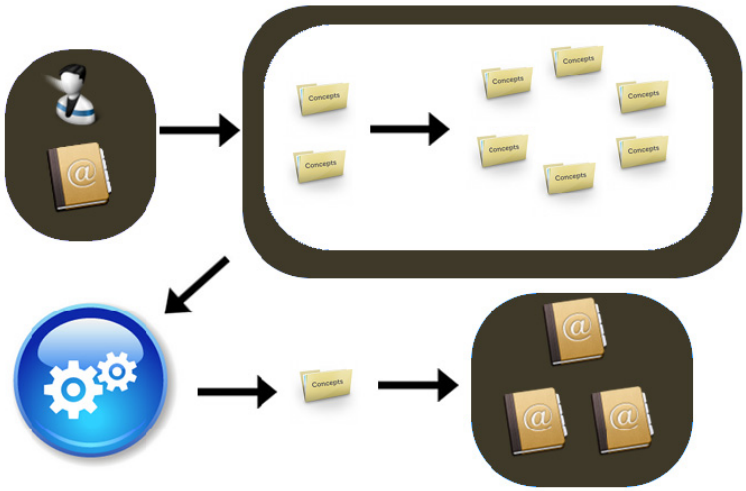


Figure 1. Recommendation process of the proposed hybrid method.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-Based Learning 2.0. In *Proc. of IFIP Advances in Information and Communication Technology*, Vol. 324, Springer, 2010, pp. 367–378.

Group Recommendation for Adaptive Social Web-based Applications

Michal KOMPAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kompan@fiit.stuba.sk

Recommender systems are an integral part of a modern – adaptive web nowadays. The need for the adaptive web increases day by day, while users are generally overwhelmed by the amount of information available. Similarly, on the other side – businesses try to increase profits and visits to web sites. Personalised recommendation is the most used approach to satisfy both – the users and the businesses.

Historically, several approaches have been proposed for the recommendation task. The content-based recommendation uses the similarity between recommended items. The similarity can be computed based on several aspects as simple text similarity, or various enhancements for specific domains as news have been proposed [1,3]. The second and increasingly used approach is the collaborative recommendation. This approach instead of content similarity takes advantage of user's similarity, which is typically computed based on user ratings. While these approaches are designed for the single-user environment, in the recent years the phenomenon of social networking and mobile devices bring us to the increasing demand for recommendation designed for groups of users. The group recommendation is usually based on the collaborative approach. The main difference of our method is that we use inter-group relations in order to provide recommendation for the whole group of users instead of only for a single-user. While the standard single-user recommendation satisfies the needs of an individual user, the group recommendation based on the used strategy and the goal of the recommendation try to maximize satisfaction of each user in the group.

Various approaches for the personalized recommendation have been proposed in the literature [2]. As the social activity over the web increases, the group recommendation becomes increasingly popular and subject of research, while the possibility of the usage of group recommendation approaches within the standard single-user recommendation was raised, but no study explored such an approach. Current approaches mostly deal with the TV or music domain, as these are activities

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp.11-12.

which are usually performed in group setting. While the TV and movies domains were logical choices at first, more and more new domains are used and researched today.

In our work we focus on exploring possibilities of using group recommendation not only in the new domains such as the research domain (scientific papers being the recommended items) or collaborative support of the learning process, but we explore the possibility of usage of the principles of group recommendation in the single-user environments. These can be used thanks to virtual groups' construction, which on the other hand can be replaced with the real or derived groups anytime.

In our experiments, we arrived at statistically significant results that support our hypothesis, that recommendation based on the group recommendation principles overcomes the standard collaborative recommendation. Correspondingly, we show that such an approach is suitable for various domains (news, movie).

Several aspects can be considered when creating virtual groups or when some aggregation is performed in generating recommendations. The satisfaction of every user is directly influenced by the actual group members, the size of the group and types of relationships within the group respectively. More over the actual satisfaction of an individual member is influenced by the relationship type and intensity of relationships between others.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Kompan, M. and Bieliková, M. Content-based news recommendation. In Proc. of E-Commerce and Web Tech., Vol. 61 of LN in Bus. Inf. Proc., Springer, pp. 61-72, (2010).

[2] Masthoff J. Group Recommender Systems: Combining Individual Models. In: Recommender Systems Handbook, pp. 677–702, (2011).

[3] Suchal, J., Návrat, P. Full text search engine as scalable k-nearest neighbor recommendation system. AI 2010, IFIP AICT 331, Springer, pp. 165-173, (2010).

Emotion Classification of Microblogs Based on Appraisal Theory

Peter KORENEK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
pkorenek@gmail.com

Services and social networks that allow their users to communicate via the Web are nowadays wide-spread among people in all age categories. One of these new forms of web communication is a microblog. It is a service that allows users to express their feelings, opinions and ideas and makes them public to everyone, who is interested in what other users write. Moreover, one of the most significant differences in comparison to other social networks is the limitation of the length of a microblog to 140 characters.

Opinions and ideas that users write in their microblogs are very valuable. For example, companies all around the world are paying large amounts of money for surveys whose aim is to retrieve opinions of customers on products of a company, to find out what they consider as positive or negative about their products or services. There is also a non-commercial usage of opinion mining, e.g., when we want to find common interests between groups of people, or just want to find a new friend who is interested in the same topics as we are.

In our work we research the field of emotion and opinion mining. We focus on utilization of Appraisal theory for emotion classification in microblogs. This theory says that emotions are consequences of how an author appraised some situation. The theory is used in emotion analysis to discover what an author feels and what the situation that caused this feeling is. Our aim is to explore suitability and applicability of the theory in relation to the specifics of user-generated microblog content.

When using appraisal theory, it is necessary to have a good dictionary of terms and phrases that are associated with keywords from this theory. We built an own dictionary using known appraisal terms and we extended them with synonyms using WordNet.

Our method for emotion classification consists of four steps that allow us to improve emotion analysis in microblogs: (i) microblog pre-processing, (ii) main target identification, (iii) emotion extraction, (iv) emotion graph composition.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 13–14.

We used syntax patterns specific to microblogs in our method to resolve which word is a target of a sentence. We created them manually according to a syntax analysis of hundreds of microblogs. After we know what the targets of a sentence are, we assign identified words from the appraisal dictionary to them. Every word, which we tag as relevant, is being searched in appraisal dictionary. According to its appraisal category, we assign a numerical value to it, which represents its attitude. Coefficients for each appraisal type were derived experimentally. These values are used in final step of our method where user target-emotion graph is created. There are two kinds of nodes – users and targets in this graph. The edges between these nodes are vectors with values:

- count of microblogs,
- dominant appraisal – appraisal type which occurred most often,
- intensity – normalized sum of classified orientation values of microblogs.

Using this graph it is easy for each user to find out the most favourite topic (determined by targets) or the most hated topic. We can also specify how much and in what way he likes / does not like the target.

To evaluate the accuracy of our method we conducted two initial experiments. Firstly, we classify 5,000 microblogs according to their polarity and compared results to manually annotated polarities. In this experiment we achieved 85 % accuracy. In our second experiment we manually annotated the main target of microblogs and theirs appraisal type. We executed all three steps of our method – including target graph creation. We achieved 70 % accuracy of assigning appraisal expressions to targets.

The results we obtained are comparable to other works [1, 2, 3]. In comparison to other works, our method is independent of topic of microblogs. We proved by the evaluation that despite microblog's weaknesses (in terms of quantity of content to analyse) our method can effectively classify emotions of random microblogs with different topics.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: *HLT '11 Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, ACL, Stroudsburg, PA, USA (2011), pp. 151–160.

[2] Bloom, K.: *Sentiment analysis based on appraisal theory and functional local grammars*. Dissertation thesis, Illinois Institute of Technology. (2011).

[3] Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: *CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM Press, (2005), pp. 625–631.

Analyzing Temporal Dynamics in Search Intent

Tomáš KRAMÁR*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kramar@fiit.stuba.sk

Web search, although fairly mature both as a research area and commercial deployment still poses as many problems as it solves. Search engines have become effective tools for information retrieval in large amounts of documents, even at Web-scale. Their main deficiency lies in their greatest strength, the ease of use and the familiarity of keyword based lookup. This query-based approach largely supersedes other complicated approaches, such as semantic search, which promises more relevant results, although at a price of forcing the searchers to use more complicated and less intuitive types of queries. The keyword-based query model dominates in simplicity, but this simplicity is diminished by several factors:

- the number of keywords is usually low, typically 1-3 keywords
- many of the words are ambiguous; a word "jaguar" can refer to an animal, a car and even has less-known meanings such as a game console or German battle tank; this is problematic not even for queries, but even for words in the document index;
- the queries are almost never accurate, they are either too generic or too specific, but almost never exactly aligned with the specific intent the user has in mind

The existence of these problems has led to a new research and application field called search personalization, a process which deals with finding the underlying search intent – i.e., the specific goal that the user pursues, the reason that he typed in the query – and biasing the information retrieval process towards this goal. This is usually implemented in two ways, either as a query modification, altering the original query to make it more explicitly express user's goal or as a ranking function modification, altering the ranking behavior to increase rank of pages more relevant to user's current goal.

Search personalization usually acts upon some kind of user model, where the interests of the user are stored. This model often contains interests built incrementally

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 15–16.

from the entire activity history of the user. Recently, a new research trend has emerged – moving from the monolithic user models to more lean models that are more focused on user's immediate interests, instead of his global and often historical needs. White et al. [2] show that personalizing search based on short-term interests significantly improves the relevance of search results, when compared to personalization based on long-term interests. The challenge that this new area of short-termness brings, is the lack of data. While the long-term user models provide sufficient data for confident adaptation of content, the short-term models may often be insufficient.

We hypothesize, that each person is having multiple personas. The term roots in psychology [1], where it denotes a "social face the individual presents to the world", it "reflects the role in life that the individual is playing". We believe that among many personas an individual can have, two should stand out: the persona related to personal life, and persona related to work life. Separating these two personas and creating a separate user model for each of them has the potential to bring a user model that is focused, similarly to the lean, short-term model, and yet has enough data to allow confident adaptation.

In our work, we analyze how well is the existence of these two personas reflected in Web search and specifically, the intent of the queries that the users type. We propose two hypotheses:

- The search intents/goals of the users vary during workweek (that is Monday through Friday) and weekend (that is Saturday and Sunday).
- The search intents/goals of the users during workweek vary during business hours (that is 9:00 through 17:00) and off-business hours (that is 17:00 through 9:00).

First hypothesis is based on the observation, that during weekend, individuals usually do not work and have the personal life persona. This observation should be reflected in the search queries, i.e., the search intent of the queries issued on workweek should differ from the intent of queries issued on weekend.

Second hypothesis is similarly based on the observation, that even during workweek, users switch between a work-related persona and personal life persona during business hours and leisure time.

To support or disprove these hypotheses, we analyze the publicly available log of search queries from an AOL search engine.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Jung, C.G.: Two Essays on Analytical Psychology. Volume 7 of Bollingen XX:7. Princeton University Press, 1966.Shin, D., Lee, J.-w., Yeon, J., Lee, S.-g.:

[2] White, R.W., Bennett, P.N., Dumais, S.T.: Predicting short-term interests using activity-based search context. In: Proceedings of the 19th ACM international conference on Information and knowledge management. CIKM '10, New York, NY, USA, ACM, 2010, pp. 1009–1018.

Search in Source Code based on Identifying Popular Fragments

Eduard KURIC*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kuric@fiit.stuba.sk

When programmers write new code, they are often interested in finding definitions of functions, existing working code fragments with the same or similar functionality, and reusing as much of that code as possible. Short code fragments, which are returned to a programmer's query, do not provide enough "background" to help them understand how to reuse the fragments. Keyword-based code search instruments (tools) face the problem of low precision of the results due to the fact that a single word of the programmer's query may not match the desired functionality. Understanding code and determining how to use it, is currently a manual and, consequently, a time-consuming process. In general, programmers want to find entry points such as relevant functions. They want easily understand how the functions are used and see the sequence of function invocations in order to understand how concepts are implemented. When programmers try to understand a program (source code), the control flow (execution of function calls) needs to be followed.

Our main goal is to enable programmers to find relevant functions to query terms and their usages. In our approach, identifying popular fragments is inspired by PageRank algorithm, where the popularity of a function is determined by the number of functions that call it. We designed a model based on the vector space model, using which we are able to establish relevance among facts, which content contains terms that match programmer's queries directly. Our method consists of two phases: processing of the source code repository, and searching for relevant functions given a programmer's query (see Figure 1).

Index creator creates document and term indexes from the source code repository. The function graph creator determines the directed graph of functional dependencies on which the PageRank algorithm is executed. The algorithm calculates a rank vector, in which every element is a score for each function in the graph.

When a programmer enters a query (1), a list of relevant documents (source code files) is retrieved (2). The list contains documents, where at least one query term occurs

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 17–18.

in each document. Similarity (3) between two documents (query q and a relevant document d_i) is calculated (4) using the cosine distance.

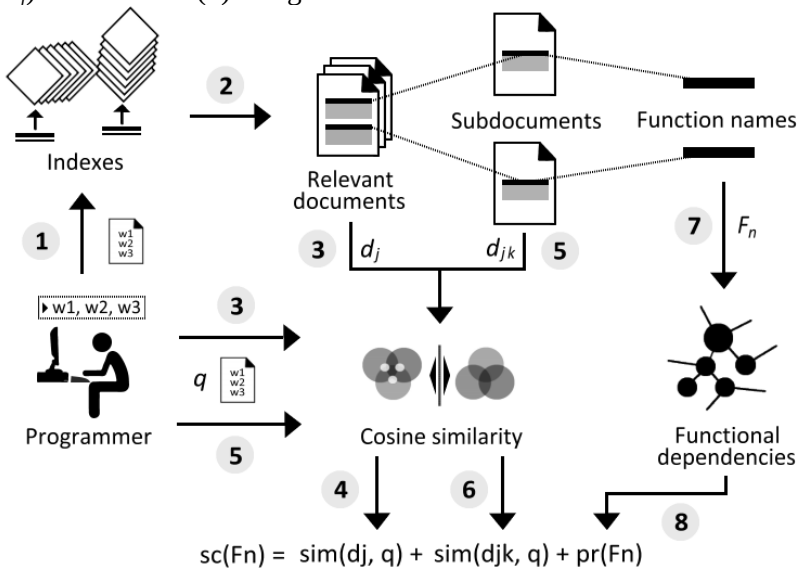


Figure 1. Searching for relevant functions.

Subdocument processing is as follows*:*

1. Each retrieved relevant document is divided into subdocuments, where each one contains only one definition of a function with surrounded comments (if any).
2. From each subdocument, terms are extracted from comments, function name and identifiers. Next, TF-IDF weights are calculated for these terms. The cosine similarity (5) is calculated between each subdocument d_{jk} and the query q (6).

Ranking of the relevant functions is as follows:

1. Names of defined functions are extracted from the relevant documents.
2. For each function name F_n, a final score $sc(F_n)$ is calculated as sum of:
   a. a similarity between the programmer's query q and the document d_j, in which is the function F_n defined (4); a similarity between the programmer's query q and the subdocument d_{jk}, in which is the function F_n defined (6),
   b. a PageRank score $pr(F_n)$ for the F_n (7)(8).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Grechanik, M, et al.: A search engine for finding highly relevant applications. *In Proc. of the 32nd ACM/IEEE Int. Conf. on Softw. Eng.*, NY, 2010, pp. 475-484.

[2] Sillito, J., et al.: Asking and Answering Questions during a Programming Change Task. *IEEE Trans. Softw. Eng.*, vol. 4, 2008, pp. 434-451.

Implicit Feedback in Recommendation

Martin LABAJ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
labaj@fiit.stuba.sk

Modern web systems are becoming more and more adaptive and they are modifying their behaviour to suit different users' needs. Recommender systems are an important part of such adaptive Web systems. They are of benefit to both users (e.g., see more items in which they are interested in without navigating through vast amount of items available) and system owners (e.g., sell more items). In technology enhanced learning (TEL) environment, it is important to aid users while learning. Some user tasks supported by current recommender systems are well explored (e.g., find good items, recommend sequence), but some can be supported better (e.g. find good pathways) [3].

In our research, we explore the use of implicit and explicit feedback in recommendation. Previously, we employed gaze position (tracked in common settings of home users) and other interest indicators in fragment recommendation on the Web. Currently, we are exploring implications of parallel browsing. In current web browsers users can open multiple new windows and tabs and switch between them at any time.

Such behaviour is often invisible to web usage mining [4], as only page loads are tracked server-side. However, if we can capture these data from common users, we can track revisitations, paths through resources, etc. more accurately. Tabbing behaviour is easily observable via client-side browser extensions or plugins, since they have access to browser information and they have been used in several studies using limited user groups, e.g. [1]. However, if we want to obtain user model from a broader group of users, we would have to maintain multiple extensions for multiple current browsers and persuade users to use them. In order to unobtrusively collect information from all users, we are restricted to client-side scripting, where the embedded scripts are restricted to actions in one page and cannot observe tabbing directly. Such actions have been previously used to observe parallel browsing only in a limited way – action of branching has been reconstructed from ordering of clicks on search engine results [2].

We proposed a model for parallel browsing behaviour, in which we cover large part of a tab life-cycle – page visit (including type-in, branching and linear browsing), page leave and switches between tabs. We reconstruct user's behaviour from page loads (PL; including referrers), focusing and defocusing of the page and page unloads

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 19–20.

(PU). Using these events, we detect actions shown in Figure 1 – O_1: opening a new page in existing empty tab, O_2: opening a new page in existing tab replacing another page, F_L: following a link linearly, F_B: following a link with branching, C_p: closing a page. We track which tab is active at any given time using focus and defocus events.

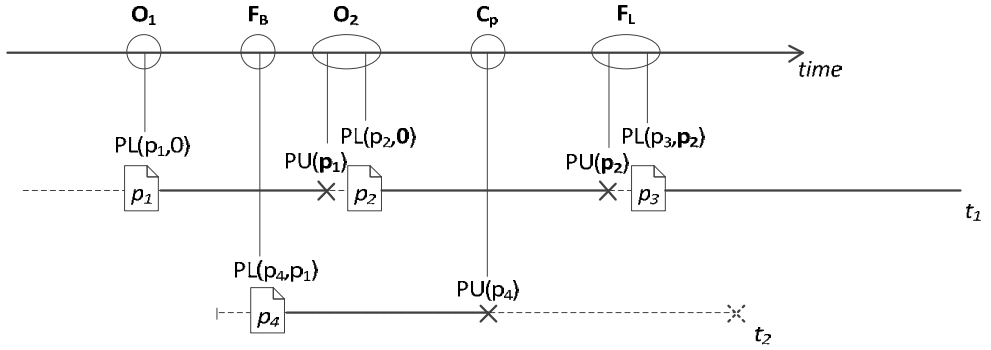


Figure 1. Sample parallel browsing session with two tabs.

By observing tabbed browsing of a user group larger than in limited studies, we know which resources are being browsed in parallel and more specifically, how users "travel" between them. For example, when users in learning system frequently branch from an exercise to a set of explanations and switch between them and the exercise, such explanations are helpful for that exercise and can be recommended to other users trying to solve it. There are also negative indicators, e.g., fast switching through the tabs. Suppose that the user opens a tab with our web application "in background" without visiting it. Later, when he switches through tabs looking for one particular tab, he may activate tab with our web application for brief moments of time and even use mouse or keyboard on it, but it is not a visit, interest, nor time spent on the page.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dubroy, P., Balakrishnan, R.: A Study of Tabbed Browsing Among Mozilla Firefox Users. In: *Proceedings of the 28th international conference on Human factors in computing systems – CHI '10*, ACM Press, (2010), pp. 673–682.

[2] Huang, J., Lin, T., White, R. W.: No Search Result Left Behind: Branching Behavior with Browser Tabs. In: *Proc. of the fifth ACM international conference on Web search and data mining – WSDM '12*, ACM Press, (2012), pp. 203–212.

[3] Ricci, F., Rokach, L., Shapira, B., Kantor, P. B. (Eds.): Recommender Systems Handbook. Springer US, (2011), pp. 842.

[4] Viermetz, M., Stolz, C., Gedov, V., Skubacz, M.: Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining. In: *2006 IEEE/WIC/ACM International Conf. on Web Intelligence (WI'06)*, IEEE, (2006), pp. 262–269.

Unified Search of Linked Data on the Web

Peter MACKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
mr.petermacko@gmail.com

Searching for information on the Web is increasingly difficult because of its enormous growth. To make matters worse, most of the data published on the Web is in unstructured format. However, more and more structured data is being published, which is also evident from the emergence of unifying initiatives like Linked Data. Structured data enables us to make web applications allowing users to search for information more comfortably. But querying this type of data is not a trivial task.

Nowadays, there are various structured data sources, but only few search engines are able to search in them utilizing the full power of the provided semantics. The majority of the search engines search for information using keywords which may not always give the users the results they desire. To utilize the full power of the structured data a special query language, like SPARQL, has to be used. However, queries in this language are not easily constructible for majority of standard users.

We would like to change this fact by creating complex search engine which could understand a pseudo-natural language of humans. In our approach, user just types in his request to the interface, interface sends his query to the server where it is transformed to SPARQL language. This SPARQL query is then passed to an ontological database.

The most difficult step in this process is the transformation mechanism which transforms user defined query to SPARQL. This is not a trivial task and we are now considering two simplifications:

1. User has a skeleton which will guide him through writing of a valid query. This is a simple way for transforming query but the user is limited in construction of the query.
2. We will employ natural language processor (e.g. Stanford CoreNLP) which will help our search engine to understand what the user wants. This is easier for the user but it is more difficult to transform the query to SPARQL.

After receiving the response from semantic database, the user will be able to edit the way how our constructor understands and reproduces the query. This will give us valuable feedback about the correctness of the transformation.

* Supervisor: Michal Holub, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 21–22.

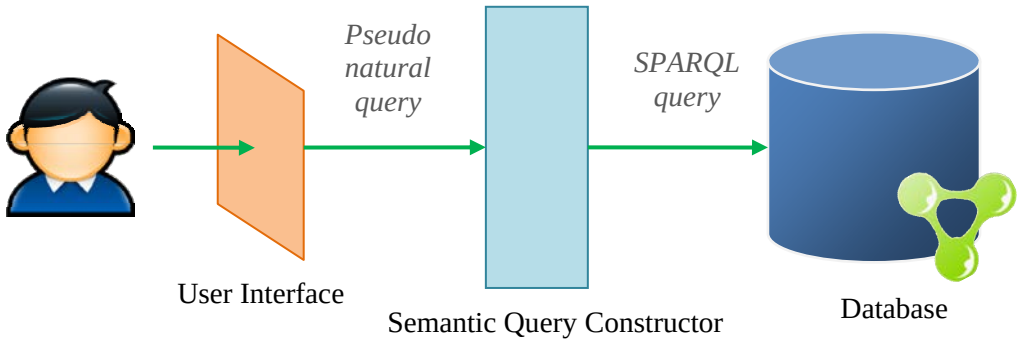


Figure 1.Schema of transformation from pseudo natural language to SPARQL query.

We would like to create combination of the two approaches described before by using suggestions. While the user types his query, suggestions will be shown next to the search box. The user will be able to select a query from the list or refine it and therefore we will guide him in writing a query which our constructor can easily understand and transform it to a SPARQL query. Constructor will then use a natural language processor to understand the terms in the query.

We will evaluate our method in the domain of scientific articles, authors and other parts of ACM, Springer and other digital libraries.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Valencia-Garcia, R., Garcia-Sanchez, F., Castellanos-Nieves, D., Fernndez-Breis, J. T.: OWLPath: An OWL Ontology-Guided Query Editor. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 41, No. 1, IEEE (2011), pp. 121-136.

[2] Tummarello, G., Cyganiak, R., Catasta, M.: Sig.ma: Live Views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web (8)*, Elsevier (2010), pp. 355-364.

[3] Kaufmann, E., Bernstein, A.: How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users? In: *Proc. of the 6th Int. The Semantic Web and 2nd Asian Conf. on Asian Semantic Web Conference*, Springer-Verlag Berlin, Heidelberg (2007), pp. 281–294.

[4] Bouquet, P., Stoermer, H., Bazzanella, B.: An Entity Name System (ENS) for the Semantic Web. In: *Proc. of the 5th European Semantic Web Conf. on The Semantic Web: Research and Applications*. Springer-Verlag Berlin, Heidelberg (2008), pp. 258–272.

Context-Aware Physical Activity Recommendation through Challenges

Štefan MITRÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
stefan.mitrik@gmail.com

The lack of physical activity is a phenomenon of this age. It negatively affects both our physical and mental health. Diabetes, heart-related diseases and cancer are just some of the diseases that are partly caused by the lack of physical activity. There already are some attempts to use smartphones and other technologies to help people exercise more [1,2]. However, more research in this field is needed.

We believe that personalized challenges can motivate people to exercise more and thus improve their health and the quality of their lives. Personalized challenges are physical activities we recommend to the user. A very simple example of such a challenge could be: "Can you walk 3000 steps in 3 hours?" or "Can you get to the nearest park in under one hour?". Every challenge has several attributes such as concreteness, dynamics, length, category or score. Example of a challenge with high dynamics is a game where user has to "catch" a running Yeti-like creature. Users can see moving image of the Yeti on the map and their goal is to get close enough and thus "catch" the creature.

We believe that there are some people who might consider such a challenge very interesting and funny but there are also others who might not like it. As we all know, people's preferences differ significantly. There are some of us who prefer a large number of shorter and more focused challenges, but others might like longer and more dynamic challenges. We assume that the challenge tailored for a specific user can achieve higher acceptance rate and thus positively affect his or her level of physical activity.

We are already able to track users' physical activity throughout their day with their smartphones, so no additional hardware is needed. Also, the recommendation and visualization of the challenges takes place in the phone application. A smartphone allows us to exploit contextual information [4], such as the user's location, agenda or physical condition, in order to recommend challenges.

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 23–24.

Most of the activities we perform during our days have repetitive character. We usually go to and back from work at certain timespan or attend our dancing lessons at specific time and day. These repetitive activities form certain patterns of the user's physical activity and interests. Thanks to these patterns we are able to predict one's movements and habits [3]. We calculate probabilities of movements from location *A* to location *B*, which subsequently could be used to tailor concreteness of the challenge. For example, if we discover that the user goes straight home from work almost every Wednesday we can recommend a challenge that takes place in a park near user's home. However, if we discover that there are many distinct places where user goes from work on Friday, we recommend a challenge that is not bounded to a certain location.

Another important thing is time for recommendation. For example, we can discover that our user leaves her office between 4 p.m. to 4:15 p.m. on Tuesday and thus assume that the ideal time for a challenge recommendation is at 3:45 p.m. just before she leaves the office.

The weather is a thing that affects our physical activity patterns a lot. However, while there are people who prefer sitting at home during the rain, there are also those who enjoy walking in rain with their umbrellas or raincoats. Our recommendation system learns these user-specific preferences and exploits that for better recommendation.

The certain category and form of the challenge is personalized for specific user according to previous implicit ratings of challenges in similar contexts.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Lin, Y., Jessurun, J., Vries, B. D., & Timmermans, H.: Motivate : towards Context-Aware Recommendation Mobile System for Healthy Living. In: *Pervasive Computing Technologies for Healthcare*, IEEE, 2011, pp. 250-253.

[2] Lim, B. Y., Shick, A., Harrison, C., & Hudson, S.: Pediluma : Motivating Physical Activity Through Contextual Information and Social Influence. In: *Human Factors*, ACM, 2010, pp. 173-180.

[3] Ashbrook, D., & Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. In: *Personal and Ubiquitous Computing*, ACM, 2003, pp. 275-286.

[4] Song, J., Tang, E. Y., & Liu, L.: User Behavior Pattern Analysis and Prediction Based on Mobile Phone Sensors. In: *IFIP International Federation For Information Processing*, ACM, 2010, pp. 177-189.

Navigation Using Annotations in Web Documents

Jakub ŠEVCECH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
sevo_jakub@yahoo.fr

We are facing the rise of services for sharing various kinds of content, whether in the form of links to interesting web pages, images, comments or multimedia information. Many applications provide tools for creation of annotations into web pages or various electronic documents, but few of them use these annotations to provide additional value to their authors. If small additional effort is made by the authors and/or the amount of created annotations reaches critical value, it can provide interesting benefits to users when they visit previously read documents.

We aim to provide a reward for inserting annotations at the time of their creation. Annotations can be used in different ways to improve information retrieval [1, 2]. We work on supporting navigation between documents using annotations. Namely we use annotations as an input for the process of creation of search query that is used in web search engine to search for related documents.

The task of search for documents relevant to source document is very similar to the task of recommending citations to academic papers. Several authors deal with this task [3] and they are searching for citations using different methods taking into account the authors, context and other characteristics of the document, and the citations graph. The search for documents that are relevant to the annotated document is in several ways similar to the citation recommendation:

- source for query creation is another document, and
- relevant documents are searched using the source document, and
- relevant documents are searched for specific parts of source document.

In search using annotations, annotations represent important source information that can enrich the generated query. Annotations highlight the most important parts of documents and they determine the specific topics for which the user wants to get more information.

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 25–26.

By annotations we mean comments, highlights in the text, bookmarks or tags that users are often inserting into documents while reading them. They are actually electronic equivalents of marginalia that people naturally create while reading books or newspapers. Annotations inserted into the document are usually describing the part of the document that is most interesting for the user. That is why another reader of the same document might use them to find documents that provide additional information.

Annotations describe users' interests in two ways:

- They describe the interest of user for whole document and they suggest certain level of quality of document. Typical representatives of annotations describing user's interest in whole document are bookmarks, where the user explicitly saves documents for later use.
- They specify parts of the document that are the most relevant for user. Using these annotations we can extract topics which are most interesting for a particular user.

Not only annotations inserted by user can be used as a source of information for the process of query generation. We use the content of the document, document metadata and annotations of other users. In addition to information related to the source document, user annotations attached to different documents can be used as well. Annotations in other user's documents provide important information about users' interests. By means of these annotations we can provide personalized search results for a particular user.

In order to collect annotations as described, we will create a tool for manual creation of annotations to web pages by their visitors. Users will be provided with facilities for adding tags, highlighting parts of web pages, attaching comments to content and bookmarking these pages. When a user is reading and annotating the document, a query will be generated and the retrieved documents will be continuously displayed. When user finishes reading the document, he will be able to adjust the generated query to better match his requirements and he will be able to retrieve related documents using this query.

The main contribution is a method for creating queries to retrieve relevant documents using content of the source document, annotations attached to the document and other user's annotations.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proc. of the 16th Int. Conf. on WWW*, 2007, pp. 501-510.

[2] S. Xu, S. Bao, Y. Cao, and Y. Yu, "Using social annotations to improve language model for information retrieval," in *Proc. of the 16th ACM Conf. on Information and Knowledge Management*, 2007, pp. 1003-1006.

[3] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proc. of the 19th Int. Conf. on WWW*, 2010, pp. 421-430.

Group Recommendation Based on Voting

Ján TREBUĽA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
trebulaj@gmail.com

Personalized recommendation is very helpful for individuals, but there are many activities which individuals perform in the group. Satisfaction of the individual group member depends on several factors as group size, its composition and type of the group (e.g. established group, occasional group, random group or automatically generated group) [1]. However, each recommendation requires knowledge about the preferences of individual group members. We are able to gather users' preferences based on their ratings of a set of items. This set consists of items that are going to be used for recommendation generation. While gaining the preferences we have to consider the weight of particular rating of an item based on user's group status.

In our work we propose voting method which explores several approaches how to process acquired users' preferences. We verify this method by implementation of software prototype of a social web application in domain of movie recommendations. Social network environment provides sufficient set of users, which are able to join the group or create their own groups and invite their friends to these groups. After generating the recommendation, users' satisfaction is observed and evaluated.

The proposed method consists of two parts (Figure 1) – the initial part is the users' ratings processing and the generation of the groups' recommendations.

For users' ratings pre-processing, our method uses a form of normalizing. We let each user a_j to submit a numerical vote $score(s_i, a_j)$ for each item s_i reflecting the preference of the particular item. These votes are given as ratings, for example 3 out of 5 stars, and normalized so that the scores given by each user sum to 1:

$$score(s_i, a_j) = \frac{rating(s_i, a_j)}{\sum_i rating(s_i, a_j)}$$

During the process of generating group recommendations, we use the following two aggregations strategies [2]:

- average aggregation strategy,
- multiplicative aggregation strategy.

* Supervisor: Michal Kompan, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 27–28.

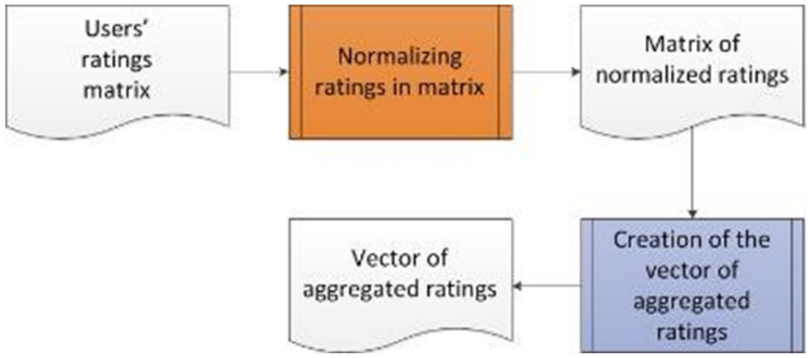


Figure 1. Sequence of steps describing proposed method.

The average aggregation strategy is based on averaging individual ratings. While, the multiplicative aggregation strategy is based on multiplies of individual ratings. Both these strategies return a vector of aggregated ratings of items. Items with the highest score will be recommended to the group. After the process of recommendation generation we seek to receive feedback by questioning the users about their satisfaction with the results of the different strategies.

The proposed method will be verified in several iterations by the web application aimed at movie recommendation for groups of users which want to watch movie together. Application is set-in directly into the Facebook [1] social network. After logging in, the user is able to join already existing groups, create his own group and invite his friends to the groups. Each group member is able to rate a list of movies, and add a new movie as well. The possibility to rate a movie is time limited. During the process of rating movies, preliminary results are visible to the user. After a timeout, possibility of voting is retracted and final recommendation is generated for the given group. Based on explicit feedback we will find out the level of satisfaction with recommendations using the particular strategies of aggregation.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Boratto L., Carta S. State-of-the-Art in Group Recommendation and New Approaches for Automatic Identification of Groups, 1-20, 2010.

[2] Ricci F., Rokach L., Shapira B., Kantor P. B., editors. *Recommender Systems Handbook.* Springer, 2011.

User Modeling, Virtual Communities and Social Networks

Integration and Adaptation of Motivational Factors into Software Systems

Pavol BIELIK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
pbielik@acm.org

Gamification defined as „applying the mechanics of gaming to non-game activities to change people's behavior", is often used in variety of areas including business services, behavior promotion, content portals or even in project management web applications. After its widespread adoption in the second half of 2010, the term itself is unfortunately now often misused for marketing purposes. Nevertheless, the idea of using game mechanics and dynamics to drive participation and engagement, mostly by using extrinsic motivation, is certainly worth further examination. Examples of game mechanics include points, levels or challenges whereas game dynamics include rewards, competitions or achievements.

The entire field however needs to be examined to determine what elements work in what situations as we do not currently know how exactly they affect our motivation, both positively and negatively, and which combination of game mechanics are suitable in given situation. Moreover, is it still not clear what effect these mostly extrinsic game mechanics have on human intrinsic motivation as "corruption effects of extrinsic incentives" [4] could overweight positive aspects. Research suggests that using an extrinsic reward have significant negative effect on our motivation as they undermine free-choice and self-reported interest in the given task [5, 6]. Recent study of badge systems [1] however suggests that negative aspects are mostly contributable to poor design of such systems.

In our work we seek to integrate and adapt game mechanics and dynamics in the domain of health promotion, specifically to motivate people engage in appropriate physical exercise. The solution will be implemented in Move2Play [3] system, which already provides required activity tracking, evaluation and recommendation of appropriate physical exercise for our purposes on Android platform. This system will also be deployed in Android Market, which should hopefully provide enough users for evaluation of our proposed method.

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 31–32.

Integration of motivational factors is based on extensive review of literature and existing applications. We have identified both common game mechanics and dynamics used in various domains, focusing mostly health promotion. As motivation varies among people we integrate various types of motivational factors, both intrinsic and extrinsic, as for a single user it is more engaging when there are multiple types of motivation throughout the day. For each user we maintain his personality model proposed by Bartle [2] using four personality types – Achievers, Explorers, Socializers and Killers. This model is used when choosing appropriate game mechanics and dynamics for users.

Adaptation consists mainly from tailoring motivational content using user interests. To obtain such user interests we will use rather simple approach consisting of analyzing popular social networks such as facebook and twitter which we believe will be sufficient for our purposes. Relevant interests include favorite music, movies, TV serial, books, sports or hobbies. When recommending appropriate content, it is equally important to consider interests of user and her friends when promoting social motivation.

Game mechanics and dynamics are placed within certain content, for example music band, a movie, book or a local or global event such as valentine. Content is important because core mechanics does not change very often and therefore we need to change content of these mechanics to keep user interested. Also the domain of physical exercise inherently contains repetitive tasks, which could become boring over time, if they are not perceived as boring already from beginning. Most of current integrations of game mechanics use content sources created by domain experts, but we plan to propose a method to obtain subset of such content automatically.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Antin, J., Churchill, E. (2011). Badges in Social Media: A Social Psychological Perspective. *Human Factors*, 2011, ACM, pp. 1-4.

[2] Bartle, R.: Hearts, Clubs, Diamonds, Spades: Players Who Suit MUDs. *The Journal of Virtual Environments*, 1996, Vol. 1, No. 1.

[3] Bielik, P., Tomlein, M., Krátky, P., Mitrík, Š., Barla, M., Bieliková, M.: Move2Play: an innovative approach to encouraging people to be more physically active. In: Proc. of the 2nd ACM SIGHIT International Health Inf. Symposium. IHI '12, NY, USA, ACM, 2012, pp. 61–70.

[4] Deci, E.: Effects of Externally Mediated Rewards on Intrinsic Motivation. *Journal of Personality and Social Psychology*, 1971, Vol. 18, pp. 105-115.

[5] Deci, E., Koestner, R., Ryan, R.: A meta analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125(6):627–668; discussion 692–700, 1999.

[6] Kohn, A.: Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes. Mariner Books, 1999.

Recognizing User's Emotion in Information System

Máté FEJES*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
matefejes13@gmail.com

Human emotions and theirs signs are innate to humans, regardless of the particular person. Thanks to that they can serve as implicit feedback from the users in information systems. Mimics of the face are unconscious signs of reaction to affect. According to a number of researches in the area of psycho feedback different movements of the face are common for all people, so we can derive reasons for these reactions – emotions. Applicable in various domains, e.g. in the case of educational systems we can estimate user's opinion about a text from his/hers face mimics, so we might be able to find out his/hers knowledge, interests, mood and other attributes that we would normally be unable to determine using only traditional feedback.

In this project we deal with gaining, representing and utilizing of user emotions while using web-based information system. To find out, what is on the subject's mind, we need to have a camera that records the user's face. For the extraction of emotions from video we can use a number of existing tools, which are able to recognize human face and its facial features. Facial features are important points on human face (e.g. border of mouth or eyes). Their locations depend on the movements of facial muscles therefore on the emotions of the user.

Our aim is to develop a method for user modeling based on emotions invoked during their work in a web-based system. Our method is going to be based on results of experiment we plan to realize in real environment with users. Within the experiment we will track the users by a webcam while they are working in the selected system. By the means of comparison of extracted emotions with users' activities we will try to explore relations between the actual activity of the user and the executed actions in the user interface (e.g. web browser). The goal of the experiment is to identify clusters of activities that are specific to users in certain emotional state. This way we will be able to find out the way emotions are invoked by a given content.

We plan to deploy our emotion recognition method in an existing adaptive web-based learning system and it will be utilized for enriching of the user model. One of the

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 33–34.

most important metadata in the user model of such a system is the user (student) knowledge level regarding different topics, so we try to contribute to estimation of these values. To achieve this we have to find the way of deriving knowledge from emotions. For this reason we prepare a set of short texts (stories, news, parts of curriculum etc.) and a sequence of questions according to each text that will be read and answered by users within the experiment. We compare users' answers with the emotions detected to identify the relationship between emotions and the level of student knowledge.

The result of our method will be a conceptual map that represents the relationship between content visited by the user, the emotions invoked by the content and the user's knowledge level. The model could be utilized for collaborative filtering using traditional methods – estimating similarity between users or items, calculating relevancy level of given content for the user. The relevancy level could be expressed either by knowledge or emotional state – for users in certain emotional state should be recommended items by similar emotional character. Another way of utilizing of the model is adaptive teaching (selecting the proper level of explanation depending on the user's reactions in real time) or adaptive testing (selecting a question of proper level depending on user's reaction to the previous question).

Thanks to the learning environment which the method will be deployed in we have a good opportunity to evaluate it. Since we want to estimate users' knowledge by the help of implicit feedback, we can easily check the precision of estimation. By asking users some questions according to the content that has been already modeled we can compare the implicit and explicit feedback.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] M. Yashar (2012): Role of emotion in information retrieval. PhD thesis, University of Glasgow.

[2] Woolf, B., W. Burleson, I. Arroyo, T. Dragon, D. Cooper and R. Picard (2009): Affect-aware tutors: recognizing and responding to student affect. International Journal of Learning Technology, 4(3/4), pp. 129-163.

[3] Grafsgaard, J., Boyer, K., Phillips, R. and Lester, J. (2011): Modeling Confusion: Facial Expression, Task, and Discourse in Task-Oriented Tutorial Dialogue. Artificial Intelligence in Education 2011. Springer. pp 98—105.

Gathering Information on User Environment

Tomáš JENDEK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
jendek.tomas@gmail.com

Nowadays people use mobile devices on daily basis. Coverage of wireless networks, mobile devices and software platforms is making mobile computing a mainstream issue. Every user is situated in a certain environment, which provides contextual information. Context represents situation or condition [2]. Context can be as simple as location, or as complex as emotions. In our work, we focus on obtaining location context as contribution to social context acquiring. We need to define an effective way for obtaining location context. We propose a method for tracking user's location in time, estimate future location of the user, and define user's important places. This information can be further used in recommendation systems.

There are existing solutions based on acquisition of location using GPS module on mobile devices. This is not very battery efficient, however, so our solution uses GSM transmitters. We map GSM transmitter towers to GPS positions, providing an energy efficient solution. There is a certain lack of accuracy in using positions obtained by using GSM transmitter, but it is sufficiently accurate for our purposes, and additionally, it can be considered better privacy and security.

There are services for obtaining user's and friend's location. They are mostly oriented for real-time tracking. That is, user needs the internet connection for obtaining the location and also a GPS module. There are cons for these solutions, for example high energy consumption due to using GPS module. The advantage is that these solutions are relatively accurate (error of less than 10 – 20 meters).

People visit various places on a daily basis and visited locations can induce connections between people/friends. People's co-presence in places provides a link between people or between places [1]. This is useful for us to determine the social context of the user. If we can determine social connections between user and his friends, we can easily suggest, for instance, meeting or any other activity using a different context such as calendar. Our concept of obtaining user's location lies in implementing a mobile application, which provides a useful service for users and also helps to obtain the location context. A user is motivated to use the application when

* Supervisor: Dušan Zeleník, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 35-36.

he/she wants a prediction of a friend's location. An estimate of friend's location and a possible meeting arrangement can be provided.

Our method is based on tracking user location and analyzing this data. First we find important places for user to filter out other unimportant places, e. g. when user is commuting. Important places are places in which the user spent certain time. The most important for us is to determine whether the user is at home or at work. Assuming that average person sleeps at night and works during day we estimate whether user is at home or at work/school. In our experiment, we estimated user being at home with 93% success rate and user being at work with 68% success rate. For predicting user's future location in time, we created a time vector and try to find appropriate similar vector in our database of user's logs.

Table 1. Time vector sample.

Minute / 60	Hour / 24	Day / 7	Week / 4	Month / 12	Year/10000
0.232	0.543	0.468	0.887	0.229	0.2012

Time vector (see Table 1) consists of five columns. Minute of the hour, hour of the day, day of week, week of month, month of year and year / 10000. This time vector is compared with history using cosine similarity to estimate future location. There are some options of adjustment for vector components, i.e. hour of the day is more important component than year or month. Day of week and hour of day are more important vector components than month or year in discovering user's behavior pattern, so we adjust cosine similarity vectors.

To evaluate our solution, we use retrospective analysis – implicit feedback. We predict location in future time and we are able to verify whether our prediction was correct or not. For verification user's home/work location we use explicit feedback. To sum up, we presented a method for location prediction, which facilitates contact between people by implementing mobile application for Android OS. This application tracks user position and estimates future position of user's friend and important locations. In future work we want to discover relationships between friends based on common location as contribution to social context acquiring.

Acknowledgement: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Konomi S.: Colocation networks: exploring the use of social and geographical patterns in context-aware services. In: *Proceedings of the 13th international conference on Ubiquitous computing*.

[2] Zeleník D.: An Approach to Context Aware Event Reminding. In *Proc. of the Information Sciences and Technologies Bulletin of the ACM Slovakia*, pages 126-130. ACM, 2011.

Trend-Aware User Modeling with Location-Aware Trends

Marcel KANTA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kanta07@student.fiit.stuba.sk

Microblog is highly significant source of social information and interaction. In our work we use Twitter as a source for user modelling and we focus on creating a user model to help people cope with the issue of information overload.

There are various approaches to user modelling in web environment for recommendation of various types of entities [1, 2, 3]. Personalized recommendation is more important than trend-aware recommendation, but integrating trend-aware and personal recommendation can improve recommender results. The aim of our work is to incorporate trend-awareness into personalisation. We also consider locations-awareness to improve the results, thus the user model will be more precise. The idea is based on the assumption that employing location of trends will improve the quality of user model. We believe that applications that will incorporate our user model will have more precise results compared with traditional non-location-aware model.

We formally define our user model as follows:

$$P(u) = \left\{ \left(c, w(u,c,l) \right), l \middle| c \in C, u \in U, l \in L \right\}, \tag{1}$$

where *c* stands for concept, *w* for weighting function, *u* for user and *l* for location. We introduce location *l*, which means that every concept and user belongs to a region and its parent regions that are represented by quadtree structure (see Fig. 1). We use TF-IDF and t-TF-IDF [2] for region and time as a weighting function *w*. t-TF-IDF is a trend-aware modification of standard TF-IDF. It uses temporal stability of concepts in a form of computing a standard deviation of appearance of concepts in time quanta.

Our hypothesis is that location-awareness will improve the quality of user model. The principle of location awareness is that user model is modelled in regions, so weighting is done with region in mind; weighting only per region. This is the key that enables location-awareness. Our location-aware model uses at most $M(\log n)$ times more data than traditional model where *n* is maximal number of regions.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 37–38.

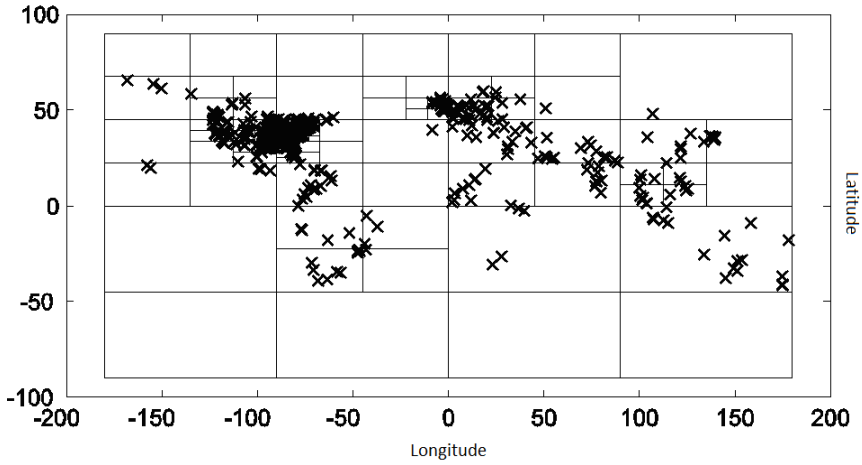


Figure 1. Our quadtree regions generated from the dataset.

Our user model will be quantitatively evaluated using a Twitter dataset. We chose Map-Reduce as a platform for evaluation. We plan a general synthetic evaluation approach used in machine learning. In testing phase we recommend top *n* items that are matched by cosine similarity and then test, if user actually posted a link that matches one of the recommended links. For evaluation of results, we will use Precision P@n measure. We will show the difference between location aware and not-aware user model. This is how we will prove the quality of location-aware user model.

This approach does not depend only on domain of microblogging, it may be used everywhere where we want to model entities with semantic information, where location matters. Also, regions of our approach do not have to be based on geographical coordinates, it may be a tree of clustered data, e.g., students in classes and recommendations in different classes and schools.

Acknowledgement: This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Abel, F., Gao, Q., Houben, G.J., and Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In *Proc. of the Conf. on User Modeling, Adaption and Personalization*, Springer, (2011), pp. 1–12.

[2] Gao, Q., Abel, F., and Houben, G.: Interweaving Trend and User Modeling for Personalized News Recommendation. In *Proc. of the Int. Conf. on Web Intel. and Intel. Agent Technology WIIAT 2011*, IEEE, (2011), pp. 100–103.

[3] Morales, G.D.F., Gionis, A. and Lucchese ,C.: From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation. In *Proc. of the Fifth Int. Conf. on Web Search and Web Data Mining* (WSDM), ACM (2012), pp. 153–162.

User Modeling Using Social and Game Principles

Peter KRÁTKY*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kratky.peto@gmail.com

Personality has a significant impact on the way users use web applications. Many systems would take advantage from adapting content or search according to personality traits. Examples of such applications are social networks that provide search for a suitable relationship or learning systems that might improve the efficiency of learning by adapting to different personalities which correspond to different learning habits [5]. Modeling user's personality traits in personalized systems is increasingly important [2]. However, this specific kind of user modeling is still less explored than the others and therefore is the subject of research in our project.

The traditional approaches for modeling user are explicit such as modeling based on questionnaire, especially when it comes to personality traits. This attitude is simple and straightforward, and typically takes the form of answer sheets, which are widely used to predict one's personality. On the other hand, explicit methods of data acquisition might be obtrusive or might seem too personal for the user [1].

In our project, we focus on designing user modeling method which collects data from the user in funny and appealing way. For this purpose we want to use benefits arising from computer games. The first benefit is the possibility to perform data collection while user enjoys playing the game, so the probability of providing data is higher than in traditional approaches. Additionally, the actions user performs during the game play reflect his/her actions in the real world. Research proves correlation between personality traits and actions done in the game [3]. However there are only few gaming actions only in specific games that are proved to map themselves to some personal traits. And this is a big issue we have to deal with.

In our work, we have defined few steps to follow, of both designing and experimental character. We have decided to integrate the process of verification in the whole process so we have adjusted the steps according to it. Firstly, we will start with designing an explicit question form for acquisition of the user's characteristics and deploy it in an information system. Further, we will design a set of actions tracked in

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 39–40.

the game and set of actions tracked in the information system. We want to prove the assumption that the actions in the information systems are insufficient to model personality profile and therefore the need for other approach arises. Finally we design an inference method for prediction of the user's personality profile.

The first step is crucial to decide which characteristics we want to model. The most widely used model in general practice is Big Five. This structure defines personality in five dimensions (Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Intellect). We can benefit from the fact that the characteristics are rather exclusive and represent personality at the broadest level of abstraction [4]. The advantage of a user model based on this well-known structure might be the possibility of reuse among different adaptive systems. We need a measurement instrument that allows us to obtain the accurate user's characteristics. For this purpose we use an official questionnaire and deploy it in an information system.

In the next step we use an instance of a game that is able to track performed actions. We take into consideration an open-source game or a game developed by students of our university. In the second experiment we will collect data about user interaction in the game and the information system. Then, we use regression analysis will to explore correlations between the actions and the personality traits. As mentioned above, we expect better results from the data provided by the game.

Using the discovered correlations we will design a method that infers user's personality from gaming actions. This method should be subsequently able to continuously model the user.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Barla, M.: Towards Social-based User Modeling and Personalization. Information Sciences and Technologies Bulletin of the ACM Slovakia. - ISSN 1338-1237. - Vol. 3, No. 1 (2011), pp. 52-60.

[2] Ho, S. Y., Davern, M. J., Tam, K. Y.: Personalization and choice behavior: the role of personality traits. SIGMIS Database, 2008, ACM, Vol. 39, pp. 31-47.

[3] Herodotou Ch., Kambouri M., Winters N.: The role of trait emotional intelligence in gamers' preferences for play and frequency of gaming. Computers in Human Behavior, 2011, Vol. 27, No. 5. pp 1815-1819.

[4] John, O.P., Srivastava, S.: The Big five trait taxonomy: History, measurement and theoretical perspectives. Handbook of personality theory, Guilford Press, New York, 1999.

[5] Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., Zancanaro M.: Modeling the Personality of Participants during Group Interactions. Conference on User Modeling, Adaptation and Personalization (UMAP 2009), 2009, pp. 114-125.

Decentralized User Modeling and Personalization

Márius ŠAJGALÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
sajgalik@stuba.sk

In our work we focus on a Web browser as a basic tool of a user. The goal is to shift adaptivity closer to the user and enable personalization of the Web content directly in the browser to improve and simplify users' work. That is the main difference compared to other solutions and also an advantage, because a user has all the data and she decides whom to trust and allow an access to it. It is not only the decentralized and distributed nature of the solution that differs, but the direct integration into the browser in the form of an extension. Using our approach, we have access to such key information, which remote servers have not (e.g., the user's detailed Web browsing history, user activity itself like working with browser tabs or mouse movements).

Our work can be divided into two basic parts. The first is the modeling of user interests based on the browsing history and user's activities within a browser. This part can also be seen as a kind of intermediate stage of pre-computation, which is later used in the second part – personalization. As a distributed and decentralized solution, it enables personalization directly on the client device with the possibility of communicating with other users, making use of knowledge and experience of more users to achieve even better personalization.

Proof that modeling the user and carrying out the personalization directly in the browser is not just possible, but even richer on captured data can be induced from [1]. The purpose of user modeling is to capture such user characteristics, which can be used further on to customize the browsed Web content. In the core of our implementation and design we focus on modeling the user interests based on the current Web browsing history, which can be later (within personalization subsystem) even more extended and customized.

The process of interests modeling is partly similar to indexing pages on the Web. Indexing involves the creation of an index of all pages that a crawler has found on the Web. In our case we have in hands a kind of an intelligent crawler, which does not visit all pages on the Web, but only those that are of interest to the user. In fact, it is the

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 41–42.

user herself who is in the role of crawler, who visits these pages and chooses them according to her own personal interests. Since we extract weighted terms representing user interests from the content of visited pages (evaluation in [2] shows it can be done with satisfying accuracy), we can say that our indexer produces an index of user interests instead of a regular index of any keywords and terms. Interesting part in our model of user interests is the ability of identifying also the local interests of user within some specific domain.

The second part of our solution deals with the personalization of Web pages in the browser. Based on the user interest model we personalize the Web content to the user so that her work in the browser is easier, more intuitive and efficient. The solution is not limited to one user, but uses the entire network of users and their relationships, thereby increasing the quality of the personalization. Moreover, communication channels allow users to help each other, share their experiences and collaborate to recommend the common interests of what they are most interested in.

The realization of our work represents a decentralized distributed collaborative platform for personalization of user interests in the Web browser. Physically, it is composed of multiple instances of the browser extension of individual users. These extensions communicate with each other, allowing multiple users to collaborate. Since the main goal is to enable personalization of content on the Web, this platform provides an interface for access to user interest model, which can be further extended.

The real personalization is done via personalization extensions, which are basically executed when a Web page is loaded and which can modify its content and thus personalize it to the user. These extensions are basically pieces of JavaScript code where besides the basic possibilities of this language also other functionality is available, like built-in support of jQuery framework and external JavaScript files inclusion, support of cross-origin requests, access to the user's browsing history and persistent database API, personalization API and communication API. Database API provides a unified interface for database access and personalization API provides access to information about user stored in user model. Communication API enables communication among users and has a form of channeled multicast in which users do not communicate directly with each other, but they use the communication channel to send the messages. Its purpose is to bring together a group of similar users based on their common interests.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Ohmura, H., Kitasuka, T., Aritsugi, M.: A Web Browsing Behavior Recording System. In: Lecture Notes in Computer Science: Knowledge-Based and Intelligent Information and Engineering Systems, Springer, Berlin, (2011), vol. 6884, pp. 53-62.

[2] Matthijs, N., Radlinski, F.: Personalizing web search using long term browsing history. In: WSDM '11: Proc. of the fourth ACM international conference on Web search and data mining, ACM Press, (2011), pp. 25-34.

Encouragement of Collaborative Learning Based on Dynamic Groups

Ivan SRBA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
srba07@student.fiit.stuba.sk

Computer-Supported Collaborative Learning (CSCL) is an approach to learning based on support of information and communication technologies. Research in CSCL can be grouped into systematic and dialogical approaches [2]. The systematic approach concerns the creating of models how the specific features of technological systems support or constrain collaboration, reasoning, knowledge representation, and structure of discourse [1]. On the other hand, the dialogical approach considers learning as a social-based activity.

We deal with the dialogical approach, especially with encouragement of students in collaborative learning by creating dynamic short-term study groups and design a collaboration platform which allows these groups to collaborate efficiently. The reason to follow this goal is the fact that we do not know what makes collaboration really effective.

The basic idea of our method for group formation is derived from Group Technology (GT) approach. According to Selim, et al. [3] GT is an approach to manufacturing and engineering management that helps manage diversity by capitalizing on underlying similarities in products and activities. GT seems to solve similar problem as we have to solve to reach our goal. Thus, we developed a new method inspired by *Group Technology* techniques.

The proposed method consists of two main processes. *Group Formation* takes different personal or collaborative characteristics as inputs and creates study groups. Personal characteristics can be student's knowledge, interests, or any other personal characteristics (e.g. age, gender). We can obtain these characteristics from many sources, such as existing user models, social networks or questionnaires. Furthermore, characteristics can be collaborative, such as friendship with other students or collaborative behaviour. *Collaboration* allows students of created groups to participate on task solving via a collaboration platform called PopCorm which provides

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 43–44.

appropriate collaboration tools together with functionality for observation groups' dynamic aspects which are used as one of inputs in the proposed method.

Evaluation of our method consisted of a long-term experiment during the summer term as a part of education on the course Principles of Software Engineering at the Slovak University of Technology in Bratislava. 106 students in total participated in 208 created groups. 3 613 activities are recorded during task solving.

Table 1. Comparison of achieved results during the second phase of the experiment.

Groups created	Average evaluation	Feedback
By the proposed method	0.459	4.01
By the reference method (k-means clustering)	0.392	3.55
Randomly	0.422	3.29

The 8-dimensional evaluation of the groups created using our method was compared with a reference method (k-means clustering) and randomly created groups (see Table 1). Groups created by our method achieved the most effective and successful collaboration in comparison with the other two types of groups. We employ ANOVA statistical model to evaluate significance of achieved results and we got p-value 0.0048. Thus, the achieved results can be considered as statistically significant. Additionally, students have provided a higher explicit feedback in these groups.

Our method is not limited only to the CSCL domain. It can be easily applied in other domains where dynamic groups should be created according to different user characteristics. We have successfully applied the proposed method during the experiment in collaborative learning by creating dynamic short-term study groups, which showed high potential of the proposed method. It would not be possible to evaluate our method for group creation without the PopCorm collaborative platform which provides students the appropriate environment for effective task solving and automatic identification of their activities.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dillenbourg, P.: What do you mean by collaborative learning?. In: Dillenbourg P. (eds): *Collaborative learning: Cognitive and Comp. Approaches*. Oxford: Elsevier, 1999, pp. 1-19.

[2] Ludvigsen, S., and Mørch, A.: Computer-supported collaborative learning: Basic concepts, multiple perspectives, and emerging trends. In *The Int. Encyclopedia of Education*, 3rd Edition, edited by B. McGaw, P. Peterson and E. Baker, Elsevier (in press), 2009.

[3] Selim, H.M., Askin, R. G., Vakharia, A. J.: Cell formation in group tech.: review, evaluation and directions for future research. Comput. *Ind. Eng. 34*, 1 (January 1998), 1998, pp. 3-20.

Feedback Acquisition in Web-based Learning

Andrea ŠTEŇOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
andrea.stenova@gmail.com

Explicit feedback from website visitors is very important and its meaning is constantly growing nowadays. For efficient web-based learning, we need information on how students interact with an educational system, which materials they find hard to learn, which are insufficiently explained, or which on the other hand they like. User's interests and opinions can be determined by user feedback acquisition, thanks to that it is possible to review, improve, recommend and personalize webpage content.

Most of the time, students are not willing to provide the feedback or they do it only when they are very satisfied or not satisfied at all. Another problem is that we should not disturb students during their studies. Inappropriate explicit feedback might bother them, however, lack of explicit feedback and irreplaceable information we receive from it is a great issue that should not be ignored.

To rate acquisition we use rating scale, the widget to express user's preferences on selected range. Different users prefer using different scales for their ratings [1] and also rating scale itself affects user's rating [2, 3]. The way rating scales are perceived by user is called rating personality. A rating scale has characteristics such as range, visual representation in the system and presence or absence of neutral value. Selection of scale depends on domain, size (texts or contents) of object, in which it is displayed, and of users' preferences. User rating process consists of these steps:

1. display of rating scale in object,
2. user rating and logging,
3. object evaluation.

Displayed scale is selected according to several rules. Its scope is determined by a range of the displayed object. For short texts we use smaller range, however, for larger texts ratings should be more accurate, so we use larger rating scale range. Displayed scale can change after user has rated an object. If the ratings are concentrated around neutral value, we change the range of the scale to leave out neutral values and force user to decide. It is possible to extend the range of the scale in cases when user ratings contain the same value very often. That means that he does not have

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 45–46.

enough options to rate as he wants. To eliminate problem with changing scale without informing the user, we only offer the change to user and he has to approve it.

After user's rating we record rating value, object, that was rated, and also rating scale, on which user rated. It may happen that one object will have ratings on different scales as scale type can change. Since with change of rating scale we do not want to throw away previous ratings, we have to recalculate them. However, it causes the problem with transformation of these ratings. The transformation may not be easy, in out method we use mathematical normalization.

We have to evaluate the results, to decide which objects in a system are good and bad. This evaluation is computed from users' ratings and we also normalize these ratings. Results are transformed according to used scale and to previous user ratings. All ratings are converted into $\langle 0,1 \rangle$ interval and then we normalize them. When transforming according to previous user ratings, we extend ratings onto whole interval. By using this technique we can effectively eliminate user overly positive or negative ratings.

We have partially evaluated transformation of ratings between different rating scales. Moreover, we found out most and least preferred rating scales. In future work we will focus on other methods of feedback. We will evaluate the proposed method in the educational system ALEF (Adaptive LEarning Framework) [4]. Collected feedback can be used to delete or improve bad learning objects, to recommend learning objects with the best results and to personalize the content according to the identified user preferences.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Amatriain, X., Pujol, J.M., Oliver, N.: I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In: *Proc. of the 17th Int. Conf. on User Modeling, Adaptation, and Personalization: formerly UM and AH*. UMAP '09, Berlin, Heidelberg, Springer-Verlag, 2009, pp. 247–258.

[2] Cena, F., Vernero, F., Gena, C.: Towards a customization of rating scales in adaptive sys-tems. In: *Proc. of the 18th Int. Conf. on User Modeling, Adaptation, and Personalization*. UMAP'10, Berlin, Heidelberg, Springer-Verlag, 2010, pp. 369–374.

[3] Gena, C., Brogi, R., Cena, F., Vernero, F.: The impact of rating scales on user's rating behavior. In: *Proc. of the 19th Int. Conference on User modeling, adaption, and personalization*. UMAP'11, Berlin, Heidelberg, Springer-Verlag, 2011, pp. 123–134.

[4] Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-Based Learning 2.0. In *Proc. of IFIP Advances in Information and Communication Technology*, Vol. 324, Springer, 2010, pp. 367–378.

Method for Social Programming and Code Review

Michal TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
michal.tomlein@gmail.com

Code quality in software projects can generally be achieved in a number of different ways. In order to create high quality, reliable and reusable code, incorporating code review into the development process is among the most effective of the available options. The use of code review is, however, not limited to the domain of software development. In programming courses, peer review has the potential to be an effective driving force behind the learning process.

However, due to the significant amount of time reviews take, whether in software development or a programming course, they cannot be and are not done thoroughly in practice. Specifically in programming courses, reviewing code requires the selection of suitable reviewers, as pedagogues and graduate students are often not available in sufficient numbers to be able to help all of the students.

For this and other reasons, peer reviews are used to complement normal reviews. In many cases, peer reviews have been found to be adequate substitutes for pedagogical reviews. Peer reviews do, however, present a few problems in their realisation. Students' ability to review code and provide useful feedback varies to the degree that it becomes very important to select suitable reviewers for the individual problems the students have.

Reviewer selection is a non-trivial problem. The relevance and quality of the resulting review is highly dependent on reviewer selection, as different reviewers may have a different amount of experience with the particular problem they are assigned to review. Research has shown that reviewers with insufficient knowledge and experience in the problem area only contribute to user confusion and do not provide the necessary motivation [1]. Because of this, there is a need for a solution, which takes such reviewer attributes, as well as the reviewers' availability, into consideration when selecting the most suitable reviewer for a given problem.

Currently, there are a number of solutions aimed at reviewing code in the form of commits in a version control system. These solutions offer a varying degree of

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 47–48.

automation of this process, ranging from simple solutions to systems such as Gerrit, which can interoperate with continuous integration systems such as Jenkins. Some solutions, such as CodeCollaborator, offer integration with development environments such as Eclipse or Visual Studio. The use of these solutions is, by their nature, generally limited to larger projects and/or requires a certain discipline to be effective.

Previous research activity in the field of code review has also focused on code review as part of programming education. For example, the Caesar system was designed with the students' lack of experience with version control in mind, which is a major obstacle in the use of common code review solutions in programming courses [2]. Caesar attempts to overcome this problem through the division of code into smaller partitions, as opposed to other systems based on incremental code review.

Reviewer selection is generally done manually or using systems with defined rules, priorities and responsibilities, which divide work evenly among the available reviewers. Social approaches have so far not been applied to reviewer selection.

We propose a solution which integrates social approaches and collaboration into the process of code review. The social approaches are based on user and reviewer modelling. We intend to model the experience and knowledge of reviewers primarily using review assessment by users. This explicit feedback will be aggregated in problem areas. The resulting model will be used to select reviewers for each new case. The user model can also be used in reviewer selection to match reviewers who have received positive feedback from users with a similar level of domain knowledge.

To evaluate the proposed solution, we intend to create a prototype of an extension for an integrated development environment used in programming education. This extension will serve to enable students to request reviews of their code and ask for help. In addition to this, a prototype of a web application will be created for reviewers to get access to pending requests for review. This application will provide the tools necessary to review the submitted code and provide feedback and help to the users.

In our work, we aim to make programming more effective through social interaction and peer reviews. We believe it is very important to reduce friction in the process, from asking for a review to getting feedback and communicating with a reviewer. To do so, we intend to integrate our solution with widely used development environments. We believe that by making it possible for students and software developers to collaborate more tightly and easily, we can speed up the development process and achieve higher quality overall.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Hundhausen, Ch.D., Agarwal, P., and Trevisan, M. (2011): *Online vs. face-to-face pedagogical code reviews: an empirical comparison.* Proceedings of the 42nd ACM technical symposium on Computer science education (SIGCSE'11), ACM, New York, NY, USA, pp. 117 – 122.

[2] Tang M. (2011): *Caesar: A Social Code Review Tool for Programming Education.* Master Thesis at Massachusetts Institute of Technology.

User Modeling in Education Domain

Maroš UNČÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
xuncik@is.stuba.sk

The trend of using web-based educational systems is progressively growing and opportunities that the Web provides are huge. Nowadays, e-learning systems offer more and richer content, enable communication and collaboration among users. This evolution is related with information overload, which caused increase of difficulty to find and classify certain information. In order to allow personalization, web-based educational systems monitor characteristics of individual users, including modeling of their skills, knowledge and/or interests. The performance of such systems is derived from a core element – the user model, which is used to minimize errors and learning time [1].

Many problems in the domain of user modeling were identified in the past. Combinations of several different inputs entering the user modeling process or the use of information about user beyond the adaptive system to enrich the user model are just two of them. Another challenge, *scrutability*, concerning the visibility of the user model to users, is also closely related. In most of the systems, the user cannot directly access the user model and cannot provide explicit feedback about him/her, which could be otherwise taken into account.

In our work we deal with visualization of user model. To visualize the user model we design an overlay user model based on lightweight domain model representation [2]. In the proposed user model, built on top of the domain model, we use domain-independent (e.g., age) and domain-dependent characteristics (knowledge and interests). The value of domain-dependent characteristics is represented by a three-dimensional vector [*level*, *confidence*, *source*]. The level of a characteristic denotes its value. It takes real values from a closed interval <0, 1>, where 1.0 denotes the maximum value. Each characteristic is assigned with *confidence* expressing a probability that a user has given value in real life. Each characteristic is determined by its source, which can be either a tool or a method.

Characteristics of knowledge and interests are directly changed by the behavior of the user, based on recorded activity in the educational system. To determine these characteristics, we can use several sources. The problem is that every single

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 49–50.

characteristic has more probability-value pairs from different sources. For practical reasons we need only one value-confidence pair, which characterize a relation between user and concept. Therefore, each pair must be appropriately combined. To this purpose we designed a method combining multiple inputs entering the user model.

Our main aim is to create a visualization of the user model of educational system, which allows us to interfere between believes modeled by the system and opinions of user herself. Visualization thus serves also as a tool to evaluate user model. We assume three main features: (1) to show the student what are the system's believes about her, (2) to provide an insight into a user model and to show its changes over time, (3) to allow students to give explicit feedback.

We decided to visualize user's knowledge as a graph. Vertices in the graph represent domain relevant terms and links among them represent relationships represented in the domain model. Each term has connections to other term if there is a relationship between them. We use two main relationships:

- prerequisites (it is presented as solid oriented narrow),
- relatedness (it is presented as dashed non-oriented line).

Another important aspect is the color scale we used in the graph. We use levels of green, red and white color. The green color is used to label the terms that user has understood. The brighter color indicates the level of user knowledge about the term. The red color shows the term with some sort of defect in the learning process

An important part of visualization is an explanation for the user, represented in a natural language as a pop-up window, which appears while pointing to the concept. The explanation gives basic information about knowledge achieved by user and is generated according to information, we had logged about a user. We use the weight of the *concept-to-learning-object* relation from the domain model to show only relevant learning objects related to chosen concept.

We believe that described method enables users to better understand system's believes about them and enhance the user-system interaction by making the user model visible. We presume that visualization is not just another feature for users of an adaptive system, but rather a tool to help in understanding of possible differences between reality and system's believes. This opens further possibilities in the use of visualization for qualitative evaluation of a user model, which is currently considered as a difficult and challenging task.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Brusilovsky, P.: Methods and Techniques of Adaptive Hypermedia. User Model. User-Adaptation Interaction, (1996), 6(2-3):87–129.

[2] Šimko, M., Bieliková, M.: User Modeling Based on Emergent Domain Semantics. In*: LNCS 6075 User Modeling, Adaptation, and Personalization*, Springer-Verlag Berlin Heidelberg, (2010), pp. 411-414.

Association Rules Mining from Context-enriched Server Logs

Juraj VIŠŇOVSKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
visnovsky.j@gmail.com

As users are browsing the Web, servers are recording millions of theirs actions to eventually offer better service. These large amounts of Web logs should be reused, otherwise their recording could be considered as a waste of resources. This field is covered by Web usage mining which discovers Web usage patterns. In our work we aim to prove the importance of context in the field of Web usage mining.

By using Web usage mining techniques we are able to discover information about users' behaviour. Resolving their habits may be used in improving personalized recommendation and targeted advertising. For our purpose we mine frequent patterns using FP-Growth algorithm and then generate association rules.

In this paper we focus on the analysis of Adaptive proxy server [2] access logs. The goal of the Adaptive proxy server is to improve user's experience by personalising Web content. Our dataset consists of more than 3 million access logs describing the activity of 77 unique users. When analysing server access logs, we have to bear in mind that these logs represent human actions and we have to consider many different contexts which could affect user's activity. Expressing what the word *context* means can be difficult and so is to define it. Probably the best definition came up from Dey [1]. According to this definition context is any information that can be used to describe a state of an entity. This entity could be a person, an object or a place that is relevant to the interaction between an application and a user.

We consider only small amount of contextual information influencing user's actions. As we are limited to use the Web access logs, we are not able to find out neither user's current mood nor his exact location. We use contexts as follows:

- Time
- Location and occupation
- Weather
- Web domain category

* Supervisor: Dušan Zeleník, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 51–52.

Our method analyses logs gathered by Adaptive proxy server and generates association rules from it. The process is performed in three steps, depicted in Figure 1.

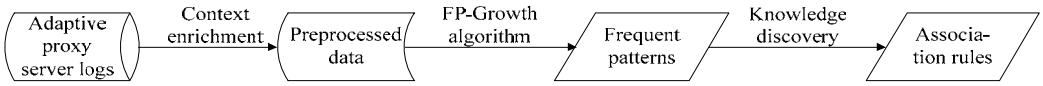


Figure 1. An overview of the method of association rules mining in access logs.

An association rule represents a correlation between two or more items which do not seem related at the first glance. An association rule can be expressed as logical implication A ⇒ B with attributes of support and confidence. Support is a probability that both A and B occur together while confidence is a probability of B occurring when A is already observed to be present.

For actual discovery of association rules we propose a modification of standard Frequent pattern growth algorithm (FP-Growth). While we are seeking to discover context-aware association rules, every node of the FP-Tree will be represented by a context-enriched record of access from access log.

FP-Growth algorithm needs two perform two scans through the access log stored in a database. At the first scan, algorithm evaluates occurrence of every context-enriched access record. Then the algorithm builds FP-Tree structure and inserts only the most frequent records as the tree nodes. Minimum support threshold defines how many times a record has to be noticed during the scan to be considered a frequent item.

FP-Growth algorithm handling large amounts of access logs may produce a high number of frequent patterns. The results may consist of strongly related association rules. These can, however, be misleading. In order to improve our results we have to get rid of misleading, obvious or redundant association rules.

To evaluate our predictions of future events, the set of access logs will be split into two parts. The first larger part represents training dataset of context-enriched access logs. Based on knowledge gathered from this dataset we generate predictions of the future events. The prediction consists of the combination of context data and access logs, as we are not interested in uninteresting, misleading or obvious statements (e.g. “It is Saturday.” ⇒ “It is raining.”). A set of logs from the second part, testing dataset, is used for evaluation of accuracy of our prediction.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dey, A. K.: Understanding and Using Context. *Personal and Ubiquitous Computing*, 2001, pp. 4-7.

[2] Kramár, T., Barla, M., Bieliková, M.: PeWeProxy: A Platform for Ubiquitous Personalization of the "Wild" Web. *Proceedings of the 19th International Conference on User Modelling, Adaptation and Personalization*, 2011, pp. 7-9.

Context Influencing our Behavior

Dušan ZELENÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
Zelenik@fiit.stuba.sk

Web and mobile devices became very important part of our lives. These are the main sources of information, self presentation, communication or entertainment. To make this everything available almost for free, we silently agreed to share our personal information which is then somehow used by companies to adapt, personalize or recommend the content for us. This respects the model of user and her interests. But we can move to the more advanced level and claim that this interest is influenced by current state of the environment or user herself. This information on the conditions is known as contextual [1].

Contextual information could be applied in many different domains to improve the quality of the model based on user model. For instance, we present AdaptiveReminder [3] as a tool which is able to dynamically adapt the plan for a day according to current conditions. AdaptiveReminder uses history of user movements and passed events to recognize the influence of specific context information on the time needed to transport (e.g., it wakes user up earlier due to traffic jams caused by fog).

We designed AdaptiveReminder as the application which tracks user location. We use SSIDs of WiFi access points which are commonly scanned through inbuilt WiFi module of every smartphone. These SSIDs are used as identifiers of a location. We do not need exact GPS coordinates as we only need to differentiate distinct locations. Sets of SSIDs are assigned to a specific location thus we are able to identify this location when user returns. Using context of location obtained in such a way is then used to track movement and time needed to transfer from one location to another. We use learnt transport duration to adaptively set reminders for upcoming events. To remind adaptively using weather context, we only enrich learnt transport durations by weather conditions. We are then able to remind according to weather, location and time. Since our application is integrated with user calendar, it is informed about events. Last thing needed to adaptively remind events is to determine where an event happens. When an event happens for the first time, we are not able to determine its location but at the time of event as it is under way we assign current location for future reference.

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 53–54.

We also analyze the impact of contextual information on trends in news reading to boost some news in order to support the interest. We have server logs of largest Slovak news portal sme.sk from which we are able to predict the interest in a particular topic. For example someone reads football news before the match to confidently place a bet. Others are reading recipes before Christmas or Thanksgiving. These behavioral patterns are known as human rituals.

Currently we work with context of time and its derivations (minute in hour, hour in day, day in week, day in moth, day in year). We also incorporated approximation of user's location (using IP address). IP addresses and timestamps are also used to determine whether user is at home, at work or somewhere else.

Another context we use is the actual content of news articles. We interpret content using combination of section and category used by authors of articles. Each visit made by readers is enriched by this contextual information. Our research is dedicated to infer contextual information which is not available directly. It means that we can infer dwelling time in cases where direct acquisition is not successful.

Another domain where our work could be applied is code review support. This helps software developers to identify bugs in the code. We monitor software developers and their contextual information while they program. We learn which context influences the quality and occurrences of bug reports. These rules are then applied to discover potential problems automatically and to mark the code as potentially wrong.

Ultimate goal of our research is to infer context in a generic way. Context inference is based on user behavior. We discovered that not only user has behavioral patterns [2] but there is a correlation among behavior of more users. We have to find behaviorally similar users what enables us to infer missing context for other users. For instance, one user does express emotions explicitly while listening to music. Another one is behaviorally very similar but he does not express emotions. We are able to group them according to their contextual history and infer missing contexts. Knowing user's emotions would help to prepare better recommendations or predict his further behavior.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Carlson, D., Schrader, A.: A wide-area context-awareness approach for Android. In *Proc. of the 13th International Conf. on Information Integration and Web-based Applications and Services*, ACM, 2011, pp. 383-386.

[2] Kramár, T., Barla, M., Bieliková, M.: PeWeProxy: A Platform for Ubiquitous Personalization of the "Wild" Web. In: *UMAP 2011. Adjunct Proceedings of the 19th Int. Conference on User Modeling, Adaptation and Personalization : Poster and Demo Track*, 2011, Girona, Spain. 2011, pp. 7-9.

[3] Zeleník D.: An Approach to Context Aware Event Reminding. In *Proc. of the Information Sciences and Technologies Bulletin of the ACM Slovakia*, ACM, 2011, pp. 126-130.

Domain Modeling, Semantics Discovery and Annotations

Validation of Music Metadata via Game with a Purpose

Peter DULAČKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
dulacka@gmail.com

Quantity of music metadata on the Web is sufficient, music recommendation and online repository systems are the proof. However it became a real challenge to keep quality of metadata at reasonable level. Automatic approaches are fast but inaccurate; the cost of human computation is too high. Games with a purpose (GWAPs) [3] try to solve this problem. Many such games have been presented, e.g., Little Search Game – a single-player game for creating folksonomy-like network relationships [1]. We took the best from all games invented yet and present a Game with a purpose called City Lights – a music metadata validation game which lowers the cost of human computation and makes validation fun.

Our game focuses on validation of existing metadata fetched from the Web. Metadata validation is as much important challenge [2] as metadata acquisition. In our game a player is given several sets of fetched annotations; each set relates to a different song. A player then hears a part of a song and has to decide, which of given sets relates to a song she is listening to. By player's decision we are able to determine the level of correctness of a song-annotation relation – *crowds validating crowds*. The player can pause or rewind a song and she is not limited by time. The interface of the game is shown at Figure 1. The level design goes as follows:

1. A player is given a game board with highlighted initial node and possible directions. Each node of the game board is related to exactly one song and its set of annotations.
2. A music player starts playback and the player is forced to explore annotations related with possible direction nodes. Based only upon these annotations, she has to decide which of the available nodes contains annotations related to the song.
3. When an attempt is made, the player chooses certainty of hers decision (effectively a bet height).

* Supervisor: Jakub Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 57–58.

4. After making a correct decision, she eventually marks incorrect tags and proceeds to the next song. Music is still being played in the background to make the decision easier.
5. Game ends when user reaches final node via pre-generated path.

Processing of annotations is based upon the following rules. Every annotation used in the game has initial value of 0. This value changes through the player's actions and every action (even no action) changes this value. When the value gets above 4 (many factors are considered such as number of options player had to choose from, level of player's confidence, etc.), we consider the annotation validated. Such annotations can then be used as last-stand hints for players requesting more annotations. If the value drops below -3 (at least 6 people marked the annotation), annotations is being considered wrong and is not used in game anymore.

We managed to evaluate an early version of the game. Our dataset contained 150 popular songs and 100 annotations for each song, which needed to be validated or removed. Annotations were fetched from public LastFM[1] database and song previews are being played from 7Digital[2] library. Because of the small amount of rounds we lowered validation levels of annotation value to 1.5 for a validated annotation and -1 for validation to be removed. After only 20 rounds played we got rid of annotations such as: *elotmbgmegamixx, nice, favorite, good lyrics, etc.* On the other hand, we validated tags as: *female vocalists, british, singer-songwriter, pop rock, etc*.

We believe that with bigger player base we can be even more successful in validating any set of music annotations. However, we have to be aware of choosing correct validation limits according to the actual player base and difficulty of game so we will not get into the state where just a couple of annotations is being validated or (worse case) annotations are being validated too fast.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Šimko, J., Tvarožek, M., Bieliková, M.: Little search game: term network acquisition via a human computation game. In *Proc. of the 22nd ACM conf. on Hypertext and hypermedia*. 2011, pp. 57–61.

[2] Turnbull, D.: Automatic Music Annotation. *Department of Computer Science and Engineering, University of California, San Diego, CA*. Research Exam. 2005.

[3] von Ahn, L., Dabbish, L.: Designing games with a purpose. In *Communications of the ACM*. Aug. 2008, vol. 51, no. 8, p. 57.

[1] http://www.lastfm.com
[2] http://www.7digital.com

Discovering Relationships between Entities in Web-based Digital Libraries

Michal HOLUB*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
holub@fiit.stuba.sk

A lot of information can be found on the Web; however, it is mostly intended for humans. If the computers could better understand the data it would enable us to make more intelligent web applications, which could extract new facts, better adapt to users' needs and make searching easier and faster. This requires representing the data together with semantics and relationships, which is the focus of our research.

In particular, we focus on the domain of web-based digital libraries, which contain much information useful for other researchers. The main entities we deal with are authors, papers, conferences and publications. The problem is that as data is scattered across many web portals we need to integrate it. This requires finding relationships between entities from different sources.

Apart from obvious relationships, e.g., person writes a paper, there are few explicitly expressed relationships between these entities. Relationships bring another dimension to data which we can use for filtering, finding similarities or differences, etc. Search engines working with the Web of Data with semantics can provide more precise results for the queries, especially when asking questions about entities.

Relationships between entities and objects are also essential for their integration and creating mashups of things. We can use it with exploratory search when we create an overview of the target domain from web objects. There are various types of relationships which we can find between entities. The most interesting relationships are created based on the interaction of users with web objects. These relationships may not have a meaningful name but can express relatedness among web objects.

In our research we focus on building a platform enabling users doing research more efficiently. The platform is based on semantic data and collaboration by its users.

The base is a domain model of digital libraries represented as RDF triples. This allows us to record relationships and their types. We crawl various web portals (ACM Digital Library, Springer, DBLP, CiteULike, etc.) and parse information about entities,

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 59–60.

which we subsequently integrate into one dataset. Apart from metadata about scientific articles this dataset also contains user-generated content like tags.

Next, we transform the data according to the Linked Data principles and we discover relationships between various entities. We use a reasoner to derive new facts.

Contribution of our approach is in deriving relationships between entities also according to the behavior of users when doing research (e.g., when searching for a paper on selected topic). We transform the actions he makes into implicit feedback and connect the entities accordingly. These relationships can vary per user so we get more configurations of the domain model which can be used for different groups of users.

The process of relationship discovery introduces new research challenges such as verification and validation of relationships, their weighting and ranking. Newly discovered relationships enrich the domain model, which later leads to improvement of search, personalization and recommendation processes.

We have done an experiment involving automatic generation of facets for filtering the entities in digital libraries represented in our domain model. Nowadays, faceted interfaces are being successfully used to browse textual data [1]. This is useful when we have a long list of search results which we want to narrow down. In this experiment we have not yet included the relationships.

We applied the information retrieval methods on the attributes of the entities (such as title of an article or name of the conference). After tokenization, stemming and stop-words removal we counted document frequencies for each term in the collection (here, a document represent an entity). As facets we used terms occurring in an interval from 10 % to 50 % of the entities.

The evaluation showed that this method is appropriate for attributes which do not have a wide range of values, e.g., names of events (workshops and conferences) where the results are satisfactory. However, it is not applicable for attributes with values from a very wide range, e.g., article titles. In this case, the precision of our method was only 25 % (we had a domain expert requested to choose the meaningful facets which were subsequently compared to the results obtained using our method).

In the future work we plan to incorporate the relationships into the process of facet generation, which was partially done in [2]. Then, we will observe the behavior of users using the faceted search and derive new relationships which will improve the domain model.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dakka, W., Ipeirotis, P.G., Wood, K.R.: Automatic Construction of Multifaceted Browsing Interfaces. In: *Proc. of the 14th ACM Int. Conf. on Inf. and Knowledge Management (CIKM '05)*, ACM Press New York, NY, USA, 2005, pp. 768–775.

[2] Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. In *The Semantic Web (ISWC 2006)*, LNCS 4273, Springer, 2006, pp. 272–285.

Augmenting the Web for Facilitating Learning

Róbert HORVÁTH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
roberthorvath89@gmail.com

Every day, users on the Web go through large amount of articles and texts while fulfilling various needs. It takes a lot of time and this time can be spent more effectively. Information technologies and text augmentation methods are able to provide user with additional information during web browsing, which is helpful in learning process, e.g., for learning new languages. Web content is typically written in natural language. This is a reason why it is not understandable for computers and, as a result, its augmentation is a complicated task. Finding methods which allow augmentation of selected parts of web documents is a research challenge in field of Technology Enhanced Learning (TEL).

The goal of our work is to devise a method for web augmentation during casual web browsing, which facilitates learning – foreign language learning in particular. The potential of this approach is supported by agreement of experts that vocabulary acquisition occurs incidentally and minimal mental processing (of presented vocabulary) can have memory effects [2]. Our method needs to represent user knowledge and its specifics (for example, the issue of forgetting) in open information spaces. It is important to take into consideration the amount of knowledge user already has and his goals, i.e., what he would like to learn.

When looking for a solution, we face the following five open problems:

- natural language processing,
- user knowledge modeling,
- augmentation of a proper part of a web page,
- identification of a proper approach and a moment for augmentation,
- problems specific for learning (remembering, forgetting…).

Existing approaches to technology enhanced learning during web browsing aim mostly at vocabulary acquisition. They can be divided into two main categories [3, 4]:

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 61–62.

- substitution of words in documents written in native language (L1) with words from target language (L2),
- augmentation of text written in target language (L2) with word translation possibility.

In our approach we plan to augment web pages written in language (L1) and offer user both word substitution and word explanation if requested. To increase chance of remembering presented words, we think of special augmentation using different colors which will be associated to special categories or sentence members (see Figure 1).

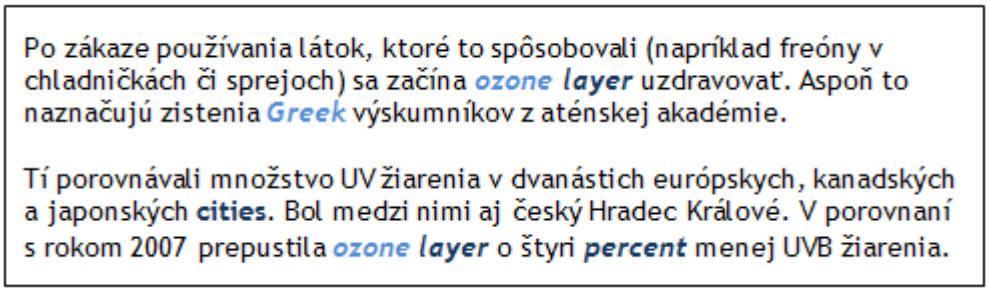


Figure 1. An example of web augmentation.

In order to evaluate our approach, we consider implementation of our method as a plugin into Adaptive proxy project [1] or a web browser extension.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Kramár, T., Barla, M., Bieliková, M.: PeWeProxy: A Platform for Ubiquitous Personalization of the "Wild" Web. In *UMAP 2011: Adjunct Proc. of the 19th Int. Conf. on User Modeling, Adaptation and Personalization*. Demo. 2011, pp. 7–9.

[2] Quirchmayr, G., Wahl, H., Winiwarter, W.: Natural language processing technologies for developing a language learning environment. In *Proc. of the 12th Int. Conf. on Inf. Integration and Web-based Applications & Services*, ACM New York, USA, 2010, pp. 381–388.

[3] Trusty, A., Truong, K. N.: Augmenting the Web for Second Language Vocabulary Learning. In *Proc. of the 2011 annual conference on Human factors in computing systems*, ACM New Yok, USA, 2011, pp. 3179–3188.

[4] Tanimura, M., Utiyama, M.: Reading Materials for Learning, TOEIC Vocabulary Based on Corpus Data. *JACET Bulletin*, vol. 42, 2006, pp. 81–96.

Discovering Keyword Relations

Peter KAJAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
peto.kajan@gmail.com

When working with keywords, there are some issues that one has to deal with. For instance, in the search process problems can be caused by: homonyms (wrong entities are found), synonyms (entities are not found) and the abstraction level (queries are too specific/general). If these relations are identified, search queries may be adapted to achieve more relevant results.

Analyses of folksonomies created in the tagging process are commonly used in the approaches revealing keyword relations [2]. But folksonomy-like data can also be obtained from analysis of the Web usage and therefore larger datasets are produced. Users visit pages, which can be described by keywords that are automatically extracted from pages content. These data (further called logs) cover more domains and are "cheaper" to acquire since document browsing is a more frequent action than tagging.

Another advantage of logs in comparison to folksonomies is a relevant timestamp attribute specifying the order of visits of documents. We propose a method based on this timestamp attribute and formulate following hypotheses for revealing relations:

- *Similarity hypothesis:* If a significant number of users visit a page with keyword A and right after they visit a page with keyword B it means that these keywords may describe similar concept.
- *Hierarchy hypothesis:* Child keyword occurs more likely after parent keyword in browsing sessions.

In terms of [1], we define browsing session as a *set of documents visited by the user with particular information needs*. Our method consists of two steps (see Fig. 1). Logs are analysed to reveal similarity and parent-child relations in the first step. First, browsing sessions have to be identified in this step. Document similarities are calculated from the browsing sessions and then, they are used for keyword similarities calculation. Last, parent-child relations are identified according to the second hypothesis. Keywords are mapped to Linked data in the second step. There is a high probability that a relation between keywords can be named if the similarity of two keywords is high. We decided to use the Linked Data for naming these relations.

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 63–64.

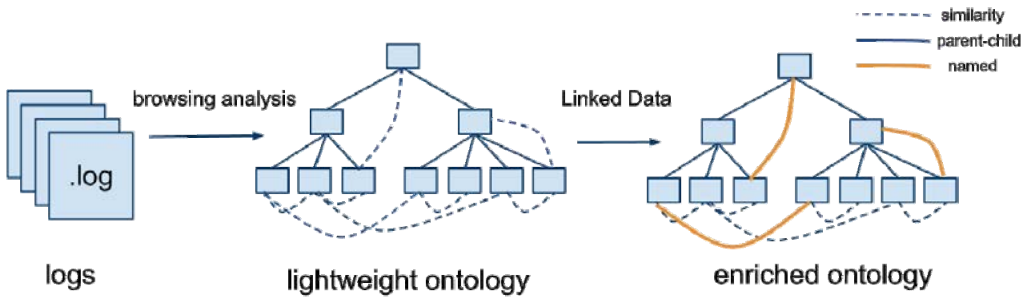


Figure 1 Method of discovering keyword relations

We evaluate the proposed method on logs of Web browsing activity from PeWeProxy[1]. The computations have been distributed using Hadoop[2] to achieve an acceptable performance. The actual prototype is able to identify keyword similarities. For illustration, chosen clusters of similar keywords are displayed in Table 1.

Table 1. Similar keywords grouped into clusters.

Cluster	First 10 keywords
1	Electrical, faculty, Electrical Engineering, Informatics, Industrial, Control, industrial informatics, Robotics, cybernetics, Technical
2	LG LCD Monitor, Sapphire Radeon HD, intel core 2 quad, Kingston HyperX XMP, Intel P45 Memory, box 2.66 ghz, ii 500w hdd, brand new price, CPU Cooler
3	Adriana Barraza, Idris Elba, Jaimie Alexander, Chris Hemsworth, Kat Dennings, Natalie Portman, Rene Russo, Anthony Hopkins, Ray Stevenson, Clark Gregg

The goal of this work is to explore the importance of information ordering for revealing relations. The idea of finding relations from document browsing seems to have a potential. In future work we plan to evaluate the results using qualitative experiment in which participants will rate discovered relations.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Gayo-Avello, D.: A survey on session detection methods in query logs and a proposal for future evaluation, In *Information Sciences, Vol. 179, No. 12*, 2009, pp. 1822-1843.

[2] Mika, P.: Ontologies are us: A unified model of social networks and semantics. *International semantic web conference*, 2005, pp. 522–536.

[1] http://peweproxy.fiit.stuba.sk/proxy/

[2] http://hadoop.apache.org/

Named Entity Recognition for Slovak Language

Ondrej KAŠŠÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
ondrej.kassak@gmail.com

Nowadays we are literally overwhelmed with information. It is impossible for us to process all the information we find. Various approaches have been proposed to address the information overload problem, such as personalized recommendations based on the content or searching methods using key entities from the texts. They assume named entities appearing in the text for their proper working as an input. Based on them, recommendation algorithms can search and work more efficiently in comparison with other methods working only with text titles or with the most frequent words in the text.

In our research we propose the method for recognizing and extraction of named entities in texts. The aim of our proposed method is to recognize entities in the text and then place them into the proper categories. We primarily focus on texts in Slovak language because a comprehensive tool for this language that would identify all entities classified according to the MUC-6 (6th Message understand conference) [1] is missing. We also describe possibilities of application for other flectional languages.

The proposed method consists of two parts – the initial part of pre-processing of the text and the recognition of the named entities.

For text pre-processing our method uses a form of stemming. We remove the word suffix caused by inflection. Suffices are identified by comparing each word with the set of Slovak word endings. The result is a form of words which is not the word formation root but, with only a few exceptions, we get the uniform forms of words, which can be used in further computation [2].

The process of named entities recognition consists of identifying potential entities occurring in the processed text, determining its scope and consequently identifying the category to which they belong. We use Slovak[1] and English version of Wikipedia to identify new entities, database for fast recognition of entity that we found before and Slovak National Corpus[2] for filtering common words with first capital letter from

* Supervisor: Michal Kompan, Institute of Informatics and Software Engineering

[1] http://sk.wikipedia.org/
[2] http://korpus.juls.savba.sk/

Spring 2012 PeWe Workshop, April 1, 2012, pp. 65–66.

entities. We propose the named entity extraction (Figure 1), which consists of several steps.

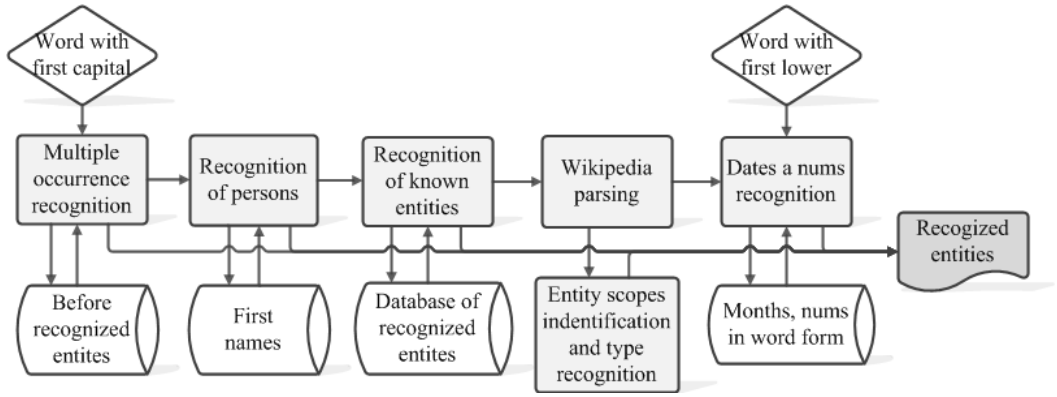


Figure 1. Sequence of steps describing the proposed method.

Entities usually start with a capital letter in Slovak. We have just to search for the capital letters in the text. If we then filter out the beginnings of sentences or quotations that are not entities at the same time, we get a set of entity beginnings. Thus we are able to identify persons, organizations, locations and miscellaneous entities.

After we find the beginning of the entity we compare it with a list of before recognized entities, so we can simply identify entities that have already been recognized before, omitting the long process of standard recognition of a new entity.

If we do not find an agreement, we compare it with the set of first names and possibly the database of recognized entities. If we still do not identify the entity we recognize its scope through web parsing. If we are able to find the scope we try to recognize entity type.

In addition to the mentioned entity types, our method identifies also numeric entities and dates. The numeric ones distinguish between money amount, percentage and a number itself. In identifying the type of numerical entities the context words are the most significant.

To evaluate proposed approach we processed 60 articles from 3 various Slovak news servers. Texts were manually annotated by a human expert. We obtained result of 79 % F-measure (84 % precision, 74 % recall). We correctly recognized 1204 entities of 1620 in total. 232 entities were identified incorrectly.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: *Proceedings of COLING '96*, 1996, pp. 466–471.

[2] Przepiórkowski, A.: Slavonic Information Extraction and Partial Parsing. In: *Computational Linguistics*, June 2007, pp. 1–10.

Building Domain Model via Game with a Purpose

Marek KIŠŠ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
keramssik@gmail.com

The Web brings many great uses for personalization. Prerequisite for a working adaptive system is the existence of the domain model to provide basis for modelling both user knowledge and semantics of domain documents. Despite the use of conventional, heavy-weight ontologies, a simple domain model represented as a concept relationship network, hence easier to create, were also proved to be efficient for this task in certain systems [1]. However, even the creation of such a simple model cannot be fully automated and it is usually a work for authors of the content.

We created a game with a purpose [4] for building domain models. Our game is based on similar principles as Little Search Game [3]. It is an online single player game working with search queries and creating a lightweight term network. Its player interface is depicted in the Figure 1. At the beginning, the game displays to a player a single term. Then he tries to find a term related to the displayed term. The number of points he obtains for his choice is based on a number of occurrences of this pair in a corpus of domain documents. Our game differs from Little Search Game in a way how the player interacts with a game. Instead of thinking out the best term and writing it to the input form, the player has to choose the best one from terms offered by the game. They appear in colourful bubbles, blowing up upon selecting, yielding the number of points received by the player by that action. This makes our game more dynamic. We believe that such game can be more attractive for players.

We test our game in the domain of Principles of Software Engineering course and as input we use its documents from the learning system ALEF comprising tens of textual learning objects (few pages each). As terms we use manually created concepts for this domain. We narrow down the number of terms, from which we randomly select terms for the game. We worked just with a subset of all terms. If many players have not chosen a certain word in the same round, the game realizes that there is probably not any kind of a strong "hidden relationship" [2] between this term and

* Supervisor: Jakub Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 67–68.

round specific term, removes it from this subset and replaces it with another one. This narrowing helps us to obtain results fast even with a small number of players.

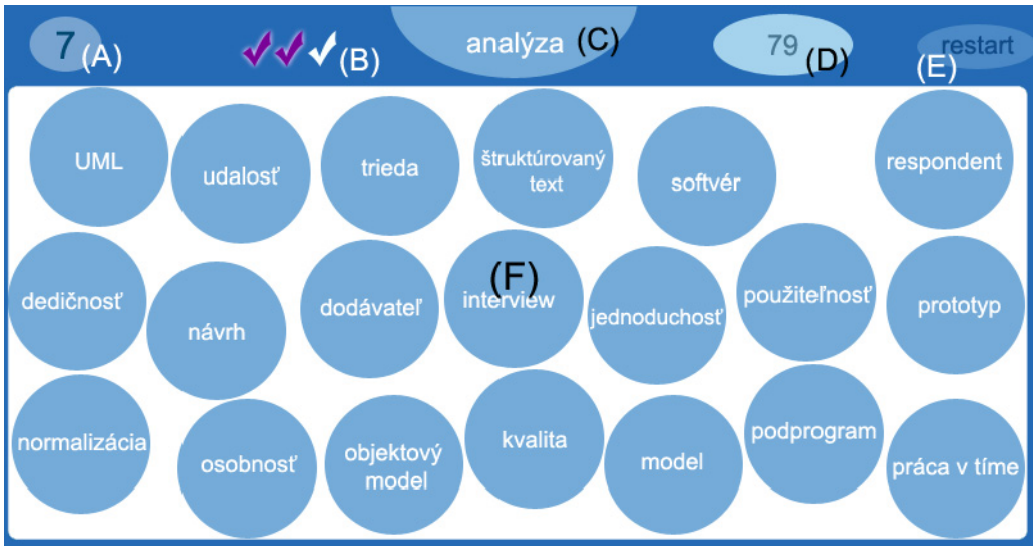


Figure 1. GUI of our game: timer (A), shot counter (B), game specific term (C), current score (D), restart button (E) and cloud of bubbles (F).

Outputs of our game are relationships between terms. It uses the "wisdom of crowds" paradigm: "If many say that A is an instance of B, A is likely an instance of B" [2]. So our game creates a relationship between a pair of terms when couple of players have connected them in a game.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Bieliková, M., Šimko, M., Barla, M.: Personalized web-based learning 2.0. In: *Proc. of the 8th Int. Conference on Emerging eLearning Technologies and Applications*, 2010.

[2] Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. In: *Proc. of IEEE Intelligent Systems*, 2008, pp. 50–60.

[3] Šimko, J., Tvarožek, M., Bieliková, M.: Little search game: term network acquisition via a human computation game. In: *Proc. of the 22nd ACM conference on Hypertext and hypermedia*, 2011, pp. 57–62.

[4] von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: *Proc. of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 319–326.

Automated Public Data Refining

Martin LIPTÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
mliptak@gmail.com

Public institutions have legal obligations to share certain data on the Web. While public registers (e.g., businesses, organizations) and bulletins (public procurements) are essential for business communication, other data increase transparency of public institutions and enable public investigation (public contracts). Despite the fact that these data are becoming publicly available on the Web, there are several problems.

The first problem is format and structure that might not be suitable for machine processing. For example some documents are published as scanned images with censored names and prices. This makes such documents difficult to investigate by a human expert and almost impossible to process with computer. For example company liquidations are published in periodic PDF bulletins as an unstructured textual content and it is difficult to reliably find out if a company is being liquidated or the liquidation is being cancelled. Fortunately the most common format is HTML, which is easy to parse and in most cases provides a structure. However, even in correctly parsed and structured data, there are various mistypings, disambiguities and duplicates. These inconsistencies are the second problem we address in this paper.

Mistypings, duplicates and disambiguities are a major problem not only in the public data domain. In fact every database possibly merged of multiple sources needs to be cleaned so that queries provide reliable results. This process is widely known as data integration, duplicate detection or record linkage [1]. M. Bilenko and R. Mooney in [2] proposed to employ learnable string distance functions for duplicate detection task. M. Hernández and S. Stolfo in [3] have developed a method for removing duplicates from databases of 100 milion to 1 bilion records in a matter of days.

We propose a duplicate detection method based on machine learning algorithms. We use a logistic regression classifier to predict whether samples are duplicates or not. The classifier trains weights of features, provided by user, for a particular datasource (e.g., Levenshtein distance of compared fields or presence of particular combination of substrings in compared fields). The user also provides a labelled set of samples that is used to train the classifier. Trained classifier can detect duplicates by predictions based on learned feature weights.

* Supervisor: Ján Suchal, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 69–70.

We have evaluated our method on a real-world database of people occurring in Business Register of the Slovak Republic provided by foaf.sk. There are many duplicates and it is difficult to determine, who exactly occurs in which company. There is a set of heuristics already detecting duplicates on foaf.sk. We have used their results for training and as a baseline for measuring precision, recall and F_1 score of our approach. We are comparing names and addresses of people. Our features are string equality (=), Levenshtein distance (L), N-Grams (NG), degree combinations and degree disjunctions. Results are shown in Table 1.

Table 1: Results

Feature set	Precision	Recall	F_1
=(names), =(addresses)	0.8777	0.9874	0.9293
L(names), L(addresses)	0.8782	0.9923	0.9318
2G(names), 2G(addresses)	0.8777	0.9874	0.9293
3G(names), 3G(addresses)	0.8777	0.9874	0.9293
4G(names), 4G(addresses)	0.8777	0.9874	0.9293
5G(names), 5G(addresses)	0.8777	0.9874	0.9293
6G(names), 6G(addresses)	0.8777	0.9874	0.9293
L(names), L(addresses), Degree combinations	0.8803	0.9622	0.9194
L(names), L(addresses), Degree disjunctions	0.882	0.9777	0.9274

From the results we can see that machine learning approach for duplicate detection yields reasonably high precision-recall values. However, data samples are simple and further research with more complicated samples needs to be done. Our results have clearly shown that we need a new data set with our own labels (created manually) instead of baseline foaf.sk labels. Besides name and address attributes, it would be reasonable to include relations of people to companies for better results.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Fellegi, I.P., Sunger, A.B.: A Theory for Record Linkage, In *Journal of the American Statistical Society*, Vol. 64, No. 328, 1969, pp. 1183-1210,

[2] Bilenko, M., Mooney, R.: Employing trainable string similarity metrics for information integration, In *IJCAI 2003 Workshop on Information Integration on the Web*, 2003, pp. 67-72.

[3] Hernández, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem, In *Data Mining and Knowledge Discovery,* Vol. 2, No. 1, 1998, pp. 9-37.

Acquiring Web Site Metadata by Heterogeneous Information Sources Processing

Milan LUČANSKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
lucansky.milan@gmail.com

We live in a world where the amount of freely accessible data increases faster than ever before. The World Wide Web is almost unlimited source of knowledge and information and every year the number of available web sites increases in millions. That introduces the demand for automatic processing of vast collection of web documents. We need to assign descriptive metadata to web pages to facilitate further processing and it turns out that keywords are suitable representation of web content. Nowadays, keywords form a basis for semantic representations as they are utilized in the field of ontology engineering [2]. The most of popular search engines are based on keyword search paradigm and keywords are even used in user modeling for adaptive web-based systems to represent the context [1]. Social services, such as Delicious[1] utilize keywords too.

We need an automatic approach to keywords acquisition. There are various approaches to automatic term recognition (ATR) in offline document collections. ATR algorithms use statistical and probabilistic features to get relevant keywords and are widely utilized for plain text document (with no internal structure) processing. If used on web documents, they could possibly benefit from hidden semantic of HTML elements used to format and style sheets to visualise text content. Our current research aims at cascade style sheets (CSS) as additional source for identifying potential keywords. The idea of utilization of CSS in co-operation with ATR algorithms is quite new and unexplored. Therefore we see a possibility to combine semantic potential of HTML tags and CSS with ATR algorithms in order to yield better results rather than using them separately.

We introduce a *TagRel*, *LinkRel* and *CssRel* coefficients that modify weight of a term obtained by an ATR algorithm. Plain text content is passed to an ATR algorithm, which extracts weighted keywords. We extract keywords from the web page formatted by selected CSS attributes, compute the *CssRel* coefficient and improve

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

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

ATR keywords weights. For the anchor texts pointing to examined web page we compute *LinkRel* and improve ATR keywords. Finally, we acquire text content from selected HTML elements, compute *TagRel* and improve ATR keywords. The final weight of extracted keywords we compute as follows:

$$w_t = w_t' + w_t'' \tag{1}$$

where w_t is final weight of a term *t*, w_t' is weight of a term *t* obtained by a ATR algorithm and w_t'' represents weight of term *t* as combination of *TagRel*, *LinkRel* and *CssRel* coefficients.

While we use three different coefficients for assigning a weight to potential keywords we need a scheme for combining them to the single measure (w_t''). A possible option is to use a weighting scheme. We assign to each type of *Rel* coefficient a multiplication number denoting probability of containing relevant keywords. The estimation of multiplication numbers that denote the probability of containing relevant keywords is a part of our research.

$$w_t'' = \alpha \cdot Link\,\mathrm{Re}\,l + \beta \cdot Tag\,\mathrm{Re}\,l + \gamma \cdot Css\,\mathrm{Re}\,l \tag{2}$$

where w_t'' is combination of *TagRel*, *LinkRel* and *CssRel* coefficients for term *t*, α is the multiplication number for *LinkRel*, β is a multiplication number for *TagRel* and γ is a multiplication number for *CssRel*.

Using equation (1) we produce a new order of extracted keywords, where the most relevant have the highest weights. In order to evaluate the proposed approach, we are currently conducting an extensive experiment on a set of randomly chosen pages from the Web. So far we performed synthetic experiment on a small set of randomly chosen web pages, in order to process visual information represented by style sheets formatting. The web pages use different style sheets formatting. We extracted all emphasized words and short terms from main textual content and tried to state either the term is relevant to the contents of web page or not. In average 38 % of extracted terms were relevant to the topic of article. Actual results seem very encouraging. In a more extensive experiment we need to examine the method on different types of web pages and to compare the results with contemporary approaches (e.g., freely available web services for term extraction as tagthe.net[2], OpenCalais[3]).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Barla, M., Bieliková, M. Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In White, B., Isaías, P., Andone, D. (Eds.): *WWW/Internet 2010*, IADIS Press, 2010, pp. 227–233.

[2] Cimiano, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, 2006, pp. 23–24.

[2] http://www.tagthe.net/

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

Personalized Text Summarization

Róbert MÓRO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
xmoror@stuba.sk

Information overload is one of the most serious problems of the present-day web. Automatic text summarization aims to address this problem by extracting the most important information from a document, which can help readers (users) to decide, whether it is relevant for them and whether they should read the whole text or not.

Conventional (generic) summarization methods summarize only the basic content of a document and do not take into account differences in users' interests, goals or knowledge. Personalized summarization, on the other hand, uses this additional information about users' characteristics to produce summaries more suitable for a particular user's needs.

We propose a method for personalized summarization based on a method of *latent semantic analysis (LSA)* [1][2]. We have chosen LSA as a basis for our approach, because of its ability to provide better results compared to the other summarization methods. We have identified a construction of a terms-sentences matrix representing the document as a step suitable for summarization personalization. In this step, terms extracted from the document are assigned their respective weights. Our proposed weighting scheme extends conventional weighting scheme based on tf-idf measure by linear combination of multiple raters, which positively or negatively affect the weight of each term (see Figure 1).

We have proposed a set of raters which can be divided into two groups:

- *generic raters* take into account basic content of the document and some additional information to adapt summarization regardless of a particular user;
- *personalized raters* consider information about the specific user and her characteristics.

Our approach is language and domain independent; however, we focus on the domain of learning and the knowledge revision scenario. For this purpose we have designed the specific raters that take into account terms relevant for the domain or the level of knowledge of an individual user; we have also proposed a method for personalized selection of documents for revision. Because annotations (e.g., highlights) can indicate

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 73–74.

user's interest in the specific parts of the document [3], we use them as another source for personalization.

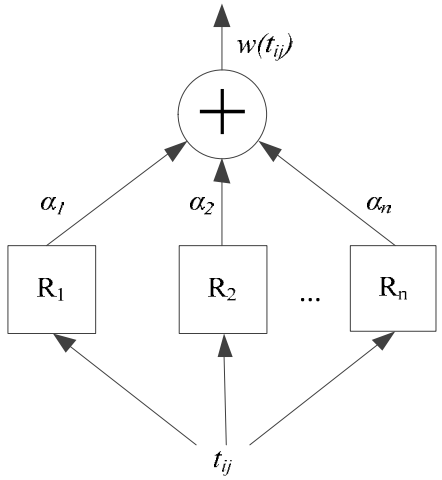


Figure 1. Term weighting as a combination of raters, where t_{ij} is a term, R_i is a rater with its linear coefficient α_i and $w(t_{ij})$ is a weight assigned to the term t_{ij}.

We have experimented with our proposed method. Our dataset has consisted of educational materials from the *Functional and Logic Programming (FLP)* course in the learning system ALEF. We have focused on evaluation and comparison of the two variants of summaries – *generic summarization* and *summarization considering the domain-relevant terms*. We have asked the FLP course students to evaluate quality of the generated summaries.

Our experimental results suggest that using the domain-relevant terms in the process of summarization leads to selecting representative sentences capable of summarizing the document for revision. In following experiments, we plan to consider users' annotations in order to further improve the summarization process.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Gong, X., Liu X.: Generic text summarization using relevance measure and latent semantic analysis. In: *SIGIR'01: Proc. of the 24th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, 2001, pp. 19–25.

[2] Steinberger, J., Ježek, K.: Text summarization and singular value decomposition. In: *ADVIS'04: Proc. of Advances in Information Systems*, LNCS 3261. Springer, Berlin, 2005, pp. 245–254.

[3] Zhang, H., Ma, Z.C., Cai, Q.: A study for documents summarization based on personal annotation. In: *Proc. of the HLT-NAACL Workshop on Text summarization*, Association for Computational Linguistics, 2003, pp. 41–48.

Metadata Collection for Effective Organization of Personal Multimedia Repositories Using Games with a Purpose

Balázs NAGY*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
chelseadrukker@gmail.com

Nowadays, an average person is overloaded with enormous amount of digital data. Besides multimedia (music, videos, images) we can mention also emails, web pages and information on social networks, blended together in a hypertext environment. For implementation of effective search and navigation in this space it is necessary to have enough *descriptive metadata* available for each resource. These can be collected automatically or manually through *crowdsourcing* [1] methods and in particular, by *games with a purpose* [2, 3].

Games With A Purpose (GWAP) refer to games that are not ordinary, but which address specific problems. The primary objective of these games is to solve problems that are unsolvable by computers. The usefulness of these games lies in the elimination of unnecessary costs for manual human labor by using voluntarily playing users.

In our research, we focus primary on image metadata acquisition. Our goal is to upgrade and extend an existing GWAP called *PexAce*, which collects useful annotations for photos and transforms them to tags. Due to lack of metadata for personal photo albums [4, 5] we want to focus on obtaining descriptive metadata for this kind of media. Using them we will be able to *query, order and filter* these enriched photo albums much better.

We want to allow users to import their photos either from online or from local storage. They will be able to create, update and remove albums, but also browse photos applying different filters with the metadata obtained via our game.

Although with our original game we have achieved remarkable results, we would like to obtain better metadata by analyzing and processing of the various logs recorded during the games. All tags will receive reliability weight on the basis of information derived from logs. We also want to assess credibility of each user examining tags obtained by logs.

* Supervisor: Jakub Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 75–76.

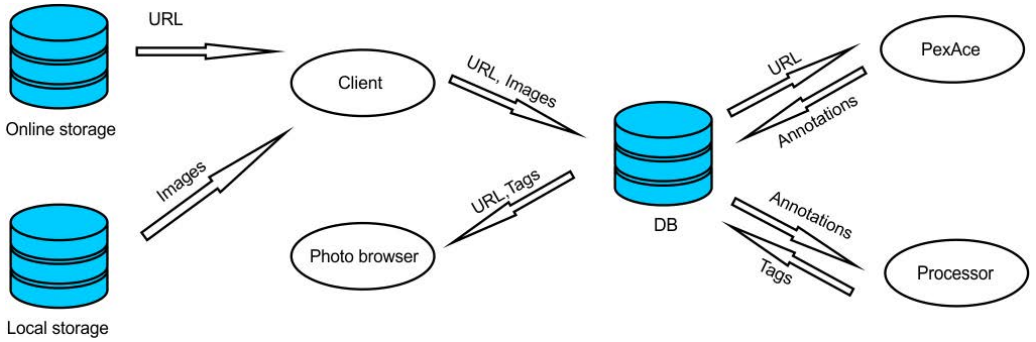


Figure 1.Components of the project with data flow between components.

Our previous experiments with the PexAce within general domain indicate that this method of obtaining metadata is effective. According to our expectations, we should get positive results also after using our method in specific area such as personal photo albums. In fact, users may be more motivated because they are annotating their own photos. Another side effect of this should be reflected also in the quality of obtained tags.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Roman, D.: Crowdsourcing and the question of expertise. *Communications of the ACM*, 52(12), 2009.
[2] von Ahn, L., Dabbish, L.: Designing Games With Purpose. *Communications of the ACM*, 2008, pp. 58–67.
[3] Šimko, J., Tvarožek, M., Bieliková, M.: Little Google Game: Creation of Term Network via Search Game. In *Proc. of Datakon 2010*, 2010 (in Slovak).
[4] Vainio, T., et al.: User needs for metadata management in mobile multimedia content services. In. *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems*, 2009, p. 51.
[5] Cunningham, S. J.: Identifying Personal Photo Digital Library Features. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 400–401.

Knowledge Tags Repository

Karol RÁSTOČNÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
rastocny@fiit.stuba.sk

Semantics in the Web is necessary for computer systems, e.g. for recommending. For this reason, they process documents on the Web and assign knowledge tags, short metadata (e.g. keywords) to documents' parts. By the means of sharing these knowledge tags a new layer of lightweight semantics over the Web can be created collaboratively. Current systems usually store these metadata inside private repositories in a form which is not understandable to others; consequently other systems cannot readily use this metadata. In this project we propose knowledge tags maintenance approach (Figure 1), which allows effective sharing of metadata.

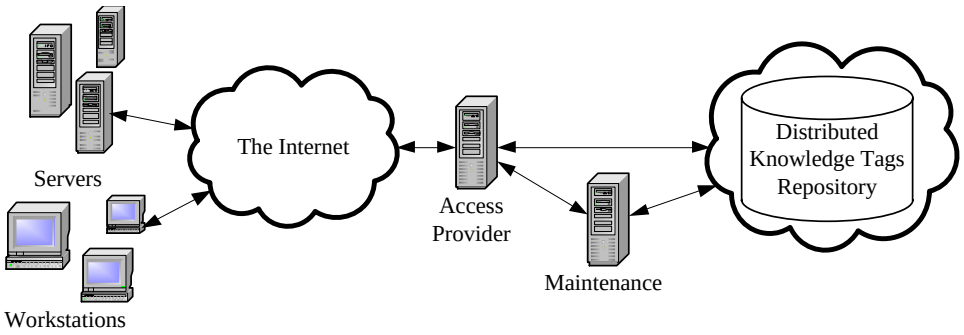


Figure 1. Architecture of proposed knowledge tags maintenance approach.

Knowledge tags repository is one of the core parts of the proposed maintenance approach. This repository stores large amount of knowledge tags in flexible open format which has to be understandable to computer systems and yet provide fast parallel access for a number of clients. The knowledge tags model is based on the Open Annotation (OA) model[1]. We decided for this model because of it is already accepted

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

[1] http://www.openannotation.org/spec/beta

by wide range of systems and knowledge tags and annotations have common characteristics. Both of them are anchored to specific parts of documents and they contain small information on these documents' parts. This model is based on RDF and it is highly recommended to implement it by RDF triple databases and support at least basic access (e.g. querying by SPARQL) with RDF serialization output format.

To provide effective and powerful data access, we analyzed standard use cases of annotation repositories and itemized a list of requirements which respect these use cases and specific requirements of OA model and maintenance part of proposed method. We have found out that manipulation with whole knowledge tags is important for almost all use cases. But this is in disagreement with RDF triple databases, which have good deduction possibilities but they have issues with efficiently obtaining the complete information about an object. In order to do so several simple queries have to be processed and each query can take several seconds in large datasets [3]. On the other hand, document databases appear suitable for our need of access to whole knowledge tags, while they store documents (objects, in general) as one item and not sparsely fragmented over several tables or collections. This allows for fast access to whole objects without the necessity of time expensive joins.

Particularly, MongoDB database system matches the requirements: it provides efficient data access (loading and updating) and supports distributed data processing via MapReduce [1]. However, it does not provide support for SPARQL query processing. This could be implemented via MapReduce, but the existing approaches are for Apache Hadoop [2], which has differences in processing of the Map and Reduce phases and works with RDF triples databases instead.

Our algorithm for distributed SPARQL query processing firstly builds optimal joining tree, to minimalize the number of necessary join operations. MapReduce uses ordered list of join attributes with their values as a key and a list of deduction objects which consists of an ordered list of joined pattern identifiers and an ordered list of attributes from patterns with their values as a result. The MapReduce phase is executed iteratively over remaining layers of the optimal joining tree. Map function of iteration emits for each result from previous iteration new result with a key from current joining attributes and same value. Reduce function creates result as Cartesian product of results with same keys. The last finalize function removes from results deduction objects, which do not have complete list of patterns mapped to join keys.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. *Commun. of the ACM*. (2008), vol. 50, no. 1, pp. 107-113.

[2] Kim, H.S., Ravindra, P., Anyanwu, K.: From SPARQL to MapReduce: The Journey Using a Nested TripleGroup Algebra. In: *Proc. of the VLDB Endowment* 4. VLDB Endowment, Inc., (2011), pp. 1426-1429.

[3] Rohloff, K. et al.: An Evaluation of Triple-Store Technologies for Large Data Stores, In: *Proc. of the 2007 OTM Confed. Int. Conf. on the Move to Meaningful Internet Systems, LNCS 4806*. Springer, Berlin, (2007), pp. 1105-1114.

Modeling a Tutor for E-Learning Support

Peter SVORADA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
psvorada@gmail.com

Peer tutoring is a process where one student supports another in the learning process consisting mostly of diversity of different tasks belonging to the certain domain (mathematics, physics, programming etc.) It has been shown [2] [4] that students involved in such a process (both tutor and tutee) show greater progress compared to studying alone.

However peer tutoring provides challenges for the pedagogue to give it a form in which it will not bear too much risk wasting teaching time. Students sometimes do not understand the problem very well and can reach impasses or get repetitively wrong results, leading to frustration and demotivation [3].

In our model we try to broaden the existing abilities of a computer tutor by shortening the feedback loop until students receive first feedback on their answer. We propose a model of a tutor capable of telling students that they are working on "an incorrect" solution (solution that does not resemble any known correct solution) while they are still working on it (even before the answer is submitted). In doing so, we created a method that can compare two pieces of source code, one of which is being written from scratch.

There are several different methods used to determine similarity between two pieces of source code. In our approach we adapted the method of Baxter et al. [1] that is used in comparing abstract syntax trees. As basis of our method we employed the formula used in [1] to calculate similarity between two parts of abstract syntax trees. We proposed a novel formula and process that account to the nature of writing source code from scratch. The reason for this is mainly based on fact that the original method and our method are both used in different situations. While the original method is used to compare two finished pieces of source code, our method is used to compare two pieces of code one of which is still incomplete due to the very essence of writing a source code from the scratch.

The exact process that we use to determine whether the currently written source code is similar to some of the known correct solutions can be divided into the following phases:

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp.79–80.

1. Code pre-processing
2. Creation of abstract syntax tree
3. Determining the similar nodes in abstract syntax trees
4. Calculating of similarity
5. Method application

Code pre-processing must be done in order to prepare code to be parsed into abstract syntax tree. The nodes we create are of different types and we use these types later to compare two nodes and their sub-trees to count similar nodes in each sub-tree. Based on actual number of similar and dissimilar nodes we calculate the actual similarity. For this calculation we used our novel formula.

Our method does not in fact tell students that they work on a wrong solution. It can only determine that the solution they work on is similar to some known correct solution. This, however, can provide increased guidance toward a good/better solution. That is why our tutor does not stop students when the level of similarity is low, it just recommends care to be taken simply because the current answer does not resemble any previous solution known thus far. Novel solutions that students submit improve the dataset the tutor is using to detect similar solutions.

This is not the only way how a method like this can be used by a computer tutor. Additionally, the tutor is able to give students automated advices depending on the current context. Tutor with this functionality can guide students through the tasks using tutorial steps and hints which are provided to the student depending on how he advances in his solution.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Baxter I., Yahin A., Moura L., Sant Anna M. *Clone Detection Using Abstract Syntax Trees.* In Proceedings of the 14th International Conference on Software Maintenance (ICSM'98), pp. 368-377, Bethesda, Maryland, November 1998.Vardhan, A.: Distributed Garbage Collection: A Transformation and its Applications to Java Programming. Master's thesis, (1998).

[2] Fantuzzo, J. W., Riggio, R. E., Connelly, S., & Dimeff, L. A. (1989). *Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: A component analysis.* Journal of Educational Psychology, 81(2), 173-177.Friedman, A.D., Menon, P.R.: Theory and Design of Switching Circuits. Computer Science Press, Inc., (1975).

[3] Medway, F. & Baron, R. (1977). *Locus of control and tutors' instructional style. Contemporary Educational Psychology*, 2, pp.298-310.

[4] Roscoe, R. D. & Chi, M. (2007*). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions, Review of Educational Research.* 77(4), pp.534-574.

Games and Crowds: Authority Identification

Jakub ŠIMKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
jsimko@fiit.stuba.sk

Some of the today's computational tasks are still subject to human labor, because computational machinery paradigms are unable to deal with them (esp. in terms of quality). These tasks include metadata acquisition and domain modeling, the two essential processes needed for enabling effective and adaptive hypermedia systems. Hence, a whole field of crowdsourcing-based (and game-based in particular [2]) approaches has emerged to do the job.

However, this field has also its own problems. It is dependent on mass participations of users and meanwhile, it is usually not very effective in using this power as it is based on redundant task solving in order to filter out incorrect task solutions. We believe that the effectiveness of crowdsourcing approaches can be improved through authority identification, i.e., identification of trustworthy contributors with more experience in particular domain, whose problem solutions should be taken with greater weights, assuming their high correctness probability [1]. Authority identification has not been sufficiently addressed yet, and within the domain of games with a purpose (GWAP), it completely absents. We aim to explore the possibilities of authority identification within crowd-based, collaborative and gaming metadata acquisition systems.

In the GWAP domain, our preliminary experiments have demonstrated the increased potential of GWAP solutions that include player domain expertise considerations. During the experiments with the PexAce game – a GWAP for image annotation – we have let the players to work over their own, personal images, rather than general ones. Besides the players were more attracted to the game, they have also been more productive in their annotation efforts. They were able to provide more valuable specific metadata, such as concrete person or event names, not just general descriptions of objects. All this was enabled solely by "smart" game content assignment.

The above approach is, however, bit supervised – the player must provide his own content, or someone else must assign, what content is best for the player to play with. To explore whether there is a way to implement a less supervised approach, we

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 81–82.

experiment with tracking of the player expertise in different subdomains of the GWAP content. We plan several synthetic experiments over the game logs collected in the *Little Search Game* and *PexAce* GWAPs, in which we want to show, that by measuring the player performance (score gain), we can subsequently build up "player models" and assign later task instances to them with respect to their fields of expertise. The overall hypothesis is, that the convergence to correct problem solutions would become faster, which will spare the human cycles to other tasks.

As for the general crowdsourcing, an analogous experiment is currently underway in the e-learning domain (a software engineering course), where students identify correct and wrong question answers. As an input to the process, we use the last year dataset of question-answer combinations created by students during semestral mini-tests. During the experiment, students are presented with question-answer combinations which they use as learning exercises. Students subjectively evaluate the correctness of answers on the scale 0 to 1. Afterwards, they receive the feedback in the form of average correctness of the answer, computed from other students' evaluations provided so far.

While the primary hypothesis is that such process would bring up correct answer evaluations (which we will validate by the existing real teacher evaluations), we also aim to prove that measurement of the student skills based on his past exercises can improve the crowd-based filtering tasks, if applied during solution voting procedures. In other words, we will measure the performance of students and compute the student rating, which will later serve as a weight of their votes for next question-answer evaluations.

We believe that expertise-aware extensions for the GWAPs and crowd-based applications could significantly improve their performance. An open issue, however, is the possible general (in)applicability of these principles for certain approaches, since the recognition of user models may fragment the crowd to possibly too small groups, where individuals share the same expertise. Then, the approaches that rely heavily on some sort of online collaboration, or cross-user artifact validation (e.g., multiplayer GWAPs) would experience only minor improvements. On the other hand, approaches less dependent on direct collaboration and possessing other means of ensuring output quality (e.g., single player GWAPs) could benefit much more.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Raykar, V. C., Yu, S., Zhao L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L.: Learning From Crowds. *The Journal of Machine Learning Research, 11*, 2010, pp. 1297–1322.

[2] Šimko, J., Tvarožek, M., & Bieliková, M.: Semantics Discovery via Human Computation Games. *International Journal on Semantic Web and Information Systems (IJSWIS), 7*(3), 2011, pp. 23–45.

Acquiring Metadata about Web Content Based on Microblog Analysis

Tomáš UHERČÍK *

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
uhercik07@student.fiit.stuba.sk

The amount of information on the Web is so huge, that searching can be done only by machines. However, information presented on the Web is intended for humans and is understandable only by humans. The Semantic Web is vision, where this problem is solved by the layer of machine-processable metadata. These metadata are not available as often as we would like. The challenge is to obtain them automatically.

Socially-oriented data are those, which are created by the activity of users. There is a lot of useful metadata within that data. Web applications for social networks allow a user to share a lot of information with others. Data created by their activity constitute a very valuable source of indirectly originated metadata.

We decided to use the microblog *Twitter* as source of metadata. We selected the URL as entity about which are the metadata acquired, because it can be unambiguously identified in the tweets' text.

We proposed a method for keyword extraction utilizing *Twitter* posts. Its flow is illustrated in the Figure 1.

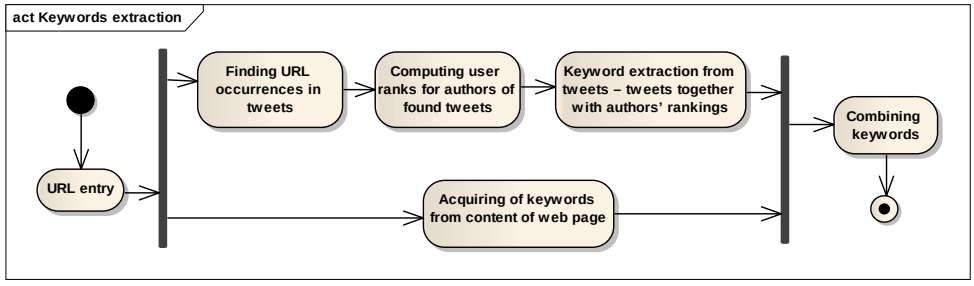


Figure 1. Activity diagram showing the process of the proposed method.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2012 PeWe Workshop, April 1, 2012, pp. 83–84.

In addition to state-of-the-art extraction methods we consider the different relevance of particular tweets depending on an author who published them. It is very effective to use this information as input of extraction method. We proposed the ranking formula, based on Tunkrank [1], considering also the frequency of user's tweets publishing as follows:

$$URank(X) = \sum_{Y=followers(X)} \frac{1+\frac{p}{log(T)}URank(Y)}{|followers(Y)|} \quad (1)$$

where *URank* is the user(author) ranking, *X,Y* are users, *followers(X)* is the set of users following *X*, *p* is convergence constant and *T* is median of time gaps between publishing individual tweets.

For extraction of relevant keywords we used TextRank algorithm [2], but we could use any extraction algorithm, which would give us keywords with their relevancies. Final *Twitter* (microblog) relevance *MRank* of keywords we obtain as follows:

$$MRank(t) = \max(URank(t)) * TRank(t) \quad (2)$$

where the *max(URank(t))* is maximum of all user ranks of all users, who are authors of tweets, which contains extracted keyword and *TRank* is the textual relevance of keyword.

For evaluation, we obtained more than 50 GB dataset from *Twitter* using *Twitter streaming API* during 10 days. In a preliminary experiment, we evaluated the results of our method for the set of 10 recent URLs from Twitter. We obtained average precision 86 %. We also measured to what extent our method enriched basic set of keywords extracted from resource content only. The Twitter keyword we consider to be important, when it is relevant and we cannot find it within keywords extracted directly from content of URL. 46 % of extracted keywords matched this condition. We consider this enrichment very reasonable and metadata coming from *Twitter* to be very valuable.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Tunkelang D.: *A Twitter Analog to PageRank*, 2009. Available at: http://thenoisychannel.com/2009/01/13/ a-twitter-analog-to-pagerank/ [cit. 2011-11-6]

[2] Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: *Proc. Of Conference on Empirical Methods in Natural Language Processing*, ACL, 2004, pp. 404–411.

Workshop Events Reports

Hadoop Workshop Report

Tomáš KRAMÁR, Dušan ZELENÍK

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
{kramar,zelenik}@fiit.stuba.sk

1 Motivation and expected outcomes

After the last year's successful PeWeProxy workshop, we have decided to incorporate another hack-afternoon/workshop into the 11th PeWe meetup. We decided to base the workshop on two events that happened in the life of our institution since the last year:

- we operate a Hadoop cluster, suitable for processing large amounts of data
- we have started a cooperation with Azet, largest Slovak internet company, with the aim of improving the click through rate (CTR) of their advertisement. For research purposes, we have available a large dataset of advert impressions and their respective clicks.

For this workshop, we decided to organize a hack-afternoon, which would help our bachelor and master students to familiarize themselves with both Hadoop platform and Azet dataset, and thus our goals were:

- introduce students to the Azet dataset and encourage them to play with it;
- introduce students to our new cluster, familiarize them with basics of the Hadoop environment, mainly the map-reduce framework and technologies based on it, such as Hive and Mahout;
- show students that they can use the cluster to process large datasets

Unfortunately, the venue of the PeWe meetup lacked a working Internet connection, so we decided to build and bring in our own small cluster of commodity PCs, just for the purposes of the workshop. We have connected 8 PCs, with a total of 28 CPUs and installed Hadoop environment on each of them. We built and configured the cluster in the lab and then transferred it to the meetup location. The connection to the cluster was enabled by a wireless access point, connected to the master node in the cluster. Given the processing power of our cluster and number of simultaneously working participants, we have decided to only work on a small part of the dataset.

The task given to the participants of the hack-afternoon session held at the PeWe meetup, given the short duration of the workshop, was to analyse the given dataset,

Spring 2012 PeWe Workshop, April 1, 2012, pp. 87–94.

find interesting facts that could help us to get an initial insight into the data and eventually help us to build a method for optimising impressions placements.

We expected not only to collect interesting ideas and results, but our aim was also to show to PeWe members how easily they can realize their ideas in the Hadoop platform and thus how easily they can use this platform as a basis for their future work. Last but not least we wanted to excite them about the available data and the task of optimising advert placement.

2 Team Reports

The hack-afternoon was preceded by an introductory pre-session few days before the actual event, where we discussed the basic concepts and ideas of the Hadoop platform. This kick-off session was supposed to make the attendees familiar with the possibilities and to provide a short overview of available data, so the participants could start hacking immediately.

Participants were supposed to work in teams, with each team having a doctoral student as a team leader. Each team was asked to come with an idea of analysis, which could be realized on the available dataset with the aim of improving CTR of the adverts. Teams were created right after the tutorial part of the kick-off session – we strived for balanced distribution of master and bachelor student across all teams.

In this section, we present ideas and results achieved by individual teams.

2.1 AdSexynessMeter

Team leader: Marián Šimko
Team members: Pavol Bielik, Peter Dulačka, Máté Fejes, Matúš Tomlein, Tomáš Uherčík

To increase profit from advertisements, advertising companies perform analysis of web pages and users of the Web representing potential customers in order to increase click through rate (CTR) for delivered ads. We believe not only proper association of an ad with a web page and potential customers (i.e., offering right product for right person) is important for achieving the goal. In our work we focused on quality analysis of advertisement content to study properties of ads that have high CTR.

We were interested in a very basic characteristic of underlying content textual description – letter occurrences. Based on an assumption that visual experience from advertisement, i.e., its lexical “attractiveness”, positively reflects into advertisement’s CTR, we performed lexical analysis of content of advertisements with high CTR. We introduced a measure representing an extent to which ad content is attractive for a user. We refer to it as *sexyness* of an advertisement. A method determining sexyness of an advertisement is based on statistical comparison of the advertisement’s letters distribution with distributions of the letters in so called sexy corpora of advertisements, i.e., those with high CTR. The comparison of letters from sexy corpora with advertisements viewed, but not clicked, is shown in Figure 1. Both distributions are compared with distribution of letters from Slovak general corpora.

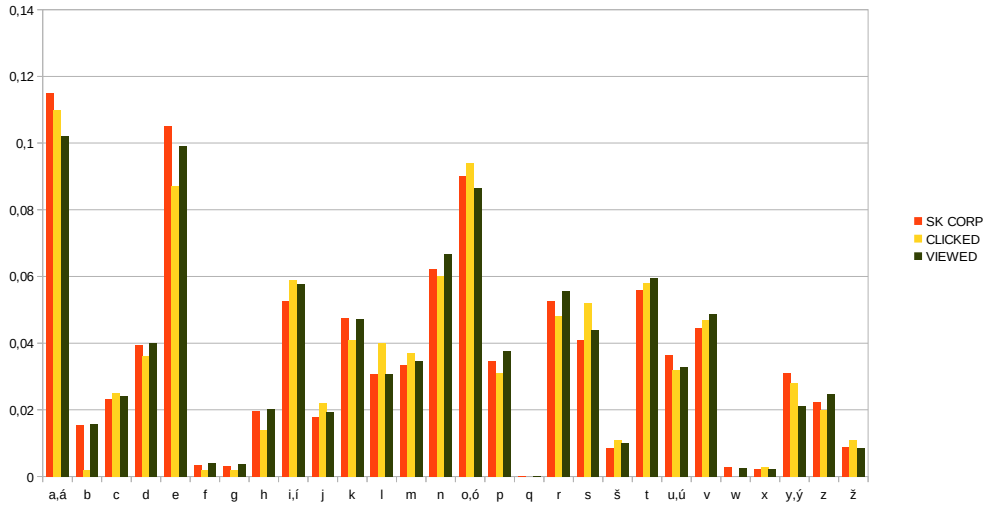


Figure 1. Letter distribution of ads with high CTR ("clicked", yellow) and low CTR ("viewed", green) compared with letter distribution in Slovak general corpora ("SK corp", red).

We greatly benefited from Hadoop architecture supporting parallel computing. In order to get statistics on letter occurrences in the provided dataset, we created map and reduce jobs to process huge number of advertisements.

However, we are aware of restrictions of the results. Since only the advertisement from short period of time (though a huge number of them) were present in the dataset, results can be biased by dominating trends popular during the period. Therefore, more comprehensive analysis of larger amount of advertising data is necessary.

2.2 The User Group Recommendation for the Advertisement

Team leader: Karol Rástočný
Team members: Roman Burger, Jakub Ševcech, Eduard Fritscher, Martin Lipták, Marcel Kanta

One way to increase the advertisement CTR is to display interesting advertisement to users. While this can be realized by employing standard recommendation algorithms, it must overcome a barrier that users are not primarily interested in advertisements (they do not want to see them) and are even more reluctant to have personalized advertisements. It can lead to an impression that users are being spied upon. To solve this issue we look on advertisement recommendation from a different point of view: we do not recommend advertisements to users, but we recommend user groups to advertisements.

We evaluated our idea by a simple method which recommends user groups by user's age and user's gender. This method works over pre-computed data – the number of clicks on advertisements with specific keywords by users in the group. We build this dataset using the set of simple Hive queries evaluated over standard logs. The recommendation algorithm then only extracts keywords from the target advertisement

and summarizes the number of clicks for each user group and each extracted keyword. After that the algorithm normalizes summarized clicks and calculates relative weight for each user group. We tried this simple method with advertisement's phrase "The sale of winter tires." and we obtained promising results, when the user group of 18-39 old males was recommended (see Table 1).

Our proposed method provides only simple recommendation and does not perform difficult heuristics or computation to increase accuracy. But this method can work with large logs distributively, it updates pre-computed values using delta increments. It also allows user groups recommendation for the advertisement at the time of annotation's entry to the system, so a recommendation does not need to be performed at each page load. This will save some computing resources and decrease an access time.

Table 1. Recommendation results of our prototype for the phrase "The sale of winter tires."

Age Group	Sex Group	Weight of User Group for the Phrase
1-13	Female	0.004
14-17	Male	0.025
14-17	Female	0.004
18-23	Male	0.211
18-23	Female	0.032
24-29	Male	0.329
24-29	Female	0.014
30-39	Male	0.343
30-39	Female	0.011
55-	Male	0.029

2.3 Correlating User's Age with Interests

Team leader: Jakub Šimko
Team members: Róbert Horváth, Peter Macko, Tomáš Jendek, Ján Trebuľa, Anton Benčič

We analyzed the click-through rate at first and projected the click-through rate for individual regions and age values. While the regional distribution appeared quite flat (the differences of click-through rates per region were marginal), the age-based distribution has shown interesting trend: the youngest users (8-11), presumably with only little Web experience tended to ad-clicking twice as much (0,0005) as their older coevals. Low rates were then flat until the age of 35 years, when they begun to rise slowly but steadily, reaching back the 0,0005 border around the age of 45 and rising even further in higher ages reaching almost 0,001 at the age of 65. The explanation for this might lie in the lack of experience with the web and lesser ad-blindness of the middle and senior generations.

Secondly, we performed frequential analysis of keywords used in ads clicked by different age groups. Not surprisingly, while youngsters enjoy the life of boundless communication (reacting especially on mobile operator offers), for middle and senior

generations different issues were identified such as financial aid offers, living or furnishing offers and medical treatment offers.

2.4 Time and Advert Type Correlation

Team leader: Michal Kompan
Team members: Peter Kajan, Samuel Molnár, Peter Svorada, Pavol Zbell

Today's web-based applications have to face up tremendous information increase. On one hand more and more users access various social portals, but on the other hand, these users produce they own web content. When mixed, this produces non trivial amount of data, which in order to create personalized content or experience have to be processed and analyzed. Moreover, not only users but businesses respectively can benefit from adverts personalization.

In order to figure out specific recommendations for the business (increase amount of click over adverts) we compared clicked and displayed adverts for specific campaigns and hours respectively. Several interesting facts were discovered. First of all, the phase of day was investigated. Our results clearly show that users tend to click on displayed adverts between 16:00 and 24:00. This brings us to the assumption that most of users are browsing the specific portals from their homes. When comparing the click rate for one specific campaign, we observed that clear peak can be identified, when the campaign is clicked. This can be used to find the specific time for campaign to display.

In order to decrease the amount of parameters logged and observed for specific advert display, we investigated the correlation between attributes. We found relatively strong correlation between attributes such as town and campaign ID (Table 1). These findings were quite unexpected.

Table 2. Most interesting and strong correlations between attributes.

Attribute 1	Attribute 2	Correlation
Town 109	Campaign ID range702 [12149.500 - 12182.500]	1.0
Town 102	Campaign ID range947 [12477 - 12584.500]	1.0
Town 94	Campaign ID range989 [12708.500 - 12787]	1.0
Town 56	Campaign ID range702 [12149.500 - 12182.500]	1.0
Zone 1065	Campaign ID range15110 [16941.500 - 16942.500]	0.868

To enhance the possible recommendations, the process of data mining and application of association rules seems to be a promising approach.

2.5 Mapping the World of Advertisement

Team leader: Michal Holub
Team members: Michal Barla, Milan Lučanský, Peter Macko, Márius Šajgalík, Juraj Višňovský

For us, the world of on-line advertisement is an unexplored area. There are many factors influencing whether the user clicks on an ad ranging from demographic data (e.g., age or region) to external context (e.g., season of the year). Therefore, we

decided to firstly map this area by summarizing the information available and creating various statistics about the users and their behavior.

We accomplished this task by executing queries on a map-reduce cluster using Hive. We were only interested in impressions that were subsequently clicked so we filtered out the unclicked ones in the preprocessing phase. Then we worked with much smaller data sample which speeded up our work.

Because the data contained 8 positions on a web page where an advertisement can be displayed we were interested whether the clicks were evenly distributed over all the positions. We found out that the first 2 positions are the best ones for an advertisement to appear, they had 10 times more clicks than positions 3-5 and 100 times more clicks than positions 6-8.

There are significant differences among the advertisers; around one third of all the advertisers did not get a single click on their advertisements. This brings an opportunity to incorporate the strategies of successful advertisers as positive examples in a machine learning process and test the results on the (yet) unsuccessful ones.

The advertisements do not attract people's attention evenly during the week. Each day there is a different advertisement which has the most clicks. One reason could be that the advertisement is only visible during one day. Another possible explanation is that on each day people look for different information (e.g., on Friday they look for tips for weekend activities). The number of clicks also differs in various hours of the day. The most successful advertisements were clicked on at 11 am (possible coffee break), 4 pm (people finishing work just before going home) and 10 pm (right before going to sleep). We also discovered that the most active group of users is aged 18-26.

These findings are limited by the size of the dataset we had. However, we think that we can improve the click-through rate of an advertisement by incorporating a simple recommender system which will select the most appropriate advertisement based on the demographic data and previous behaviour of similar users. We can also improve the advertising strategies by learning from successful advertisers.

2.6 Finding Characteristic Patterns of Behaviour in Online Advertisement Clicking

Team leader: Eduard Kuric
Team members: Róbert Móro, Balázs Nagy, Ondrej Kaššák, Martin Konôpka

Online advertisement is a main source of income for many portals and services on the Web nowadays; they offer their content for free for all users to consume, while placing the advertisement in the form of banners on their web pages.

The advertisers pay for the number of clicks on the adverts, not the number of times the adverts are shown (i.e., impressions). Therefore *click through rate* (CTR, number of clicks divided by the number of times the advert is shown) is used as a measure for evaluating whether an online advertisement campaign is successful or not.

While the average CTR on the Web is reported to be between 0.2 to 0.3 per cent [1], our dataset shows CTR to be only about 0.03%. We have focused on finding typical patterns of behaviour in users' clicks on the online adverts. The identified patterns could be used to target adverts to specific segments of users, thus effectively raising the CTR and income both for the advertiser as well as the web page or portal.

We were interested in whether there is any difference in clicks between days in week and during weekend. Our hypothesis was that during weekend users prefer advertisements about leisure activities or holidays, while during the week they are more inclined to click on work-related adverts. We have found out that most clicked adverts were the same for week as well as for the weekend. However, there have been differences in less clicked adverts which suggest that our hypothesised pattern exists. For example, during the weekend there were more frequently clicked adverts targeted at household services.

We were also interested in whether we could find typical patterns based on users' age. We have found out that there are some interests which do not change with age. On the other hand, there are specific adverts targeted at a particular age group (see Table 3). For example, users' interests in an advert of a cellular company were relatively uniform across all ages while adverts targeted at lending money were more interesting for users between age 30 and 50.

Table3. Examples of age-specific words appeared in clicked adverts.

Users' age	Words appeared in a text of an advert
0-19	O2, free, SMS
20-29	O2, eshop, price
30-39	O2, loan, sale, tyres
40-49	O2, loan, repayment, house, heating
50-59	O2, liver disease, kidney, body, house
60-69	O2, family, traveling, holiday

2.7 Demographics, First and Repeated Clicks and Portals in the Azet Ads

Team leader: Martin Labaj
Team members: Ľuboš Demovič, Marek Láni, Ivan Srba, Andrea Šteňová, Maroš Unčík

In the available pre-processed logs consisting of ad impressions with matched ad clicks, we focused mainly on the demographic data of users in relation to clickthrough rates (CTR) on various portals. As tools we used Apache Hive for data retrieval and basic processing on the Hadoop cluster and local statistical and spread sheet applications for further analysis.

First we considered user's gender. Impressions-wise, the gender A sees 56% and the gender B sees 44% of all impressions, however the gender A has overall CTR of 0.028%, while the gender B has CTR of 0.032%. While the data did not include information on gender-to-ID mapping, by considering portals often visited by different genders (women's magazine versus information technology news), we estimated that women users have the ID which we address here as A.

Next, we analysed when users do click for the first time and how they click repeatedly. The ads come from various campaigns and in each campaign several banners may exist. The data suggest that when a user has not clicked on one banner or on any other banner from one campaign after first 10 to 15 impressions of given banner or campaign, the ad space can be better used for another campaign as user is very less likely to click after further impressions (power law distribution with long tail). Also

very few users click again in the same campaign, therefore after a first click in a campaign, a place can be used to display other campaigns.

Different age groups were clicking the most on different campaigns with average age per campaign being from 20 to 40 years. Most users are 15 to 40 years old, and there are very few users older than 65 years (and many of such users may have false demographic data, since there are also users reported to be older than 100 years). The CTR is rising slowly with the age of a user. After the boundary of 70 years, the CTR rises to very high values (3.5%), however there are too few users in this age group with even fewer clicks to draw assumptions. Next, the CTR is different across various portals, with the lowest being on a dictionary site and the highest being on a chatting portal. While the most visited is the aforementioned chatting portal and therefore the most impressions come from it in all age groups, the clicks (CTR) are varying across different age groups across different portals (e.g., in a group of 30 to 40 years old users, the CTR is highest on a technology news site). Having established relationships between age groups, campaigns, CTR and portals, it is evident that for each campaign a target age group should be determined or evaluated and then the campaign should be displayed more on selected portals.

We are aware of limitations of our study. The pre-processed data, while large in volume (9M impressions with demographics), spanned only limited time intervals, which may skew presence and behaviour of various age groups. The portal information, represented as an ID, but having varying URL host data, is also uncertain in the sample data. Therefore results in this abstract should be taken as preliminary and they should be further evaluated.

3 Summary

We were very pleased with the course of the whole event, the motivation and enthusiasm of all participants. The presented ideas and results form a base to work upon and many of them will be further expanded. To illustrate the hard work of our participants, we present some statistical evidence: we have collectively burned 10.46 CPU hours of our cluster, used 485 GiB of RAM, and used 4.59kWh of electricity. We also appreciate the quality of final presentations of the projects, given by team leaders.

Acknowledgement. This contribution is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

The authors would like to thank all participants for excellent work and ideas. We would also like to thank Professor Mária Bieliková for her help with organisation of the event and motivating all participants.

References

[1] Stern, Andrew: *8 Ways to Improve Your Click-through Rate*. [Online; accessed April 5, 2012]. Available at: http://www.imediaconnection.com/content/25781.asp

SeBe 6.0: Beer Distribution Issues and Solutions

Marián ŠIMKO, Jakub ŠIMKO, Michal BARLA

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
{simko,jsimko,barla}@fiit.stuba.sk

The Beer Driven Research (BDR) phenomenon has had a glorious existence of over past years and has revolutionized the way we work and live [1, 2, 3]. Its major aim and mission is to foster searching for a peak, when research skills exceed average value (see Figure 1). Being inspired by preceding research in the programming field [4], we now focus not only on finding, but also on preserving a desired level of skill of research. Our initial experiments comprise research in the field of software engineering in particular.

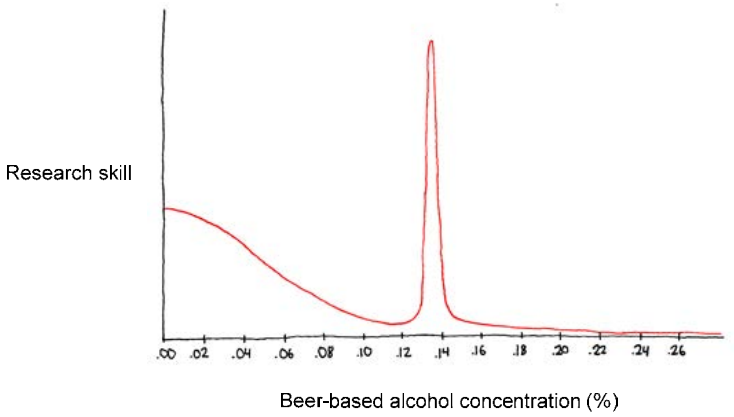


Figure 1. Research skills peak. Inspired by [4].

Traditionally, BDR researchers and practitioners, unified under the banner of beerlightment, conduct their semi-annual Semantic Beer meetings and the spring of the year 2012 was no exception. This time, the main topic discussed was the role of ancient brewing traditions in shaping our modern savouring paradigms. For this sake, participants of the workshop submitted a wide variety of raw data, which was then systematically consumed applying also new distributed savouring methods.

Spring 2012 PeWe Workshop, April 1, 2012 pp. 95–96.

The participants submitted to the following tracks:

- ancient treasures,
- colonial era brews,
- beer moderna – the story of the 20th century,
- new-age streams – freshest trends in the beernovation field.

In this edition of SeBe we focused on acquisition of temporal metadata of beers. We follow the efforts to assembly and populate a beertology introduced at SeBe 4.0 [2] and continued at SeBe 5.0 by acquiring spatial metadata [3]. The submissions did not disappointed expectations.

In order to recognize the most valuable submissions, we rewarded authors with the most significant impact on the ontology being acquired. Maté Fejes and Ivan Srba earned a prize for the most ancient submissions and Michal Holub together with Andrea Šteňová were awarded for the biggest spread in the temporal variance of the submissions.

As the tradition is, a new game-with-a-beer (GWAB), this year called the Tap and Reduce, was also played by workshop participants. The game was particularly aimed at cultivation of beer savouring. Each participant was offered a mug of beer with no information about its brand (the 'Tap' phase). The participants were instructed to find all other participants offered with the same brand (the 'Reduce' phase). However, they could do so only by discussing the taste and experience of the beer. Eight brands of different tastes were distributed in the crowd. To help pioneers in the field, each mug was assigned a part of a phrase popular in the SeBe community. By combining mugs of common taste, participants could assembly correct form of phrases, helping them to find their beer-fellows.

There indeed is a great potential in the SeBe community. Soon after the start of the game, a winner had emerged, the group of very brave minds and tongues: Karol Rástočný, Samuel Molnár, Márius Šajgalík, Ján Trebuľa and Jozef Tvarožek.

At SeBe 6.0, experts and young researchers in Beer Computer Science and allied fields researched beer distribution issues and solutions. SeBe pushed further the BDR phenomenon by contributing to the beer-awareness of crowds. Although many techniques and methods were devised, there is still much in crowdsavouring that has to be discovered to address the challenge of her highness, the beer.

Acknowledgement. This work has been totally supported by our never ending desire and thirst for knowledge.

References

[1] Šimko, M. Barla, M., Šimko, J. Zeleník, D. SeBe 3.0: Means of Beernovation. November 2010, Modra, Harmónia, Slovakia (2010).
[2] Šimko, M. Barla, M., Šimko, J. SeBe 4.0: Towards Ubiquitous Savouring. In Proc. of 9th Spring 2011 PeWe Workshop: Personalized Web – Science, Technologies and Engineering. Viničné, Galbov Mlyn, 2011, pp. 91–92.
[3] Šimko, M. Šimko, J., Barla, M. SeBe 5.0: Mug-Centered Design. October 2011, Modra, Slovakia, 2011.
[4] Munroe, R. Ballmer's peak. In XKCD. Available at: http://xkcd.com/323/

Index

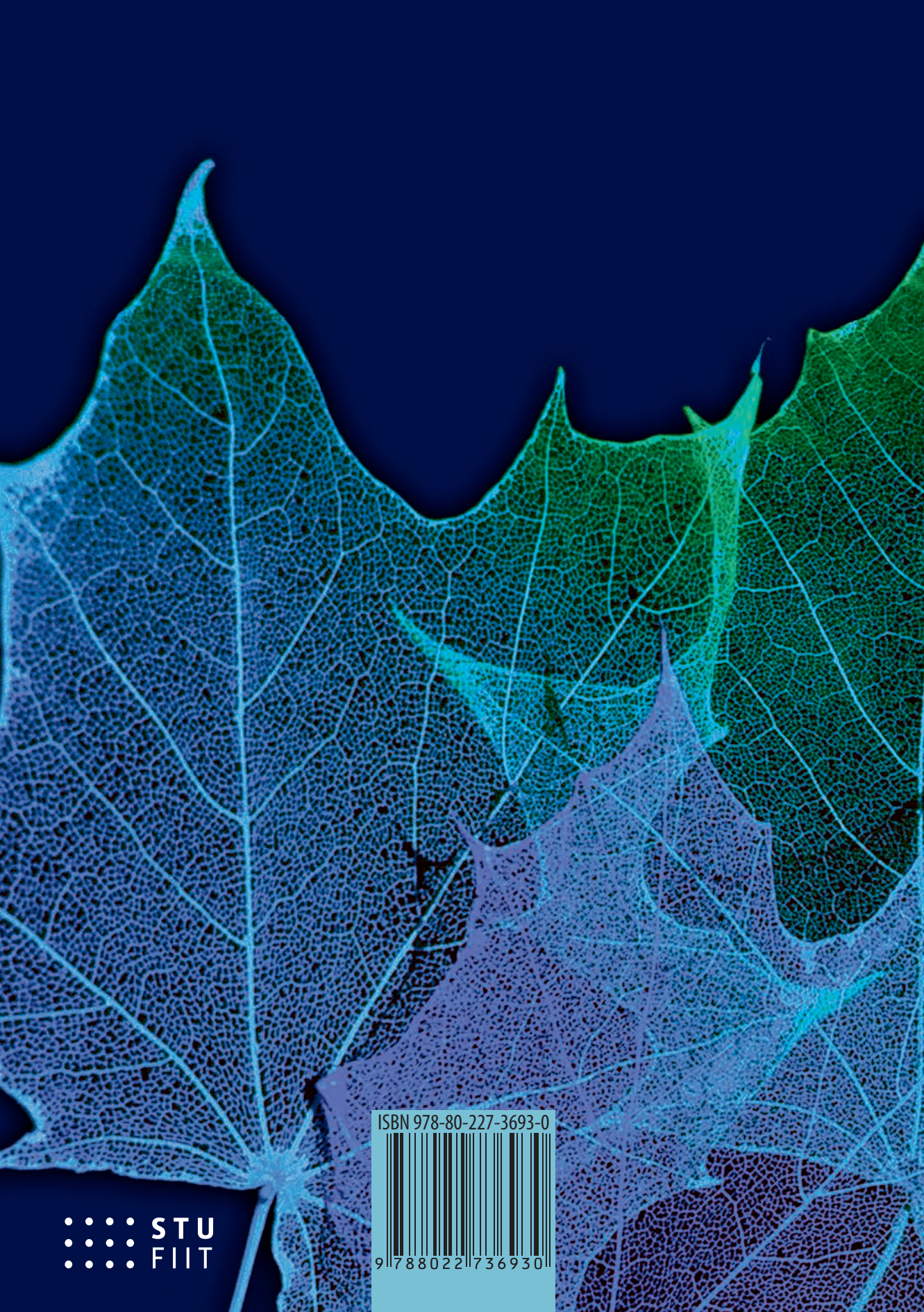
Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko, Jozef Tvarožek (Eds.)

Personalized Web – Science, Technologies and Engineering
11th Spring 2012 PeWe Workshop

1st Edition, Published by
Slovak University of Technology in Bratislava

114 pages, 50 copies
Print Nakladateľstvo STU Bratislava
2012

ISBN 978-80-227-3693-0



ISBN 978-80-227-3693-0

STU FIIT