Personalized Web - Science, Technologies and Engineering

13th Spring 2013 PeWe Ontožúr,
Gabčíkovo, Slovakia, April 2013
Proceedings

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

Proceedings in
Informatics and Information Technologies

Personalized Web – Science, Technologies and Engineering

13th Spring 2013 PeWe Workshop

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

Personalized Web – Science, Technologies and Engineering

13th Spring 2013 PeWe Workshop
Gabčíkovo, Slovakia
April 5, 2013
Proceedings

Slovakia Chapter
PeWe Group

STU
FIIT

SLOVAK UNIVERSITY OF
TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS
AND INFORMATION TECHNOLOGIES

Proceedings in
Informatics and Information Technologies

Personalized Web – Science, Technologies and Engineering
13th Spring 2013 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, 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 contracts No. APVV-0208-10, APVV-0233-10

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 978-80-227-3914-6

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. Topics related to the Web each year attract many students, which results to a number of interesting results achieved by enthusiastic and motivated students.

This volume is entirely devoted to students and their research. It contains short papers on students' research projects presented at the 13th PeWe (Personalized Web Group) Workshop on the Personalized Web, held on April 5, 2013 in Gabčíkovo. All papers were reviewed by the editors of these proceedings. The workshop 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:

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

The projects were at different levels mainly 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 one of our bachelor students who take an advantage of our research track offered within study programme Informatics – *Ľubomír Vnenk.* Ľubomír was just about to start his bachelor project.

Bachelor projects:

- *Dominika Červeňová:* Emotion-Aware Movie Recommender Based on Genre Impact Analysis

- *Peter Demčák, Ondrej Galbavý, Miroslav Šimek, Veronika Štrbáková:* Improving Speech Therapy by Motivational Home Exercises
- *Marek Grznár:* Adaptive Feedback in Web Systems
- *Matej Noga:* Recommendation based on Difficulty Ratings
- *Matej Marcoňák:* Querying Large Web Repositories
- *Martin Markech:* Semantic Wiki for Research Groups
- *Samuel Molnár:* Trending Words in Navigation History for Term Cloud-based Navigation
- *Pavol Zbell:* Concept Location Based on Programmer's Activity

Master junior projects:

- *Karol Balko:* Keeping Information Tags Valid and Consistent
- *Ľuboš Demovič:* Linked Data on the Web in order to Improve Recommendations
- *Peter Dulačka:* Combining the Power of Crowd and Knowledge of Experts in GWAP
- *Eduard Fritscher:* Group Recommendation of Multimedia Content
- *Martin Gregor:* User Modeling for Facilitating Learning on the Web
- *Ondrej Kaššák:* Group Recommendation for Smart TV
- *Martin Konôpka:* Software Metrics Based on Developer's Activity and Context of Software Development
- *Jakub Kříž:* Context-based Improvement of Search Results in Programming Domain
- *Marek Láni:* Acquisition of Learning Object Metadata Using Crowdsourcing
- *Martin Lipták:* Researcher Modeling in Personalized Digital Library
- *Martin Plank:* Extracting Word Collocations from Textual Corpora
- *Ondrej Proksa:* Discovering Links between Entities on the Web of Data
- *Michal Račko:* Automatic Web Content Enrichment Using Parallel Web Browsing
- *Richard Sámela:* Personalized Search in Source Code
- *Andrea Šteňová:* Browsing Information Tags Space
- *Matúš Tomlein:* Modelling the Dynamics of Web Content
- *Juraj Višňovský:* Context-Aware Recommender Systems Evaluation

Master senior projects:

- *Miroslav Bimbo:* User Interest Modelling Based on Microblog Data
- *Roman Burger:* Personalized Web Documents Organization through Facet Tree
- *Máté Fejes:* Facial Expression Recognition for Semantic User Modeling
- *Róbert Horváth:* Augmenting the Web for Facilitating Learning
- *Peter Krátky:* User Modeling Using Social and Game Principles
- *Jozef Lačný:* Personalized Recommendation of Learning Resources
- *Peter Macko:* Preprocessing Linked Data in order to Answer Natural Language Queries
- *Štefan Mitrík:* Discovering and Predicting Human Behaviour Patterns
- *Balázs Nagy:* Metadata Collection for Personal Multimedia Repositories Using GWAP

- *Jakub Ševcech:* Web Navigation Based on Annotations
- *Michal Tomlein:* Method for Social Programming and Code Review
- *Matúš Vacula:* Information Retrieval Using Short-term Context
- *Petra Vrablecová:* Relationship Discovery from Educational Content

Doctoral projects

- *Michal Kompan:* User's Satisfaction Modelling in Personalized Recommendations
- *Tomáš Kramár:* Multiple Sources of Search Context, Their Influence and Applicability
- *Róbert Móro:* Exploratory Search Using Automatic Text Summaries
- *Eduard Kuric:* Activity-Based Programmer's Knowledge Model for Personalized Search
- *Martin Labaj:* User Feedback in User/Domain Modelling and Adaptive Evaluation
- *Ivan Srba:* Reciprocity as a Means of Support for Collaborative Knowledge Sharing
- *Márius Šajgalík:* Using Site Specificity to Build Better User Model from Web Browsing History
- *Michal Holub:* Building a Domain Model using Linked Data Principles
- *Karol Rástočný:* Information Tags Maintenance: Anchoring
- *Jakub Šimko:* Crowdsourcing in the Class
- *Dušan Zeleník:* Beyond Code Review: Detecting Errors via Context of Code Creation

Considerable part of our research meeting this year was devoted to *discussion club* activity followed by several experiments. Discussion club was chaired by Jakub Šimko and Dušan Zeleník. Our aim was to engage all participants into a discussion on topics related to research methods (in particular the experimentation) and technologies. We organized the discussion club as a two-hour session. During this session eight groups of 5-6 workshop participants discussed various topics to which they have been assigned. The groups have been composed of students of different levels of seniority (ranging from senior doctoral students who prepared the topic to junior bachelor students) to maximize knowledge exchange potential. At the end of the session, each group presented resumé of their discussion.

One of the main goals of this year workshop was to help PeWe members with evaluation of their research. The members of the PeWe group helped their peers by providing not only constructive criticism and advice for experiment design; they also actively participated on the announced experiments. Conducting experiments even in our small group is always useful, in particular for initial evaluation, to verify preconditions of proposed methods, or locate errors in the design that helps in the end to achieve better results. Experimentation session of 9 experiments that together required active participation of 2 000 minutes resulting in 50 minutes on average for each workshop participant. The participants in many cases took part on more than one experiment; they played a game, navigated in the digital library, annotated and organized their documents, learned new vocabulary, enriched wiki with semantics, provided feedback or ordered named entities. The session lasted for several hours and officially ended at 1am with some of experiments continuing even after that.

Our workshop hosted for the eighth 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, Róbert Móro and Michal Barla. The main topic of this year workshop was the unity of science and art as presented in visual aspects of beer-human interfaces. The focus was on artistic features of beer etiquettes and bottles that are frequently overlooked in the-state-of-the-art research, but are nevertheless important as a means for visual navigation and shelf search, driving our decision-making during travelling through beer-selling spaces.

As an accompanying event of our workshop, we conducted the Gold Rush competition. This competition was inspired by popular competitive style TV shows. We particularly copied the principles from well-known Czech and Slovak TV show called Riskuj, broadcasted in late 90's. To make Gold Rush personalized for our research group we prepared 210 questions regarding the knowledge base needed to accomplish second degree in software engineering and information systems. We arranged 7 rounds for 8 teams. Every round represented the fight between two teams which were competing to achieve new round. One team managed to win all the rounds. This team was awarded together with other awards within the Sebe Workshop.

More information on the PeWe workshop activities including presentations is available in the PeWe group web site at pewe.fiit.stuba.sk. Photo documentation is available at mariabielik.zenfolio.com/ontozur2013-04.

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 and main actors in the workshop show for contributing interesting and inspiring research ideas. Special thanks go to Katka Mršková for her effective organizational support of the workshop.

April 2013

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

Table of Contents

Students' Research Works

Personalized Search and Recommendation

Personalized Navigation

User Modeling, Virtual Communities and Social Networks

Domain Modeling, Representation and Maintenance

Semantics Discovery

Accompanying Events

13th PeWe Ontožúr participants

5.4.2013 Gabčíkovo, Slovakia

Scenzo

Peťo

Dušan

Ivan

Peťo MiB

Michal

Juri

Ondrej

Palo

Personalized Search and Recommendation

Emotion-Aware Movie Recommender Based on Genre Impact Analysis

Dominika ČERVEŇOVÁ*

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

Recently users’ mood has shown up as an important context feature, relevant to making recommendations and it has become an object of interest for many researchers. The interactive web radio Musicovery of Shi et.al [2] is only one of the many works that take users' mood into account.

Our context-aware knowledge-based movie recommendation method is based on assumption that there is a relationship between users' current mood and movie genre suitable for her at the specific moment. With the knowledge of how exactly specific genre influences emotions and provided that we have the information about users' current mood, we are able to determine which genres are the most suitable and make the decision which movie to watch much easier for user.

The method uses postfiltering of data from a metadata-based recommendation service developed at our faculty as a team project. It recommends user a list of movies that might be interesting for her in general. However we try to identify what the user might find interesting at the moment, to make the recommendation even more personalized and this is where the emotions help us. A schema of our recommendation method is in the Figure 1.

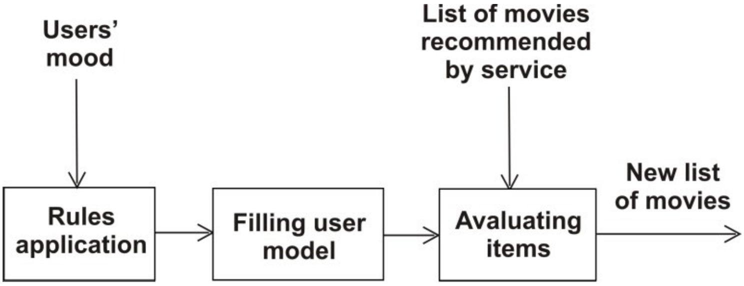


Figure 1. A schema of recommendation method functionality.

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

After retrieving the list of recommended movies from the service, we take information (gained explicitly) about users' current mood into account. Then based on defined binding rules between genre and mood, the method transforms the list of movies into a new list that we recommend to user.

To acquire the rules for recommendation we used several approaches. We started with simple manual rules based on our opinions, consulted the movie-mood relationship with a psychologist and interviewed 10 randomly selected web users of various age and interests. As the second step we tried to mine association rules from the LDOS-CoMoDa [1], a dataset of users, movies and other contextual information (including users' feelings before, during and after watching a movie).The result of rules extraction process was a table of percentual occurrence of each genre in each mood (positive, negative, neutral). It showed up, there are some genres basically independent on users’ emotions (e.g. Crime), but in many cases there were observable differences between frequency of choices in different mood. For example Drama appeared to be more wanted by negatively tuned people; on the other hand Comedy is more preferred by people in positive mood. Final created binding rules are represented by a binding matrix, where *value[i; j]* means desirability of *genre[i]* in *mood[j]*.

Our user model contains a relevancy of each genre in the context of current mood represented by a value computed according to the binding values. It is updated every time we get new information from a user about his/hers emotions. Then, we evaluate each movie from the list recommended by the service, by calculating an average value of desirability values for genres in the movie. In the end, the movies are sorted descending into a new list and the items with value greater than -0.8 are shown to user as the most proper recommendations.

Our recommendation method is currently being implemented and we already made some experiments with explicitly acquired context, using the LDOS-CoMoDa dataset that proved our hypothesis. In addition we are about to make some qualitative experiments with real users. A comparison between items recommended without our method and the resulting list after postfiltering applied and also a following feedback from users can fully confirm the relevancy of our method.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 409-410.

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

References

[1] Košir, A., Odić, A., Kunaver, M., Tkalčič, M., Tasič, J.F.: Database for contextual personalization. In: Elektrotehniski Vestnik 78(5), 2011, pp. 270-274.

[2] Shi, Y., Larson, M., Hanjalic, A.: Mining mood-specific movie similarity with matrix factorization for context-aware recommendation. In: Proceedings of the Workshop on Context-Aware Movie Recommendation - CAMRa ’10, 2010, pp. 34-40.

Improving Speech Therapy by Motivational Home Exercises

Peter DEMČÁK, Ondrej GALBAVÝ, Miroslav ŠIMEK, Veronika ŠTRBÁKOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
Icup2013@googlegroups.com

The ability of speech has an important place in human lives. It is closely connected to other human skills like thinking and perception, learning, writing and reading. Nowadays, many people and especially children suffer from speech disorders. It is essential that children perfect their ability to speak correctly before they begin attending primary school, because after reaching the age of seven, the correction of habitually incorrect speech becomes lengthy and rather difficult.

Many parents with children with speech disorder attend speech therapy. Although speech therapy may show the wanted results, about 40% children in Slovak and Czech primary schools still suffer from some kind of speech disorder. In secondary schools, this persists in approximately 10-15% cases. These numbers were striking enough for us to start finding ways in which we could support the process of speech therapy by modern technologies. The resulting solution we are introducing is called Speekle.

The speech therapists need to prescribe children various types of exercises:

- In many children the speech disorder originates from weak tongue muscles.
- There are many phonemes (speech sounds used in spoken language) that children find difficult to pronounce. Generally, teaching them to pronounce them correctly requires repeating the sounds many times, often continually and with use of over-pronunciation.
- Children who have difficulties with spotting difference between different phonemes usually need exercise on listening words with the help of their parent.

In reality, many children refuse to exercise and view the exercises as punishment because it is tiring and takes up a lot of their time. Parents also tend to underestimate the importance of exercising, or they may lack the needed time and energy to help their children to exercise. Although speech therapists can usually see whether the child has been exercising or not, they lack the means to ensure exercise is not neglected and to monitor and evaluate its correctness.

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

Our solution, which we propose to solve this situation, aims to improve home exercises. It stands on two main cornerstones:

- We created specialized methods for controlling a computer that can be used to support speech therapy – to emulate speech exercises in our application. We currently have at our disposal two types of special controls:
  - Tongue tracking – developed using Kinect sensor. This technology enables us to track the position of the tongue, so the patient can use it to control the computer. This way we provide means to exercise oral-motor skills of the patient.
  - Phoneme recognizer – made by real time sound analysis and extraction of certain problematic phonemes, which are characteristic for the local language. We concentrate on recognition of continually pronounced phonemes as they are pronounced during ordinary speech exercises.
- Each child who suffers from a speech disorder is its own case and has special needs. The human factor represented by a speech therapist cannot be replaced by any software system. This is why Speekle is more than just an application for children. It represents a whole platform which enables a speech therapist to review the information about children's exercising progress and configure the exercises to help solving complex problems related to speech.

Our solution is aimed at supporting speech therapy via reinvented practice of speech exercises. To build on our proprietary means of control, we created the client game application TalkLand. From the child's point of view, TalkLand is a game world where they participate in adventures by helping various characters in this world. They help them by playing minigames, which are designed to emulate a specific speech exercise. Minigames in TalkLand utilize not only ordinary controls but also tongue tracking and phoneme recognizer. Our special controls are naturally integrated into gameplay where they also support game immersion and provide instant feedback via gameplay to ensure the child exercises correctly.

Another part of Speeke – the Speekle web application enables speech therapists to monitor the progress of their patients and customize their therapy according the patients abilities and needs. The application also provides the speech therapist with "key moments", which are the most interesting recordings taken during the exercising session, accompanied by their success rates.

We deployed the first iteration of Speekle in speech therapy center ASOBI in Bratislava where children were amazed by playing with TalkLand and the speech therapist reviewed and approved of the session results.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 1-6.

Acknowledgement. This work was partially supported by the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava.

Group Recommendation of Multimedia Content

Eduard FRITSCHER*

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

In our times it is very important that web pages and applications, not only store information, but also that the page and application communicates with the user appropriately. The amount of information which is stored in online space has increased due to the growth of the web. To solve the problem of this information burst, recommendation techniques and methods are invented, but as the world changes, the access to the internet also changed. People collaborate more often with each other. The most visited pages in the world are social network pages hence group recommendation methods are needed to adapt to these new trends.

In order to make accurate recommendations we need data that describe users' interest and taste for a given domain. Considering that most information about users is stored in social network applications, we propose a method for recommendation that extracts information about users from social networks. Social networks offer a great opportunity for user information extraction. People willingly provide information about their taste and interest. Usually in social networks applications we have two types of data: structured and unstructured data [1]. Both types can be useful for recommendation but in case of using unstructured data we need to pre-process it. We need to extract knowledge from the data. After the data extraction we have the information about the user that we can use to make recommendations.

The main problem in group recommendation is the implicit heterogeneity of groups, i.e. that the generated recommendation has to satisfy requirements of all the members in the group. Users have different tastes and interests. To solve this problem we are propose aggregation strategies for the collected data from the users. These aggregation strategies are able to determine the common interest and taste of the group. The most common approach to solve this problem is to create a virtual user for each group, and the recommendation is generated for the virtual user [2]. But there are other approaches like generating recommendation for each user and filtering the result by some group descriptors that describe the group [2]. The scheme of the proposed recommendation method is in Figure 1.

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

The last step will be devising a recommendation technique for the domain where the proposed methods will be tested. We need to be able to recommend to individuals before we can recommend anything to groups. The proposed recommendation method uses graph algorithms to predict the recommended content for the user. The domain where the evaluation will take place is recommendation of multimedia content, more exactly recommendation of movies.

The functionality will be integrated into the application Televido, which is an application that is able to generate personalized movie recommendations. Televido is available as web and mobile application so for the mobile we can integrate even the context for the recommendation. Televido uses graph databases and algorithms for its core functions. The proposed method will be tested on the Televido user base.

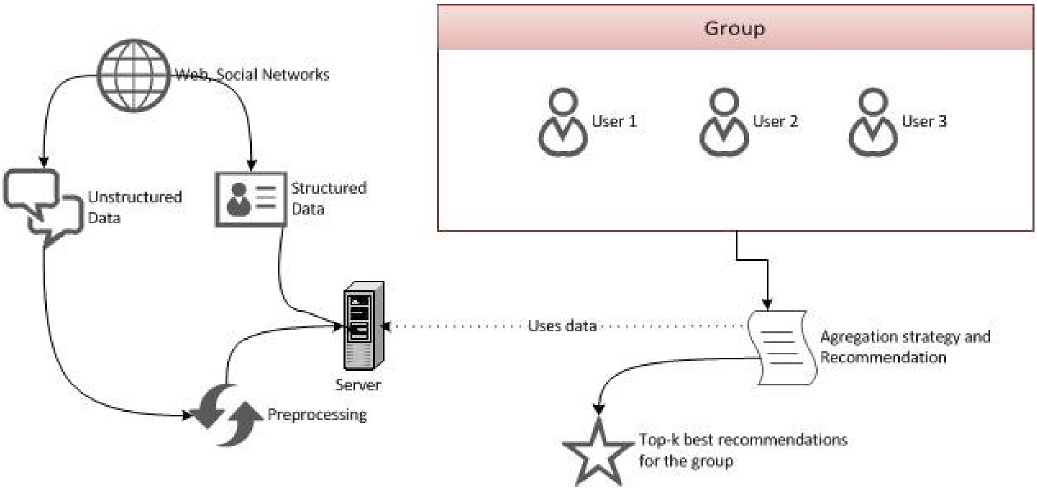


Figure 1. Scheme of the group recommendation method.

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

References

[1] Esmaeili, L., Nasiri, M., & Minaei-bidgoli, B.(2011). Personalizing Group Recommendation to Social Network Users

[2] Gartrell, M., Xing, X., Lv, Q., Beach, A., Han, R., Mishra, S., & Seada, K. (2010). Enhancing group recommendation by incorporating social relationship interactions. *Proceedings of the 16th ACM international conference on Supporting group work - GROUP'10*, 97. doi:10.1145/1880071.1880087

Group Recommendation for Smart TV

Ondrej KAŠŠÁK*

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

Watching TV is one of the activities that people often do together. According to the surveys [2], cases watching television the most are: parents with children, groups of siblings but also friends. Each of us, however, is having different interests and not everyone can be equally satisfied. At a time when there are 'smart' televisions that can offer a specific program from large databases, we can increase the satisfaction of users with recommendations of contents that is the most attractive opportunity for specific group and opportunity. When recommending to groups, we cannot recommend an individual list of programs to each user from the group. It is necessary to choose to all together and try as much as possible to satisfy the whole group. There, as with many other activities, we show that we are sociable individuals. We are often willing to offer a compromise and would rather watch a program which is not quite optimal together, then to watch a program that is more attractive individually.

In the domain of group recommendation, a group is made of individuals who collectively perform an activity. We always recommend a common content to the whole group. Therefore, we need to know which users are currently in the group. We also need to know what content individual users like.

The existing recommendation approaches store information about an individual user in the user model. In a model we represent information about what users are interested in and what content they prefer. Based on the models, we can recommend a relevant content to the users and subsequently to the groups.

In basic terms, the user profile is made up of a set of pairs [category, value], where value is a number from the interval <0, 1>. The size refers to the interest of the user for a category of items (such as comedies). Based on users rating and the time in which user views the items of a given category we can determine her interest in this category, and thus gradually model her profile [3].

A standard approach aggregates user profiles into a group profile, and make a common recommendation for the group represented by a single model. Another approach is to make recommendations for each user, and only then aggregate the results into a common group list [1, 3].

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

There are three common approaches for results aggregation. The first approach consists of techniques based on the majority, where the component most preferred by users is selected. There may be a disadvantage of ignoring minorities especially in large groups. The second approach is represented by methods based on consensus among the group members. Here we take into account all members and the recommended item is chosen from the average of all preferences. The third approach includes border strategies. These include dictatorial method in which we choose only by one dominant group member, least misery method where we select the solution with the smallest dissatisfaction of members or vice versa method with greatest pleasure for group members.

In our work we focus to achieve satisfaction of all group members (to maximize) in the group recommendation process. On the other hand, we believe that maximizing satisfaction for group as a whole is not enough. If a group is not completely homogeneous, it would ignore satisfaction of members whose view is in minority [1]. It would discourage them from further use of the service. Conversely, if users will be satisfied with the recommended content and they will see satisfaction of other members of their group too, there is a greater chance that they will use the service using this recommender in the future. To achieve this goal we propose a hybrid group recommender, which adapts to the actual group structure and social relations e.g. by changing the aggregation strategy dynamically.

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

References

[1] Kompan M.: Group Recommendations for Adaptive Social Web-based Aplications. Doctoral report. FIIT STU. 2011.

[2] Masthoff, J.: Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. In: User Modeling and User-Adapted Interaction. Vol 14. Kluwer Academic Publishers. 2004, Feb, pp. 37-85.

[3] Senot, C., Kostadinov, D., Bouzid, M., et al.: Analysis of Strategies for Building Group Profiles. In: User Modeling, Adaptation, and Personalization. Vol. 6075. Springer Berlin. 2010, pp. 40-51.

User’s Satisfaction Modelling in Personalized Recommendations

Michal KOMPAN*

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

Thanks to the huge information increase in the last years, the domain of personalized recommenders became intensively studied. Two basic approaches for the personalized recommendation have been proposed in the literature. The collaborative filtering approach uses similar users’ preferences to predict ratings for unrated items. On the contrary, the content based recommendation uses the content similarity with existing items to predict these ratings. These two approaches are often mixed in order to bring better results and to minimize the shortcomings of each approach, such as cold-start and sparse ratings.

One of the proposed enhancements to the recommenders’ approaches is considering of the user’s context. In this case the final predicted rating is not based only on the user and the item, but also his/her context is considered.

The problem of the satisfaction modelling [1] is well known in the domain of group recommendation. Users who experience the content within a group are influenced by the other users and thus the predicted rating can be dramatically influenced. For the user’s satisfaction modelling, we propose a novel method, which considers user’s context during the recommendation process and reacts to actual user’s circumstances by adjusting the predicted ratings for items (Figure 1). Proposed approach is based on the assumption that actual user’s ratings are influenced by the previous experienced content and actual user’s situation. This computation reflects to the user’s feelings intensity in the history and contributes to the actual predicted rating. Our method for satisfaction modelling consists of three steps:

1. Predict ratings for unrated items
2. Spread activation through user’s item specific influence graph
3. Combine user’s ratings history and result of influence graph

An influence graph is constructed for every user and the every predicted item’s rating. In this graph the vertices represent the user’s context (e.g. mood, day type) and

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

predicted item rating, and edges model the context influence (based on the assumption that a context can be strengthened by another context). Next the spreading activation is applied, which results to the adjusted rating based the actual context. As the graph consists of the context and item the performance decrease (comparing to standard approach) is minimal. The proposed approach considers the user's rating history as well, while the previous ratings are combined with the adjusted rating from spreading activation. This is done by weighting the history ratings and combine in the ration 1:2 with the adjusted rating. In this manner we are able to adapt to various user's contexts and domains.

Experiments support the hypothesis that the proposed approach outperforms standard prediction (average difference MAE-0,28, RMSE-0,30), and with incorporating into a recommender system it brings a recommendation improvement.

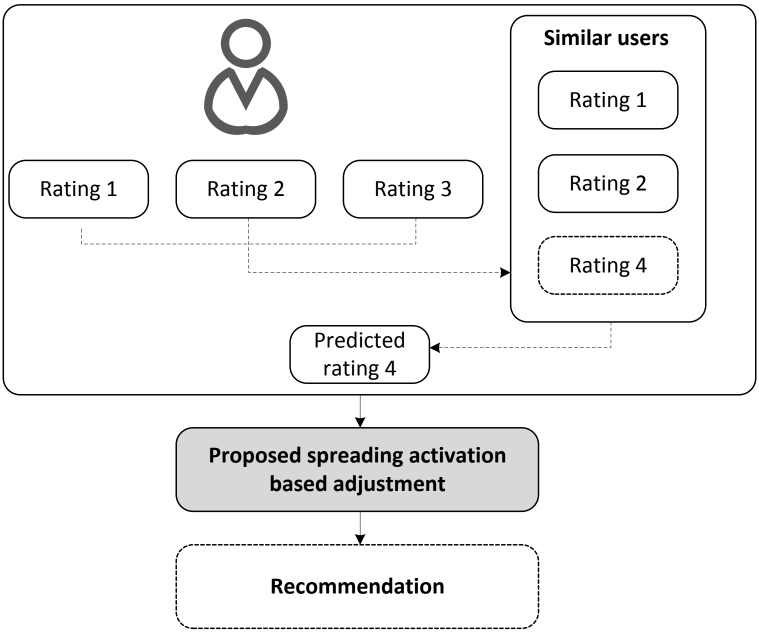


Figure 1. Collaborative recommendation process, enhanced by proposed satisfaction modelling (gray box).

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 67-73.

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

References

[1] Masthoff, J., Gatt, A.: In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Modeling and User-Adapted Interaction* 16, 3-4, (2006), pp. 281-319.

Multiple Sources of Search Context, Their Influence and Applicability

Tomáš KRAMÁR*

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

Web search begun as a relatively simple process, where the person types in the query in form of keywords, the underlying database of documents is searched for a match using the given model (e.g., vector space model) and the relevant documents are returned. The important concern that is not addressed by this process is the actual underlying goal that the person is trying to fulfil by issuing the query. With semantic search not yet widely used in production, a whole generation of people has been trained to express their information needs in form of keywords. Expressing the information goals in keywords may not be trivial; e.g. [2] shows that an average length of query is just 2.2 words, which is not enough to express complex goals.

It has been recognized that Web search needs some form of implicitly acquired information that would help to understand the underlying intent. That information is collectively referred to as a search context [1]. There is an important difference between the concept of context in the area of recommender systems and in the area of personalized search on the Web. In recommender systems, the context is viewed as a set of external attributes of the environment that have impact on user's immediate preferences. While many of these external factors arguably impact the perceived relevance of Web search results, it is mostly in cases where the search engine is used as a recommender system, e.g. when searching for a restaurant, user's location in combination with the weather can play a large part in the relevance judgment.

Traditionally, in Web search the context describes any information that can be used to infer the specific goal that the searcher wants to fulfil by issuing a query. The concept of context per se is decoupled from its representation. The only important aspect of a search context is that it reveals information about the underlying search goal that can be used in the ranking phase of the search process to rank documents matching the goal higher.

Over the years many sources of search context have been identified and studied. They vary in many aspects, but their differences mostly draw from the character of the

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 13–14.

particular source and from the temporal characteristics of the particular source. There are sources, which produce long-running context, such as context of seasonality and there are sources which produce short-term context, such as activity-based context. Considering the multitude of available sources of contexts, there can be many sources at any single moment that provide some search context. Most of the research in the area of search context focuses on studying novel sources of context or improving existing sources of contexts. However, given the many available sources of context, it intuitively makes sense to combine them, but the effects of such combination have not been studied yet. A search system that leverages multiple sources of search context can run into two situations: when only one source is able to produce a search context – in this case, the single source of context can be used as usual and no problems arising from combination of multiple sources of context arise; and when multiple sources are able to produce a search context, which opens new research questions of which source of the context is the best, or how to combine multiple sources.

It has been shown that a single source of context can dramatically improve search relevance, but whether multiple sources of context can further improve relevance has yet to be shown. Many questions need to be answered before multiple sources of context can be reliably used in Web search, e.g.:

- Is there a single best source of context, so that it does not make sense to use other sources?
- Does the best source of context depend on characteristics of the user? Are there users for whom one source of context would bring most benefits?
- Can we combine the evidence from multiple sources of search context and improve the relevance of search results?
- Does the feasibility of the source depend on some query features, related to the source of context?

In this work we analyse three sources of context: a long-running context of seasonality, a short-term activity-based context and a social-based context. We monitor user’s clicking activity, extract many features from each search and analyse which features influence which context. Our goal is to tell whether the feasibility of a particular source depends on some of the external search features and to show whether a combination of multiple sources can outperform a single source of context.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 155-162.

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

References

[1] Lawrence, S.: Context in Web Search. IEEE Data Engineering Bulletin, 2000, vol. 23, no. 3, pp. 25-32.

[2] Jansen, J., Spink, A., Saracevic, T.: Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. Information Processing & Management, 2000, vol. 36, no. 2, pp. 207-227.

Context-based Improvement of Search Results in Programming Domain

Jakub KŘÍŽ*

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

When programming the programmer encounters many difficulties he cannot instantly overcome. These difficulties might be rather trivial, like errors in compilation or more complex, like tasks he cannot solve or tasks he does not remember the solution of. When solving one of these problems the programmer usually uses a web search engine to look for help. However, the search results might not be effective – the user usually enters a query consisting of a few words only. This query often does not describe the matter properly and so the search results might end up being inaccurate.

The programmer does many other things along with writing the actual source code. He opens applications and other source codes, searches them, writes notes and, last but not least, he searches the internet for solutions of problems or errors he might have encountered. He does all these tasks in a context or in order to do something, usually to solve a problem. Via identification of this context we can understand the meaning behind the programmer's actions. This understanding can be used to make the web search engine to be context-aware – to make web search results more accurate and relevant to the current task. In this work we analyse the existing methods used to mine the context in other domains and analyse their usability in the programming domain.

Context-based recommender systems or search engines are quite common on the web nowadays. When working with the web search engines, the context is often understood to be the search history of a particular user in the current session and the visited documents from the searches performed. The data used to create the context model with is gathered from the metadata from visited documents with various approaches. For example, White et al. [3] use document categorization whereas Kramár et al. [1], among other methods, use document content analysis, namely keyword extraction.

The most important source of contextual information in programming domain seems to be the source code the programmer is currently working on. The first goal of this work is to build a contextual model based on the metadata extracted from the

[*] Supervisor: Tomáš Kramár, Institute of Informatics and Software Engineering

source code. Metadata extraction from source code files is an area which has not been thoroughly researched in other works. Therefore we intend to use approaches used for metadata extraction from general documents and modify them, like standard methods for keyword extraction.

Using general statistical keyword extraction methods like TF-IDF to extract keywords from source codes has been evaluated before and showed promising results. A modification of the algorithm was proposed and evaluated in. Ohba et al. [2]. Authors modify the standard TF-IDF algorithm to extract what is called conceptual keywords. This method is designed to help programmers with reading and understanding source codes previously unknown to them. Its goal is to mine the keywords, which express helpful, key concepts for understanding of the algorithm. The method proved quite effective and it should, after some adjustments, also prove effective in order to build the conceptual model.

To help the programmer with errors and specific tasks we also need to find out what technologies he currently works with. Therefore we intend to modify the method to also extract technical keywords. A very important part of keyword extraction via TF-IDF is choosing the right corpus. When it comes to source code files, multiple choices exist – we can use all files written in the same programming language, all files in the same project, or, when extracting keywords from a fragment of the source code, only the current file. We intend to evaluate multiple approaches in order to determine the best approach to mining metadata from source code files.

We can combine this method with methods generally used to create context models, like the analysis of visited websites and with other, probably much less significant sources of context, like the keywords extracted from notes.

To actually improve the search results we can use standard methods, like search query expansion, which is often used with great success [1]. Our methods will be experimentally evaluated on a large dataset of logs made from programmers' activities.

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.

References

[1] T. Kramár, M. Barla, M. Bieliková. Personalizing Search Using Metadata Based, Socially Enhanced Interest Model Built from the Stream of User's Activity. In *Journal of Web Engineering*. Vol. 12, No. 1&2, pages 65-92. 2013.

[2] M. Ohba, K. Gondow. Toward mining "concept keywords" from identifiers in large software projects. In *Proceedings of the 2005 international workshop on Mining software repositories*, MSR '05, 2005.

[3] R. W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information re-trieval*, SIGIR '09, pages 363-370, New York, NY, USA, 2009. ACM.

Personalized Recommendation of Learning Resources

Jozef LAČNÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
lacny08@student.fiit.stuba.sk

Personalized recommendation plays an important role in wide variety of fields nowadays. Its main purpose is to deliver the most relevant content to each user in specific scenario. The library users would like to get recommended books according to their taste, the researchers would like to get papers in their field of study and the shoppers like to be offered goods to buy according to their actual shopping purpose. It has been done a lot of research in this field including various recommendation techniques – content-based recommendation, collaborative recommendation and many others approaches combining these two or inventing other new methods [1].

An interesting field for personalized recommendation arises in the field of education, where it is important to recommend study materials to achieve better study results and enhance whole learning process. This domain is very specific mainly regarding the process of choosing relevant study materials for recommendation, because the main focus here is not to fulfil individual satisfaction of the user but to help him to learn more effectively and achieve better results in a shorter period of time. To take it even further, it is very challenging to recommend for groups considering collaborative learning. Collaborative learning helps students to better understand the subject of study by letting them to interact and share their thoughts. In collaborative learning it is easier to cover bigger areas of study and advance faster as in individual learning, because diversity of individuals' knowledge in the group makes it necessary to discuss various matters and therefore enhance the whole group's understanding.

We propose a method for group recommendation in educational systems utilizing users' learning style in the process of group creation and recommendation itself. We use layered user model reflecting user's knowledge and knowledge spreading proposed in [2] and we enhance it with user's learning styles defined in [3] that we get from the questionnaire. The learning styles represent the user's cognitive style in four dimensions: perception (sensing/intuitive), input (visual/verbal), processing (active/reflective) and understanding (sequential/global). The learning styles are then

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

used to prioritize student's preferences in the process of calculating learning resources' suitability for recommendation where we take into account the fact that various learner types have various tolerance for difficulty, concept skipping and repetition. The suitability of a learning object is calculated as a minimum of three measures:

1. suitability of concept for further recommendation according to its difficulty and prerequisites fulfilment,
2. suitability of learning object difficulty,
3. suitability of repetition of learning objects.

In the end, we employ a hybrid approach to aggregate single students' recommendations for group recommendation. This approach combines the least distance and average value measures when aggregating single user's recommendation according to the variance of recommended items.

Because communication in collaborative learning is very important, we take this matter into account and facilitate communication between students within their group online directly in the educational system. We plan to evaluate our method in live setting on real users in learning system ALEF [4].

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 37–42.

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. In *Information Retrieval and Mining in Distributed Environments*, pp. 1-20, (2011).

[2] Unčík, M.: Visualization of User Model in Educational Domain. In *Proc. of Informatics and Information Technologies IIT.SRC 2012 Student Research Conference*, pp. 199-204, (2012).

[3] Silverman, L. K., Felder, R. M.: Learning and teaching styles in engineering education. In *Engineering Education*, vol. 78, pp. 674-681, (1988).

[4] Šimko, M., Barla, M., Bielikova, M. 2010. ALEF: A Framework for Adaptive Web-based Learning 2.0. In Reynolds, N., Turcsanyi-Szabo, M. (Eds.). *KCKS 2010, IFIP Advances in Information and Communication Technology*, Vol. 324. Springer, pp. 367-378, (2010).

Preprocessing Linked Data in order to Answer Natural Language Queries

Peter MACKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
pmacko@outlook.com

Nowadays, the Web contains a lot of webpages providing information intended to be processed by humans. Many of these pages need data storage with dynamical data loading and for these reasons they use object-relational databases. But what they don't have is fully linked content. The Semantic Web is based on a different concept. In the world of sematic datasets there are many semantic databases which are linked together. Therefore, we can use these databases for answering more complex queries than using traditional keyword-based search engines. In this case, the easiest way is to ask for information in natural language. Remember, how many times you have written a sentence like "How to do something". This type of query is now rare on the Web.

There already has been some research done in the field of querying data using natural language in classical databases [1] and semantic databases [2, 3].

Pre-processing is a key part of the natural language interface. In our approach, w e scan the whole dataset and create two lexicons: classes and properties lexicon, and values lexicon.

The first lexicon is based on structural part of our dataset. It consists of the names, labels etc. of all classes and properties. Next, all structural parts are decorated with synonyms from WordNet, which allows us to formulate query using different words than the ones used in the dataset. We call these alternative names *descriptors* and we provide ranking based on their source.

The second lexicon, using of which is completely new in our approach and none of the examined methods uses it, consists of property values that were obtained during the pre-processing phase. When the user types a value to his query, this lexicon can navigate us to an object type, which contains this value.

One of the main points of our method is processing of transformation to onto-dictionary which is shown in Figure 1. In this part we use the preprocessed lexicons for transformations. Then we use transformation rules to convert a user request to

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

SPARQL language. In this process we identify modifiers in the user query and add them to the SPARQL request.

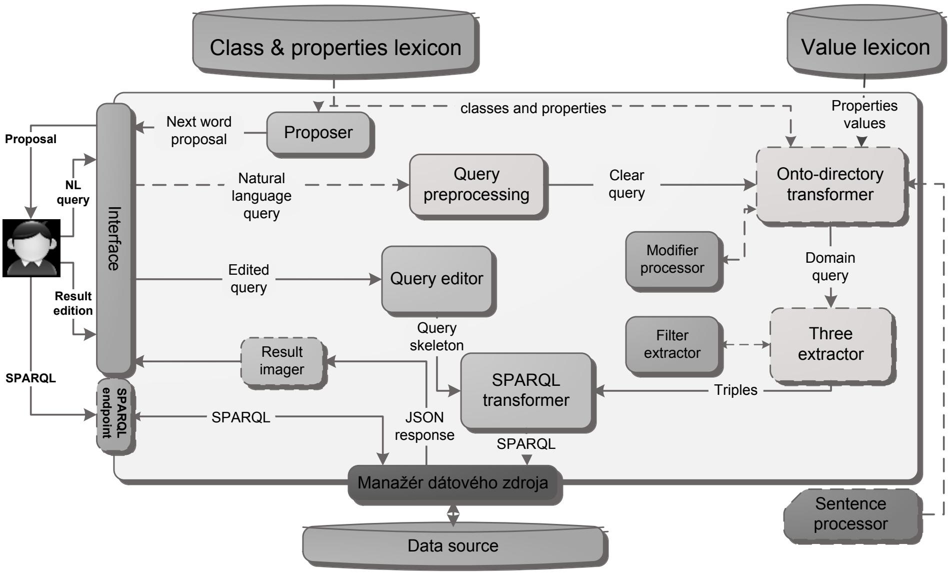


Figure 1. Query processor schema.

Next, we plan to evaluate our method with experiment using Annota Firefox extension. We will enhance the ACM digital library web site with our own search engine. We fill our dataset with data produced by Annota which currently has metadata from various digital libraries (ACM, IEEE, etc.) and store them in an ontological dataset.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 131-136.

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

References

[1] Owda, M., Bandar, Z., Crockett, K.: Conversation-Based Natural Language Interface to Relational Databases. In: *Proc. of 2007 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology - Workshops*. IEEE Computer Society (2007), pp. 363-367.

[2] Valencia-García, R. et al.: OWLPath: An OWL Ontology-Guided Query Editor. In: *Systems, Man and Cybernetics, Part A: Systems and Humans*. IEEE Systems, Man, and Cybernetics Society (2011), pp. 121-136.

[3] Wang, C., Xiong, M., Qi Z., Yong Y.: PANTO: A Portable Natural Language Interface to Ontologies. In: *4TH ESWC, INNSBRUCK*. Innsbruck, Springer-Verlag (2007), pp. 473-487.

Discovering and Predicting Human Behaviour Patterns

Štefan MITRÍK*

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

The fast development of advanced mobile technologies opens up new possibilities for analysis of humans' behaviour. Location-acquisition technologies such as GPS in combination with intelligent mobile applications allow us to collect huge spatio-temporal datasets of human mobility. These datasets contain trajectories that are performed by individuals during the day. Each trajectory is determined by sequence of visited geographical points and corresponding timestamps.

These datasets give us the opportunity to discover movement behavior and form users' behavior patterns. Each pattern consists of visited locations and routes among them. It also contains time and distance annotations that describe users' behavior in the more detailed manner. A pattern location is enriched by additional semantics information that is acquired from *Foursquare Venues API.*

The behavior patterns are naturally being performed in repetitive manner. We utilize this to predict the actions of the users in the future. The ability to predict users' actions is crucial in fields such as physical activity recommendation, where we need to recommend the activities in advance so the users can adjust their schedules and plans.

We utilize existing pattern mining techniques [2,3,4] and introduce their enhancements. Even though people naturally repeat similar behavior patterns over and over again, the humans' behavior changes over the time. It can be caused by different year season or change of the timetable at the university. Our method takes that into the consideration and uses *time degradation* parameter that determines how quickly method adapts to the recent behavior of the user.

Our solution is integrated into the fitness app called Fitly *(formerly known as Move2Play)* [1]. Discovering and predicting human behavior can be used as a basis for physical activity recommendation. Figure 1 depicts system overview of our prototype. It consists of client mobile application that is responsible for activity tracking, pattern construction and prediction. The activity recommendation module identifies transitions that are suitable for physical activity. Server side is responsible for collaborative

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 21–22.

features that help overcome problems with cold start so the user can receive meaningful recommendations from the beginning.

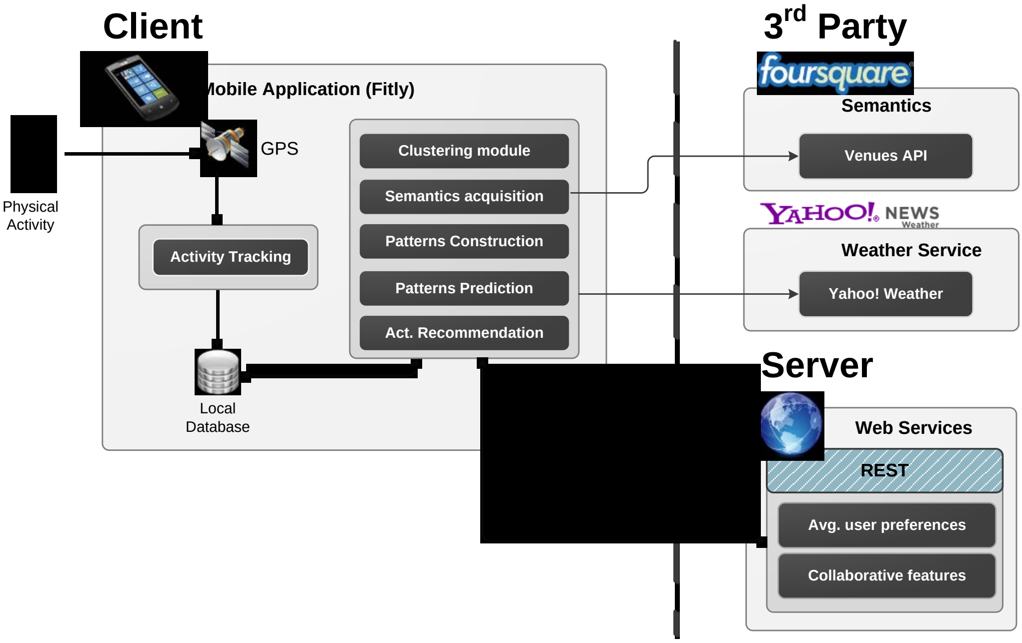


Figure 1. System overview.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 43-48.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

[1] 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 Int. Health Informatics Symposium (IHI '12)*. ACM, New York, USA, (2012) pp. 61-70.

[2] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Mining sequences with temporal annotations. *In Proc. of the 2006 ACM symposium on Applied computing (SAC '06).* ACM, New York, NY, USA, (2006) pp. 593-597.

[3] Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. *In Proc. of the 13th ACM SIGKDD international conf. on Knowledge discovery and data mining (KDD '07)*. ACM, New York, NY, USA, (2007), pp. 330-339.

[4] Pei, J., Han J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth. *In Proc. of the 17th Int. Conf. on Data Engineering.* IEEE CS, Washington, DC, USA, (2011) pp. 215-224.

Recommendation Based on Difficulty Ratings

Matej NOGA*

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

Our main goal is to design and create a recommendation method that will make recommendations based on the actual user knowledge and his/her rating of learning objects difficulty. We try to recommend examples, solving which would present a challenge to the student and he/she would rate the examples with our predicted rating. This kind of example has the greatest value for the student, because it is adequate to his/her knowledge and he/she can learn a lot of from it. The example should not be too easy, otherwise the student might start getting bored and he/she would acquire no knowledge from it. Then again, the example should not be too difficult, because it could discourage the student from continuing with the learning process [1].

Our method is designed for application in systems, which have same domain model as ALEF [2]. ALEF has a domain model split into two parts [1]: learning objects and metadata about them. Individual concepts in metadata part are matched with the learning objects by weighted relationships. The weight shows the significance of the concept for the learning objects. On the other hand, user model [3] contains the level of mastering individual concepts by the given user. Due to these data we can use our method in ALEF system.

Our method utilizes the level of user's knowledge, in other words – the level of mastering the individual concepts. Then, it looks up exercises with similar weight value of the selected concepts and selects the most suitable exercises. For these exercises, a group of users who are able to master the selected concepts and are similar to the current user is subsequently looked up. These users have already evaluated the selected exercise. Next, it is identified whether the users rated the example as too difficult or too easy, neither of which is desired in order for the example to be recommended. We select the most appropriate examples from the examples that met these requirements.

The method will be incorporated in RECO [4] recommendation system and interconnected with the ALEF educational system. It will provide recommendations to

[*] Supervisor: Martin Labaj, Institute of Informatics and Software Engineering

ALEF users. RECO system offers many additional services through which we can further improve our method later on.

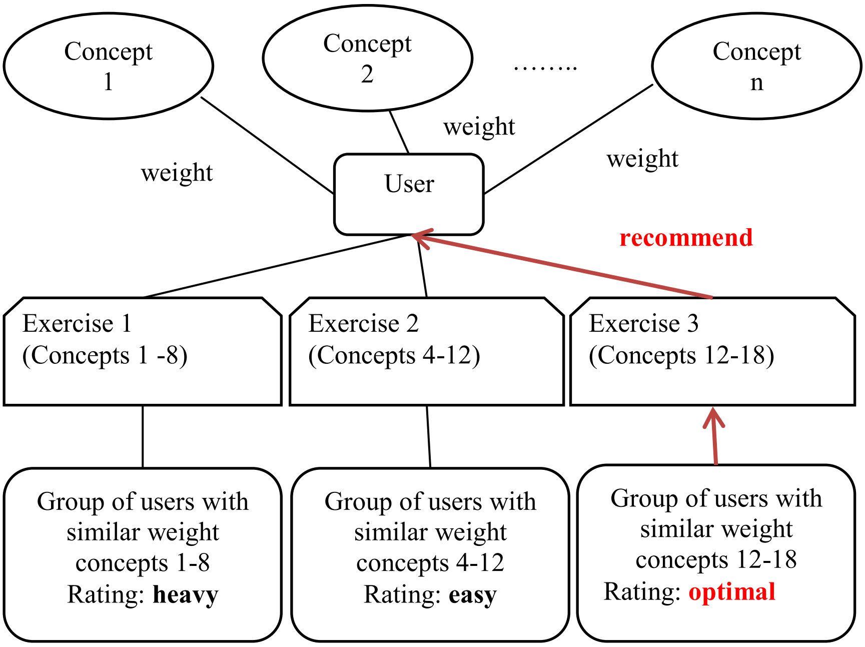

Concept 1
Concept 2
……..
Concept n
weight
weight
weight
User
recommend
Exercise 1 (Concepts 1 -8)
Exercise 2 (Concepts 4-12)
Exercise 3 (Concepts 12-18)
Group of users with similar weight concepts 1-8 Rating: heavy
Group of users with similar weight concepts 4-12 Rating: easy
Group of users with similar weight concepts 12-18 Rating: optimal

Figure 1. Overview of the proposed method.

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

References

[1] Michlík, P., Bieliková, M. Exercises Recommending for Limited Time Learning. In *Procedia Computer Science*. Vol. 1, Issue 2, Elsevier, ISSN 1877-0509, pp. 2821–2828 (2010)

[2] Bielikova, M., Šimko, M., Barla, M., Chuda, D., Michlik, P., Labaj, M., Mihal, V., Unčik, M. ALEF: Web 2.0 Principles in Learning and Collaboration. In *Proc. of the 6th E-learning Conference: E-learning and the Knowledge Society*. Riga: Riga Technical University, pp. 54–59 (2010)

[3] Unčík, M. *Modelovanie používateľa v doméne webovo-orientovanej podpory vzdelávania*. Diplomová práca (2012)

[4] Kompan, M., Ševcech, J., Bieliková, M. RECO – Experimental Personalized Recommendation Framework. In *Proc. of IADIS Int. Conference WWW/ INTERNET 2012*. IADIS Press, pp. 117–124 (2012)

Personalised Search in Source Code

Richard SÁMELA*

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

The Web is a very huge information channel, which provide a lot of text based web pages, pictures, videos or sounds. For finding information about these types of content, users use web search engines. Parts of these users are programmers, which search for content related to software development. This content can be in the form of source code. The Web contains many free source code repositories as well as private repositories accessible via web interfaces.

Naturally, with webification of software development, programmers have more choices where to find inspiration, advice or some other solution of their development problems. Controversially, as a result, when programmers are trying to help themselves via Web-based resources, it might take a lot of time. This spent time can be influenced by quality of personalisation. Having more information about a programmer, means better search results and less time spent by this programmer by searching[1].

In our project, we would like to collect as much information about programmer as we can. In order to create high quality user model, we need to resolve, which fact about programmers are relevant for us.

Every piece of source code is referring to a programmer, who wrote it down, which is a valuable source of knowledge about this programmer. Except of getting source codes of a programmer, we can identify technologies, which a programmer has learnt and worked with. Next, we can evaluate knowledge score from specific domain model for all of the programmers [2].

Programmer's user model should contain these attributes:

- ranked experiences
- queries posted to search engines related to software development along with ranked search results
- programmer's activity within development environment
- programming languages which a programmer is able to understand

[*] Supervisor: Eduard Kuric, Institute of Informatics and Software Engineering

In this project, we will analyse various methods for creating a user model of a programmer, which model knowledge, experience and abilities to write a source code in a programming languages, well-known by the programmer. Next, we will analyse various approaches for personalisation within a source code, based on user model principles. We propose a method for creating programmer's user model automatically, for the purpose of a personalised search in the source code.

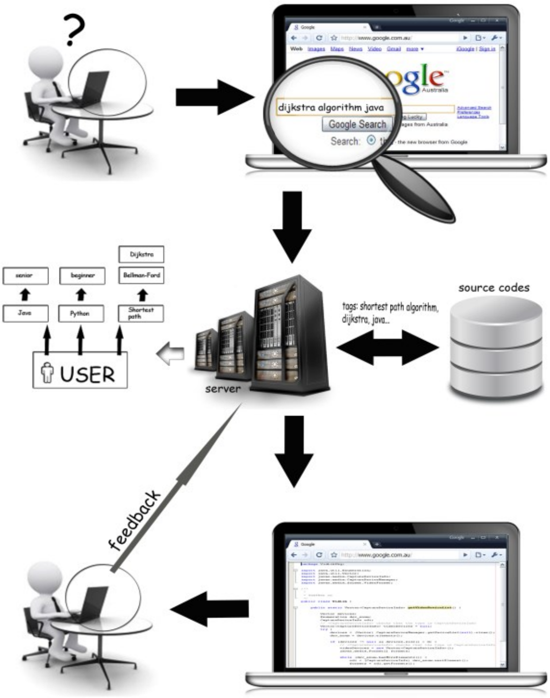


Figure 1. Process of obtaining user information.

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.

References

[1] Zigoris, P., Zhang, Y.: Bayesian adaptive user profiling with explicit & implicit feedback. In: *Proc. of the 15th ACM int. conf. on Information and knowledge management*. ACM, New York, USA, 2006, pp. 397-404.

[2] Bauer, T., Leake, D., B.: Real time user context modeling for information retrieval agents. In: *Proc. of the 10th int. conf. on Information and knowledge management*. ACM, New York, USA, 2001, pp. 568-570.

Context-Aware Recommender Systems Evaluation

Juraj VIŠŇOVSKÝ*

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

The amount of data provided by information systems constantly increases and we need to offer a filtered set of items personalized to user's needs. Recommender systems suggest a set of items that a user might be interested in or might find them useful. Basically, accomplishing recommendation task consists of two steps. At first recommender system has to collect information about user's activities and construct user model, which represents his preferences. Second step is to apply an algorithm thar uses the user's model to generate a set of items to suggest.

However, we may improve the quality of the set of suggested items by including information describing user's environment or state in the user model. The value of the information for improving recommendation may vary depending on the domain of proposed recommender and type of the information (e.g. user's wealth may not be relevant while recommending a movie to watch, however it is relevant while suggesting an item to buy in an e-commerce system). Including context in the process of recommendation matters, because there is a correlation between user's behaviour in certain situations and contexts as Riboni et al. proved in [2].

The question now arises is how to evaluate the quality of suggestions provided by context-aware recommender system? As we mentioned before, context-aware recommenders consider various context information and include them in the user models. On one hand this allows recommenders to generate highly specific suggestions. On the other hand it makes the evaluation of the recommendations slightly more complicated.

The simplest thing to do, in the process of context-aware recommender evaluation, is to evaluate every recommendation only if all environment conditions given by recommendation are matched with real contexts. This approach is precise, however may be, and usually is, very costly (e.g. In the middle of the summer we want to evaluate recommendations suggesting user's actions in snowy weather.).

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

The problem of costly evaluation may be solved by using so-called hypothetical situation [1]. In this case the evaluation process does not depend on the real contexts and thus is far more effective. The principle of this evaluation method is simply to ask a set of users how would they act in a given, hypothetical, situation. The basic premise is that some users may be more open-minded than others and their potential should be developed, as they may be used to evaluate recommendations dedicated to other users as well. On the other hand we must bear in mind the less open-minded users too.

Our first goal is to classify evaluating users depending on their answers to classification questions and in this manner we create their user model. Thanks to knowledge stored in the user models, we decide how they would be used in the process of context-aware recommender system evaluation. Depending on the user models we generate evaluation questions, which are presented to a proper user in a proper manner. By proper user we mean a user able to handle a set of recommendations to evaluate (e.g. an empathic user is more capable of evaluation recommendations dedicated for other users, however an apathetic user is more likely used only to evaluate recommendations dedicated to himself). By proper manner we mean personalized stylization of evaluation questions. If we reveal our purpose some users may be negatively influenced, because they tend to think they would act differently in a supposed situation. However this may be untrue. On the other hand for some users, the knowledge of our purpose might be helpful and therefore we should describe them supposed situation in as many details as possible.

To evaluate our method we use a set of movie ratings with context information (e.g. user's mood before and after watching a movie, his physical condition, etc.). At first we generate recommendations based on training set of ratings from this dataset. To achieve this, we prepare a simple context-aware recommender. This recommender system consists of the most used recommendation algorithms. Then we apply our evaluation method to generate evaluation questions for a collection of users. Finally we compare answers gained in the process of evaluation with action took by real users in the past.

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

References

[1] Ono, C., Takishima, Y., Motomura, Y., Asoh, H.: Context-aware preference model based on a study of difference between real and supposed situation data. In *User Modeling, Adaptation, and Personalization '09*. Heidelberg, Germany, Springer-Verlag, 2009, pp. 102-113.

[2] Riboni, D., Bettini, C.: COSAR: hybrid reasoning for context-aware activity recognition. In *Personal and Ubiquitous Computing*. London, UK, Springer-Verlag, 2011, vol. 15, no. 3, pp. 271-289.

Concept Location Based on Programmer's Activity

Pavol ZBELL*

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

Search in source code is a necessary part of the daily work of most programmers. Programmers often search and explore source code to enrich their existing knowledge of the workings and functionality of a software system, or to get answers to questions about software evolution tasks they are currently working on, or they search for source code fragments that they could reuse. Research in this area focuses mainly on concept location – a process of identifying an initial location in the source code that implements functionality in a software system. Existing techniques were recently summarized by Dit, B. et al. [1].

Built-in search tools available in IDEs which programmers use usually return a list of functions relevant to specified queries and programmers try to locate desired concept by jumping through functions based on the function calls they see. This approach is generally considered ineffective as it is usually resource and time consuming.

In our work we focus on searching the source code in terms of concept location. We propose a method which takes changes over time to fine-grained elements (functions) of the source code into account. We assume that changes made at a particular time are related, and may represent a concept of the software system. Our project is part of the research project PerConIK (Personalized Conveying of Information and Knowledge, perconik.fiit.stuba.sk) [2].

Our approach to concept location is illustrated in Figure 1. First we watch how programmers modify the source code in IDE and how they contribute these modifications to the RCS system. Then the document creator takes activity logs containing sequences of changes over time made to the source code elements and maps them to corresponding commits obtained from the RCS system. The document creator then produces documents representing source code elements in particular time. The documents containing commit metadata, the original source code and the nearest comments are indexed by the search engine. As long as searching in the source code is

[*] Supervisors: Mária Bieliková, Eduard Kuric, Institute of Informatics and Software Engineering

not the same as searching in plain text, proper indexing is achieved by extracting identifiers from the source code and tokenizing them using the INTT (Identifier Name Tokenization Tool) which is further described in [3]. Above mentioned steps are continually repeated during programming.

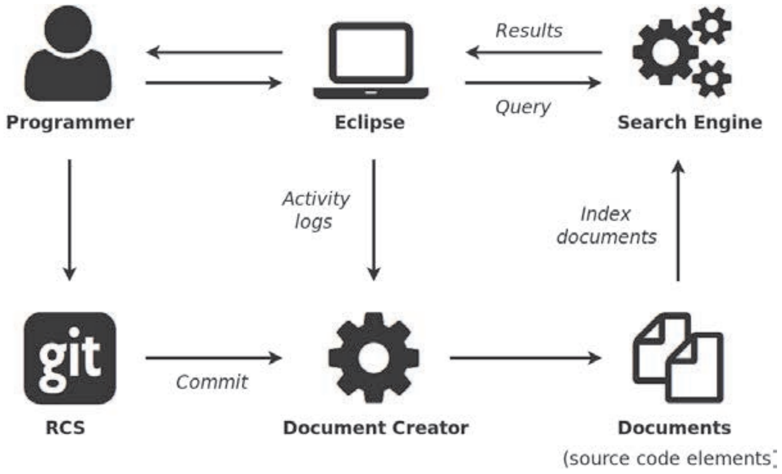


Figure 1. Proposed approach to concept location diagram.

Programmers are able to search in and explore the documents directly from the IDE’s interface. The performed source code search is based on simple vector space model, TF-IDF weighting scheme and cosine similarity between terms from programmer's query and terms extracted from the documents – source code elements (identifiers and comments).

We believe that developers using our method will gain a better tool for exploring the source code, and a better overview of the source code evolution and concepts of the software system. They will not need to jump between function calls because they will be able to see the whole evolution of the desired concept.

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.

References

[1] Dit, B. et al.: Feature location in source code: a taxonomy and survey. *In Journal of Software Maintenance and Evolution: Research and Practice*, 2011.

[2] Bieliková, M., Návrat, P., Chudá, D., Polášek, I., Barla, M., Tvarožek, J., Tvarožek, M.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. *In Procedia Technology*, 3rd World Conference on Information Technology (2013).

[3] Butler, S. et al.: Improving the Tokenisation of Identifier Names. *In Proceedings of the 25th European conf. on Object-oriented programming*, 2011, pp. 130–154.

Personalized Navigation

Personalized Web Documents Organization through Facet Tree

Roman BURGER*

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

With vast amount of accessible and relevant information and resources through the Web, one may start to seek for effective archiving and organization of resources. Most existing solutions though support only very specific use case scenarios and are not generalizable to the broad public.

There has been extensive research done such as [1] in identifying main strategies commonly used in personal information management. Most of the works share the idea of a spectrum, where on one side are strategies that rely on almost context free archiving. This strategy tends to be easier to use at first (with reasonable library sizes), but can radically impact effectiveness in large libraries. Resources tend to be harder to find and it is not uncommon to lose resources (completely forgetting about it).

On the other side of a spectrum are strategies that rely on punctual structuring of the personal library. Advantages of fully structured library are in better transparency. In addition, this strategy is less prone to errors and resources losing. An obvious drawback is that it is harder to create functional state of library and more time is required for maintaining the library. One of the latest researches [2] identified three basic strategies (or roles) that most users can be mapped onto:

1. piling strategy,
2. filling strategy,
3. structuring strategy.

Piling strategy is located on the context-free side of the spectrum and structuring strategy is on the context-full side. Filling strategy is located somewhere in the middle of the spectrum. Filling strategy is though not about using average amount of context to describe resources. Filling strategy is more about a combination. Some parts of the personal library are in context-free zone, having stacks or piles of resources that a user wants to dig in later (or never). Other parts of the library are reasonably structured,

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

giving the user an option to fill in new resources that are in great importance to the user.

We propose our organization method based on facet filtering. Facet filtering allows us to construct various views on the same subject (our personal resources library). In our domain it means constructing different context views, specifying particular collections of resources.

Users normally work with state-full personal organization structures (i.e., a structure maintains its internal state until explicitly updated). This is in contrast with typical facets methods, which usually look upon facets as querying framework. Therefore, we utilize in our design a new concept of facet tree originally proposed in [3]. Facet tree maintains its state and can be easily dynamically adjusted. Individual facets in chained facet tree can be removed or added creating context views on demand (and can still perform as a search tool). The prototype of facet tree interface is shown in Figure 1. The example shows chained facets *Color* and *Author* and the respective dynamically generated hierarchical tree.

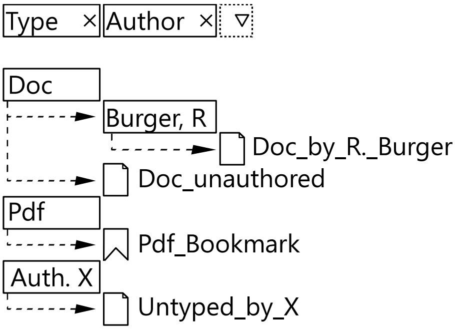


Figure 1. Facet tree with chained facets Color and Author.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 125-130.

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

References

[1] Abrams, D., Baecker, R., Chignell, M.: Information archiving with bookmarks: Personal Web Space Construction and Organization. In *Proc. of the SIGCHI conf. on Human factors in computing systems, CHI'98*, ACM Press, pp. 41-48 (1998).

[2] Henderson, S.: Personal document management strategies. In *Proc. of the 10th Int. Conf. NZ Chapter of the ACM's SIG on Human-Computer Interaction, CHINZ'09*, 2009, pp. 69-76 (2009).

[3] Weiland, M., Dachselt, R.: Facet folders: flexible filter hierarchies with faceted metadata. In *Proc. of the 26th annual CHI conf. extended abstracts on Human factors in computing systems, CHI'08*, ACM Press, p. 3735 (2008).

Trending Words in Navigation History for Term Cloud-based Navigation

Samuel MOLNÁR*

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

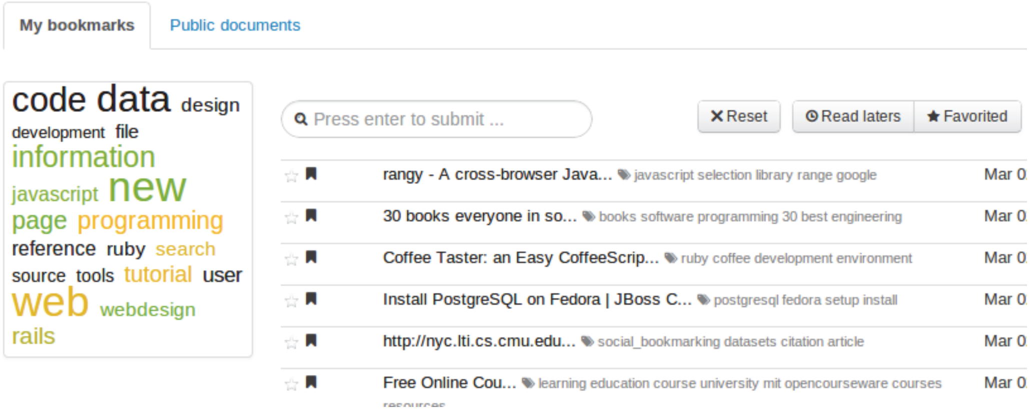
Nowadays, the amount of information available on the Web is making navigation difficult. Many approaches leave users to rely solely on result lists provided by the keyword-based search engines. Thus, over the past years several novel approaches were presented as an alternative solution to navigation support in search engines, such as tag clouds. Tag clouds focus mainly on exploiting different visual features of words like font size, colour or justification to emphasize their relevance. A tag cloud usually consists of keywords extracted from documents or user-added tags that represent documents on more abstract level. Tag clouds not only represent documents, but they also serve as a navigation tool. By employing visual features of tag clouds we aid user's navigation with the knowledge of how large is the information space behind the specific word or how the word is relevant to a user's current context. The navigation support provides the users a convenient way to refine their queries and discover new topics that are similar to their information need.

Gwizdka, et al. [1] introduced a novel method for tag cloud navigation by taking history into account. Their approach used pivot browsing, so in each step of navigation the content of a tag cloud is adapted to a current user's query. By highlighting co-occurrences of tags, the authors demonstrated coherence in navigation and similarities between the words in the user's query, but their enhancement was used only for purpose of visualization of user's current navigation history.

We proposed a method for term cloud navigation which exploits navigation history as a source of metadata for personalized browsing of information. By using this approach we utilize the users' interests in specific period of time to personalize their navigation in the domain.

In order to represent documents, we choose to use tags created as folksonomies and keywords extracted from documents which we both denote as terms. The relevancy of a term is determined according to the number of times the term occurs in documents specified by the query (the sequence of words selected by user).

[*] Supervisors: Mária Bieliková, Róbert Móro, Institute of Informatics and Software Engineering


My bookmarks
Public documents
code data design development file information javascript new page programming reference ruby search source tools tutorial user web webdesign rails
Press enter to submit ...
✕ Reset
Read laters
★ Favorited

rangy - A cross-browser Java...	javascript selection library range google	Mar 0
30 books everyone in so...	books software programming 30 best engineering	Mar 0
Coffee Taster: an Easy CoffeeScrip...	ruby coffee development environment	Mar 0
Install PostgreSQL on Fedora | JBoss C...	postgresql fedora setup install	Mar 0
http://nyc.lti.cs.cmu.edu...	social_bookmarking datasets citation article	Mar 0
Free Online Cou...	learning education course university mit opencourseware courses resources	Mar 0

Figure 1. Prototype of our cloud navigation in Annota.

History records from a specific period of time are used for adapting the content of a cloud by choosing similar queries that the user navigated within the particular period of time. By exploiting user’s history, we provide query refinement that helps users to customize their last query with words already used in the specific period of time. Our approach for visualization of history in a cloud exploits the time of the last usage of the words from cloud in user's queries by using different color according to last time of usage as shown in Figure 1.

Our contribution towards cloud navigation is in exploitation of history in a period of time to provide personalized content of term cloud with color-based visualization of history. We extend the content of term cloud with words that are similar to current user's query by exploiting queries from history containing the words from current query. The different shades of colors distinguish last usage of word in history relative to the current query.

We evaluate our approach in the domain of digital libraries. We implemented our proposal as a module into a system for web page annotating – Annota [2], which is being developed by several PeWe group members (annota.fiit.stuba.sk).

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 107–112.

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

References

[1] Gwizdka, J. et al. Tag Trails: Navigation with Context and History. In *Design*, vol. 69, no. 2, pp. 4579-4580 (2009).

[2] Ševcech, J., Bieliková, M., Burger, R., Barla, M.: Logging activity of researchers in digital library enhanced by annotations. In *Proc. of 7th Workshop on Intelligent and Knowledge oriented Technologies*, pp. 197-200 (2012). (in Slovak)

Exploratory Search Using Automatic Text Summaries

Róbert MÓRO*

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

Nowadays, keyword search is a prevalent search paradigm on the Web. This approach works reasonably well for simple information retrieval tasks such as fact finding. Selecting the relevant links and navigating among the documents can, however, be an uneasy task if the information seeking problem at hand is more complex and requires exploring multiple sources to find relevant information, such as researching a new domain. For these types of searches the term *exploratory search* was coined by Marchionini [1]. Exploratory search is characterized by information seeking problems which are open-ended and start with ill-defined information needs; and by processes which can span over multiple search session and require employing different search strategies [4].

In order to support the exploration of the domain by the users, we propose a method of navigation using automatic text summaries. The summaries consist of sentences conveying the most important information of the document; they can reduce information overload by helping the users to decide whether the document is relevant for them and they should read the whole text or not. For the purpose of summary navigation we applied the method of summarization proposed in [2] utilizing the term frequency rater (using tf-idf) and the location rater which prefers terms in the first and the last sentence of the document. We do not summarize the whole contents of the papers from the ACM Digital Library, only their abstracts, thus constructing short generic summaries indicating what the main idea of the paper is.

The actual method of navigation is based on enriching the individual document summaries with the navigation leads; we enrich the summaries in two ways. Firstly, users can manually select a word or a phrase in a summary; popup appears with the possibility to filter the search results by the current selection. After that an explicit link is created; it is visualized in the summary the next time the user comes across the document and can be used in the future searches to filter the results of the user's query.

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 37–38.

Moreover, we exploit social relations and recommend potentially valuable navigation leads followed by other users based on the collaborative filtering approach. The user can follow the recommended navigation lead, or reject it, thus providing us with an implicit (in the first case), or explicit feedback (in the second case) which we use in computing similarity between users as well as in evaluating candidate lead words.

We realized our proposed method of navigation in the bookmarking system Annota [3] on the dataset consisting of 566 scientific papers that have been bookmarked by the users at the ACM Digital Library over the period of five months.

In order to find out what words users choose to navigate and explore their differences, we conducted a qualitative experiment with five participants. We described them three situations that motivated their search and presented them with the corresponding search results for the initial query. Their task was to select the words they deemed useful for the further exploration of the topic. We verified the hypothesis that the users would select similar words from the summaries if their work task context was similar and yet could benefit from recommending the leads followed by others.

In the follow-up questionnaire the participants independently agreed that the main advantage of our approach is the easier and smoother navigation which requires less cognitive load, because they can see the words in their context and instantly refine their query to explore the lead.

We plan to extend our approach by recommending the whole navigation trails using the leads in the summaries that will support different exploratory search strategies. Because it is crucial what information is selected into the document summaries we will experiment with our proposed method of summarization to achieve optimal settings of raters utilizing additional information about documents such as the keywords or highlights added by the users.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 163–169.

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

References

[1] Marchionini, G.: Exploratory search: from finding to understanding. *Communications of the ACM*, vol. 49, no. 4, pp. 41-46 (2006).

[2] Móro, R., Bieliková, M.: Personalized text summarization based on important terms identification. In *DEXA '12: Proc. of the 23rd Intl. Workshop on Database and Expert Systems Applications*, IEEE CS Press, pp. 131-135 (2012).

[3] Ševcech, J., Bieliková, M., Burger, R., Barla, M.: Researcher activity tracking in digital library of scientific resources enriched by annotations (in Slovak). In: *WIKT '12: Proc. of 7th Workshop on Intelligent and Knowledge Oriented Technologies*, STU Press, pp. 197-200 (2012).

[4] White, R.W., Roth, R.A.: Exploratory search: beyond the query-response paradigm, Morgan & Claypool, (2009).

Web Navigation Based on Annotations

Jakub ŠEVCECH*

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

We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Web or when reading electronic documents. We use these annotations to highlight important parts of documents and to mark our thoughts in the margin of the document. User created annotations are commonly used to support navigation, e.g., for text summarization [2] or search [1]. We proposed a method for searching related documents to currently studied document using annotations created by the document reader. We perceive annotations as indicators of user's interest in particular parts of the document.

The proposed method uses text to graph transformation to preserve document content and its structure and spreading activation algorithm to identify most important words in the text of studied document. The text to graph transformation step transforms words from the document to graph nodes and creates edges using word neighbourhood in the source document. We use annotations highlighting parts of the document as well as annotations attaching additional information to insert initial activation into the document graph. Annotations highlighting parts of the document insert activation to nodes representing highlighted words and annotations inserting additional content extend document graph by new nodes and edges and introduce activation to these nodes. The initial activation is spreading from nodes with attached annotations and concentrates in most important words. The proposed method extracts words, which are important for annotated parts of the document, but it also extracts globally important words that are important for the document as a whole. The portion of locally and globally important words can be controlled by number of iterations of the algorithm. We use these words as a query in retrieval of related documents. We determined the right number of iterations and the right amount of activation introduced by various types of annotations into the graph using simulation based on user created annotations.

To evaluate proposed method we created a service called Annota (annota.fiit.stuba.sk), which allows users to insert various types of annotations into web pages and PDF documents displayed in the web browser. We analysed properties of various types of annotations inserted by users of Annota into documents and we

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

derived probabilistic distributions of annotation attributes such as note length, number of highlights per user and per document or probability of comment to be attached to text selection. Figure 1 presents an example of derived distribution of number of highlighted texts per document that follows logarithmic distribution.

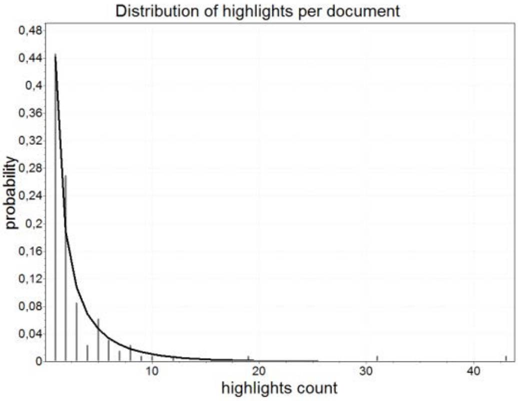


Figure 1. Logarithmic distribution of highlighted texts number per document.

Using these distributions we created a simulation to find optimal weights for various types of annotations and number of iterations of the proposed method, where we optimized query construction for document search precision. We used dataset created from Wikipedia pages to create index of documents. We created query from source document and annotations generated using parameter distributions. We compared relevancy of documents retrieved when searching within created index using this query and when searching using query created by TF-IDF-based method.

The proposed method outperforms compared method when generated annotations were used as user's interest indicators and it received better results when no annotations were used and when whole document sections were annotated. Proposed method outperformed compared method even though it is not using information about other documents from a collection unlike the TF-IDF-based method.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 143–148.

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

References

[1] Golovchinsky, G., Price, M. N., Schilit, B. N.: From reading to retrieval: freeform ink annotations as queries. *SIGCHI Bulletin*. ACM, pp. 19–25 (1999).

[2] Moro, R., Bieliková, M.: Personalized Text Summarization Based on Important Terms Identification. *23rd Int. Workshop on Database and Expert Systems Applications*, IEEE, pp. 131–135 (2012).

Browsing Information Tags Space

Andrea ŠTEŇOVÁ*

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

Software systems process information whose amount are growing exponentially. We found out, that storing the data is not enough, because we should be able to organize and simply and effectively browse stored information. Therefore, we use metadata that can describe structure of data and give us additional information about users that created them and how they were created.

Adding metadata to our content has various advantages. Their importance lies in ability to store, organize and provide information about the data, to that they are assigned. Metadata can also allow a better understanding of the data and display the data changes over the time. They provide us information about users, who read these data, work with them or create them.

One type of metadata is information tags that contain structured information associated with a particular piece of content, such as a number of clicks on a link on the page, keywords characterizing a paragraph in an article or a number of lines of a method's source code.

Metadata are usually generated and processed by machines, and their amount and structure make it difficult for people to effectively read and understand them. There are various problems with comprehensibility of the data by humans. E.g., when there is too much of the data, we cannot display them at once, because they would be not easily legible. However, if we do not display all of the data, we need to ensure that users are able to get the information they are looking for. When users search in the data, there might be a problem with lack of knowledge of a particular dictionary. Or users do not have to know exactly what they are looking for and which sources are most relevant for them [1].

However, the metadata can contain valuable information about the content, as well as about the users working with it, it is important to enable the understanding and analysis of metadata by humans. To ensure readability and understandability we can use navigation of users through the metadata with the help of their visualization. Currently, it helps us to understand the huge amount of the data and allows us to more

[*] Supervisor: Karol Rástočný, Institute of Informatics and Software Engineering

easily navigate within them. Research methods are therefore dealing with how to handle metadata, visualize metadata and how to navigate users through them.

Navigation can be resolved by using graphs, namely using zoom techniques, displaying only the selected subgraph, clustering of graph nodes, or other techniques to clarify content without losing context [2]. Furthermore, we can use the tables to navigate users, faceted browsers, etc. Each of these techniques has its advantages and disadvantages, and its use is appropriate for solving various problems.

In our work we propose a method that helps users browsing information tags space. We focus on structured metadata about user activity, as well as the structure and content of the data. One of problems of metadata is their change in time. Only updated data, or their combination with their changes in time, can be presented to users.

Another problem is computational complexity of algorithms to display data. If we do not display to users all of the data and calculation of their representation is challenging at same the time, we need some way to predict users' actions to optimize the computational time.

In our solution we will use graph visualization to browse metadata. Graph nodes will be clustered according to their common characteristics. Users will be able to display the change of metadata over period of time, which should help them to understand metadata better. We will adapt the size of the displayed graph for users to support readability by humans. Since the algorithms necessary to display sub-graph are time consuming, we will predict users' actions, which will accelerate browsing.

Our solution must be understandable and easily readable by users. To support that, we will use some of the existing visualization solutions. We plan to verify our solution in domain of the project PerConIK [3]. Therefore, our solution will be focused on a user role researcher, who uses these metadata for his own research and looks for relationships between metadata and inconsistencies in them.

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.

References

[1] Katifori, A. et al.: Ontology visualization methods - a survey. In *ACM Computing Surveys*. 2007. Vol. 39, no. 4.

[2] Herman, I. et al: Graph visualization and navigation in information visualization: A survey. In *IEEE Transactions on Visualization and Computer Graphics*. 2000. Vol. 6, no. 1, pp. 24-43.

[3] Bieliková, M., et al.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. In: Procedia Technology, 3rd World Conf. on Inf. Tech., to appear.

User Modeling, Virtual Communities and Social Networks

User Interest Modelling Based on Microblog Data

Miroslav BIMBO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
bimbo08@student.fiit.stuba.sk

This paper is focused on extracting information about users who write microblog posts and creating their user model. User model can be used with advantage to recommend or filter content for a particular user, helping him to overcome information overload and allowing him to focus attention on relevant information resources.

In many works the interests of a particular user are extracted only from posts written by the user. Interesting way of improving the user model is microblog post enrichment, where interests are extracted from documents, which are related to original user posts. Abel et al. proposed method, where user interests are extracted from news articles, what produced fuller and richer user interest model [1]. Bernstein et al. proposed method for extracting topics of interests leveraging Yahoo search, achieving better results compared to extracting the topic from a post itself [2].

In our work we create the user model inspired by approaches [1, 2]. However, rather than focusing on one enrichment method, we build on the intuition, that we can create better user model by aggregating results from several different enrichment methods.

The proposed method for user model creation can be described as follows:

1. Classification: Classifier computes *interest relevance i* of posts. It is trained by a supervised machine learning algorithm. We create train set (pairs <post, interest relevance of post>) for classifier by manual annotating of posts, and train the classifier using several features of posts.
2. Enrichment: To enrich posts by external documents, we employ several methods: Hashtag (documents are microblog posts, which contain same hashtag as given post), Tagdef (documents are descriptions of hashtag found on Tagdef service[1]), URL (document is text of URL included in given post), News (document is the most similar news article, method proposed in [1]), Youtube (documents are descriptions of Youtube videos, retrieved after transformation of posts to queries).

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

[1] http://tagdef.com

For some of these methods, we are able to compute *confidence of relation c* between posts and documents.

3. Interest extraction: We extract interests (represented as semantic web entities) from given documents using the OpenCalais web service (web service returns *weight w*).
4. Weighting: We compute importance of each particular interest for a user as:

$$score(user,\ interest,\ post,\ method) = w^{\lambda_1} * c^{\lambda_2} * i^{\lambda_3}, \lambda_i \in \{0,1\}$$

5. Aggregation: In situation, when one interest is found in one post by more methods, or when one interest is found in more posts several aggregation methods, we proposed several aggregation methods.
6. Filtration: We filter out low score interests and mostly repeated repeating false positive interests.

For evaluation of our method, we used UMAP 2011 dataset[2]. We divided posts of one user into 5 equal groups. Four of these groups are used to create a model, last part is a test set – viewed as text representation of its posts. Then, we can compute precision (how many of interests from model have its text representation in the test set) and recall (how many of words from test set are found in model). We applied 5-fold cross validation, so the final results are averaged from the 5 partial results.

The preliminary results shown, that according to F1 measure the method aggregating Youtube, News and Tagdef method performs better than baseline method (i.e., no enrichment). This supports our hypothesis.

Our second result is that filtering out all interests found in documents with *confidence c* lower than some threshold can improve results. In contrast, we found that same filtering based on *weight w* is not useful.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 119–124.

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., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. In: *Proc. of the 8th extended semantic web conf. on The semanic web: research and applications*, Springer-Verlag, pp. 375-389 (2011).

[2] Bernstein, M. S., Suh, B., Hong, L., et al.: Eddi: interactive topic-based browsing of social status streams. In: *Proc. of the 23nd annual ACM symposium on User interface software and technology*, pp. 303-312 (2010).

[2] http://wis.ewi.tudelft.nl/umap2011/

Facial Expression Recognition for Semantic User Modeling

Máté FEJES *

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

Human computer interaction covers the methods of information exchange between man and computer. The interaction is typically used to obtain explicit user commands and/or to collect implicit feedback, which is the more problematic of the two. The observations of implicit actions may be ambiguous in that we try to guess what the user is thinking without actually knowing it explicitly. In this paper we explore detecting user's facial expressions/emotions – which ultimately serve as a vehicle for better user modeling – as one of the possible ways to obtain implicit information from the user beyond the scope of the typical human input devices allow.

Basic human emotions and their expressions are innate generic reactions, constituting an implicit way of communication. Implicit signals like tone of voice, gestures and facial expressions are applied in verbal communication and often have non-trivial power of expression, which can confirm, refute or totally alter the meaning of the verbal part of communication. Analogously in the task of information retrieval, informational need affects emotional need, and vice versa [2]. Our project is based on this influence of user's informational and emotional needs, ultimately aiming to enrich the user feedback with them.

In this paper we describe the stages of our research. We propose a method for recognizing facial expressions/emotions of a human subject based on a sequence of images (frames) of the subject's face. In order for the recognition method to provide feature rich input for subsequent machine learning-based method of user modeling, we recognize lower level facial features that can be effectively used to build up the higher level emotions. Most existing recognition systems consider the discrete representation of the six basic emotions: joy, sadness, anger, disgust, surprise and fear [3], while others represent the extracted information in two-dimensional space (positive - negative and active - passive). Our experiments have shown that the facial expression of these basic emotions can be more complex; consequently we decided to recognize facial features with lower granularity. The output of our method is a set of small atomic

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

movements of the facial muscles – so called Action Units [1] – which are, or are not present in the input image. Due to their physical nature, the complex facial expressions consist of the simple movements. Using this representation we obtain a more accurate description of user's emotional state. Our approach is based on several similar implementations [3, 4] which are realized using Support Vector Machine (SVM) learning.

In the final stage, we propose a user modeling method that uses the emotional states for user/student modeling in a personalized information system, which, in our case, is an online web-based learning environment used by hundreds of users in teaching of programming. Our method determines the relations between the emotion recognition output and user activities within the information system. The ultimate goal is to anticipate user's (student's) immediate action based on previous activities and emotional state.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 25-30.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

[1] Ekman, P., Matsumoto, D. R., Friesen, W.V.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). New York. Oxford University Press. 1997.

[2] Moshfeghi, Y.: Role of emotion in information retrieval. PhD thesis. University of Glasgow, 2012.

[3] Kotsia, I., Pitas, I.: Real time facial expression recognition from image sequences using support vector machines. IEEE International Conference on In ICIP 2005. 2005, Vol. 2 (2005). pp. 966-969.

[4] Shan, C., Gong, S., McOwan, P. W.: Facial expression recognition based on Local Binary Patterns: A comprehensive study. Pattern Recognition and Image Analysis. 2009. Vol. 17. pp. 592-598.

User Modeling for Facilitating Learning on the Web

Martin GREGOR[*]

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

History of e-learning begins in 1800 under the name distance learning in process of teaching villagers which had no opportunity to go to school. The evolution of technologies and especially information technologies pushed the envelope of distance learning to the form of e-learning and web learning as we know it now. Web learning brings a lot of new opportunities how to enhance learning, but also we have to realize that there are many problems with adaptation of learners and teachers to new trends. The main advantage is asynchronous learning that allows a learner to learn anytime and anywhere and also a teacher to teach in the same way. But that brings also disadvantages. For example, complicated scheduling of learner's life, because variable learning time can replace user free time, which is so important. Also chance to cheating is higher than in a classroom teaching. According to mentioned pros and cons the web learning is suitable for working people or higher educating future professionals [1].

Everyday web browsing is the normal routine of each of us and through experiences that users have, it can be used as an effective way to support learning. For example, personalized text enrichment with the potential to improve access to information is an easy way to obtain new knowledge. In many cases, web learning is supported with an adaptive learning system. Adaptive learning system necessitates the knowledge about a user and his skills, characteristics, preferences and so on. The best way how to store this knowledge is usage of user model. We have options like individual or stereotype user model, we can visualize model like vector, network, key-set or ontology [2].

Human brain's forgetting property is considered in user modeling only occasionally. In [3], user model is split into a hierarchical structure of three models: Active model contains active knowledge about user, while archive model contains knowledge which user has not used for long time. Time decay of information moves information from higher level of hierarchical model to lower. The last level of hierarchical user model is deleted model. Knowledge in the last model is used for

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 49–50.

statistics and aggregation of user knowledge. The potential approach to support users in handling forgetting is to employ user model scrutability which allows a user to control own model. It also increases user confidence to web learning mechanism.

User skills, personal characteristics or preferences can be captured from obtained feedback from user web browsing behavior. We have several options how to capture user browsing behavior: using web browser extension, where we can detect tab usages, search on the page, usage of bookmarks and so on; employing proxy-based methods, and injecting JavaScript to every visited webpage; implementing own method in the stand-alone adaptive web system.

We can diversify feedback into implicit and explicit based on user action, and into internal and external based on source of information. Implicit feedback consists of every user action during web browsing like clicking, scrolling, copy&paste, searching on page and especially time spent on the page. Explicit feedback is requested from user in form of questions. Explicit feedback helps to assure correctness of information gained by implicit feedback. On the other hand, a difference between external and internal feedback is in a source of information of feedback. Internal feedback emerges in user cognitive processes after completion of learning task. External feedback comes from another external teaching medium which evaluates user performance and progress in learning process after completion of learning task [4].

Our aim is to gain a feedback from every possible action of the web browsing, transform this feedback to the knowledge about the user, store the knowledge to user model and finally enrich web pages user browses according to the knowledge calculated from the user model. We aim to consider forgetting and support user model scrutability. User modeling in adaptive learning web-based systems covering various domains with an emphasis on user behavior tracking on the web poses many research challenges to solve.

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

References

[1] Neal L., Miller, D. The basics of e-learning: an excerpt from handbook of human factors in web design. *eLearn,* Vol. 2005, Issue 8, p. 2 (2005)

[2] Brusilovsky, P., Anderson, J.ACT-R electronic bookshelf: An adaptive system for learning cognitive psychology on the Web. In *Proc. of WebNet'98, World Conf. of the WWW, Internet, and Intranet*. AACE. pp. 92–97 (1998)

[3] Barua, D, Kay, J., Kummerfeld, B., Parism, C. Theoretical foundations for user-controlled forgetting in scrutable long term user models. In *Proc. of the 23rd Australian Computer-Human Interaction Conf.* (OzCHI '11). ACM, New York, NY, USA, pp. 40–49 (2011)

[4] Narciss, S. Feedback strategies for interactive learning tasks. In J.M. Spector, M.D. Merrill, J. van Merriënboer, D.M. Driscoll (ed.), *Handbook of research on educational communications and technology*, Routledge, pp. 125–144 (2008)

Adaptive Feedback in Web Systems

Marek GRZNÁR*

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

One of the main processes in the adaptive web systems is the communication between the system and its users. User evaluates the presented information, whether it was useful or helpful. Based on this information, the adaptive system provides recommendations of other content useful for the user. The processes of feedback collection and result presentation used in the adaptive methods could also be adaptive themselves.

Users are often not willing to rate the learning object which they use. They provide ratings especially if they are motivated [1], or, alternatively, if they are very satisfied or very unsatisfied [2]. There is also another problem that the act of rating of an object may disturb the user during her interaction with the object.

In our research we combine implicit and explicit feedback to increase quantity and quality of ratings in an educational web system. We aim at difficulty ratings of learning objects. We monitor user behaviour by the means of implicit feedback during question/exercise solving. We monitor mouse cursor movements and keyboard actions.

We collect learning object ratings with explicit feedback. We use a difficulty scale with six values for this purpose.

We detect four situations (see Figure 1) when explicit feedback could possibly appear:

1. The user has finished working with the learning object – she has solved the problem and submitted a solution.
2. The user tries to solve the problem but loses interest in the exercise. E.g., she starts to scan the sidebar widgets.
3. The user does not know the correct answer and asks for a hint. The system then shows a request for learning object rating. This request also appears when she finally solves the problem.
4. The user starts solving the exercise, but after some time we do not detect any user activity. In order to check for user's presence, we ask her to rate the learning object.

[*] Supervisor: Martin Labaj, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 51-52.

We calculate the user inactivity time as average time needed to solve the exercise.

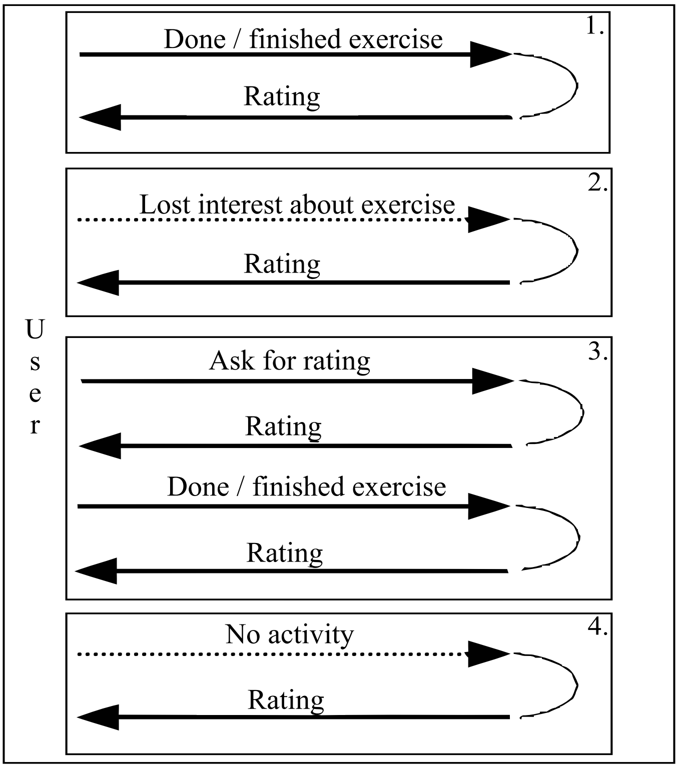


Figure 1. User interaction with system.

The proposed method will be evaluated in the ALEF educational system (Adaptive LEarning Framework) [3]. The collected feedback can be used to recommend most appropriate learning objects.

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

References

[1] Carenini, G., Smith, J., Poole, D. Towards more Conversational and Collaborative Recommender Systems. In *Proc. of the 8th Int. Conf. on Intelligent user interfaces*, IUI'03, pp. 12–18 (2003)

[2] Hu, N., Zhang, J., Pavlou, P. A. Overcoming the J-shaped Distribution of Product Reviews. In *Communications of the ACM – A View of Parallel Computing*, vol. 52, no. 10, p. 144 (2009)

[3] Šimko, M., Barla, M., Bieliková, M. ALEF: A Framework for Adaptive Web Web-based Learning 2.0. In Reynolds, N., Turcsányi-Szabó, M. (Eds.): *KCKS 2010, IFIP Advances in Information and Communication Technology*, Volume 324. Springer, pp. 367–378 (2010)

User Modelling Using Social and Game Principles

Peter KRÁTKY*

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

User modelling is an essential part of adaptive systems because the user model represents the data relevant for adapting/personalizing the user experience. In our project, we are interested in a generic type of user modelling based on personality traits. Our goal is to identify user's personality within information systems in general, and computer games in particular. Classic methods based on questionnaires hinder smooth user experience especially in games that should provide entertainment. In our research, we explore to what extent the personality-based user modelling can be conducted unobtrusively in computer games (information systems). Games are interaction rich artefacts, and identifying player's personality while he/she enjoys the game provides significant factor for further personalizing the gaming experience according to player's psyche. Games are different, and various game mechanics can work differently with different players' personality profiles [2, 3].

In order to study effects of player's personality on games in general we have designed a feature-rich causal browser game in which different game mechanics, functional components of a game (points, leaderboard, challenges, timer) [1] can be turned on/off based on the user experiment design. The game is tracking both the user interface actions and game actions, providing a complete footprint of user's personality in terms of the manifested game play. Correlating the activity logs with different personality measures (Big Five and Index of Learning Styles, in our case) reveals the relationships between player's personality and game play.

The core principle of the game is to collect reward by eliminating shapes placed in a grid using the mouse cursor (mouse move over the shape). The shapes appear on the grid in different sizes and different rewards. The whole game consists of several short levels lasting 30-60 seconds, which difficulty rises with its number. Other gaming features, which correspond to the game mechanics we are studying in our research, enrich this core game play. Points are awarded for eliminating shapes so that bigger shapes receive lower rewards compared to the smaller shapes. The leader board

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

displays current position of the player in comparison with few other players. Challenges game mechanic is in the form of objectives to fulfil, specifically, eliminating specific sequences of shapes. Skills game mechanic is integrated as characteristics that player can improve in (fast hits and sniper hits).

The aim of our experiment is to collect relevant data from game play, process and analyse it and evaluate model of predicting player's personality. We have designed several experimental groups each having game with different game mechanisms enabled. Logged raw data (e.g. position of a new shape in the grid, type, position of eliminated shape and duration in the grid) is aggregated into numeric indicators that are then correlated with the personality measures (Big Five and Index of Learning Styles) obtained by questionnaires. The first groups of indicators hold characteristics of mouse controller usage (average movement speed, percentage of time inactive, etc.). The second group of characteristics deals with game mechanisms (number of points, challenges, skills per minute, etc. and availability of mechanisms coded 0/1).

We have conducted our study in a web-based platform Peoplia, and have collected data on 65 players (university students), totaling 601 game levels played. We have examined how the Big Five factors influence interface indicators using correlation analysis and the results based on 12 users who reached at least Level 9 show high correlations of neuroticism with mouse speed (0.53), path efficiency (-0.56) and hits success (-0.53) and high correlations of agreeableness with inactivity (-0.62) at the significance level 0.10. We constructed a regression model for predicting traits based on interface interaction can be used to determine neuroticism and agreeableness with the coefficient of determination R^2 0.49, 0.40 and p-value 0.13, 0.03 respectively.

To summarize, in our work we seek possibilities to predict player's traits according to game log by exploring influence of personality traits of the player on his/her game play. We have designed a browser game which tracks both interface and game play characteristics. We found interesting correlations of interface actions and personality traits. We continue with analysis of game features influence on game play of players having different personalities.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 31-36.

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

References

[1] Hunicke, R., Leblanc, M., Zubek, R.: MDA: A formal approach to game design and game research. In: *Proceedings of the Challenges in Games AI Workshop*, Nineteenth National Conference of Artificial Intelligence. 2004, pp. 1-5.

[2] Khan, I., Brinkman, W.: Measuring personality from keyboard and mouse use. In *(ECCE '08)*. 2008. no. 1984.

[3] Lankveld, G. Van et al.: Games as personality profiling tools. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*. 2011. pp. 197-202.

Activity-Based Programmer's Knowledge Model for Personalized Search

Eduard KURIC*

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

Every day, a programmer needs to answer several questions for the purpose of finding solutions and making decisions. It requires the integration of different kinds of project (software system) information, as well as, it depends on the programmer's knowledge, experience, skills and inference.

The main advantage of search-driven development should be that programmers save time and resources by reusing (external) source code (components) in their software projects. To support search-driven development it is not sufficient to implement a "mere" full text search over a base of source code. In [2], authors investigated what functionality programmers would ideally like to have in a source code search engine. They included characteristics as: *review by other independent programmers – programmer's feedback*, *reviews/ranking for the source code*, *satisfaction level of other programmers*, and *number of programmers actively using it*.

When a programmer reuses source code he has to trust the work of external programmers that are unknown to him. It is possible to solve using a trustability metric so that the programmers assess the quality of source code search results. Existing approaches are often based on collaborative filtering of programmers' votes and project activity of programmers [1]. It would be helpful for programmers to see not only some activity statistics about software project (components) but also a karma value of each author (programmer). If a target programmer will easily see in the search results that an experienced programmer with good reputation has participated in writing the source code (component) then the target programmer will be more likely to think about reusing. Thus, author's reputation can provide information to support programmers' decisions.

Reputation ranking can be a plausible way to rank source code search results, i.e., if we determine programmers' karma values, we can prefer software components based on reputation of their authors. To model programmer's reputation (to calculate programmer's karma value), we need to investigate software components which he

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

created. We propose programmer's knowledge model and methods for its automatic retrieving. Currently, we focus on two factors, namely, calculation of programmer's know-how about used technologies and calculation of programmer's karma based on importance of components.

Our focus is to automatically find out, using source code, which technologies (libraries) programmers really use, when they used them and their depth (degree) of usage. Based on the result, we find out the programmer's know-how with the specific technology in the comparison with other programmers who used (use) this technology as well (in a software system).

Let's consider the programmer who knows the platform Java which includes various technologies. The programmer's know-how about used technologies is represented with so-called *know-how score* in the interval $\langle 0,1 \rangle \cup \{-1\}$, where 0 means that the programmer does not have any know-how (experience) with the selected technology comparing to other programmers, and 1 means that the programmer's know-how about the technology is the highest in the comparison with other programmers. The value -1 means that the programmer used the selected technology as the only one and therefore we are not able exactly to express his relative experience.

The method for automatic calculation of programmer‘s know-how about used technologies consists of these steps: (1) identification of the author of the source code and technologies which were used, (2) construction of the programmer's model of used technologies and the calculation of the know-how score for each used technology.

The next factor is calculation of programmer's karma based on importance of components which he authored/co-authored. We suppose that the more source code fragments of a program (project) refer to a method then its importance is greater. For the calculation of method importance, we inspired by PageRank algorithm, where the importance of a method is determined by how many methods call it.

The method for automatic calculation of programmer's karma consists of three main steps, namely, (1) construction of a graph of method dependencies from source code of a project and calculation of PageRank score for each method; (2) construction of an index which contains a list of all the methods, the number of their Logical Lines of Code, calculated PageRank score, authors with determining their degree of authorship; and (3) calculation of programmers' karma.

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.

References

[1] Gysin, F., S., Kuhn, A.: A trustability metric for code search based on developer karma. In: *Proc. of ICSE Workshop on Search-driven Development - Users, Infrastructure, Tools and Evaluation*, ACM, New York, 2010, pp. 41-44.

[2] Sim, S., E., et al.: How well do search engines support code retrieval on the web? In: *ACM Trans. Softw. Eng. Methodol.,* ACM, New York, 2011, pp. 1-25.

User Feedback in User/Domain Modelling and Adaptive Evaluation

Martin LABAJ*

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

User and domain models are essential components of adaptive web-based systems, as well as the evaluation of such systems. In our research, we focus on user feedback used as a source for user and domain modelling, specifically on the tabbed browsing (also called parallel browsing). We also work on adaptive evaluation of adaptive web-based systems.

The tabbing is currently established as a more accurate representation of user browsing activities than the previous linear model [1]. We model the user tabbing behaviour from events sourced from a browser agent (extension) or scripts included in a page, recognizing sequences of events (e.g., pageload of a page P with referrer R, not preceded by page unload of R, and followed by blur of R, focus of P) as user actions (e.g., the user has opened the link in a new tab, then switched to it) [2]. From these actions, we discover *tabbing scenarios* (e.g., keeping a tab opened as a reminder) modelled after *reasons for using tabs* [3]. Various tabbing scenarios in which the tab participates are tracked per each tab during its life, effectively putting opened tabs into groups with various current or future levels of user's interest and various user tasks and goals. These data serve as a basis for stereotype-based user model of tab scenarios usage and overlay user model of interests, as well source of relations for domain model augmentation.

Another area of our research is the user-centred evaluation of adaptive web-based systems. We ask evaluation questions (EQs) during the user's typical work in the system. The questions are adapted for the user and their actions and are asked at appropriate moments using the evaluation engine. In this way, evaluation feedback is collected even from users who otherwise would not actively seek to provide the feedback, e.g., in post-session questionnaire. Moreover, the data is more accurate as the users are asked and they answer right when they are working with relevant parts of the system.

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

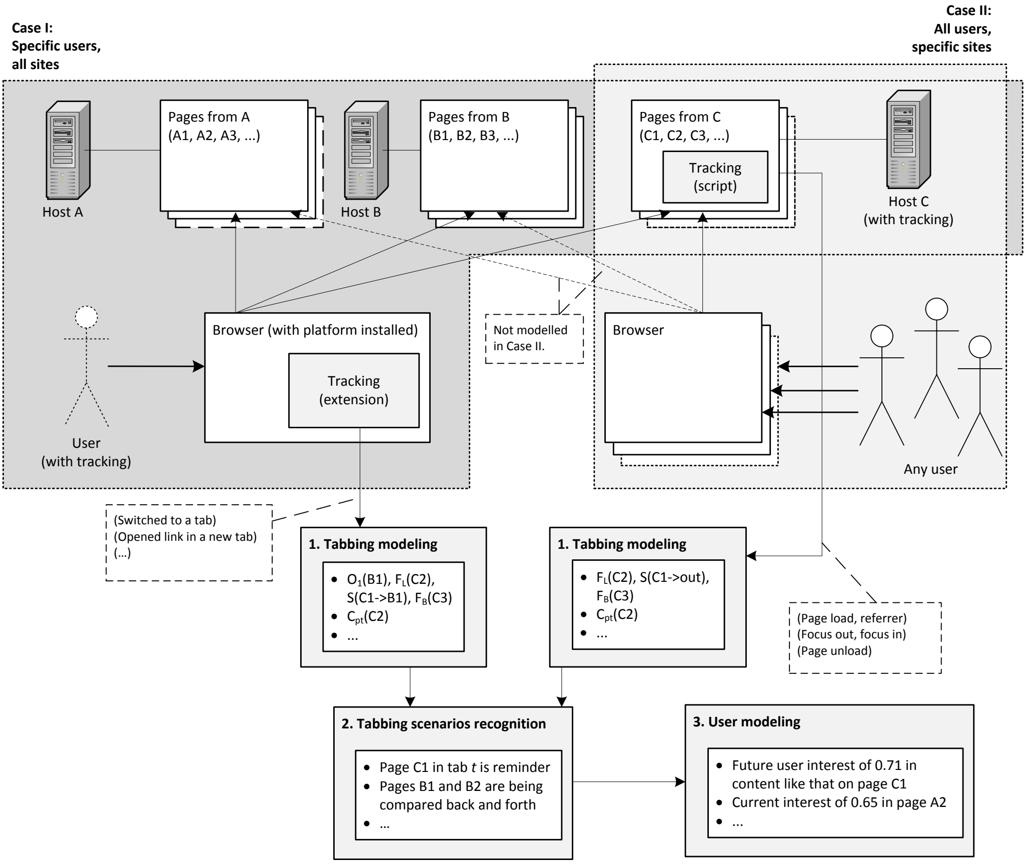


Figure 1. Overview of our user modelling method based on tabbing behaviour of the users.

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

References

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

[2] Labaj, M., Bieliková, M. Modeling parallel web browsing behavior for web-based educational systems. In *2012 IEEE 10th Int. Conf. on Emerging eLearning Technologies and Applications (ICETA 2012)*, IEEE, pp. 229–234 (2012)

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

Researcher Modeling in Personalized Digital Library

Martin LIPTÁK*

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

Researchers use digital libraries to either find solutions to particular problems concerning their current research or just to keep track with the newest trends in areas of their interest. However, the amount of information in digital libraries grows exponentially. This has two serious consequences. Firstly, many interesting works go unnoticed. Secondly, researchers spend too much time reading articles that turn out to be low-quality, unrelated to their current research or unrelated to their other interests. These kinds of problems are nowadays solved with recommendation systems or more effectively with personalized recommendation systems.

The core of every personalized system is its user model. User model is built from user data and can be used to personalize any feature of the personalized system. Model creation process and representation depend on availability of user data and requirements of personalized features [1]. They also depend on domain of user modeling. For example, user knowledge is essential in educational domain [2], but in domain of digital libraries, other characteristics of the user like interests can become more important.

We propose a user model, which is based on data analysis from Annota digital library organization service[1] [3]. The model will leverage

- articles the user has read
- tags the user has used
- folders the user has used
- terms the user has searched for
- search results the user has read

The model enables to add personalization to or improve existing personalization services in multiple features of Annota system such as:

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

[1] http://annota.fiit.stuba.sk

- searching articles,
- article recommendation,
- organization of articles in folders, and
- article summarization.

Based on the available user data and evaluation options, we will seek for suitable representation and creation process of researcher (user) model in domain of digital libraries.

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

References

[1] Peter Brusilovsky, Eva Millán. User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: *Brusilovsky, P.; Kobsa, A.; Nejdl, W. (eds.): The Adaptive Web*. Springer Berlin Heidelberg. Berlin. 3-53. 2007.

[2] Peter Brusilovsky. Adaptive Hypermedia for Education and Training. In: *Durlach, P., Lesgold, A. (eds.): Adaptive Technologies for Training and Education*. Cambridge University Press. Cambridge. 46-68. 2012.

[3] Ševcech, J., Bieliková, M., Burger, R., Barla, M.: Logging activity of researchers in digital library enhanced by annotations. In: *Bieliková M., Šimko, M. (Eds.):* 7th Workshop on Intelligent and Knowledge oriented Technologies, (2012), pp. 197-200. (in Slovak)

Reciprocity as a Means of Support for Collaborative Knowledge Sharing

Ivan SRBA*

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

Knowledge management systems provide organizations with many progressive ways how to create, improve and mainly share knowledge. These systems can be divided into three categories according to their perspective of knowledge [3]: knowledge as object, knowledge embedded in individuals and knowledge embedded in communities.

We consider knowledge management systems focused on knowledge embedded in communities as interesting and promising area for further research. This perspective views knowledge as collectively owned and maintained by the whole community. All members of a community are involved in dynamic process of knowledge obtaining, exchange and continuous improvement. One type of these knowledge communities is called *knowledge building communities* [2] which focus not only on knowledge sharing but also on learning new valuable practices. It is possible to identify knowledge building communities in many areas. The most common examples are classrooms, academic research teams or workplace teams.

Employing computer support in knowledge sharing process provides many advantages [1]:

1. *A representation advantage*, where technology provides an innovative ways of information presentation with the aim to support knowledge sharing.
2. *A process advantage*, where technology can support users in their activities, i.e. technology can provide scaffold for a novice user to find required information.
3. *A social context*, where technology can be used to shift the social context in which users share knowledge, i.e. by providing users with a possibility to communicate anonymously.

On the other side, technology caused that users' with high diversity of knowledge and specialization are supposed to collaborate together. According to several researches, reciprocity is one of the most important motivator for knowledge sharing because users with different specialization and knowledge expect that they will get back from the

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

community the same amount of knowledge as they give to community. However, this symmetry of knowledge is not usually achieved. Therefore, we decided to propose an innovative model of adaptive web-based system for collaborative knowledge sharing which will be aimed to support symmetry in activities of receiving and producing knowledge among all members of particular knowledge community (see Figure 1).

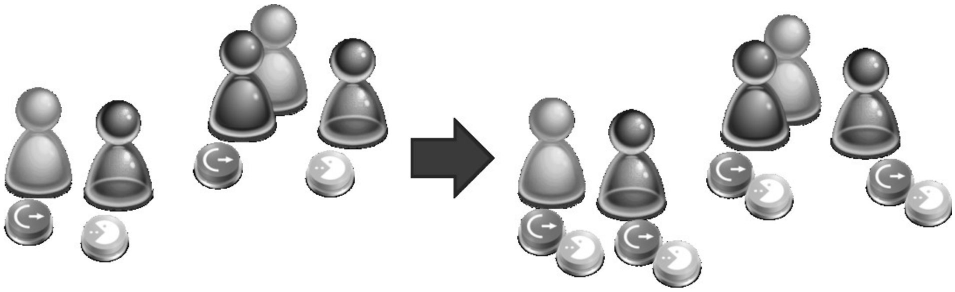


Figure 1. Symmetry in providing and receiving knowledge is not usually achieved. The goal of our project is to propose a model of system for reciprocal collaborative knowledge exchange.

There are several types of knowledge management systems based on perspective of knowledge embedded in communities, e.g. electronic discussion groups, chats or community question answering systems. We focus on *community question answering* systems which are very popular recently (i.e. StackOverflow or Yahoo! Answers). These systems scaffold users' collaboration using three common approaches: expert finding, question routing, and question-answer topic modeling.

We plan to achieve symmetry between provided and received knowledge by employing *question routing* approach. The concept of question routing refers to routing newly posted questions to potential answerers. The appropriateness of particular answerer is typically calculated on the basis of his knowledge and specialization. We propose to include additional aspect of reciprocity in this step: the questions will be routed to users who mostly received knowledge and they will have an opportunity to return the acquired knowledge back to their community.

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

References

[1] Hoadley, C.M., and Kilner, P.G.: Using technology to transform communities of practice into knowledge-building communities. *ACM SIGGROUP Bulletin*, 25(1), 2005, pp. 31-40.

[2] Scardamalia, M.: Collective cognitive responsibility for the advancement of knowledge. In *Liberal education in a knowledge society*. 2002, pp. 67-98.

[3] Wasko, M.M., and Faraj, S.: "It is what one does": why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, 9(2), 2000, pp.155-173.

Using Site Specificity to Build Better User Model from Web Browsing History

Márius Šajgalík*

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

Extracting user interests from user browsing history has been already studied by several researchers. We present a novel method to enhance the quality of user interest extraction by harnessing the site specificity. We follow the idea that some sites are more generic and not so relevant to real user interests, whereas sites that are more specific are more relevant to the user interests. If there are multiple topics within a single website, it is more likely that user is just not interested in all of these topics, but chooses to read only a few of them. In order to infer an interest of the user in a site, we propose to calculate the site specificity. The less tightly related topics are contained within a site, the more specific it is and more probable is the higher significance of the discovered topics for user interests. To be able to discover some measurable features, which might influence the overall site specificity, we need some additional knowledge about the web content. However, there is still not enough explicit semantic information of sufficient quality included in the webpage content, which forces us to incorporate some kind of ontology to understand the content of the "wild web" better [1]. Therefore, we use WordNet [3], which can be considered a lightweight ontology.

The basic idea of our approach is to compute the specificity of just a single webpage. To generalise it to an arbitrary set of webpages, we simply concatenate them and calculate the specificity over the union. To compute the website specificity, we choose several subpages within it and concatenate their content into a single piece of text. As we are focusing on enhancing a user model within the web browser, we choose only those subpages of the website, which are present in user's browsing history.

At first, we extract the main text of an article and choose only the noun terms as keyword candidates. Then we take all the noun synsets of WordNet, which contain at least one of these feasible terms. We call these synsets the basis synsets. Then we create the concept graph *G = (V, E)*, where vertices *V* are all the basis synsets plus those reachable by following hypernym or holonym relations. This aims to influence

[*] Supervisors: Michal Barla, Mária Bieliková, Institute of Informatics and Software Engineering

also the more general concepts (WordNet synsets) to get to the broader topics discussed in the extracted article. After we have built this concept graph, we perform page ranking algorithm to infer the relevance of individual concepts inspired by [4]. We do a two-pass ranking. In the first pass, we propagate the authority of a synset to all hypernyms and holonyms to obtain the most probable word senses. Apart from [4], we consider also the information content of single concepts. Additionally, we consider collocations and link the neighbouring terms in the second-pass page rank to support the collocated word senses and thus, get the key concepts. We adapted this idea from TextRank [2].

To calculate the site specificity, we applied various measures of concept similarity to measure the topic diversity or more specifically, the semantic coherence of the topmost concepts contained in the concept graph based on [5]. We considered only the concepts ranked at the top after inferring the ranking algorithm, since those are the most probable to be the most relevant representatives of the covered topics.

We evaluated four possible measures to measure the semantic coherence; however, there was no single winner method. We believe we could enhance it further by constructing a probabilistic model (Bayesian network) based on different features of the constructed concept graph. We could train the model parameters corresponding to different features using some dataset of categorised websites. Using such model, we would be able to set the values of observed variables and do the inference to obtain the conditional probability of the website being specific. We also plan to use the results presented in this paper to devise another method, which we believe will build a better user model having taken the website specificity into account.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 186-193.

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

References

[1] Bieliková, M., Barla, M., Šimko, M.: Lightweight Semantics for the "Wild Web". Keynote. In: WWW/Internet 2011, Proc. of the IADIS Int. Conf., IADIS Press, (2011), pp. xxv-xxxii.

[2] Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In Conf. on Empirical Methods in Natural Language Processing (2004), pp. 404-411.

[3] Miller, G. A.: WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38, No. 11, (1995), pp. 39-41.

[4] Ramakrishnanan, G., Bhattacharyya, P.: Text Representation with WordNet Synsets Using Soft Sense Disambiguation. Ingénierie des systèmes d'information, vol. 8, (2003), pp. 55-70.

[5] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proc. of the 14th int. joint conf. on Artificial intelligence (IJCAI'95), Chris S. Mellish (Ed.), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, vol. 1, (1995), pp. 448-453.

Method for Social Programming and Code Review

Michal TOMLEIN*

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

In order to create high quality, reliable and reusable code, incorporating code review into the development process is amongst the most effective of the available options. Due to its effectiveness in ensuring the quality of software, code review has seen wide acceptance in the industry.

However, in addition to being an important quality assurance method, it is also a powerful learning tool. We believe it can effectively be used to significantly improve the outcomes of the learning process and its overall quality. Furthermore, its adoption in the learning process in the form of peer reviews can serve to prepare students for code review in development practice, which is also highly desirable.

In recent years, social approaches to software development have changed the way we look at code sharing, collaboration and the development process. With the advent of social programming and code sharing services such as GitHub, there have been sweeping changes to the way we perceive and expect development of open source software components to work. Unfortunately, social programming has not made its way into programming courses to the same extent. In our work, we intend to combine the benefits of both code review and social programming, to improve the quality of the courses to provide additional development skills.

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 the most 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

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

motivation [2]. Because of this, there is a need for a solution, which takes such reviewer attributes into consideration when selecting the most suitable reviewer.

Our approach is based on real-time code review using a shared live view of the code being reviewed, which the reviewer can see and comment on. This is in contrast with more traditional code reviews, which are asynchronous and therefore do not require the reviewer to be available at the same time as the author of the code.

Secondly, the method incorporates aspects of social programming to create a sense of community within the programming course. The intent is to raise the students' awareness of the whole group's progress and their place within the group using progress visualisation and to encourage them to participate in code review to help their fellow classmates complete all of the programming assignments.

We are currently evaluating the outcomes of an initial experimental study. We have collected data on 172 students in an introductory programming course at the faculty. The data contains observations of student work and describes the relationship between the personality traits and characteristics and the reviewing abilities. We designed the experimental study as an ordered sequence of 5 highly interdependent programming assignments on introductory cryptanalysis. The functionality necessary to support the live reviews and automatic reviewer assignment was built into the existing web-based learning platform deployed within the course, called Peoplia.

Using the data obtained from our experiments, we intend to determine and employ the correlation between personality traits and the abilities to deliver and receive help to be able to select a suitable reviewer with the highest probability of success in helping a particular user.

By exposing the students to other students' code through peer code review, we aim to inspire them to improve their own code, their ability to read and understand other code, to learn about different ways of looking at the same problem, and last but not least, to train them to be able to provide feedback, which is in itself an exercise of learning by teaching. Learning by teaching is a well-known practice effective especially in the long term, as it prepares students to learn new concepts later [1].

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 257-262.

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.

References

[1] Biswas, G., Leelawong, K.: Learning by teaching: A new agent paradigm for educational software. Applied Artificial Intelligence, (2005), pp. 363-392.

[2] 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.

Information Retrieval Using Short-Term Context

Matúš VACULA*

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

The content of the Web is continuously expanding. Several years ago it might have been problem for a user to find desired information on the Web because the information might not have been published online. Nowadays user experiences slightly different kind of problem. There is too much information and it is problematic to find the most relevant piece of information for user's interest. All the modern search engines have to accommodate to this.

Search engine developers are trying to make searching more efficient by involving the context. The first to define the term "context-aware" were Schilit and Theimer [3] who defined it as location and identity of nearby persons or objects and their changes. Dey [1] defined context as any information that can be used to characterize the situation of an entity. Entity is a person, place or object which is relevant to user's interaction with application. Utilisation of the context is usually done by taking into consideration the geo-location of the user, his language or previous web-searches. But there are also other forms of context that are not utilized as much as the mentioned ones. It is possible to take into consideration also information such as the time of the day, time of the week, user's personal data, social aspects of context and his activity. Although the resources are available and numerous methods how to implement the context into the information retrieval are known, search engines are still not using these possibilities at their full potential.

In our work we focus on utilising the context from the aspect of the user's momentary activity. Information about user's current activity could lead to achieve more accurate search queries or at least more unambiguous meaning of his demand. Such information is relatively easy to obtain from the web browser. We developed our method based on the hypothesis that user's web-search relates to his current activity and this activity is reflected in the web-browser in the form of open websites.

We propose a method that uses the context of user's activity to enhance the web-search. This method consists of extracting and selecting the proper keywords from

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

browser's open websites and modifying the search queries with use of them. The process of keyword selection is based on scoring each keyword based on these criteria:

- proximity of keyword to search-query if found in the content of open websites
- number of appearances
- significance – which HTML is the keyword from (more significant are keywords from headlines and highlighted texts)

We expect this will result in providing more accurate and relevant search results to the user. The aim is to raise the user's satisfaction and shorten the time while he obtains the desired information. We are aware that the proposed method will only help to improve those web-searches which would be performed with sufficient contextual data. This will be mostly applied to users who are accustomed to browse more than one website at the time. That means using tabs in web-browser.

We decided to implement the method in the form of browser extension (Chrome) which gives us wide access to data describing user's activity. The architecture of this solution is based on the concept of meta-search engine [2]. Information retrieval itself is provided by Google search engine and the browser extension is responsible for optimizing the queries.

We also propose the quantitative experiment which will evaluate the success rate of the proposed method. The principle is to provide the search results from enhanced and unenhanced web-search mixed together in random order in the way that user would be unable to identify which search result come from which web-search (enhanced or unenhanced). This way the chances of bias are reduced. During the experiment we will monitor the user's choices of search results and the time spent browsing the websites navigated through the selected search result. More clicks on search results from enhanced web search and more time spent on websites recommended by enhanced web search should indicate that the proposed method helped to make the search result more attractive to user and helped him to find more relevant information.

In order to evaluate the effect of the proposed method more precisely we suggest to consider only those searches which were performed with the use of sufficient amount of contextual data. This will provide the evaluation of effectiveness on the selected behavioural group of test subjects who tend to use more than one browser tabs at the time.

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

References

[1] Dey, A.K., Understanding and Using Context. *Personal and Ubiquitous Computing*, Volume 5, Issue 1, pp. 4–7 (2001)

[2] Glover, E. J., Lawrence, S., Gordon, M. D., Birmingham, W. P., and Giles, C. L. Web Search Your Way. *Communications of the ACM*, 44, pp. 97–102 (2001)

[3] Schilit B. N., Theimer M. M. Disseminating active map information to mobile hosts. *IEEE Network*, 8(5), pp. 22–32 (1994)

Domain Modeling, Representation and Maintenance

Keeping Information Tags Valid and Consistent

Karol BALKO*

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

Information tags as type of descriptive metadata with sematic relation to tagged content represent an opportunity to enrich objects (tagged content) with additional information and relations between objects in wide informational space such as the Web [1]. This allows us to form a space of structured information about contents of webpages and to describe relations between these contents. These information tags can be further used to make software systems cooperative so this topic is very potential for the future of the Web [1].

Information tags contain structured information understandable by software systems. Most of them are created by machines from content and user behaviour analysis. However, we must introduce mechanisms that monitor validity of information tags, e.g. it may happen that data referenced by information tags disappear or have been changed, what makes information tags no longer valid [2].

The basic principle of keeping information tags valid and consistent is identifying topicality of information tags and detecting whether information tags are still valid. Only after this, it is possible to repair and maintain information tags.

Since many information tags contain data in form of URL addresses that can reference to resources with similarly tagged objects, it's possible to check these URL addresses, if they still exist. As HTML code of webpages provides us a lot of information, we can compare suitably selected keywords from information tags and webpages, in order to determine whether these information tags are still valuable to referenced content [3].

The problem of information tags maintenance can be related to a problem of information tags organization. Increasing number of information tags increases disorganization of these tags, too. However data contained within these information tags can be useful in their categorization. The base of categorization can consist of creating main categories (optionally with more subcategories) and subsequently use keywords extracted from information tags. For keyword extraction we can used

[*] Supervisor: Karol Rástočný, Institute of Informatics and Software Engineering

projects like Metallurgy that deal with problem of metadata extraction and capability of extracted keywords. Then it can be possible to insert information tags to correct category. In our work we deal with methods of keeping information tags valid and consistent, focused on information tags that enrich content by resources from wide informational space (for example the Web, or PerConIK project [4]). We want to suggest methods of automated detection of topicality and consistency of these information tags. For example in scope of document metadata we can experiment with documents parsing and try to compare documents' properties with properties included in documents' metadata (for example a number of pages, a number of words). We can use tools like smart tokenizer or powerful tools like GATE. In scope of metadata, that describe relations between the Web's objects, we can analyse webpages included in triplets of these metadata and verify URIs that describe source and targets of triplets. For this verification we can use HTML tags like meta-tags for extracting and analysing data from these sources.

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.

References

[1] Rástočný, K., Bieliková, M.: Maintenance of Human and Machine Metadata over the Web Content. In *Current Trends in Web Engineering*. 2012. p. 216–220.

[2] Haslhofer, B.; Simon, R.; Sanderson, R.; Van de Sompel, H. The Open Annotation Collaboration (OAC) Model. Multimedia on the Web (MMWeb), *Workshop on Digital Object Identifier.* 2011

[3] Sanderson, R.; Ciccarese, P.; Van de Sompel, H. Open Annotation Core Data Model, 9. 5. 2012 *http://www.openannotation.org/spec/core/*

[4] Bieliková, M., et al.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. In: Procedia Technology, 3rd World Conf. on Inf. Tech., to appear.

Building a Domain Model using Linked Data Principles

Michal HOLUB*

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

The Linked Data principles are being used in many datasets published on the Web. The aim of our work is to use Linked Data in order to create models describing (1) the domain of software development, and (2) the domain of research in the field of software or web engineering. We use these models to represent the knowledge of IT professionals (analysts, programmers, testers), as well as the research interests of researchers in the respective fields. Using Linked Data allows us to connect entities in our models with the ones published on the Web and get more information about them.

Linked Data principles are being used in various datasets, which form a Linked Data Cloud. In the center of this cloud there are two large datasets: DBpedia [1] and YAGO [2]. Both use Wikipedia as their primary source of information, they extract it from infoboxes and categories. These datasets define as many entities as possible, so that other datasets can link to them.

We propose a method for automatic construction of a concept map serving as a basis of our domain models. For this purpose we use unstructured data from the Web, which we transform to concepts and links between them. Using a concept map we describe the knowledge of software developers as a set of technologies and principles they are familiar with. We also use a similar concept map to describe research areas, problems, principles, methods and models studied by researchers at our faculty.

Concepts are linked together using relationships like *part of*, *is a*, *is written in*, or *uses*. Due to the relationships we can later do reasoning, e.g., deduce that when a programmer knows jUnit (a testing framework for Java), he also has to know a bit of Java (a programming language), since there is a relationship stating “jUnit uses Java”.

As a source of information for the concept map population we use free online encyclopedia Wikipedia. We analyze the textual content of its articles to learn new concepts and their instances.

We use the existing concepts as a seed in the task of ontology learning. We search all Wikipedia’s articles for occurrences of concepts from our concept map. Take

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

“programming language” as an example of a concept. Article about it also links to a “List of programming languages”, from which we can extract additional concepts, which are subclasses of “programming language”.

When the ontology is ready, we populate it with instances, which we do as follows:

1. Select an article from Wikipedia containing a particular concept in its text.
2. Find the first sentence containing the verb *is* followed by one of the concept types.
3. Convert the title to a new concept instance (if it is not present) and create *is a* relationship between the instance and the concept.

Using this process we not only populate the domain model with particular technology, we also find all terms which can describe a technology used when developing software.

There can be other words following the verb *is* in the article not matching any concept from our map. These could express properties of the technology and we might enhance the ontology.

The ontology can be used in a system for gathering the knowledge of programmers. Let us assume the user adds “Java EE” to his skills. We identify it as a platform in the ontology. It is related to “programming language” by *uses* link. We can generate question “Which programming language used in Java EE are you familiar with?” This way we can get more skills from the user.

The domain models we create can be used as a basis in two adaptive systems. The first aims at capturing IT professionals’ knowledge and skills, deduce further technologies they might know and enables users to search for a suitable candidate for a certain task or project. The second one allows users to bookmark, annotate and collaborate over research papers in digital libraries, as well as other Web documents. Here, we also use the model in order to answer queries in pseudo-natural language.

We evaluate the models and methods of their creation directly by comparing them to existing ones or by evaluating facts from them using domain experts. Moreover, we evaluate the models indirectly by incorporating them in adaptive personalized web-based systems and measure the improvement in the experience of users (i.e., they get better recommendations, search results, etc.).

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 415–416.

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

References

[1] Auer S., Lehmann J. What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In *The Semantic Web: Research and Applications*, LNCS Vol. 4519. Springer, pp. 503–517 (2007)

[2] Suchanek F.M., Kasneci G., Weikum G. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: *Proc. of the 16th int. Conf. on World Wide Web*, ACM Press, pp. 697–706 (2007)

Augmenting the Web for Facilitating Learning

Róbert HORVÁTH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
robertorvath89@gmail.com

We spend large amount of time browsing the Web and we come across a lot of documents. We believe that this amount of time can be spent more effectively due to integrating text augmentation methods into web browsing. Providing user with additional information can help in education process such as foreign language learning.

Our main goal is to create a method for web augmentation for facilitating foreign language learning. We enrich web content by replacing appropriate words during web browsing, maintain user knowledge and preferences, while considering specifics of learning process like forgetting. The method brings together process of web browsing and vocabulary learning. Potential for this approach is supported by advances in technology-enhanced learning and computer-assisted language learning. It was shown that learning occurs even unintentionally and with minimal mental processing [1].

Existing approaches to enhancing webpages to help user with learning foreign language unintentionally are mostly implemented as web browsers extensions. Analysis showed they avoid user knowledge modelling. It leads to random presentation of foreign vocabulary [2, 3]. In contract to them, Duolingo provides learning platform based on user model and approach which considers specifics of learning process. Studies show that its effect on language learning is comparable to school classes [4].

Our method provides user with opportunities for vocabulary learning with no intention of studying. Our aim is to find appropriate terms for learning in webpage content user is going to read and replace them with their translations the way user is still able to understand meaning of content and remember the vocabulary. Our method consists of three main steps executed for every visited webpage (see Figure 1):

1. *Text analysis and pre-processing* – Removal of unnecessary information, webpage translation and identification of learning candidates.
2. *Personalized text augmentation* – We replace and highlight words on webpage to present new vocabulary, while preserving original meaning.
3. *User model update based on user activity monitoring* – User activity is monitored and based on his/her behavior user knowledge model is updated.

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

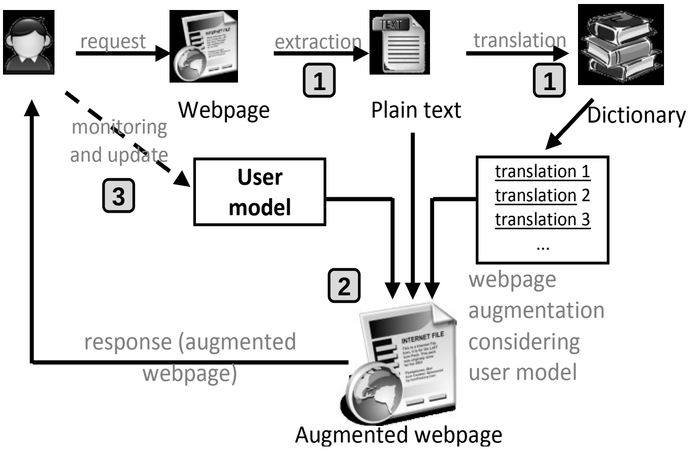


Figure 1. Process of personalized webpage augmentation.

In order to evaluate our method we have created web browser extension for Google Chrome that augments Slovak webpages with English vocabulary, which is derived from user knowledge model. For evaluation purposes extension gather both implicit feedback from monitoring user activity and explicit feedback from regular vocabulary tests. To find the effect on the learning process we propose two main hypotheses:

1. Augmentation improves foreign language vocabulary size.
2. Time spent with reading augmented webpages will increase insignificantly.

We have already conducted small supervised experiment to evaluate effect of text augmentation of reading speed. The results show that augmented webpage slows reading speed down on average by approximately 7%. We find these results very reasonable with a great potential to support the second hypothesis. However, we need to conduct further experiments using larger data set to obtain more significant results.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 417–418.

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

References

[1] GrootT, P. Computer Assisted Second Language Vocabulary Acquisition. *Language Learning & Technology*, pp. 60–81 (2000)

[2] Streiter, O, Knapp, J., Voltmer, L., Zielinski, D. Browsers for autonomous and contextualized language learning: tools and theories. In *Proc. of 3rd Int. Conf. on Information Technology: Research and Education*, ITRE'05, pp. 343–347 (2005)

[3] Trusty, A., Truong, K. N. Augmenting the Web for Second Language Vocabulary Learning. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, ACM, pp. 3179–3188 (2011)

[4] Vesselinov, R. *Duolingo Effectiveness Study*. City University of NY, USA (2012)

Acquisition of Learning Object Metadata Using Crowdsourcing

Marek LÁNI*

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

In past years the Web began to be used largely for education purposes. There are many Technology Enhanced Learning (TEL) or Question Answer (QA) portals and web sites which are being used to gain knowledge and information. Many times these systems are designed not only to provide benefits for users but also to benefit from their users. We can say that it is a win-win relationship. It is because the content of these systems is often crowdsourced, i.e., generated by users themselves. Hence these systems build up or enlarge their knowledge base due to the crowd power. A typical example of such a system is a QA community portal StackOverflow[1].

Since it is not guaranteed that people who create content within these systems are experts in a specified area, it is necessary to evaluate quality of their contributions. This quality validation should be automated for reason of its time consumption. But the question is: How?

Some approaches are trying to analyze semantics of questions and corresponding crowdsourced answers in order to determine whether the answer is relevant or not.

We think that there exists better method to determine quality of user generated content and it is usage of "second level" crowdsourcing, where not only the content is being generated by users, but also the act of evaluation is outsourced to the users. The evaluation should be done in form of ratings. However, it is not as simple as it seems to be. Many research works defined a problem, that non-filtered rating evaluation is not giving us satisfying results [1, 2]. The obtained raw rating data also suffer from a certain degree of imperfection based on non-expert contributing. The usage of next iteration of crowdsourcing would be meaningless so we have to analyze, filter and weight obtained rating data.

In our work we are focused on QA component of TEL system. In this system we have set of question-answer pairs which were collected during the examinations during the academical year at subject Principles of Software Engineering. Every QA object

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

[1] www.stackoverflow.com

has assigned teacher's evaluation. This QA component allows students to study before examinations in form of reviewing the QA pairs and assigning the rating to answers. We want to take this data and analyze them in order to determine if the crowd is able to agree with teacher on assigned rating.

There are several approaches how to filter and weight user ratings. The aim of our work is to take, combine and modify some of them to achieve satisfactory results in evaluation of answers. We plan to use some of the following approaches:

- *Spam filtering*

  Agichtein [1] defined two main reasons why there is spam within the rating data. One reason is insufficient knowledge and the second is malicious activity. In our case, within the collected ratings, there can be outbound ratings obtained from users, who have only a little knowledge or no knowledge about the questioned subject at all. As for the every rating the user is awarded by activity points, there can be produced also malicious rating activity in order to gain points. We want to detect some general patterns of named behaviors and ignore the ratings which were produced by behavior like this.

- *Weighting of ratings and expert determination*

  According to Chen et al. [2] a vote calibration can bring large improvement in answer relevance evaluation. We aim to analyze users' profiles in order to determine their expertise level. This will help us to estimate probability of correctness of user's ratings. We will also take into account the frequency and count of user's votes when designing the weighting and calibration mechanism.

We want to experiment with these weighting parameters and filtering in order to achieve the best results in crowdsourced answer relevance evaluation and to determinate whether the crowd is capable of self-evaluation or not.

To enlarge our dataset and mainly the count of ratings, we want to create interface of our QA component for mobile devices. We also think about usage of probabilistic approach of inferring the missing ratings via Probabilistic Matrix Factorization.

We plan to evaluate our methods of filtering and weighting of ratings produced by users, by comparing them to the ratings assigned to the answers by experts – teachers.

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

References

[1] Agichtein, E. A Few Bad Votes Too Many? Towards Robust Ranking in Social Media. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, ACM, pp. 53–60 (2008)

[2] Chen, B.-C., Dasgupta, A., Wang, X., Yang, J. Vote calibration in community question-answering systems. In *SIGIR'12: Proc. of the 35th int. ACM SIGIR conf. on Research and development in inf. retrieval*, ACM, pp. 781–790 (2012)

Querying Large Web Repositories

Matej MARCOŇÁK*

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

The Web is one of the largest sources of information in the World. It is necessary to efficiently store and process data from the Web for their further use. Unfortunately, large amount of information on the Internet achieved level, when we are not capable to process this amount of data on a single machine/server. It is necessary to look for other options and approaches how to process large data. One of solutions to the problem is based on parallel data processing on clusters of computers. The most known parallelism-based solution is programming model MapReduce [1]. The main idea of the model is to hide details about parallelism and to allow programmers to focus on data processing.

Because the amount of data is increasing, the effective approach for information search is needed, too. With these thoughts, an idea of integrating some mechanism for recognizing relationships among information on the Web is coming. From this requirement for better machine processing of information, a trend to apply semantics to the Web has been emerged [3]. Semantics allow us to create webpages or documents that are more intended for machine processing. This kind of data are often represented as RDF triplets of subjects, predicates and objects (e.g., John, is friend, Mathew) and organized in ontologies. Ontologies and RDF data are standardly queried by SPARQL and its extended forms.

The main objective of our work is to explore possibilities of SPARQL query language and MapReduce techniques with a goal to purpose methods of querying big RDF repositories. It is very important to obtain required information from large RDF repositories as quick as possible. To retrieve the data quickly, it is important to choose a suitable data storage. Therefore we store domain specific data in NoSQL database MongoDB, because data structure aspect is more preferable for our use. The next factor to retrieve information is advanced features of SPARQL (e.g., filter). It is in contrast with our storage, because NoSQL databases do not support querying by SPARQL. Therefore we have proposed MapReduce algorithm for evaluation of SPARQL and its advanced features.

[*] Supervisor: Karol Rástočný, Institute of Informatics and Software Engineering

Most of existing solutions to co-operation of SPARQL and MapReduce are focused on optimizing graph pattern, but our main goal is an optimal strategy for the implementation of SPARQL's advanced features. In the implementation of advanced features we are going to use different techniques on different levels of MapReduce programming model:

- function *Map*
- function *Finalize*
- combination of above two methods

Implementation of our approach within function *Reduce* is not suitable. In function *Map*, we can reduce amount of data, that are required for further processing, but unfortunately this phase is restricted only for executing simple operations.

On the other hand, function *Finalize* processes data to final state and in contrast with Map is independent from SPARQL operators and has whole information about results. We would like to connect benefits of both methods for better performance.

One of the most used operators in SPARQL is *Filter* [2] and its implementation from the view of its functionality is appropriate for our work. Expressions in operators of *Filter* can be pre-processed on smaller parts for its application in function *Map*. This decreases an amount of data for further processing and transmitting. In the next step function *Finalize* filters remaining results by the most complex filters that can be applied only at completed information about results (mapped and reduced).

We will evaluate our method on the domain specific data in NoSQL database MongoDB. Testing will be conducted on a sample of non-trivial SPARQL queries over these data. The number of these data in MongoDB will be gradually increased for comparing efficiency of our method on different levels of programming model MapReduce.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

[1] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, (2008), vol. 51, no. 1, pp. 107-113.

[2] Picalausa, F., Vansummeren, S.: What are real SPARQL queries like? In: *Proc. of the International Workshop on Semantic Web Information Management - SWIM '11*. ACM Press, New York (2011). p. 6.

[3] Shadbolt, N., et al.: The Semantic Web Revisited. *IEEE Intelligent Systems*, (2006), vol. 21, no. 3, pp. 96-101.

Semantic Wiki for Research Groups

Martin MARKECH*

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

The content published on the Web is constantly growing and the Web is becoming difficult to process. Aim of the Semantic Web is to solve these problems.

Linked Data initiative can be described as a set of the best practices for sharing and publishing information on the Semantic Web. This is in particular relevant to researchers who can more effectively interconnect if they publish their content semantized. One of the paths to semantics utilization is the replacement of traditional wiki systems with semantic wiki systems.

The first semantic wiki system was created in 2004 [1]. Many new semantic wiki systems were created since then, but there are still some open issues, which we are facing [2]. Not all semantic wikis allow RDF import, so ontologies cannot be edited by a user. Semantic wikis use URI of a page as dereferenced URI, which cannot be modified in the future. It is not problem on websites of encyclopedic type, but it is problem when we want to create deeply nested menu structure with more pages for one entity. Although many semantic wikis try to help user with content creation, neither seems to assist with semantic extraction from text.

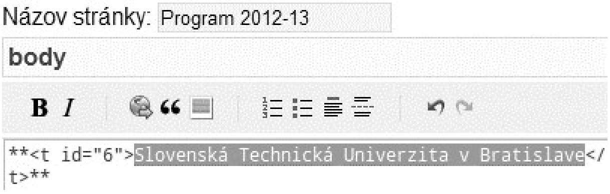
Our motivation is to improve existing wiki at our university taking into account specific needs of academic research groups, especially our group – PeWe. PeWe uses this wiki for presentation purposes and also to self-organize members.

Our method consists of several parts. At the beginning we analyse the structure of wiki content. Then we propose templates, which help user with writing repeated blocks of text with similar structure. Templates are divided by topic and by granularity from simple to advanced templates. Filling the templates helps user to keep the same text structure and to auto-generate semantics.

It ensures that the created text has properly defined semantics, because values filled in fields are inserted in accordance with pre-defined ontologies. Our method uses a triplet editor to give user the ability to change or add new semantics. Each semantic triplet has an attached ID. We use this ID in markup to connect text with semantic triplets. These triplets are stored outside of the markup – in semantic database Sesame. Since each wiki page has many revisions and allows creating multiple page parts, for

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

the key value of context triplet field our method uses concatenation of page ID with revision ID and page part ID. The application for browsing semantics is independent and thus, we are not facing issues when the URI of wiki page is simultaneously dereferenced URI.



Názov stránky: Program 2012-13

body

```
**<t id="6">Slovenská Technická Univerzita v Bratislave</
t>**
```

Figure 1. Markdown semantic marks.

Second part of our method is generating semantic bibliography reference, because creation of correctly ISO 690 reference is not an easy task. Digital libraries usually do not offer ISO 690 reference string. Mostly they offer BibTex format, which is parsed by our method. The scenario is as follows:

1. User fills in some information about publication – DOI, authors, or title
2. Webservice sends query to Google in format *site:<digital library site name> <searched string>*
3. Then it takes the first result, expecting that it is the result with the best relevance
4. Webservice loads the page in background and downloads BibTex data about publication
5. Then it converts BibTex format to ISO 690 reference string.
6. User confirms the correctness.

We evaluate the application with qualitative methods. Our hypothesis is that usability of the application with new semantic extensions increases. We analyse the trade-off between the state without semantics – copypasting markdown – and with new extension – templates with semantics.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 427–428

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

References

[1] Bry, F., Schaffert, S., Vrandečić, D., Weiand, K. Semantic wikis: Approaches, applications, and perspectives. In *Reasoning Web. Semantic Technologies for Advanced Query Answering*, Springer Berlin Heidelberg , pp. 329–369 (2012)

[2] Maalej, W., Panagiotou, D., Happel, H. J. Towards effective management of software knowledge exploiting the semantic wiki paradigm. In *Software Engineering*, 121, pp. 183–197 (2008)

Metadata Collection for Personal Multimedia Repositories Using GWAP

Balázs NAGY*

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

With the increasing number of personal albums and photos in them, users have increasingly more problems with their organization [4]. This is due to lack of descriptive data; however, their amount mostly depends on the willpower in the image owners. Tools for metadata creation are available – the main problem is with the user motivation: because tagging and annotating of photos is usually a boring activity and its execution takes extended time periods. Other methods for obtaining metadata to general images also exist (automated methods, crowd-based, games with a purpose [1]), but these are unable to deliver specific metadata needed for personal imagery.

Our aim is to create a method to enrich personal photo albums with keywords and also named entities. To achieve this, we use tool that imports personal albums, allows creating annotations using a GWAP, extracts keywords and named entities from annotations and allows browsing in albums using obtained metadata.

Earlier, we devised a game with a purpose called PexAce [2, 3] for harvesting textual annotations to general images. Earlier experiments showed that people playing with their own photos are more engaged to game and also interested in creating annotations. By merging our game with automatic approach of metadata acquisition from game-produced annotations, we found an appropriate solution for creating metadata for personal photo albums. The main contribution of this work is a framework for processing annotations written to personal photos. This framework is based on metadata extraction modules called extractors and is extensible from this perspective. Inputs of all extractors are following five entries: photo, album to which photo belongs, user who created the annotation, annotation, and timestamp. In addition to this data, extractors have access to all logs saved during games and use them to refine results of annotation processing. Extractors are divided to two groups depending on their output. While the first group includes extractors extracting keywords without specifying their types or any other information about them, the extractors in second group are exact typed keywords and also named entities such as persons, geographic

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

locations, events or holidays. To aggregate outputs of extractors we designed two types of aggregators for each type of extractor. These contain information about credibility of extractors and use it for aggregation of the results. Credibility of an extractor depends on the particular method (pre-processing, candidate selection or comparison method) used by the extraction (e.g. results provided by a particular tag extraction API can rank higher than results of another one).

To evaluate our method, we implemented different parts of our solution in particular order. In the first stage we re-implemented and re-designed the PexAce game, which is now more user friendly and as a web application it can be run on multiple platforms. Then we implemented importing tools to transfer photos into our database. After this we had the first opportunity to evaluate its functionality and realized a qualitative verification in the form of an interview with a small number of users. In second phase we implement extractors with aggregators (processing annotations written to photos) and photo gallery (exploiting keywords extracted with our methods). The photo gallery offers the opportunity to verify usability of extracted keywords based on user feedback.

For quantitative verification of our method we created a tool which enables users manually annotate personal photos. Comparing these annotations with our results we can determine the precision and recall of the method we designed.

Analyzing the current state of solutions which are oriented to obtain descriptive data to photos, there is absence of solutions for personal photo albums. Because of this, and the lack of descriptive metadata in these albums we decided to use our existing method for metadata acquisition for general photos, and redesign it for obtaining metadata for personal photos.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 431-432.

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

References

[1] von Ahn, L., Dabbish, L.: Designing games with a purpose. Communications of the ACM 2008. Vol. 51, no. 8, pp. 57.

[2] Nagy, B.: Acquisition of Semantic Metadata via Interactive Games, Proceedings of IIT.SRC 2011. Vol. 1, pp. 9-15.

[3] Simko, J., Bielikova, M.: Games with a Purpose: User Generated Valid Metadata for Personal Archives. Semantic Media Adaptation and Personalization (SMAP), Sixth International Workshop, 2011. pp. 45-50.

[4] Vainio, T. et al.: User needs for metadata management in mobile multimedia content services. Proceedings of the 6th International Conference on Mobile Technology, Application & Systems. ACM, 2009. pp. 51.

Automatic Web Content Enrichment Using Parallel Web Browsing

Michal RAČKO*

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

Creating links between resources on the Web is now an acute problem due to volume and diversity of the content. In the past, only the authors of the webpages could create their content, but now in the times of Web 2.0, any Internet user can add content themselves. This is the main cause of improper structure of such content and of weak or no links between similar resources. Creation of a clear and long-term sustainable structure of the Web is important for easier navigation when browsing or searching. Most of the existing solutions addressing the aforementioned problem are inadequate or methods that have been chosen to solve this issue are not very suitable. Therefore it is necessary to propose a method able to provide us with additional information about web resources and thus allow easier creation of links between those sites.

Web content enrichment can be done in several ways. One way is to use the way the users browse sites in browser tabs. Especially promising is the use of parallel browsing as a source of information on relations between the content as currently, there is a large number of web browsers allowing tabbed browsing. Various researchers have found that users use tabs in variety of ways, most of which were not planned [2]:

- temporary lists,
- list of search results or
- way to return to the previous page.

For the purpose of creating links between Web resources, it is necessary to provide a model for specific domain, which represents the semantic description of relevant resources and relations between.

Domain modelling is directly related to user modelling. There are two basic types of user models [1]:

- Stereotype based models
- Overlay models

[*] Supervisor: Martin Labaj, Institute of Informatics and Software Engineering

Stereotype based model categorizes individual users to default groups of similar users based on their characteristics (demographics, knowledge level, etc.).

Overlay model uses properties described in domain model. For each property, the user is assigned a value expressing a degree to which is the selected feature true for him or her. Such models are more dynamic than stereotype based models.

The process of creating the domain model often starts with identifying and visualizing the conceptual classes of domain. They represent a smaller part of the knowledge contained in the domain. Then we add relations between them and add attributes.

The aim of our project is the relationship discovery between sites using user visits with multiple tabs. We are interested in finding what are the relations between them, and whether they can be linked together based on the way they were opened in the browser – also taking into account how were the individual pages in tabs created or used and then adjusting the strength of relationships between resources.

Parallel browsing is a type of implicit user feedback, using which we can create links between pages, which are not yet connected directly. We focus on developing of a domain model creation method based on parallel browsing user behaviour and which we will also use to identify which pages are relevant for other similar readers. The method will reflect the way the user works with tabs in the browser environment.

We will evaluate our proposed method using web usage logs from brUMo browser extension[1] and ALEF adaptive learning framework [4]. During evaluation, we are going to consider all important aspects contained therein, such as keywords added by authors or various other elements describing a page [3].

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

References

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

[2] Dubroy, P., & Balakrishnan, R. (2010). A study of tabbed browsing among mozilla firefox users. Proceedings of the 28th international conference on Human factors in computing systems - CHI '10, 673.

[3] Jing Li and C. I. Ezeife. 2006. Cleaning web pages for effective web content mining. In Proceedings of the 17th international conference on Database and Expert Systems Applications (DEXA'06), S. Bressan, J. Küng, and R. Wagner (Eds.). Springer-Verlag, Berlin, Heidelberg, 560-571.

[4] Šimko, M., Barla, M., Bieliková, M.: ALEF : A Framework for Adaptive Web-Based Learning 2.0, Advances in Information and Communication Technology, vol. 324, pp. 367-378, (2010).

[1] brumo.fiit.stuba.sk, Browser-based User Modelling and Personalization Framework

Information Tags Maintenance: Anchoring

Karol RÁSTOČNÝ*

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

Current content processing and presenting systems create a lot of different metadata that contain valuable information, for example logs about users' behavior or derived concepts. These metadata are closely related to their resources – data in repositories of information spaces. But these data are not static and all their modifications affect validity of metadata, so metadata have to be maintained. Because several types of metadata exist and probably each type needs specialized maintenance approach, we have aimed to information tags (descriptive metadata with semantic relations to a tagged content) and we are working on a proposition of automatic information tags maintenance approach and information tags representation which is suitable for effective maintenance.

Due to structural similarity of information tags to annotations, we based information tags model on widely accepted Open Annotation model. Open Annotation model allows complex structures and it is proposed for RDF repositories, so we have lightened the Open Annotation model and we have redesigned it to an object model which can be stored to fast and scalable MongoDB repository.

Problem of metadata maintenance has not any sufficient solution. But this problem can be divided to two partly indifferent sub-problems. The first is maintenance of anchoring, which can be solved by accurate robust position descriptor [3]. The second problem is maintenance of bodies of metadata. This problem is not solved in current approaches of metadata maintenance.

In case of our research project PerConIK [1] we utilize information tags mainly for tagging parts of source code by behaviour information about programmers and information about source code's features. To deal with problem of anchoring in source code we proposed an approach, which works with source code files like with sequences of textual elements (lines or words). Our approach uses location descriptor within a target (source file or an AST element of a source file), which consists of two partial descriptors with different scope of use cases:

1. *Index-based location descriptor* – the index-based location descriptor contains indexes of the first and the last letter from a target text in a target.

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

2. *Context-based location descriptor* – the context-based location descriptor is used as the robust location descriptor. A context-based location descriptor contains information about:
    a. *Tagged text* – sequence of textual elements, that has been tagged in a target;
    b. *Context before and after tagged text* – minimal unique sequence of textual elements that are directly before and after tagged text;

We interpret context-based location descriptors as sequences of textual elements. The approach gives us opportunity to break up problem of time and memory complexity of approximate string matching [2] to two smaller parts – comparing textual elements and local sequence alignment that are processed separately:

1. *Compare textual elements* – unique textual elements of the source code are compared to unique textual elements of a context-based location descriptor;
2. *Local sequence alignment* – locations of a tagged text are searched as a sub-sequence of the source code via the Smith-Waterman local sequence alignment algorithm. In this step we also calculate scores for each matched location.
3. *Calculate scores of matches of contexts* – scores of contexts before and after tagged text are calculated for each possible tagged text's alignment. Confidences are calculated via the Smith-Waterman algorithm whose scoring function uses as aligned sequences connected to alignment of tagged texts.
4. *Calculate confidences of matched locations* – confidences of matched locations are calculated as linear combination of scores for each possible alignment.

We have processed preliminary evaluations of proposed anchoring approach. In the evaluations, we proved that the approach is useable for real-time processing and that it is enough accurate for anchoring information tags in source code. We performed these evaluations with several configurations, while optimal configuration (from tested values) seem to be usage of the Jaro-Winkler string similarity algorithm for comparing textual elements, while if our anchoring approach has to be used with a longer source code, usage of lines as textual elements ensures usability for real-time anchoring.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 82-89.

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.

References

[1] Bieliková, M., et al.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. In: Procedia Technology, 3rd World Conf. on Inf. Tech., to appear.

[2] Fredriksson, K., Navarro, G.: Average-optimal Single and Multiple Approximate string matching. *ACM J. of Experimental Alg.*, (2005), vol. 9, no. 1.4, pp. 1-47.

[3] Phelps, T.A., Wilensky, R.: Robust Intra-document Locations. *Computer Networks*, (2000), vol. 33, no. 1-6, pp. 105-118.

Crowdsourcing in the Class

Jakub ŠIMKO*

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

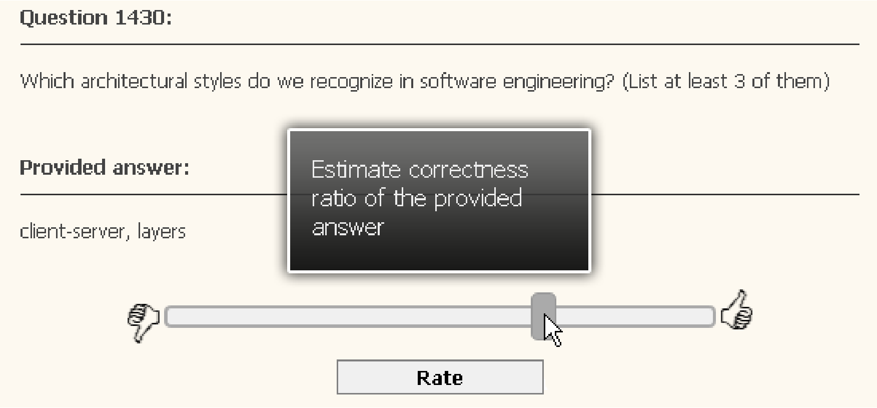
With the emergence of technology enhanced learning, we witness paradigm shift in learning, especially when considering web-based learning environments. Benefiting from concepts introduced by Web 2.0, students become more autonomous and less dependent teachers. She is provided with more competences since she can tag, rate, share, and collaborate during learning. She becomes an active contributor rather than a passive consumer of learning content [1]. The activity of a learner is "boosted" not only in relation with a learning system, but also when considering collaboration during learning and content contribution [3]. This effectively involves a crowdsourcing process into the technology enhanced learning.

In our work we contribute to the domains of technology enhanced learning and crowdsourcing [2]. Our method assesses the information about correctness of combinations of answers and questions. It is motivated by the possible reuse of existing (but not evaluated) student-written free text answers to questions, which originated as the by-products of previous learning activities (e.g. exercises, exams) and are commonly available. The goal of the reuse scenario is to provide interactive exercise to the students with automated feedback using these questions-answer learning objects (QALOs). In it, the student is presented with the question and an answer to it (possibly correct or wrong) and has to decide about the correctness of the answer, after which he receives the feedback about the true correctness.

To enable this, we must know about the QALO correctness prior to the exercise. In case of exam questions, this information is available after teacher's evaluation. However, evaluating additional exercise (training) answers might become an extensive task. Because the automated evaluation of correctness of free text answers is not yet possible, the only option to substitute the teacher is the crowd - of students themselves.

In a simple scenario (see Figure 1), the student pulls the QALO and reviews it. After that, he sets correctness to it according to his opinion. Then he retrieves a feedback based on the previous answers of other students on this same QALO (the crowd answer). His own decision is then integrated into "crowd answer", eventually modifying it.

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



Question 1430:

Which architectural styles do we recognize in software engineering? (List at least 3 of them)

Provided answer:

client-server, layers

Estimate correctness ratio of the provided answer

Figure 1. Interface of the exercise application, consisting of question, answer, correctness and correctness estimation slider.

We deploy the method within an existing learning framework ALEF, perform live experiments and show that using such bootstrapping crowdsourcing approach, the method is able to correctly evaluate majority of the QALOs, leaving only ambiguous or controversial answers for evaluating by experts. Our aim in this work is to extend the approach:

Student domain expertise detection and use – we bootstrap the information about individual student level-of-expertise in the course domain and use this information for weighting student estimations to prefer more "skilled" users.

QALO correctness estimation revision – we extend the original user application giving student a chance to revise his answer correctness estimate after he is confronted with the crowd feedback and the discussion. We aim to investigate, how do students change their opinions and whether the revised crowd answer is better?

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 194-201.

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

References

[1] Downes, S.: E-learning 2.0. eLearn magazine. ACM, (2005), No. 10, p. 1.

[2] Quinn, A.J., Bederson B. B.: Human computation: a survey and taxonomy of a growing field. In Proc. of the 2011 annual conf. on Human factors in computing systems (CHI '11). ACM, New York, NY, USA, (2011), pp. 1403-1412.

[3] Stahl, G., Koschmann, T., Suthers, D.: Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences. Cambridge, UK: Cambridge University Press, (2006) pp. 409-426.

Modelling the Dynamics of Web Content

Matúš TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
matus@tomlein.org

Web content has a very dynamic nature, it frequently changes and spreads across various information channels on the web. The behaviour of web content can be observed and analysed, however it requires a lot of archived data to be able to draw any conclusions. It can also be a challenging task to efficiently analyse the large amounts of data and recognise the ways in which the content changes and spreads across websites.

On the other hand, the knowledge of the behaviour of web content is useful in many areas of software engineering. It can improve and optimize search algorithms for web content. Predicting when a website will change can be useful in web proxies that provide a cache of web objects. It can provide some basis for recommending similar content and also for prefetching websites.

The web content consists of different kinds of information that need to be recognised in order to process them. Some are less important and provide a lesser value to the user yet some are interesting and provide a real value. Both of these kinds of information can be present on a single website at the same time.

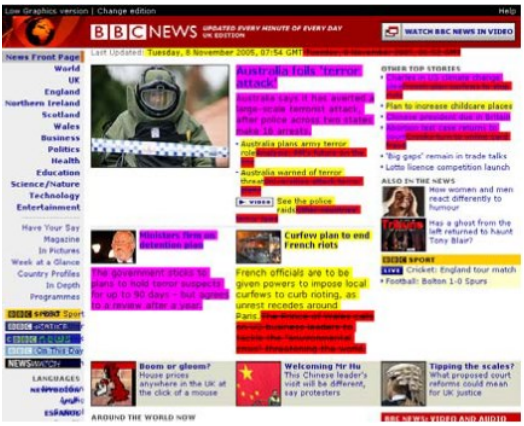
The user gets the most value out of information that creates the corpus or the main part of a website. That could be an article, a list of links to websites or any kind of relevant information. Other computer generated content, like the number of visitors to a website, the current date and time or advertisements provides a less significant value to the user. In most cases it is a good idea to filter such content when analysing a website.

Distinguishing these kinds of information is also important when tracking changes of a website. Computer generated content and advertisements tend to change frequently, however these changes are usually not interesting to the user or to the analysis of the website.

Apart from changes in the main content of a website, there is another kind of content that is potentially interesting to observe, and that is visitor-generated content such as discussions or polls. This content can add additional information to the website that might be worth analysing.

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

In terms of analysis of changes of web content, research has been done to recognise differences between two subsequent versions of an HTML document [2]. HTML represents elements as nodes in a hierarchy and although traditional algorithms for differencing text documents might work in HTML as well, it is advisable to employ an algorithm that can recognise changes in the hierarchy using tree comparison. Such algorithms can recognise insert, delete and move operations by differencing two HTML documents as can be seen on figure 1. To make the analysis aware of how content spreads across documents, further improvements need to be made. This is an open challenge to come up with a method that can effectively detect the attributes of a dynamic content.



Last Updated: Tuesday, 8 November 2005, 07:54 GMT

Australia foils 'terror attack'

Australia says it has averted a large-scale terrorist attack, after police across two states make 16 arrests.

- Australia plans army terror role
- Australia warned of terror threat

VIDEO See the police raids

Ministers firm on detention plan

The government sticks to plans to hold terror suspects for up to 90 days - but agrees to a review after a year.

Curfew plan to end French riots

French officials are to be given powers to impose local curfews to curb rioting, as unrest recedes around Paris.

- Plan to increase childcare places
- Chinese president due in Britain
- 'Big gaps' remain in trade talks
- Lotto licence competition launch

How women and men react differently to humour

Has a ghost from the left returned to haunt Tony Blair?

Cricket: England tour match

- Football: Bolton 1-0 Spurs

Boom or gloom? House prices anywhere in the UK at the click of a mouse

Welcoming Mr Hu This Chinese leader's visit will be different, say protesters

Tipping the scales? What proposed court reforms could mean for UK justice

Figure 1. Highlighting different kinds of changes on a website using the VDiff algorithm for differencing HTML content. Source: [2].

Analysis of the dynamics of web content can also be based on tracking terms and their use across websites [1]. Terms can define a temporal content that can appear on websites for a short time, for example based on current news. They can also describe a seasonal content that appears periodically in conjunction with some other recurring events. Observing the use of terms on websites can provide a basis for making connections between them and detecting the flow of information on the web.

The flow of information can also be observed in sharing multimedia content, such as videos or photos. These connections can be represented in a graph to show the flow of content across the web.

In our work, we aim to design and implement a method to effectively process Web content in order to be able to observe and analyse its behaviour in an archived data set. We plan to use the method on a sufficiently large data set of websites from various sources (e.g. blogs, social networks or news portals) to draw useful conclusions about the dynamics of such content.

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

References

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas (2009): The web changes everything: understanding the dynamics of web content. Web Search and Data Mining 2009, ACM, pp. 282-291.

[2] Rimon Mikhaiel and Eleni Stroulia (2005 Accurate and Efficient HTML Differencing. In Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP '05).

User Modeling, Virtual Communities and Social Networks

Combining the Power of Crowd and Knowledge of Experts in GWAP

Peter DULAČKA*

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

There are many problems in all areas of science needing human interaction in order to be solved – such problems cannot be solved accurately solely by computers. Games with a purpose (GWAP) are great tools for making use of the crowd for a very low cost. With fun as a motivation factor, people are willing to interact with the system and (even unknowingly) generate necessary data.

After a relatively successful game City Lights[1] which was able to rule out wrong user-created music tags, we realized that utilization of expert players is being suppressed in favour of crowd judgement which might be very costly when it comes to effective data acquisition. Experts are able to generate data which most users in the crowd are simply not able to generate, but when compared with the crowd's opinion, expert's opinions just seem to be wrong and are ruled out.

In our work we would like to tackle this problem and focus on:

- Discovery of experts in crowds in order to acquire accurate metadata.
- Realization of game with a purpose focused on music using expert's opinion to detect unseen – not obvious according to metadata – connections between songs.

In song relation discovery (hence also in music recommending systems) there are three types of discovery being used (and combined) [1]:

1. *Usage-based discovery* considering only listening habits of monitored user.
2. *Social-based discovery* considering listening habit and recommendations of user's socially close people.
3. *Content-based discovery* considering only data generated from audio analysis and valid metadata.

In the past people were dependent on record store clerks and radio DJs in order to discover new songs they might like [1]. After huge growth of digital libraries people

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

[1] http://citylights.rootpd.com

moved on the Web and depend mostly on online recommendation. By discovering experts in crowd, we would like to restore mentioned "musical authorities" and facilitate them into the online world.

The expert discovery in our game will be based purely on questioning and player's listening history. Players will be listening to a stream of music – an Internet radio – which is currently still very important source of music discovery among people [2]. We would like to target a group of users who listen to radio while working and persuade them to play from time to time when song they like is being aired. If we can sufficiently determine listener's knowledge domain, we could encourage him to play if the song from his knowledge domain is being aired – which could be motivating as player should collect more points than usual.

Players will be given tasks relevant to the song being streamed. The game will prepare questions for players (e.g. picking correct album cover, ordering of lyrics, rhythm repetition) which do have only a single correct answer – hence we should be able to determine the answers' correctness very accurately. Evaluating the answers/actions (and comparing with other players) in real time, we should be able to determine the player with the best musical knowledge playing the game at a time.

In addition to fact questioning, players will have an option to choose (enter) a song/artist/album they think is related to currently playing song. Non-expert players decisions will be used to confirm strength of relation between songs; expert players actions will be used to discover non-obvious song relations which can be later offered to non-expert players for confirmation. Our method depends on assumption that experts are able to create non-obvious relation with high success rate – so there is no need to focus on its validation.

However knowing that musical taste among people differs, some form of validation/selection has to be implemented anyway. A form of betting system could solve the issue and be a motivation/fun factor as well.

By letting players create musical playlists we not only might discover relations between two songs, but also between ordered set of songs and (if implemented correctly) also be able to describe (name) the relation as well. Expert finding is field mostly used in enterprise systems and to our knowledge had not been applied in games with a purpose yet. In the near future we would like to analyse approaches used in enterprise knowledge systems and try to incorporate them into our game.

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

References

[1] Celma, Ò., Lamere, P.: If You Like Radiohead, You Might Like This Article. *Association for the Advancement of Artificial Intelligence*, pp. 57-67, 2011.

[2] Komulainen, S., Karukka, M., Häkkilä, J.: Social music services in teenage life: a case study. *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, no. April, pp. 364-367, 2010.

Linked Data on the Web in Order to Improve Recommendation

Ľuboš DEMOVIČ*

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

Currently, the Web provides a large amount of knowledge, and also has the potential to become the largest source of information in the world. Data published on the Web is largely unstructured, intended for people, without a clear definition of entities, their meaning or relationships between them. Presently, a number of researchers are dedicated to processing unstructured data in the form of facts about entities from selected domains, which knowledge bases arose from, such as DBpedia, EntityCube, ReadTheWeb, YAGO-NAGA and others [3].

Linked Data principles describe a method for publishing structured data on the Web so that they are connected to each other in a way that makes them useful. In addition, Linked Data contains various variants of links between entities that make it possible to create graphs describing the selected domain. Promoting the importance of data represents the next stage of Web development referred to as Web 3.0.

One of the biggest challenges in the field of intelligent information processing is using the Web as a platform for integrating data and information [1]. By intelligent processing and linking of data on the Web, we could create useful new datasets, such as: quick search, translation, personalization, recommendation and user navigation.

In our work, we focus on machine-automated identification of new entities on the Web and discovery of relationships between them. A key technology to achieve this is known as extracting information [2]. This provides us with models, algorithms and tools for transferring web content, text and unstructured data sources into a comprehensible form. Commonly used methods are based on rules and patterns, natural language processing and statistical machine learning. However, existing methods do not guarantee the accuracy and relevance of the results. In addition, in our research work we want to focus on processing data in Slovak language, which brings a lot of obstacles and requires a different approach.

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

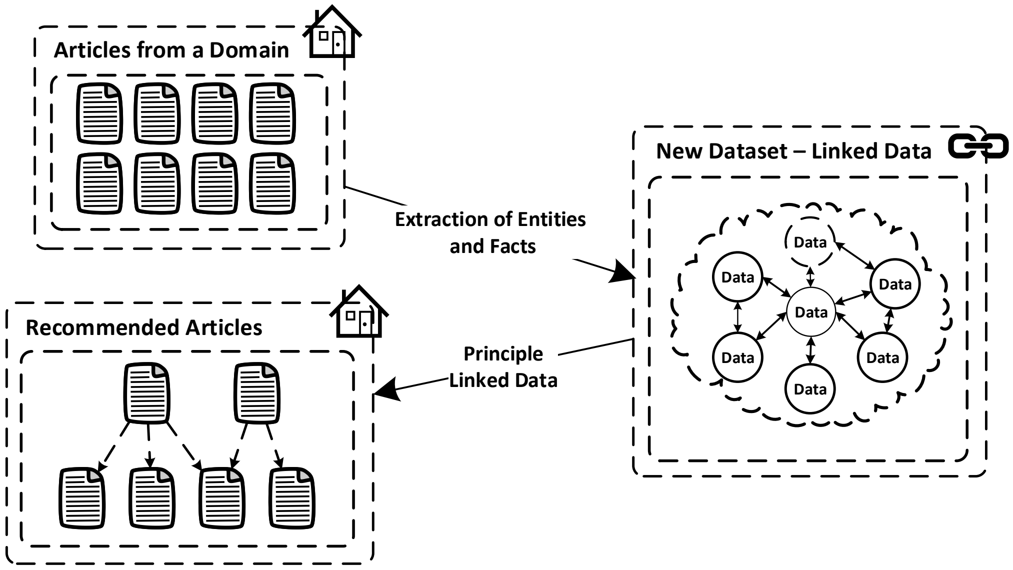


Figure 1. Using Linked Data on the Web in order to improve recommendation.

We deal with the analysis of automated machine processing of data on the Web in order to identify and extract entities and facts from Web content. We also deal with exploring the possibility of creating automated datasets obtained from the extracted entities and facts, using the principles of Linked Data.

The aim of our work is to propose a method that allows automated identification and extraction of entities and facts about them using lightweight semantics. Obtained facts will be used to create a dataset describing Linked Data from knowledge in the selected domain. We will verify the proposed method experimentally, by implementing a software tool that will exploit the knowledge base for recommendations. Figure 1 shows how a recommendation uses the principles of Linked Data.

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

References

[1] Auer, S., Lehmann, J., Ngomo, A. C. N. Introduction to Linked Data and Its Lifecycle on the Web. In *Reasoning Web: Semantic Technologies for the Web of Data*, LNCS 6848. Springer, pp. 1–75 (2011)

[2] Sahuguet, A., Azavant, F. Building Intelligent Web Applications Using Lightweight Wrappers. *Data & Knowledge Engineering*, Vol. 36, No. 3. Elsevier Science, pp. 283–316 (2001)

[3] Weikum, G., Theobald, M. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. In *Proc. of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Indianapolis, Indiana, USA. ACM Press, pp. 65–76 (2010)

Extracting Keywords from Educational Content

Jozef HARINEK*

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

When considering social educational systems, educational content has one advantage compared to other types of documents. With the emergence of e-learning 2.0 [2], it often has user annotations connected with it. These annotations are created by users who want to help themselves when interacting with the document. By processing these user created annotations assigned to the documents we can improve results of base Automatic Term Recognition (ATR) algorithms.

Base ATR algorithms give us results which are not as good as they could be. It means that there is still possibility to improve these results. The user-created annotations provide us with useful and personalized semantic information about the documents.

We propose a method for relevant domain terms (RDT) extraction based on user-generated annotations processing. We consider three basic annotation types (tag, comment and highlight). We compute the final term weight by combining relevant domain terms weights obtained from the individual annotation types and those obtained from the text.

Our method consists of the following steps:

1. Document and annotations pre-processing
2. RDT extraction from text and annotations
3. Combining the results from both sources

The final weight of the term is computed according to the following formula:

$$w_{\text{final}}(t,d) = (1 - p) * w_{\text{ATR}}(t,d) + p * w_{\text{annot}}(t,d) \qquad (1)$$

where the final weight w_{final} is computed as a combination of term weight computed only from the text of the document (w_{ATR}) and the term weight acquired from annotations for that particular term (w_{annot}).

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

Document pre-processing consists of extracting plain text from the document, lemmatization and stop-words removal. The annotations pre-processing is similar, but before lemmatization and stop-words removal we have to prepare "extended document of annotations".

This extended document of annotations consists of all the annotations connected with the particular document. We also take into account user proficiency level. The user ranking method which was used for the first experiments is a simple one which we are planning to substitute in our future work. It divides users into four groups, according to the number of annotations they added. It is based on assumption that the more user interacts with the documents the higher his level of knowledge is. Based on the group which the user falls into we add his or her annotations as one to four times more relevant to the extended document of annotations.

We have experimented with our proposed method. Our dataset consisted of 180 learning objects (documents), about 170 annotations per document and 1000 users. The dataset was taken from Principles of Software Engineering course presented in the learning system ALEF [4].

Our experimental results show that the annotations help in RDT extraction accuracy improvement. The first results showed that the most promising annotation types are tags (19.8 % improvement) and highlights (12.5 % improvement). Our final method yields improvement of 22.6 % in RDT extraction.

In further experiments we also want to take into account content of the comments. Our plans are to filter out irrelevant parts of the document by finding comments with such content. Next plan is to substitute the user evaluation method with a better one, based on HITS [3] or PageRank [1] algorithm.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 7–12.

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

References

[1] Brin, S., Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN systems,* 30(1), pp. 107–117 (1998)

[2] Downes, S. E-learning 2.0. *eLearn magazine*. Issue 10. ACM, p. 1 (2005)

[3] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM* (JACM), 46(5), pp. 604–632 (1999)

[4] Šimko, M., Barla, M., Bieliková, M. ALEF: A Framework for Adaptive Web Web-based Learning 2.0. In Reynolds, N., Turcsányi-Szabó, M. (Eds.): *KCKS 2010, IFIP Advances in Information and Communication Technology*, Volume 324. Springer, pp. 367–378 (2010)

Software Metrics Based on Developer's Activity and Context of Software Development

Martin KONÔPKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
martkono@outlook.com

Software development is extensive process which needs to be monitored and evaluated. It is important to evaluate it from the perspective of software product development as well as from the perspective of project management. The goal behind the monitoring of software project is to assess reaching desired qualities and attributes of the process and resulting product. The process of software development is very complex. It consists of various activities difficult to monitor resulting in ineffectiveness of the project management. Therefore, special metrics were developed to evaluate software project and to identify the problems easily [1].

We may take various points of view for monitoring the project, varying from planning to resource management, or mostly the work done by the developers. It is the quality of development which affects external product attributes like fault-proneness, maintainability, security, reusability or other external attributes of resulting software product [1]. The success in reaching the desired values of these attributes are affected by the quality of the source code produced in development process.

Software product is changing during its development process. Changes in source code bring not only new functionality or bug fixes but are also source of new bugs. Software fault-proneness is mostly caused by faults of human factor. By studying the periodicity of changes in source code we may predict which modules are fault-prone, even using primitive metrics (the higher the number of changes, the higher the chance of fault-proneness) can be effective in this matter.

Another approach is to evaluate software project with internal product metrics [1], namely the size, structure and resources. Even though these product metrics evaluate software product on its lowest level, they suffer from the ambiguity in interpretation exclusively for every project. It is up to managers to find out what values of these metrics they want to observe and how they relate to the product attributes. High number of source code lines or high coupling between objects may be required for one project but for other it may be viewed as bad approach.

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

Because of these problems with existing software metrics we intend to take into consideration developer's activity performed during the development process and the context of software development to find the connection with attributes of created product. Developer performs various activities during the development process:

- Activities associated with development directly, e.g., programming or modelling of components,
- Activities associated with development indirectly, e.g., searching for information, studying documentation or communicating with team members,
- Activities not related to development and context of the developer, e.g., emotional state or state of the environment.

Our project is part of the research project PerConIK (Personalized Conveying of Information and Knowledge, perconik.fiit.stuba.sk) [2]. In PerConIK project we log these activities during the software development and allow developers and managers to tag the source code with information tags. We intend to use these logs and tags to evaluate how they relate to the attributes of the created software product. Activities like frequent browsing of the Web during the development process, copy-pasting code instead of sharing it between components, developing during night hours and other activities and context falling into mentioned categories may have impact on the resulting attributes of the software product. Our prediction may be anchored to the source code with newly created information tags and used by the developers and managers, e.g., to locate problems in the source code.

We plan to take results of existing product metrics based on the source code and manual reviews of the source code for evaluation of our approach of evaluating software product.

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.

References

[1] Fenton, N.E., Pfleeger, S.L.: Software Metrics: A Rigorous and Practical Approach (2nd Edition). PWS Pub. Co., Boston, MA, USA, 1998, Feb., 24, 640 p.

[2] Bieliková, M., Návrat, P., Chudá, D., Polášek, I., Barla, M., Tvarožek, J., Tvarožek, M.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. In: Procedia Technology, 3rd World Conference on Information Technology, 2013.

Extracting Word Collocations from Textual Corpora

Martin PLANK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
plank09@student.fiit.stuba.sk

Natural language is the main way of communication between people. They use it for asking and answering questions, expressing opinions, beliefs, as well as talking about events, etc. And they communicate in natural language on the Web, too. However, the simplicity of creating the Web content is not only the advantage of the Web, but also its disadvantage. It is expressed in natural language, which means that it is usually unorganized and unstructured. This makes processing of the Web content expressed in the natural language difficult.

Difficulties in natural language processing are often connected with ambiguity of the language. Some words have specific meaning, when they are used together in one sentence. This raises the problem of collocation extraction. Detection of collocations is important for various tasks in natural language processing (word sense disambiguation, machine translation, keyword extraction, etc.). Many statistical methods, as well as other natural language attributes (e.g., part of speech) are used to resolve this task.

Pecina [3] argues that natural language cannot be simply reduced to lexicon and syntax. Individual words can be combined in various ways. This fact is common for most natural languages. The term collocation has several definitions. Choueka [1] defines a collocational expression as "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components".

During the last 30 years, several association measures were proposed for automatic collocation extraction. The most of the methods are based on verification of typical properties of collocations [3]. It is possible to mathematically describe these properties and determine the degree of association between the components of a collocation. These formulas are called association measures. They compute association score between all collocation candidates in a corpus. The score indicates the likelihood that a candidate is a collocation. These measures can be used for candidate ranking or for classification (if there is a threshold).

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

The advantage of these methods is that they can be combined together and final association score can be computed using several measures. Pecina [1] compares 84 statistical measures for collocation extraction. The best results are achieved by the pointwise mutual information. He proposes also method, which finds linear combination of selected methods that improves the performance significantly.

Other approaches employ methods based on the linguistic properties of collocations (e.g., [4]). Manning and Schütze [2] describe the three properties:

- Non-(or limited) compositionality. The meaning of a collocation is not a straightforward composition of the meanings of its parts. For example, the meaning of 'red tape' is completely different from the meaning of its components.
- Non-(or limited) substitutability. The parts of a collocation cannot be substituted by semantically similar words. Thus, 'gut' in 'to spill gut' cannot be substituted by 'intestine'.
- Non-(or limited) modifiability. Many collocations cannot be supplemented by additional lexical material. For example, the noun in 'to kick the bucket' cannot be modified as 'to kick the {holey/plastic/water} bucket'.

Wermter and Hahn [4] present method based on the non-(or limited) modifiability. The method is built on assumption, that context of a collocation is particularly characteristic. They experiment with this method and compare it to the basic statistical methods. Proposed method significantly outperforms these statistical methods.

In our work we focus on extracting collocations in the Slovak language. We analyze several methods for collocation extraction. Our goal is to adapt or improve existing methods and explore collocation properties in the Slovak language. The important choice is whether to focus on statistical methods measuring co-occurrence between word n-grams or linguistic methods. The statistical methods might be simpler, but one the other hand, they are not based on linguistic collocational properties. Wermter and Hahn [4] show that exploring these properties is a reasonable approach, too. In addition, these methods are able to outperform statistical methods.

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

References

[1] Choueka, Y. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In *Proc. of the RIAO*, CID, pp. 609–624 (1988)

[2] Manning, C. D., Schütze, H. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA (1999)

[3] Pecina, P. An extensive empirical study of collocation extraction methods. In *Proc. of the ACL Student Research Workshop*. ACLstudent '05, Association for Computational Linguistics, pp. 13–18 (2005)

[4] Wermter, J., Hahn, U. Collocation extraction based on modifiability statistics. In: *Proc. of the 20th int. conf. on Computational Linguistics.* COLING '04, Association for Computational Linguistics (2004)

Discovering Links between Entities on the Web of Data

Ondrej PROKSA*

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

A few million unique websites appear on the Web every day. Information on them is usually published in an unstructured format. Linked Data is structured data which contains entities and relationships between them, which are available on the Web. Some datasets are made via automatized processing of freely available data. These are useful for personalization, web search or for knowledge deduction. One of the main problems is the conversion from various unstructured datasets to a uniform format and the linking of the data to existing datasets.

The ontologies behind Linked Data sources, however, remain unlinked. They describe an extensional approach to generate alignments between these ontologies [1]. They present an extension of the YAGO knowledge base with focus on temporal and spatial knowledge. YAGO contains nearly 10 million entities and events, as well as 80 million facts representing general world knowledge [2]. The goal is to automatically construct and maintain a comprehensive knowledge base of facts about named entities, their semantic classes, and their mutual relations as well as temporal contexts, with high precision and high recall [3].

In this work we analyze the issue of mining structured data from various sources available on the Web and the issue of linking the mined data in order to create a domain knowledge base. We analyze various approaches to automatized dataset creation, gathering information about named entities and linking of the entities and integration of new datasets with the existing ones. We propose a method to automatically process chosen sources of unstructured data and create a structured knowledge base, which is based on the Linked Data principles.

The designed method is experimentally evaluated on data from a chosen domain by implementing a software prototype, which uses the knowledge base for a chosen problem from the field of Web personalization – search, navigation, recommendation based on relationships between entities. We validate the created knowledge base by comparing it to other existing knowledge bases.

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

We divide our work into the following parts:

1. Creating structured data – selection of data source, discovering facts about entities
2. Creating a dataset – identifying relationships, elimination of duplicate entities, linking entities in created dataset, linking dataset with selected existing datasets
3. Verification – evaluation of facts about entities, automatic answering of search queries

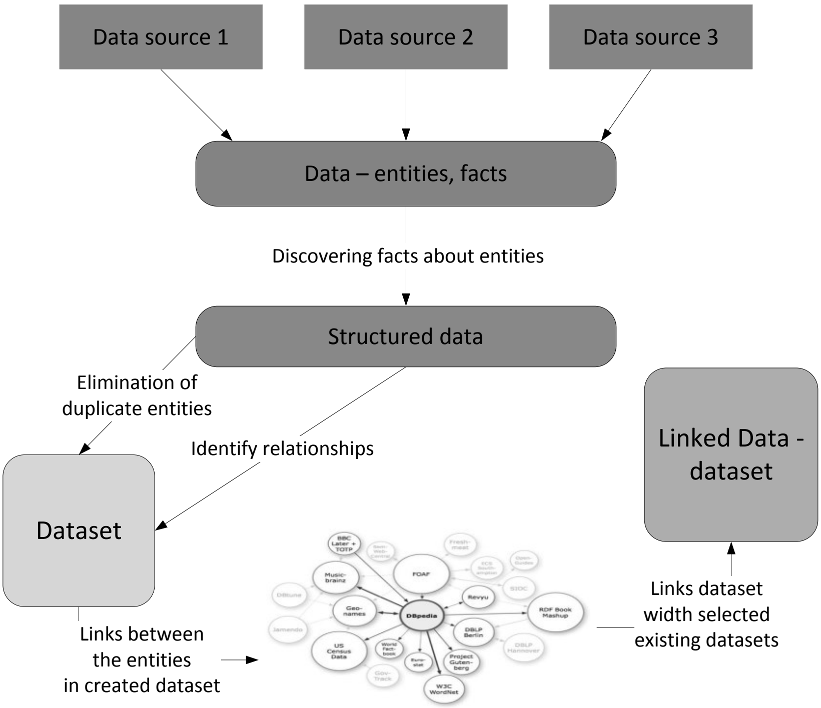


Figure 1. Our work divided into parts.

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

References

[1] Parundekar, R., Knoblock, C. A., Ambite, J. L. Linking and building ontologies of linked data. In *Proc. of the 9th Int. Semantic Web Conf. on The Semantic Web, Vol. Part I* (ISWC'10), Springer-Verlag, Berlin Heidelberg, pp. 598–614 (2010)

[2] Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., Weikum, G. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proc. of the 20th Int. Conf. Companion on World Wide Web* (WWW '11). ACM, New York, NY, USA, pp. 229–232 (2011)

[3] Weikum G., Theobald, M. From information to knowledge: harvesting entities and relationships from web sources. In *Proc. of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS '10). ACM, New York, NY, USA, pp. 65–76 (2010)

Relationship Discovery from Educational Content

Petra VRABLECOVÁ*

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

The domain model is an essential part of the adaptive learning system. It expresses the semantics of educational content in the form of metadata. We consider it to be a lightweight ontology, i.e., a set of terms and term relations. Manual domain model building is a challenging task for teachers, hence there is an effort to automate it. We propose a method for automated acquisition of metadata from educational content, aimed at relationships discovery between terms.

There are many generic methods for automatic metadata acquisition from text, which are based on natural language processing. We decided to explore the statistical approach and its advantages, like no need for syntax knowledge and language independence. The uniform vocabulary of educational texts indicates better results of statistical methods. Only few works deal with automated domain model acquisition for adaptive systems and course authoring support. These consider another specific of educational content – its structure allowing also the usage of graph algorithms.

The content of adaptive learning system is a set of learning objects (LO) – mainly text documents, formed into a hierarchy (a tree or a book structure). They are linked through LO-LO relationships implying their relatedness. The purpose of our method is the automated creation of a lightweight ontology which will describe our set of LOs. It consists of a set of relevant domain terms (RDT) and relationships of different types between them. RDTs are assigned to LOs through an RDT-LO relationship that implies the semantic link between them (e.g., the term describes a learning object).

Our method is preceded by the LO preprocessing, e.g., normalization, removal of stop words, lemmatization. Besides preprocessed texts, a set of RDTs and RDT-LO relationships are needed as input. They can be extracted from LOs (e.g., by tf-idf). RDTs have to occur in the text of LOs, because the first step of our method is the application of Latent Semantic Analysis (LSA) on the preprocessed texts and RDTs. LSA produces relationships based on similarity of words surrounding RDTs in texts. The output is a net of related RDTs, i.e., set of RDT-RDT relationships. The second

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

step is the discovery of hierarchical RDT-RDT relationships which comprise the core of the domain model. We propose two methods for their determination.

The first one is based on term subsumption [2]. The sets of LOs, which have assigned RDTs from an LSA relationship identified in the first step, are compared. The sets also contain LOs of RDT's neighbors from the LSA net. If sets' intersection is not empty then the RDT belonging to the bigger set is marked as superordinate (Figure 1).

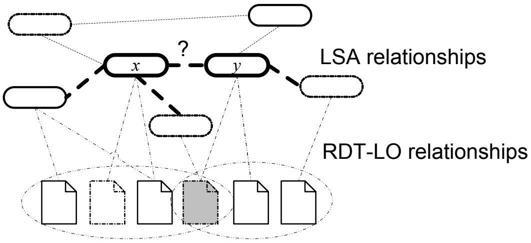


Figure 1. Determination of relationship type between terms x and y.

The second method is based on the Semantic Growback algorithm [1]. For two RDTs from an LSA relationship, are lists of top ranked RDTs created. The lists are produced by application of the PageRank algorithm with priors on the graph of LO-LO and RDT-LO relationships. If both RDTs are in both lists then the RDT which is on a higher position in both lists is marked as superordinate.

At the moment, our work is in the phase of evaluation. The goal of the evaluation is to find out whether the domain model built by our method is on the level of the manually built domain model. We perform tests on the Functional and Logic programming course. We experiment with various setups of the method and look for the optimal combinations. The preliminary results suggest that the most valuable contribution of the method is that it yields different kinds of relationships that cannot be discovered by applying linguistic approaches. The evaluation process also contains the integration of our method into an educational content management system.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 55-60.

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

References

[1] Diederich, J., Balke, W. The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In *Proc. of the 11th European Conf. on Research and Advanced Technology for Digital Libraries,* LNCS 4675. Springer, Berlin, pp. 1–13 (2007)

[2] Sanderson, M., Croft, B. Deriving concept hierarchies from text. In *Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, ACM, pp. 206–213 (1999)

Beyond Code Review: Detecting Errors via Context of Code Creation

Dušan ZELENÍK*

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

Every human has his own patterns in behavior. Everything we do has its reason. We react to different situations, different states. Therefore has the environment, in which we exist, huge impact on our outputs. When we talk about outputs, we have to mention our productivity and efficiency in work. It is definitely affected by our current situation and surroundings. Every human who is trying to accomplish something has to be focused on the task. Usually only experts repeating the routines do not need to focus so much that they are able to work in any state of their environment as it was discussed by Milton [1].In our work we focus on software developers and their productivity and effectiveness which is influenced by surroundings. Programmers do lot of mistakes when they are not in the shape. We are going to prove this assumption by analyzing their behavior while developing software. In our work we focus on two main aspects which affect the quality of programmer's outputs.

- *Continuity of work*. This means that programmer is working in continuous time and he is not interrupted by external happening. Interruptions could cause problems with refocusing on the task and reconstructing the situation. This leads to loosing time and loosing context and eventually leads to mistakes and incomplete tasks.
- *Stereotyped work*. When programmer works in common situations or in the common environment, he is used to conditions what positively affects his outputs. This also means that programmer needs some sort of stereotype in work. Loosing the stereotype brings anomalies in his behavior that causes anomalies in his outputs. Anomalies in outputs could emerge into mistakes.

Measuring the quality of programmer's outputs is not trivial issue. There are many different metrics which we could use. However most of them are not very precise and mostly incomparable among different programmers. Simple metric would be calculating the number of lines of operations in code which was created during specific

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

amount of time. However, this only calculates the quantity but not quality of outputs. We assume that we should use two metrics.

- *Commits.* Commit is an action of programmer which is done regularly to submit some part of the work. Commits submitted into source control system are treated as logical end of programmer's effort. Commits are usually executed when something is accomplished. We consider number of commit as metric which shows the success.
- *Bugs.* Even if programmer has done something, we could cause some mistake which has, however, been revealed later. We are counting commits which are fixing bug associated with previous commits. Such a bugfix commits are negative metric which enables us to express low quality of programmer's work.

We are going to support the process of code reviewing by identifying part of codes which were created in bad conditions and are probable to contain some mistakes. Our approach lays in determining the context of programmer which could be associated with lower quality of outputs. We analyze programmer's activity by tracking his behavior. Next step is to detect parts of code which were created in this context.

In the work by Czerwinski et al. [2] authors analyzed behavior of workers. The nature of tasks of these workers was rather parallel. Multitasking suggests that these workers cannot work continuously on single task. They switch among tasks. This leads to frequent interruptions. The point of their work is to show how workers react to these interruptions. They want to design software for task management which is fond of bad habits caused by interruptions. Workers observed in this study are programmers. They work in Matlab but their tasks are usually fragmented due to secondary activities such as reading emails or preparing presentations. The study also showed that only 18% are dedicated to projects and productive tasks. The rest of the activities are secondary. 7% of total tasks are those which were interrupted by other tasks. 40% of tasks were self initiated, what means that user was not interrupted by eexternal influence but on his own. Rest of the interruptions was external.

Our method is based on detecting interruptions and anomalies in programmer's coding manners. Knowing when the programmer was interrupted and when he was working in inappropriate states we mark code which was produced. We mark this code with probability calculating while discovering interruptions and anomalies.

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.

References

[1] Milton, J., Solodkin, A., Hluštík, P., Small, S.L.: The mind of expert motor performance is cool and focused. *NeuroImage*, 2007, vol. 35, no. 2, pp. 804-813.

[2] Czerwinski, M., Horvitz, E., Wilhite, S.: A diary study of task switching and interruptions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '04, New York, NY, USA, ACM, 2004, pp. 175-182.

Workshop Events Reports

Experimentation Session Report

Róbert MÓRO, Ivan SRBA

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

1 Motivation

In research it is not enough to come up with a brilliant idea and implement it. The proposed method or solution has to be verified, measured and/or compared to existing solutions. Evaluation is, therefore, perhaps the most important phase of every research. It requires much effort: researchers must formulate good hypotheses that are testable, verifiable and relevant, then design an experiment (or a series of experiments), find willing participants and evaluate results.

One of the main goals of our research group PeWe is to help its members with their research and its evaluation. The members of the group, many of them struggling with evaluation of their own research ideas, help their peers by providing not only constructive criticism and advice for experiment design; they also actively participate on experiments of others. Finding participants for experiments on our own can prove very hard in our conditions, therefore, to have colleagues willing to help is one of the most valuable benefits of being a member of PeWe.

Of course, we should in any case try to evaluate our proposed solutions with broader spectrum of participants that have ideally no previous knowledge of our research or use a golden standard (if available) in order to compare results of our methods with already published ones. Nevertheless, conducting experiments even in our small group can be very useful for initial evaluation, to verify preconditions of our methods or locate errors in our design that help us in the end to achieve better results.

2 Experiments

Experimentation session during the *Spring 2013 PeWe Workshop* consisted of 9 experiments that together required active participation of 2 000 minutes resulting in 50 minutes on average for each workshop participant. The participants in many cases took part on more than one experiment; they played a game, navigated in the digital library, annotated and organized their documents, learned new vocabulary, enriched wiki with semantics, provided feedback or ordered named entities. The session lasted for several hours and officially ended at 1am with some of experiments continuing even after that.

Spring 2013 PeWe Workshop, April 5, 2013, pp. 113–119.

2.1 Trending Words in Navigation History for Term Cloud-based Navigation

Experimenter: Samuel Molnár, Experiment supervisor: Róbert Móro

The key part of our experiment was to evaluate if history based query refinement provided by our method is beneficial for user's navigation in information space. Secondly, we observed how users interact with our navigation interface and how they navigate when their information need is more general. Our method takes advantage of history visualization based on different shades of colour to emphasize last usage of words in history, but since users tend to choose elements that are visually distinctive, we decided not to exploit colour in the cloud in order to get more relevant results.

We asked 4 participants to navigate for 20 minutes by our method integrated into Annota (see Fig. 1). Their task was to navigate in research articles concerning given topics they were not experts in. We observed that every participant chose to use fulltext search with given topic and then navigate in results by cloud.

Navigation queries constructed during sessions were not sufficient for query refinement evaluation, but we inquired users during discussion to evaluate relevance of the resultant articles and whether they suggest any improvements for navigation interface. Since the dataset from Annota for given topics was not large, users evaluated the same articles as relevant. They also proposed few notable improvements for the interface such as more detailed description of the article.

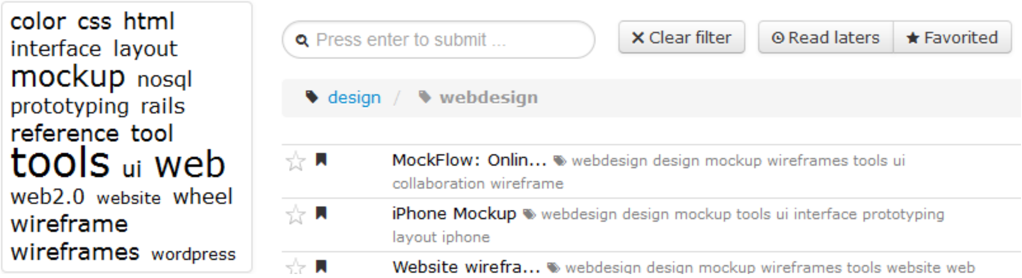

color css html
interface layout
mockup nosql
prototyping rails
reference tool
tools ui web
web2.0 website wheel
wireframe
wireframes wordpress
Press enter to submit ...
✕ Clear filter
Read laters
★ Favorited
design / webdesign
MockFlow: Onlin... webdesign design mockup wireframes tools ui collaboration wireframe
iPhone Mockup webdesign design mockup tools ui interface prototyping layout iphone
Website wirefra... webdesign design mockup wireframes tools website web

Figure 1. Word cloud navigation in Annota.

2.2 Personalized Web Documents Organization through Facet Tree

Experimenter: Roman Burger, Experiment supervisor: Mária bieliková

Our experiment was the first session out of two evaluating facet tree as a new concept of organizing personal information libraries. The goal for the users in the experiment was to search for article in digital library and then store it in the web bookmarking service Annota. All participants were asked to archive articles relevant for selected topic using traditional hierarchical approach and also using facet tree method. We measured the time needed to archive an article.

Our hypothesis is that facet tree method is more effective in storing and retrieval of information when working with personal information libraries. We also suppose that recall of information shall not be worse.

Due to design of storing information in facet tree method, this way was always faster (and thus more effective). In the second session of the experiment participants will be asked to retrieve stored information. Second session will show if storing and retrieval combined is still more effective. Preliminary observations also showed that originally planned feature of facet colour will not suffice. We have redesigned this facet to be keywords instead.

2.3 Related Document Search

Experimenter: Jakub Ševcech, Experiment supervisor: Mária Bieliková

We have performed an experiment to evaluate the method for related document search using user created annotations as user's interest indicators. The experiment consisted of two parts: questionnaire about user's habits when annotating documents and annotation of chosen document followed by rating of relevancy of retrieved documents.

The experiment was conducted by 7 participants of whom the majority is using annotations while reading printed or electronic documents. Based on usage of annotations we can divide the participants into two groups: those who use annotations for summarizing and describing studied documents and those who use annotations as means to store their thoughts about the studied document.

We presented to participants two sets of retrieved related documents and we asked them to evaluate their relevancy. Both sets were retrieved by the proposed method, one using annotations attached to the document and one without these annotations. Participants annotated source document and evaluated relevancy of retrieved related documents for ten source documents. We found out that evaluated method retrieves more documents related to the source document when using annotations and the found documents are more relevant to the theme of the source document the user is most interested in.

2.4 Games, Motivation and Personality

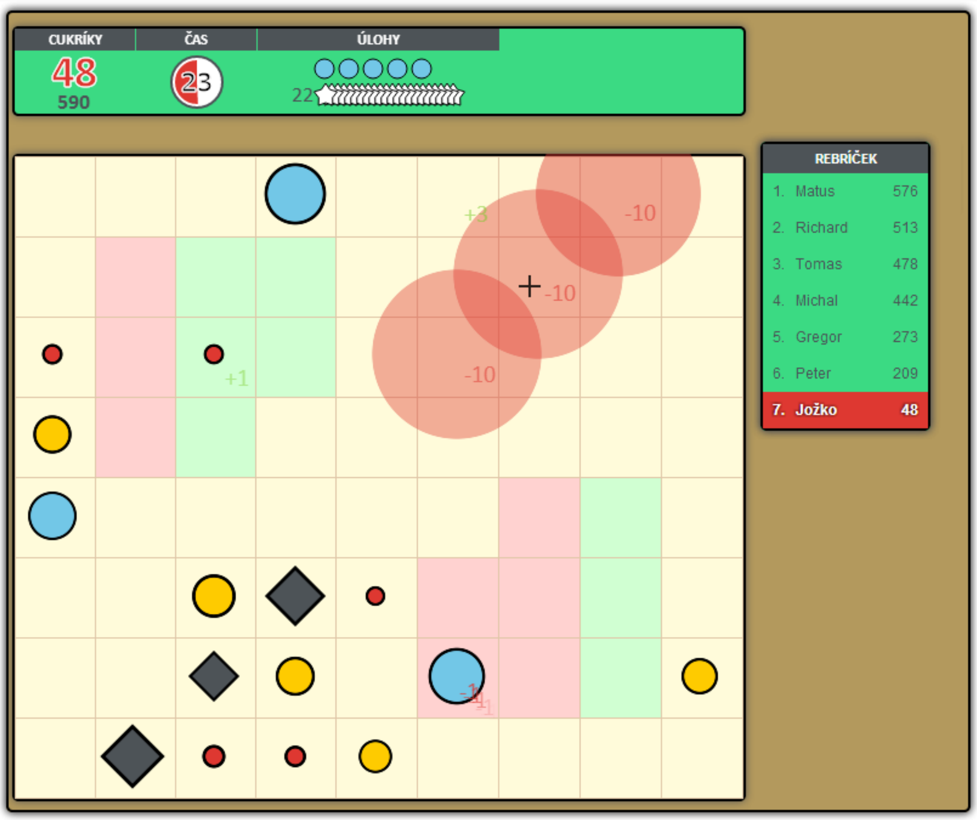
Experimenter: Peter Krátky, Experiment supervisor: Jozef Tvarožek

This experiment was a part of evaluation in our work named *User Modeling Using Social and Game Principles*. In this work, we are interested in personality model construction based on player's activity in a computer game. The goal of our experiment was to explore how user interface actions are influenced by player's personality traits and how various game elements (points, challenges and more) work with different personality profiles.

We implemented a casual browser game (see Fig. 2) to track both user interface actions and game actions. Participants were able to connect to the game server via Wi-Fi using their own computers anytime during the experiments session. They were divided into four groups having game with different game elements turned on/off what was preparation for further analysis of elements' impact on game engagement. All of the participants were also asked to fill in Big Five personality questionnaire.

We collected non-trivial dataset from 22 participants. Each of them devoted around 30 minutes to our experiment (including 10 minutes for questionnaire). During their game play we also asked about their motivation in the game and we found diversity. Answers included top position in the leaderboard, completing challenges or exploring what is in the next level.

In the next stage of the experiment we will process the data into numeric form and we will analyse relationship between personality traits and both user interface actions and game play style using linear models.



1.	Matus	576
2.	Richard	513
3.	Tomas	478
4.	Michal	442
5.	Gregor	273
6.	Peter	209
7.	Jožko	48

Figure 2. Browser game used for evaluation.

2.5 Influence of Emotions on User Actions

Experimenter: Máté Fejes, Experiment supervisor: Jozef Tvarožek

Goal of our project is to propose a method for user modelling based on users' emotions. We observe the user via webcam during work in a given software or information system. Processing of the obtained sequence of images is done by our tool for facial expression recognition. This tool is designed for determining emotion state of subject by the help of facial expression analysis. The recognition output is a representation of the actual emotional state of the user.

As an initial evaluation we performed a controlled experiment. We observed users while playing the same computer game as described in section 2.2. We hypothesise that a correlation exists between users' performed actions and emotional states. We used data obtained in experiment to map each emotional state to a domain specific activity. In this domain, we mapped emotional states to score, number of played games and other relevant information. Our goal was to obtain a model that expresses correlation between emotional states and success in the game.

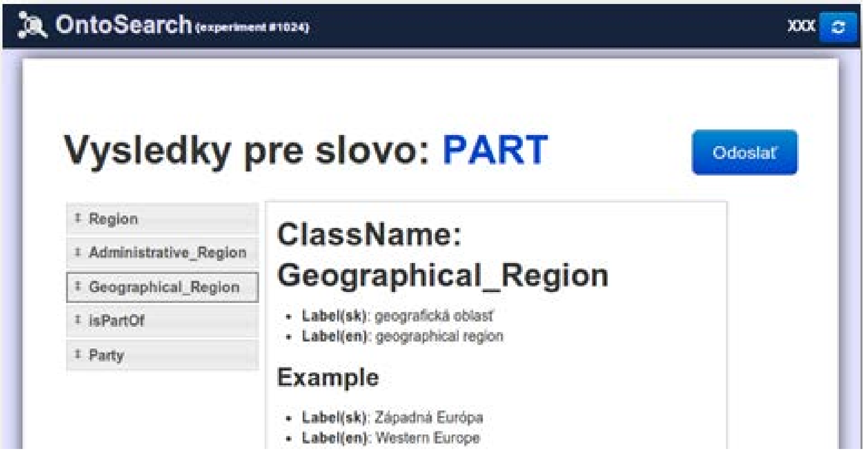
For technical reasons, we chose frequency of capturing at 500 milliseconds. Since the obtained dataset contains images of bad quality and also empty frames, we were able to detect face from 11285 out of 32858 images. We plan to use the mentioned mapping of the collected data for training the system to be able to predict or evaluate users' activities depending on his/her actual emotional state. We expect that a set of various behavioural patterns depending on emotional state will be shown.

2.6 Weighting System for Querying Dataset with Natural Language

Experimenter: Peter Macko, Experiment supervisor: Michal Holub

In our project we propose a method for finding alternative names for entities and relations in ontological database. These descriptors come from dataset definitions and WordNet database. Because of many different alternative names, we proposed a weighting system which we wanted to evaluate. We prepared a small experimental web page (see Fig. 3) where users were able to order descriptors for one entity. Users had to order up to four words for one entity name in right order. The users submitted 336 orderings for 30 terms. 89.3% orderings were in the right order; others had some mistakes in order.

We prepared the second experiment as well in which we used terms describing more than one entity. Then we showed users a term and entities which are linked to it. We showed users five terms and for every of them four classes which they were supposed to order. In this experiment, users managed to do 85 orderings, 68% of which were in the right order. We had in both experiments the same user group containing of 17 students. The terms form the first and the second experiment were shuffled, therefore the users did not know that they were taking part in two experiments.



Vysledky pre slovo: PART

ClassName: Geographical_Region

- Label(sk): geografická oblasť
- Label(en): geographical region

Example

- Label(sk): Západná Európa
- Label(en): Western Europe

Figure 3. Experimental interface for entities ordering.

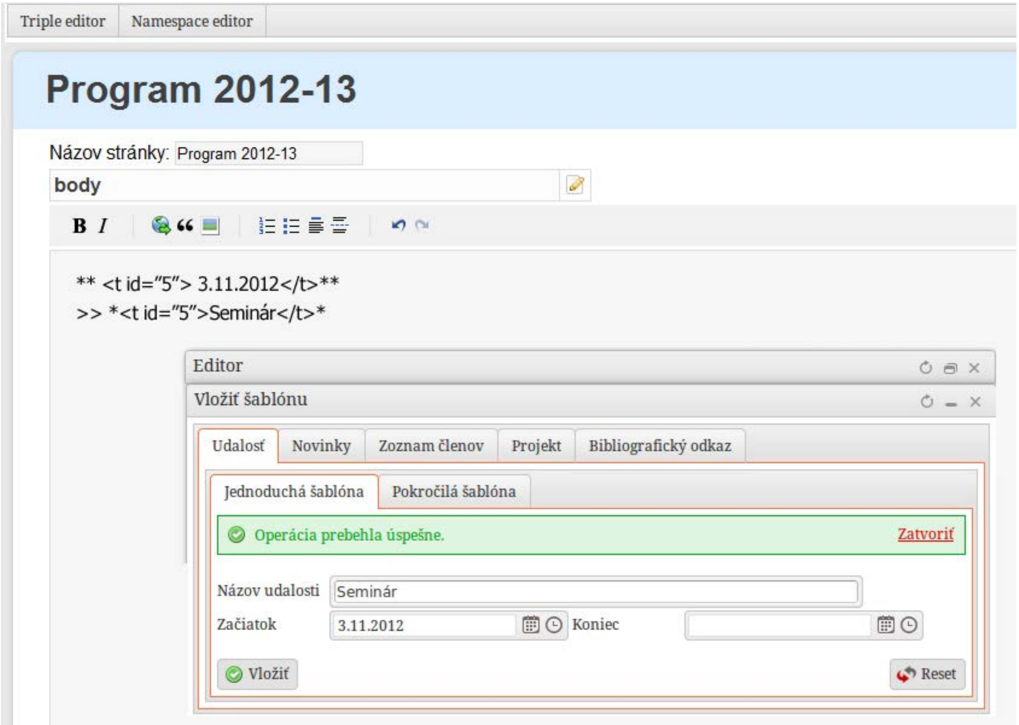
2.7 Semantic Wiki Testing

Experimenter: Martin Markech, Experiment supervisor: Jakub Šimko

Aim of our experiment was to confirm or reject our hypothesis and test usability of our semantic extension (see Fig. 4). Our hypothesis says that with our semantic extension, usability of application and readability of markdown code does not get worse too much.

We used qualitative methods in experiment. Each of our two participants spent with us about 50 minutes. During this time they gave us valuable feedback how to improve UI to be more intuitive and more usable. They answered that the markdown without our extension is more readable – which is expected - but also that syntax highlighting and grey colour of our triplet marks help the enriched code to be readable and they are able to identify raw text in the markdown editor. It supports our hypothesis that readability of markdown code does not get worse too much.

Ideas how to improve usability of our extension are very valuable for us, because we found out that in the current state our extension was not too intuitive for users. In the beginning the participants were a little distracted. Ideas which they gave us are mostly easy to implement; therefore we will try to implement most of them.



Program 2012-13

Názov stránky: Program 2012-13

```
** <t id="5"> 3.11.2012</t>**
>> *<t id="5">Seminár</t>*
```

Operácia prebehla úspešne.

Názov udalosti: Seminár

Začiatok: 3.11.2012

Koniec:

Figure 4. Using semantic extension to edit markdown.

2.8 Adaptive Feedback in a Web System

Experimenter: Marek Grznár, Experiment supervisor: Martin Labaj

The main purpose of our experiment was to find out whether adaptive feedback increases quantity of object ratings. In our experiment, we focused on difficulty ratings for learning objects. Our method of adaptive feedback is based on combination of explicit and implicit feedback. With implicit feedback, we track the user and evaluate his/her behaviour and at the right moment we request difficulty rating for the object (explicit feedback).

The participants solved exercises and questions in C course and evaluated their experience with these learning objects. In our experiment, we divided participants into two groups. One group solved the tasks with adaptive feedback and the other solved the task with passive explicit feedback. Secondly, in adaptive feedback, we looked for actions which triggered the explicit rating of the object and whether the rating was rejected or not after that particular action.

16 participants attended our experiment. They solved exercises and questions for 12 minutes in ALEF (Adaptive LEarning Framework). After 12 minutes we asked the participants several question about their ratings. We observed that participants with adaptive feedback evaluated more learning objects. Participants with adaptive feedback rated the objects most of the times after requesting the rating from them. On the other hand, participants with passive explicit rating rated object difficulty rarely.

2.9 Augmenting the Web for Facilitating Learning

Experimenter: Róbert Horváth, Experiment supervisor: Marián Šimko

We conducted an experiment in which participants were asked to browse any news on a selected domain. Webpage content they saw was augmented with French words in order to memorize them. In this experiment we tried to evaluate our method used for personalized augmentation and its impact on learning foreign language vocabulary and web browsing experience.

We were very pleased by participants' willingness and thanks to it we gathered twice as many results as we had expected. Every participant was asked a set of questions after their short web browsing session. Thanks to their answers we were able to evaluate our method and we obtained very useful feedback. We got not only positive feedback to our method, but it seems that data gathered during this short experiment gave us much more information than we already had from other long term experiment.

Positive finding was that users were able to learn a few new words and they were able to translate them correctly in a vocabulary test. Another important finding was that even though they spent more time on an augmented webpage in comparison to non-augmented, they still found it tolerable and they enjoyed browsing the Web.

3 Summary

We were positively surprised by the enthusiasm of all the participants and by their stamina that allowed them to take part on experiments even long after midnight. Whether they participated for 20 minutes or two hours, everyone helped, providing together 30 hours' worth of experiments. This spent time is much appreciated and we believe that it will result in better bachelor and master theses of all the experimenters as well as in internationally recognized publications in some cases. We also hope that this kind of collaboration will continue in the future.

Acknowledgement. The authors wish to extend their thanks to all the experiments' participants. Special thanks go to SeBe organizers for providing motivation in the form of prizes for the most active participation. Lastly, we would like to thank our professor Mária Bieliková for organizing this whole event; without her effort and motivation it would not be possible.

Discussion Club at PeWe Workshop

Jakub ŠIMKO, Dušan ZELENÍK

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

1 Motivation

To engage all participants into discussion in this edition's PeWe workshop we have incorporated a two-hour session called *discussion club* into the programme. During this session eight small groups of workshop participants discussed various topics to which they have been assigned. The groups have been composed of students of different levels of seniority (ranging from senior doctoral students to junior bachelor students) to maximize knowledge exchange potential. The discussed topics were related to research methods (in particular the experimentation) and technologies. At the end of the session, each group presented resumé of their discussion.

2 Team reports

2.1 Implicit and explicit user feedback

Team leader: Martin Labaj

The discussion group focused on implicit and explicit user feedback starting with quick overview of user feedback and means for collecting it – from various interest indicators in implicit feedback, to types and sizes of scales in explicit feedback, questionnaires, etc. We selected three domains of interest – adaptive e-learning systems, programmer support and evaluation in a software company, and a movie recommendation. In these domains, implicit and explicit emotional/state evaluation (implicit webcam based emotion recognition and explicit slider-based mood input) was selected as a topic for further discussion. Here, it was also discovered that participants engaging in "casual" domains (movies) preferred explicit feedback whereas participants doing research in domains like learning and programming tend to rely more on implicit feedback.

We then discussed user incentives to provide or not to provide feedback in these domains and reliability of such feedback. It was agreed that implicit feedback in the domain of movie recommendation is somewhat limited due to the fact that playing or even replaying the video does not simply imply that the user likes it. Even camera-based emotion recognition is of limited use as the user can like a romantic movie, which, on purpose, induces sadness. On the other hand, the user has little motivation to provide false explicit ratings and can only get better recommendations from honest ratings. In the domain of e-learning and programming, however, the user can have strong motivation to evaluate difficulty or quality of exercises or program code in a way to look better in front of teachers, employer, or colleagues. Even when the user tries to evaluate honestly, the rating is dependent on their current knowledge, status, etc. The implicit feedback, including exercise solving in an adaptive learning system or activity (i.e. effort) while programming, is more reliable in these cases, as the user always tries to be successful.

2.2 Seeking the proper definition for implicit and explicit feedback

Team leader: Róbert Móro

In the discussion concerning the explicit and implicit user feedback we focused on the domain of learning, namely on learning of foreign language vocabulary, considering only user feedback collected for the purpose of modeling users' knowledge. Explicit feedback is more precise, but sparse; on the other hand, it is easy to collect implicit feedback, but its interpretation is problematic. As an example we discussed different interpretations of the user's click on a word in our considered use case. It can mean that the users do not understand the foreign word, however, other interpretations are possible, e.g. the users know the word, but want to assure themselves, or the translation is incorrect in the given context.

We proposed two ways of dealing with the problem, namely letting the users explicitly determine their intention and using data mining from the logs in order to learn how to interpret users' actions. Because asking the users for their intention every time they click on a word can be tiresome for them, it should be used sparsely only to build sufficient labeled dataset for supervised learning from the data. Thus, it is reasonable to use combination of the two proposed approaches.

During the discussion we have also found out that the understanding of what can be considered explicit and what implicit feedback is ambiguous. Trying to distinct the two intuitively can be misleading; e.g. to give a like on Facebook seems an explicit feedback, however it is an integral part of the system's functionality and the users' goal is not to give Facebook information that they like the content, but rather to share it with their friends. Another example is rating movies; this is clearly explicit form of feedback, but if we use these ratings for inferring other unrelated information (e.g. age or sex of the users), it becomes an implicit factor. Thus, we can differentiate between two aspects of user feedback: (1) what is the users' motivation (i.e. whether they are aware that they provide us with feedback or not) and (2) how it is used and interpreted.

2.3 Entity similarities

Team leader: Dušan Zeleník

The group discussed the similarity as an important factor in many systems for adaptation or personalization. Main aim of the group was to focus on multi-attribute entities and their comparison. The most important aspect is effectiveness of the calculation. Thus we proposed graph representation to store relations among entities. Graph representation for multi-attribute entities promises that distance in graph somehow reflects the similarity. However this similarity representation might be biased by frequent attributes (e.g. genres for movies, categories for articles). In proposed representation we represent each attribute as separate node. Aforementioned genre would connect many entities causing unwanted impact on similarity. For this reason we also discussed algorithms which are suitable for calculating the similarity among items using graph storage.

We discussed mainly activation spread as the most suitable and easy to implement approach. Its complexity is low in both calculating the similarity among two entities and discovering the most similar items. We do not have to operate with all nodes in graph. We only examine the adjacent nodes and energy which spreads in the graph. This algorithm than effectively faces the aforementioned drawback with biased similarity due to frequent attributes. These nodes are simply splitting the energy among many edges thus making this connection almost irrelevant. Another advantage of graph representation is the ability to express the structure of the entity. We are able to represent sentences of an article or the plot of a movie.

2.4 Similarity in our projects

Team leader: Michal Kompan

The main focus of the second "similarity" discussion group was to explore types of the similarity in the group members' projects. Three basic areas were discovered. Firstly the multimedia domain was discussed. Various metadata for items are often available – movie is described by the genre, actors, language, duration or the storyline. The similarity here can be defined as the "content" (movie to movie) similarity. This is often computed based on the keywords vectors, while additional information can be included (genre etc.). Moreover, participants concluded that the bi-directional similarity is beneficial, while the most similar and most opposite items are interesting from the computational point of view respectively.

Second domain for the similarity application was the e-learning system. Here, not only standard keywords based similarity can be considered, but thanks to the user model (knowledge based), the real user knowledge level (with the respect to the course concepts) can be accounted.

Finally we discussed the similarity in the source code domain. The source code similarity computation is often based on the abstract syntax trees, while the similar part of codes can be easily found in this manner. Moreover, several extensions were proposed, while the code "neighborhood" can improve the similarity search (code comments etc.). The group concluded that the approach for the similarity search have to be chosen with the respect to the domain characteristics and particularly to the domain dynamics.

2.5 Logging in experimental applications

Team leader: Michal Holub

In this discussion we focused on the problem of logging in experimental applications. First, we briefly discussed purposes of logs: 1) to check if the application is running, 2) to check whether it works as expected, 3) to see how users use it, and 4) to collect data for experiment evaluation. From this point we oriented our discussion towards point 4.

Various quantities needed for experiment evaluation can be measured directly (e.g. time, number of clicks) or can be derived from primitive ones (e.g. interest, level of knowledge). We also need to think about the precision, e.g. do we need time in seconds or days? The amount of logged data is limited by technology limits like available memory or required response time of the application. There are also ethical rules to be concerned, e.g. not to log users' passwords.

The resume of our discussion is a guideline for how to choose what to log. The answer is: log everything (but wisely). If the experiment is not well designed, it is reasonable to log more data because we might need it later. We introduced an equation which can be used to determine what data should be logged:

$$L = \{M\} \bigcup \left\{ \forall_f : \frac{u(f)}{p(f)} > \varepsilon \right\} \setminus \{P\}$$

where *L* is the set of quantities to be logged, *M* is minimum what we need to log for the experiment, *P* is everything concerned with privacy not needed for the experiment (e.g. passwords), *u(f)* is usefulness of data about feature *f* and *p(f)* is price for implementing logging of feature *f*. This means that for every feature we should calculate how useful could its logs be versus how expensive is the implementation of logging. If this value is larger than a minimal threshold we should implement the logging, because we might need it in the future.

2.6 How to formulate a good hypothesis

Team leaders: Karol Rástočný, Ivan Srba

Our discussion club started with hard questions: How should we formulate hypothesis? What is a hypothesis? When is a hypothesis good hypothesis? In short time, we made an agreement that a hypothesis is an assumption which is postulated after an analysis of a problem domain. This assumption has to be formulated as if A than B statement, while B is the assumption of result if A is fulfilled. So the part A part has to contain all (only that) attributes of the domain's model, from that the model's attributes in the part B are dependent. But is it enough? No, as we know, hypotheses have to pose some new ideas and we have to be able to evaluate postulated hypotheses. This brings us to another assumption: We have to be able to set up attributes in the part A. In other words, these attributes have to independent. In addition, dependent attributes (attributes in the part B) have to depend only to independent attributes in the part A, because if there is another attribute from that has an influence to dependent attributes, we do not be able to set its value and we could not test our hypothesis correctly.

2.7 Landmines of innovative user interfaces in student projects

Team leaders: Eduard Kuric, Márius Šajgalík

Good user interfaces are crucial for good user experience. It doesn't matter how good an application is - if designers don't manage to make user interface as intuitive and attractive as possible, the application will hardly reach a breakthrough. To gain the interest in a new product, users need to understand its advantages or find themselves impressed or involved. Innovative doesn't mean usable and usable hardly means innovative. As usual, it's necessary to find an optimal trade-off.

In the discussion group we focused on visual design, interaction design and usability testing. We discussed problems of innovative user interfaces that were designed by students for their applications, namely, OntoSearch, Semantic Wiki and Facet Tree. Each of these applications provides some specific novel functionality which requires innovativeness of user interface in some reasonable measure. But despite the innovativeness emerging from functionality novelty, there is still need to include some common UI controls, which would provide more intuitive navigation. We identified positive and negative aspects of discussed user interfaces. We proposed solution for each identified problem in order to support the intuitiveness of navigation whilst retaining the original innovativeness. Aside from user interface design solutions, we proposed several ways of usability testing.

2.8 Long tails in your data

Team leader: Jakub Šimko

In the discussion group devoted to long tail phenomenon, we firstly mentioned many different types of datasets, where it long tail may occur. We identified the user related data (ranging from raw user logs to inferred user models) as the most troublesome type. This is because they often source from the quantity of user activity in the examined system, which is usually distributed by power law: few users do a lot of activity and many users do very little. For some system types, like crowdsourcing portals, this imposes no problem as they are able to optimize user work allocation, but portals that demand a certain amount of activity from users in order to function (e.g., for building user models) are hampered by the cold start problems with many users (with little activity). In some cases even, the most active users with much activity are also problematic, e.g., in case of news-recommender systems a hyper-active user crawls the whole relevant information space, leaving no items for recommendation.

We then discussed options of mitigating the negative effects of long tails in our data and ways how to take an advantage from them. To identify long tails and to prevent unsuccessful experimentation due to the “long tail user activity waste” we recommend using two phase experimentation. With first phase being a pilot for estimating future dataset properties, the second can be adapted from unsupervised system deployment to a more controlled environment examination where data can be collected in more uniform quantities. We also stressed the importance of explorative data analysis, where long-tailed visualizations of data could easily point-out extreme cases of users (or other entities) which could be subsequently analyzed qualitatively and give additional interpretation to quantitative experiment results, especially when they are not completely satisfying.

SeBe 8.0: Beer – Unity of Science and Art

Marián ŠIMKO, Jakub ŠIMKO, Róbert MÓRO, Michal BARLA

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

The Beer Driven Research (BDR) is a lively fast-evolving field with results that have direct tremendous impact on our everyday lives. No wonder that it has attracted much attention from the researchers' community over the past years [1-5]. We examine the methods of stimulating research abilities and their synergic effects with emphasis on the overall sustainability (i.e. it is equally important to stay at the peak of one's capabilities as it is to reach the peak [4]).

In pursue of achieving the kalokaghatic ideal of a researcher as defined in [5] the main topic of the traditional Semantic Beer workshop in its 8th edition was the unity of science and art (see Fig. 1) as presented in visual aspects of beer-human interfaces.

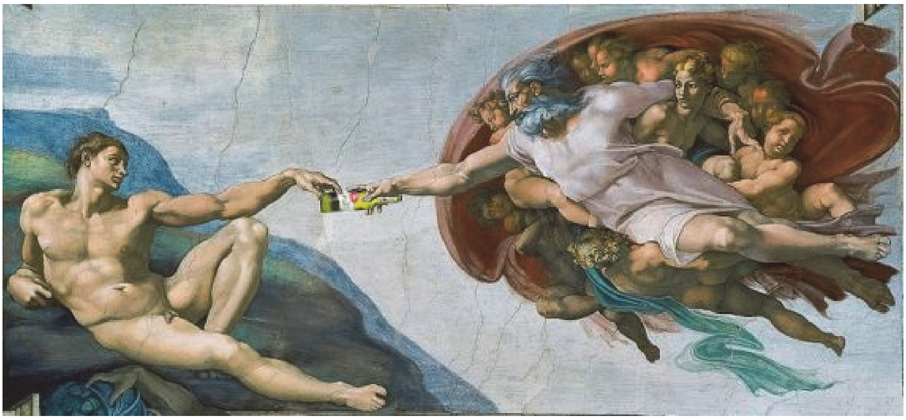


Figure 1. The beer at the center of interest of artists and scientists.

The participants submitted to the following tracks:

- figural art,
- sigil traditions,
- plantae et animalia,
- non-convex approaches,
- post-modern etiquettes.

The focus was on artistic features of beer etiquettes and bottles that are frequently overlooked in the-state-of-the-art research, but are nevertheless important as a means for visual navigation and shelf search, driving our decision-making during travelling through beer-selling spaces. We also followed the efforts of completing the grand challenge of populating a beertology introduced at SeBe 4.0 [1], continued by acquiring spatial and temporal metadata at SeBe 5.0 [2] and 6.0 [3] respectively.

We were satisfied with the overall quality of the submissions. After rigorous and systematic evaluation we rewarded Pavol Bielik, Matúš Vacula, Juraj Višňovský, Samuel Molnár and Ivan Srba as authors of the best submissions (one in each track). In addition, we awarded a special price to Jakub Ševcech in recognition of his original research that resulted into creating a series of bottles with customized etiquettes.

Because there were many researchers struggling with their experiments, we decided in this edition of SeBe to motivate members of BDR community to participate in these experiments by providing tempo-spatial resources. Participants were given a stamp for each five minutes of their activity. In the end, we rewarded the most active ones: Peter Dulačka, Michal Holub, Ondrej Galbavý, Matúš Vacula and Jozef Lačný. The absolute winner became Martin Lipták with 19 collected stamps meaning that he spent more than hour and a half participating in experiments. Altogether, the participants provided approximately 30 hours' worth of experiments in one night proving to be a vibrant community of committed (beer-driven) researchers.

The collabeeration of researchers nurtured at the workshop once more managed to push the frontiers of current knowledge by examining new methods, tastes and visual phenomena lying within the scope of beer science. The results presented at the workshop will undoubtedly soon find practical applications in universities and research facilities. There are still open problems left unexplored for the future, because the grand concept of beer and its instances with their many beneficial attributes will remain a constant inexhaustible source of inspiration and research challenges.

Acknowledgement. This work was partially supported by the SeBe Coin initiative.

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] Šimko, M. Šimko, J., Barla, M. SeBe 6.0: Beer Distribution Issues and Solutions. In Proc. of 11th Spring 2012 PeWe Workshop: Personalized Web – Science, Technologies and Engineering. Modra-Piesok, Slovakia, 2012, pp. 95–96.

[5] Šimko, M. Šimko, J., Barla, M. SeBe 7.0: Healthy Body, Healthy Mind, Healthy Research. October 2012. Modra-Piesok, Slovakia, 2012.

Gold Rush at Pewe Workshop

Jakub ŠIMKO, Dušan ZELENÍK

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

As an accompanying event of our workshop, we conducted the Gold Rush competition. This competition was inspired by popular competitive style TV shows. We particularly copied the principles from well-known Czech and Slovak TV show called Riskuj, broadcasted in late 90's. In the show three participants answered questions categorized in various categories with specific values. The winner was chosen as the one with the biggest cumulative score.

To make Gold Rush personalized for our research group we prepared 210 questions regarding the knowledge base needed to accomplish second degree in software engineering. We actually used many forms of hints. The most typical hints came from the category of the question. For instance, we named a category "Colors in Software Engineering". Answers were then somehow related to colors (e.g. Black Box, Gold Standard, Gray Code).

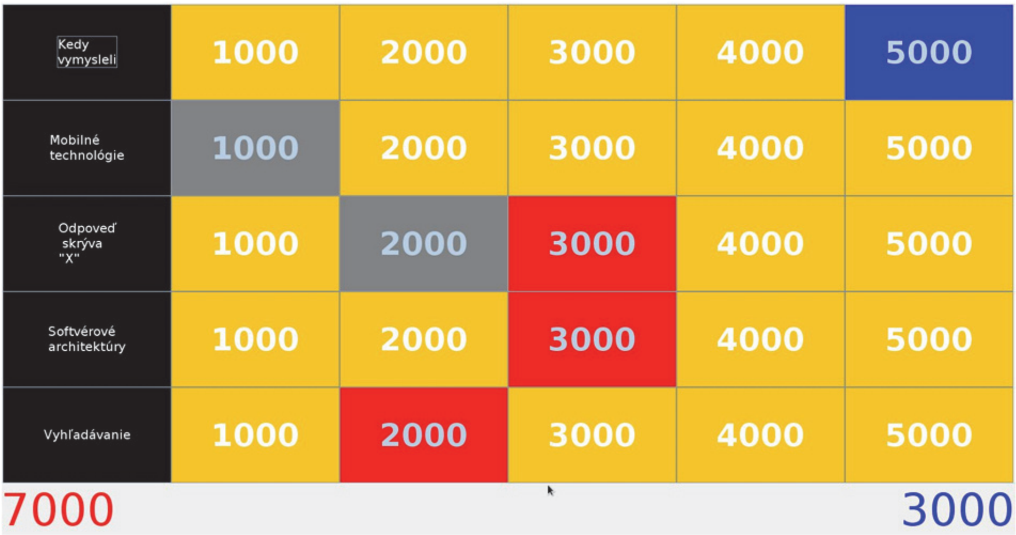
We arranged 7 rounds for 8 teams. Every round represented the fight between two teams which were competing to achieve new round. Every team had around 5 team members who were representing both degrees. Each round took 12 minutes and every team had at least one chance to compete in the first 4 rounds. Teams were choosing questions in the manner of round robin to prevent total domination and avoid the embarrassment. Each question could be answered in 5 seconds correctly or incorrectly. Value of the question was then added or subtracted respectively.

In order to entertain the audience and keep them "in the game" we decided to use bonus questions in every category which could be answered by the team which obtained most of the points available in this particular category. Correct answer meant an advantage in the form of joker. Joker could be used to ask someone from the audience or to source the crowd. To apply the joker, team had to decide in the time limit of 5 seconds which option they want. Asking someone showed to be very inefficient so teams emerged to use another option crowdsourcing. This option, in our case also known as screams from the audience harvested the power students and doctors in the crowd.

To sum up this event, after seventh round, one team managed to win all the required rounds. This team was then awarded in the following event of the SeBe Workshop.

Spring 2013 PeWe Workshop, April 5, 2013, pp. 127–128.

To visually entertain the audience, we implemented the same visual assistant as it was used in original TV show. The interactive application is presented in the following picture (see Fig. 1).



Kedy vymysleli	1000	2000	3000	4000	5000
Mobilné technológie	1000	2000	3000	4000	5000
Odpoveď skrýva "X"	1000	2000	3000	4000	5000
Softvérové architektúry	1000	2000	3000	4000	5000
Vyhľadávanie	1000	2000	3000	4000	5000

7000 3000

Figure 1. Wall with the questions in categories. Each brick in the wall worked as button which opened up the question to be answered.

Index

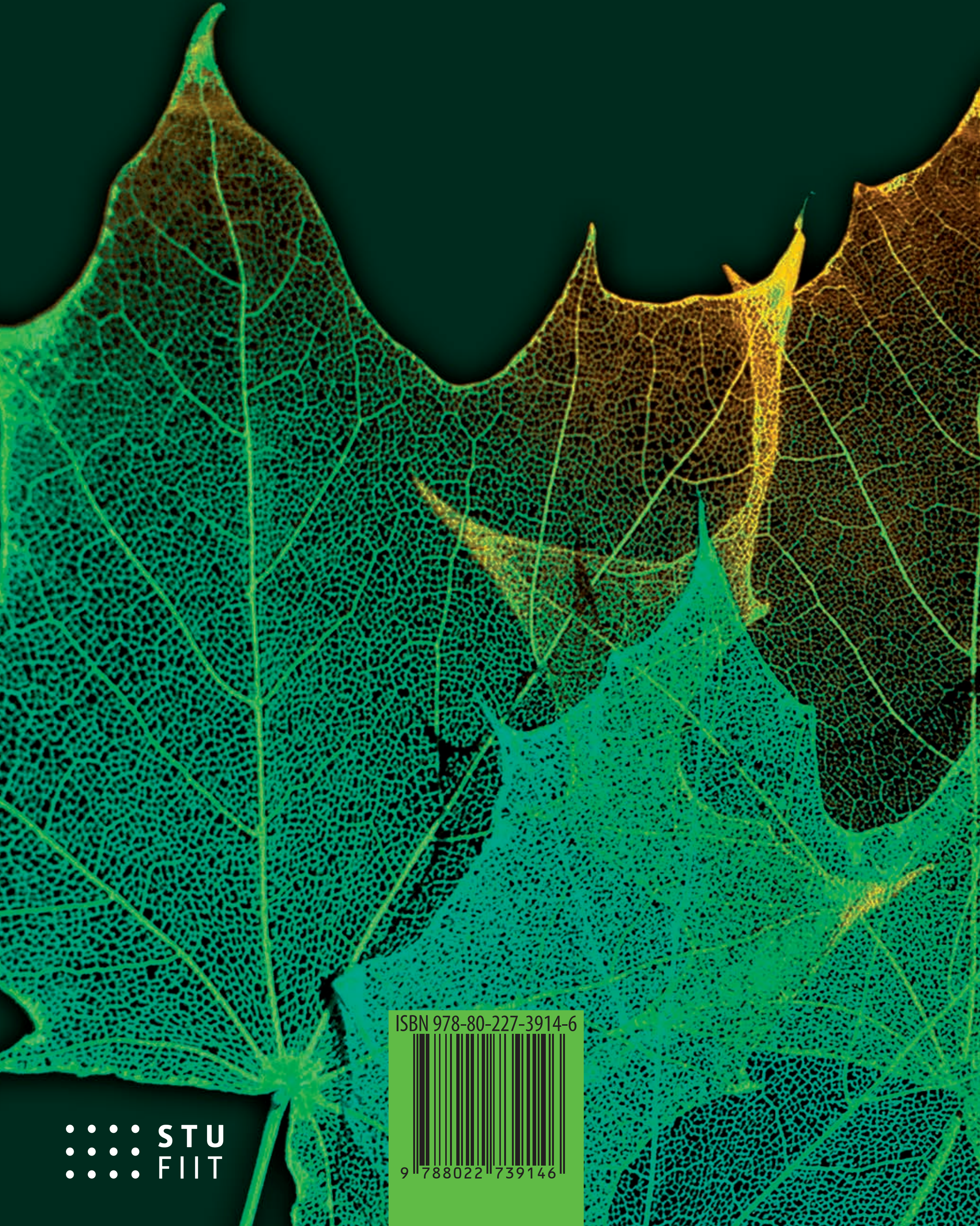
Mária Bieliková, Pavol Návrat,
Michal Barla, Marián Šimko, Jozef Tvarožek (Eds.)

Personalized Web – Science, Technologies and Engineering
13th Spring 2013 PeWe Workshop

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

146 pages, 50 copies
Print Nakladateľstvo STU Bratislava
2013

ISBN 978-80-227-3914-6



ISBN 978-80-227-3914-6

9 788022 739146

STU
FIIT