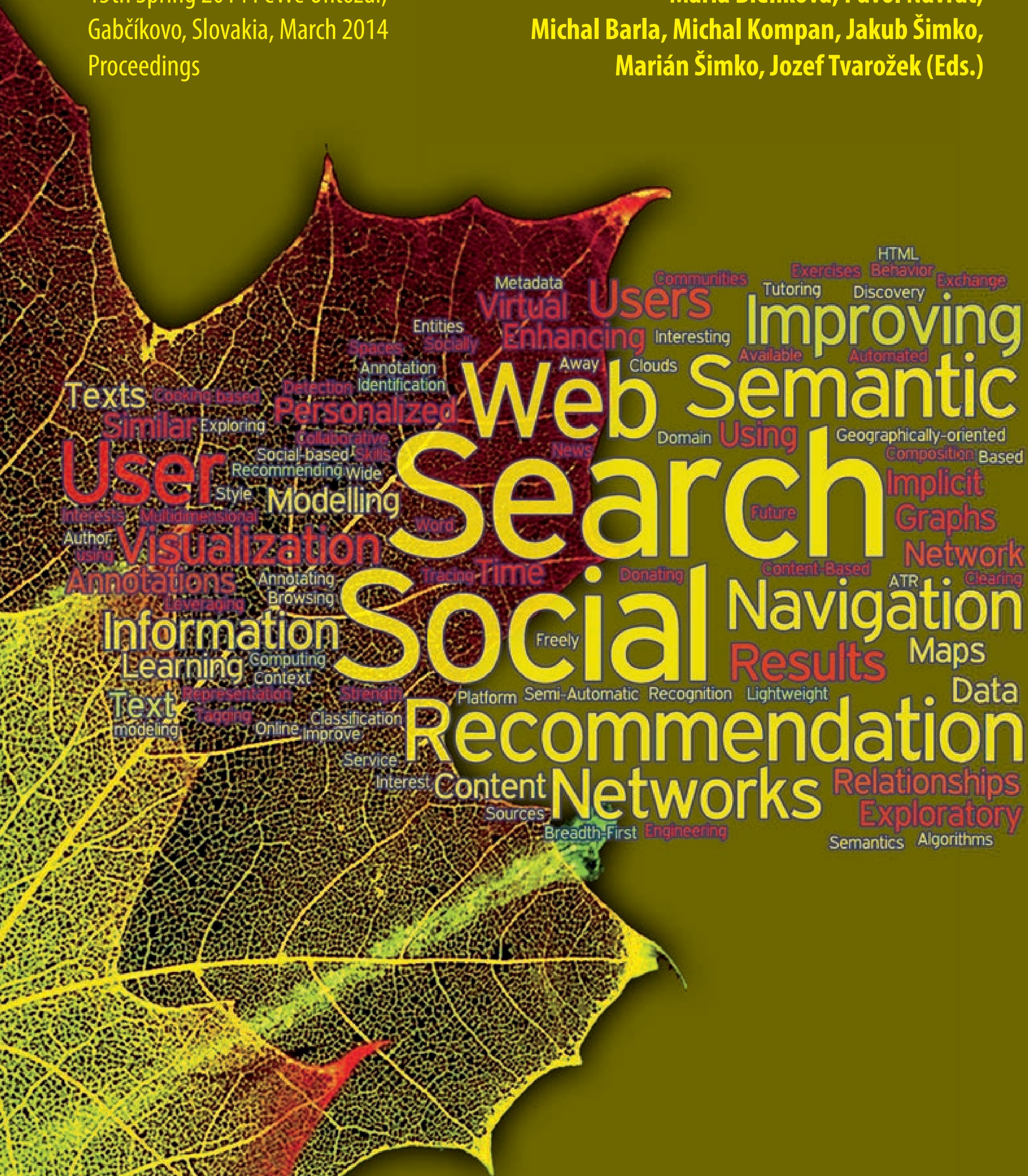
Personalized Web – Science, Technologies and Engineering

15th Spring 2014 PeWe Ontožúr,
Gabčíkovo, Slovakia, March 2014
Proceedings

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



Personalized Web – Science, Technologies and Engineering

15th Spring 2014 PeWe Ontožúr,
Gabčíkovo, Slovakia, March 2014
Proceedings

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

Proceedings in
Informatics and Information Technologies

Personalized Web – Science, Technologies and Engineering

15th Spring 2014 PeWe Workshop

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

Personalized Web – Science, Technologies and Engineering

15th Spring 2014 PeWe Workshop
Gabčíkovo, Slovakia
March 21, 2014
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
15th Spring 2014 PeWe Workshop

Editors

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

Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia

Production of this publication was partially supported by

- the Gratex IT Institute, within the framework of GraFIIT Foundation (www.gratex.com)
- the Scientific Grant Agency of Slovak Republic, grants No. VG1/0675/11, VG1/0971/11
- the Slovak Research and Development Agency under the contract No. APVV-0208-10
- the Cultural and Educational Grant Agency of Slovak Republic, grant No. 009STU-4/2014.

All contributions reviewed by the editors and
printed as delivered by authors without substantial modifications

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

Cover Designer: Peter Kaminský

Published by:
Nakladateľstvo STU
Vazovova 5, Bratislava, Slovakia

ISBN 978-80-227-4159-0

Preface

The Web influences our lives for nearly 25 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 15 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 15th PeWe (Personalized Web Group) Workshop on the Personalized Web, held on March 21, 2014 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 were 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. We organized the proceedings according PeWe research subroups, which were established in beginning 2014 considering research topics and research projects we are involved in:

- Personalized Recommendation and Search (PeWe.REC),
- Traveling in Digital Space (PeWe.DS),
- User Experience (PeWe.UX),
- Software Development Webification (PeWe.PerConIK),
- Technology Enhanced Learning (PeWe.TEL).

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 three of our bachelor students who take an advantage of our research track offered within study programme Informatics – *Peter Dubec*, *Ladislav Gallay* and *Jakub Mačina*. They were just about to start his bachelor project. Peter is interested in semantics acquisition for the Electronic Program Guide. His aim is improvement of personalized recommendation and presentation of TV shows and movies. Ladislav deals with utilizing vector model of words for text processing on the Web. His aim is to improve lemmatization process in any given language by utilizing Word2vec tool developed by Google. Finally, Jakub is about to start Imagine Cup project in either World Citizenship or Innovation categories.

Bachelor projects:

- *Tomáš Brza*: Student Motivation in Interactive Online Learning
- *Rastislav Detko*: Users' Relationship Analysis
- *Richard Filipčík*: Gamification of Web-based Learning System for Supporting Motivation
- *Peter Kiš*: Analysis of Interactive Problem Solving
- *Matej Kloska*: Keyword Map Visualisation
- *Adam Lieskovský*: Query by Multiple Examples Considering Pseudo-Relevant Feedback
- *Viktória Lovasová*: Recommendation in Adaptive Learning System
- *Filip Mikle, Matej Minárik, Juraj Slavíček, Martin Tamajka*: Low-Cost Acquisition of 3D Interior Models for Online Browsing
- *Jana Podlucká*: Assessing Code Quality and Developer's Knowledge
- *Martin Svrček*: Collaborative Learning Content Enrichment
- *Tomáš Vestenický*: Using Tags for Query by Multiple Examples
- *Ľubomír Vnenk*: Web Search Employing Activity Context

Master junior projects:

- *Dominika Červeňová*: Automated Syntactic Analysis of Natural Language
- *Peter Demčák*: Methodics of Game Evaluation Based on Implicit Feedback
- *Marek Grznár*: Adaptive Collaboration Support in Community Question Answering
- *Jozef Harinek*: Natural Language Processing by Utilizing Crowds
- *Patrik Hlaváč*: Analysis of User Behaviour on the Web
- *Matej Chlebana*: Source Code Review Recommendation
- *Martin Janík*: Web Applications Usability Testing by means of Eyetracking
- *Róbert Kocian*: Personalized Recommendation Using Context for New Users
- *Samuel Molnár*: Activity-based Search Session Segmentation
- *Miroslav Šimek*: Processing and Comparing of Data Streams Using Machine Learning
- *Veronika Štrbáková*: Implicit Feedback-based Discovery of Student Interests and Educational Object Properties
- *Martin Toma*: Using Parallel Web Browsing Patterns on Adaptive Web
- *Máté Vangel*: Determining Relevancy of Important Words in Digital Library using Citation Sentences
- *Pavol Zbell*: Modeling Programmer's Expertise Based on Software Metrics

Master senior projects:

- *Karol Balko*: Keeping Information Tags Valid and Consistent
- *Ľuboš Demovič*: Linking Slovak Entities from Educational Materials with English DBpedia

- *Peter Dulačka*: Finding and Harnessing Experts for Metadata Generation in GWAPs
- *Eduard Fritscher*: Group Recommendation of Multimedia Content
- *Martin Gregor*: 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 and Determination of Correctness of Answers in Educational System Using Crowdsourcing
- *Martin Lipták*: Researcher Modeling in Personalized Digital Library
- *Martin Plank*: Collocation Extraction on the Web
- *Ondrej Proksa*: Discovering Similarity 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*: Method for Novelty Recommendation Using Topic Modelling
- *Juraj Višňovský*: Evaluating Context-aware Recommendation Systems Using Supposed Situation

Doctoral projects

- *Michal Holub*: Utilization of Linked Data in Domain Modeling Tasks
- *Peter Krátky*: User Identification Based on Web Browsing Behaviour
- *Eduard Kuric*: Modeling Developer's Expertise
- *Martin Labaj*: Observing and Utilizing Tabbed Browsing Behaviour
- *Róbert Móro*: Using Navigation Leads for Exploratory Search in Digital Libraries
- *Karol Rástočný*: Employing Information Tags in Software Development
- *Ivan Srba*: Adaptive Support for Collaborative Knowledge Sharing
- *Márius Šajgalík*: Exploring Multidimensional Continuous Feature Space to Extract Relevant Words
- *Jakub Ševcech*: Anomaly Detection in Stream Data

Considerable part of our research meeting this year was devoted to *data analysis* workshop followed by experimentation session. Data analysis workshop was chaired by Jakub Šimko and Michal Kompan. It was focused on the analysis of a data sample containing time frame of four months of TV "set-top-box" logs acquired from one of the major cable TV providers in Slovakia. The sample was extensive: the logs themselves comprised over 10 mil. of rows (each representing a time period for which a particular user has watched a particular program), with over 10 thousands users and up to 200 channels. For more metadata (and prior to the workshop), we mapped these logs to program entries of electronic TV program guide (EPG), which we obtained from the provider as well. The research aim of the workshop was the all-purpose analysis of the sample. As for pedagogical aims, we wanted to give the participants an opportunity to work with real (and appropriately extensive) data. We also aimed to strengthen the bonds and collaboration potential of PeWe group members, especially vertically (i.e. to mix junior and senior students) and enable them to share their experience. The participants were split into ten groups, each having at least four members of mixed seniority. Their short reports are available in these proceedings.

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 consisted of 8 experiments. The experiments were very diverse: the participants evaluated quality of their research interests model, learned new vocabulary, played a game, were recommended news or movies, searched for external sources enriching the learning content, linked Slovak entities with English DBpedia or solved tasks in a learning system while being monitored by an eye-tracker. The session lasted for several hours and officially ended at 2 a.m., which greatly exceeded our expectations.

Our workshop hosted for the tenth time recessive activity organized by the *SeBe (Semantic Beer) initiative* chaired by Marián Šimko, Róbert Móro, Jakub Šimko, and Michal Barla. It was aimed this year at educational activities supported by a new academic institution – Academia SeBeana. Our ambition was to teach our prospective students all the knowledge and skills necessary for them to be successful in this highly competitive beer market as well as prepare them for their future research careers. We prepared for our students new course in our Adaptive Learning Framework ALEF, which covered various aspects of beer industry from beer elements and types, beer history, beer production including known Slovak and Czech breweries and best known beer brands to beer with semantics.

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/ontozur2014-03.

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 2014

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

Table of Contents

Students' Research Works

Personalized Recommendation and Search (PeWe.REC)

Traveling in Digital Space (PeWe.DS)

User Experience (PeWe.UX)

Software Development Webification (PeWe.PerConIK)

Technology Enhanced Learning (PeWe.TEL)

Accompanying Events

Workshop participants

Gabčíkovo, 21. marec 2014

Majka

Ľuboš

Edo

Čubo

Marek

Mišo

Kubo

Martin

Marián

Viki

Janka

Dominika

MiB

Jojo

Maťo

PeťoK.

Ondrej

Miro

Rišo

Gregor

MarekG.

Ivan

Tomáš

Veronika

Paťo

Samo

Spring 2014 PeWe Ontožúr Participants

Personalized Recommendation and Search (PeWe.REC)

Users’ Relationship Analysis

Rastislav DETKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xdetkor@is.stuba.sk

Nowadays, many people are actively using social networks services. Usually these social network services are used to share ideas and to communicate with other people. These information about the user’s interaction can be useful in various tasks of Web to user’s experience improvements (as the recommendation, information filtering etc.)

In our work we aim to analyze data containing information about users and their activities in social network service and propose method, which will set the likelihood of influence between users. In this method we are focusing on to taking public action log which contains interactions one user with subnetwork (this subnetwork contains user’s neighbors) and finding out which other users was reacting to his public activities and find out which users influence other users in the network.

Our method is based on the computing likelihood of influence between two users, from user’s statistics evaluation (count of neighbors, count of publications, count his public reactions to others) and as we mentioned mainly his public action log. We are also considering time as one of the part that likelihood depends on. We persuade that as time flows from last made action, the probability of influence will be lower. Next part in our method is the order of the reactions, which were made for public action. This information provides us create a model of possibilities, by whom could be influenced user in the network [1] (in Figure 1 we can see an example of posting status and next reposting this status in order of action was made, and directed pipes means possible spread of influence in the network).

Subsequently, we are modeling influence flow in the network, what help us to verify, that our proposed method are correct, by simulating real event. This event development will be supposed by using our method to evaluate edges in the network, which is used in model to monitor influence range and user activities. There are two options, how we can model the spread of influence. Linear threshold model is the first one model, which we can use to model spread of influence in the network. This model is based on simulating users activation, by adding each user threshold and when sum of user‘s edges, which are connected to activated neighbors, is greater than user’s threshold, user is activated. Second model is independent cascade model. This model is

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

based on evaluating edge probability, which expresses probability that already activated user will activate other user directly connected with evaluated edge. We used both of these models for showing different aspects. First model show us, when user is activated from network view and second model we use, when we want see user activation from user to user view.

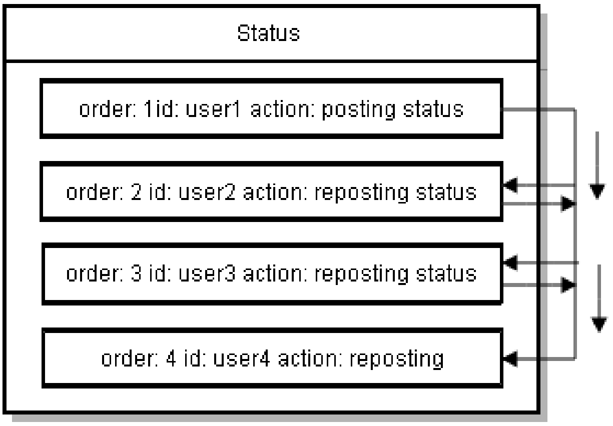

Status
order: 1id: user1 action: posting status
order: 2 id: user2 action: reposting status
order: 3 id: user3 action: reposting status
order: 4 id: user4 action: reposting

Figure 1. Influence model in public action called “posting status”.

Next step is maximizing influence in network by using one of models we already mentioned before. Maximize influence means minimize activated users at the beginning of the simulation and trying to maximize the number of activated users at the end of the simulation. Also we are monitoring, whether activation flow goes through the whole network or stops in step of simulation.

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

References

[1] Tang, J., and col.: Social influence analysis in large-scale networks. In Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York:ACM New York, 2009

[2] Kempe, D., and col.: Maximizing the spread of influence through a social network. In proc. of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM New York, 2003

Group Recommendation of Multimedia Content

Eduard FRITSCHER*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
eduard.fritscher@gmail.com

Because of the growth of the Web, the amount of information which is stored in online space has increased. Nowadays we are literally overwhelmed with the amount of information available. To solve the problem of this information burst, recommendation technics and methods were invented, but as the world changes, the access to the internet also changed. People collaborate more often with each other. In times where the most visited pages in the world are social network pages the recommendation technics have to adept these new trends. Which is mainly collaboration between users. The answer to this need is group recommendation.

To create accurate recommendations we need data that describe the users' interest and taste in the given domain. Nowadays the most information about users are stored in social networks applications. Therefore the aim of our proposed method is to enhance group recommendation with personality awareness and the power of graph traversal algorithms. In our method we build a group recommendation that uses personality models created by using data extracted from social networks. Social networks offer a great opportunity for user information extraction. People willingly provide information about their taste and interest [2].

To correctly be able to model a user's personality we need several dimension that describes a user's personality in a social crowd. Therefore we have chosen the use of the Big Five personality model in our method because of its popularity and wide scale of usage. Our method consists of 4 basic steps which you can see on Figure 1. The first step of the method is data extraction. We need to extract information from social networks to be able to create a personality model of a user. After this we can proceed to the next one, which is building to Big Five personality model of a user. When proposing the method of the personality model creation we used the discovered personality patterns of Facebook users [1]. After we have created the personality models needed to create our personality aware group recommendation we apply our proposed aggregation strategy. The strategy modifies the weights of the starting nodes in our recommendation algorithm, which is based on a traversal energy spreading algorithm with the use of the users' personalities.

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

The proposed personality model generation and group recommendation method was implemented and tested in the movie recommender system Televido. The results of the experiments shown that with the use of the proposed method we can reach higher precision than with aggregation strategies that do not include personality awareness.

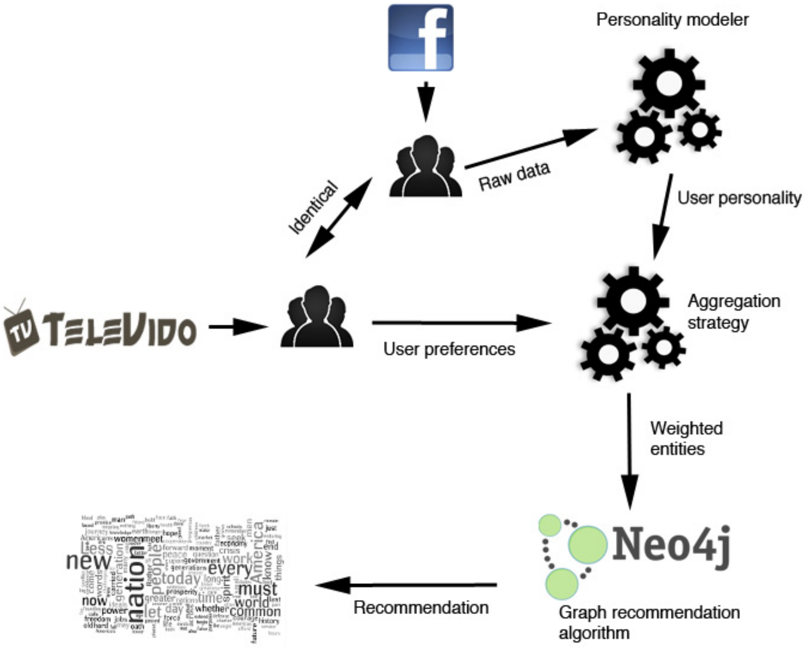


Figure 1. Scheme of the group recommendation method.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 577-578.

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

References

[1] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. 2012. Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference* (WebSci '12). ACM, New York, NY, USA, 24-32.

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

Group Recommendation for Smart TV

Ondrej KAŠŠÁK[*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
ondrej.kassak@stuba.sk

Domain of multimedia content belongs to the popular fields of human interest. Multimedia content brings us knowledge, fun, relax etc. Average people spend daily multiple hours by watching TV or videos on the internet. Thus it is very important what people the watch because it constitutes a significant part of their free time and the daily information income. Due to the big interest for multimedia content, there is a big offer too. Nowadays is in TV or on the internet available huge number of data. This situation brings to people information overload problem which still grows.

There were involved multiple ways how to help users with described problem. One of the most effective ways is the personalized recommendation, which is based on automatic selection of most interesting content individually to each user. There are multiple known types of personalized recommendation: collaborative, content based, demographic, knowledge based etc. These pure approaches are fairly good known and there exist many experimental researches describing their strong and weaker features. A good overview can be seen, for example, in work of O'Connor et al. [3].

Personalized recommendation is, in all of its variations, based on some kind of user model. User model represents content or its features, that user interacted with in past. Domain of multimedia content is in common considered as well structured, so we can fairly good describe items by its metadata. In the field of personalized recommendation however, exists along well known pure recommendation approaches a fewer scouted part, named hybrid recommendation. Hybrid symbolizes, in this meaning, any method combining by some way multiple pure recommendation methods. These can use different ones or even similar recommendation approaches [1].

In our research, we focused on mixed hybrid method combining collaborative and content based approach. We choose this combination due to its complementarity. For new user faces pure collaborative approach to the cold start problem, because it cannot find the appropriate similar users. Pure content based approach achieves the lower precision and cannot help user to find something without his/her majorly preferred domain. In combination, can these two approaches achieve the result, which is better than results of methods used alone.

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

Our method is composed form two steps. In first of them are suitable items chosen by collaborative method, which uses to recommend the power of user community. This is, in common, known as the very effective way of finding items to recommend. The main advantage is that quality content was filtered by other users who have similar preferences to target user, so there is a low chance to choose some uninteresting content. Collaborative recommendation reaches in common good results, when there are enough users in system and we know target user preferences well [4].

In the second step our main process we change the order of results from collaborative approach based on similarity level of these items to items which user watched before. Similarity level is determined on the basis of content based approach results. Advantage of content based recommendation is the, that we can determine the similarity measure between some item and user preferences based on his/her past activities [2]. This is different to collaborative recommendation, where we cannot say anything about the suitability of some item to user preferences.

Proposed hybrid recommendation method reaches for single users the higher precision, compared to pure collaborative and content based methods. The highest difference can be seen on top positions of results. This means that idea of reordering the results works. This is in fact quite important in domains such as watching movies or TV, due to the long duration of its items. There is common that user watches only one or few items, so is better to offer him 1 very suitable item that the 15 tolerable.

Our next aim is to evaluate proposed recommendation method on groups of multiple users. In domain of movies, TV or another multimedia content is common that multiple users consume the same content together in the same time. We are social beings and for this reason will we rather tolerate content that does not completely suit individually to us, but which is acceptable for whole group. This trait of humanity deserves to be supported by personalized recommendation.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 62-67.

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

References

[1] Burke, R.: Hybrid web recommender systems. The adaptive web, pp. 377–408, Springer Berlin Heidelberg., (2007).

[2] Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content based recommendation system using social network analysis. In Proc. of the 17th Int. Conf. on World Wide Web (WWW '08). ACM, New York, NY, USA, (2008).

[3] O'Connor, M., Cosley, D., Konstan, J. A., Riedl, J.: PolyLens: a recommender system for groups of users. In Proc. of the 7th Conf. on European Computer Supported Cooperative Work, pp. 199-218, Kluwer Academic Publishers, Norwell, MA, USA, (2001).

[4] Ungar, L.H., Foster, D.P.: Clustering Methods for Collaborative Filtering, In AAAI Workshop on Recommendation Systems, Menlo Park, CA, (1998).

Personalized Recommendation Using Context for New Users

Róbert KOCIAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xkocianr@is.stuba.sk

Each web application using personalized recommendation encounters problems when we do not have historical data about new user. If a new user comes to the page we have no information about him, and therefore we cannot adequately recommend relevant content. Of course we can extract data from extensive registration forms, but this approach usually transfers complexity to the new user. Moreover it is not always possible to obtain data directly from the user during registration, and that is why it is necessary to reveal users' preferences and tastes automatically. The aim of our work is to design approaches for reduction of cold-start problem.

The very first users' interaction within a personalized system is crucial from the recommender system and user modelling point of view. This activity is critical because in these moments a user creates a relationship and opinion, which are important for user to the next use of the system.

Two approaches for personalized recommendation are researched – content based recommendation or the collaborative filtering. Difference between them lies in the fact that the recommendations based on the content analyses the items from the similarity point of view and assumes that users likes similar items. The collaborative recommendation is based on the assumption that similar users liked similar content, and thus the users' activity is considered instead of items content. Often the combination of both approaches are used in order to reduce the problems connected to each approach [1].

Collaborative recommendation is more used in electronic trading against recommendations based on the content which is used in social networks. In collaborative recommendation are new users asked to evaluate a number of items that have no related link with interest new user. The aim of evaluation is finding the most appropriate items that will be shown the new user [2].

Developers of recommendation systems try to solve the new user problem by obtaining the context of new user profile or different questionnaires at registration.

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

They tried to find the demographic data that were considered different questions focused on the interests that may be similar to other users. Methods based on this approach need more user effort and is not easy to determine appropriate valuation of the common interests of users, because even if a user chooses the same area of interest as another, so one can give a higher priority to this interest and the other does not.

Some approaches try to limit or completely eliminate participation of users. This means that the user not need fill questionnaires but system chooses the convenient data about user, such as age, location of the user and data allowing the grouping users into classes. In the "cold start" domain some of these attributes may be missing but it can be estimated by the induction, which counts positional vector according to other users in the same community [2].

There are also hybrid approaches, which combine collaborative and content-based recommendation approach. This method is based on the principle of items aggregation according to matrix and then uses the clustered results and content of the items to obtain a decision tree for associating new items with existing [3].

In today’s systems, users are spited into groups according to what interests, work or relationships users have. This attribute we can use for recommendations to new users respectively. We can obtain a huge amount of information from related or similar users that can be used for increasing the quality of recommendations for the new user. We can also consider the social context obtained from other systems and applications.

In our project we want to design recommendation method which will be recommend related information for the new user according to user’s context from related and similar users. We want to test our approach on the faculty systems like ANNOTA.

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

References

[1] Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor (Eds.): *Recommender Systems Handbook*. Springer, 2011

[2] Nguyen, An-Te, Nathalie Denos, and Catherine Berrut. “Improving new user recommendations with rule-based induction on cold user data.” *Proceedings of the 2007 ACM conference on Recommender systems - RecSys ’07*. ACM Press, 2007. 121-128.

[3] Dongting Sun; Zhigang Luo; Fuhai Zhang, "A novel approach for collaborative filtering to alleviate the new item cold-start problem," *Communications and Information Technologies (ISCIT)*, 2011 11th International Symposium on , vol., no., pp.402,406, 12-14 Oct. 2011

User Identification Based on Standard Input Devices Usage

Peter KRÁTKY*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
kratky@fiit.stuba.sk

Every person is unique in the way he/she uses a computer, an operating system, programs and even input devices such a keyboard or a computer mouse. Patterns produced by usage of standard input devices could be utilized to identify an unauthorized user. Especially, adaptive and personalized systems accessed by multiple users might benefit from this feature as tailoring the content heavily depends on previous activity of the user. In our research, we focus on patterns in browsing the web, a very common activity nowadays. The goal of our work is to design a method to identify persons based solely on their behaviour rather than machine or browser they use to access websites.

Construction of a user model representing input device usage patterns has been a problem discussed by many researchers. The common protection of a user account by password is enriched with mechanism of patterns comparisons when accessing the account. Apart from comprehensive studies of keyboard usage dynamics, numerous papers focus on computer mouse. One of the first works [1] has proven the possibility of identification by characteristics such as distance, duration and angle of the mouse movement path. Comprehensive list of spatial and temporal characteristics of mouse movement strokes (curvature, acceleration, angle velocity, etc.) have been published and studied in [2]. Analysis of changeability of characteristics according to the environment (display resolution, mouse sensitivity) was described in [3]. We focus on identification which is harder task, but the previous research shows it is a possible one.

The process of user identification consists of three stages – (I.) acquisition of the data while browsing, (II.) construction of a user model and (III.) the identification itself. We designed a system consisting of three modules corresponding to these stages.

The first part of our system is logger. This module is responsible for collecting data from users in implicit way, meaning that harvesting of the data does not bother users at all. Four computer mouse events are tracked: mouse movement, single click,

[*] Supervisor: Daniela Chudá, Institute of Informatics and Software Engineering

mouse wheel movement, scrolling of a page. Each event record is assigned an id of a user, type of the event, time, x and y coordinates of the cursor.

Extractor module converts raw data into meaningful information. Extracted characteristics form the user model that represents user's mouse usage patterns. We calculate tangential velocity, acceleration, angle of the path tangent, angular velocity and curvature of the stroke, a sequence of mouse movement events separated by clicks.

Matcher module provides the identification task itself. The user model constructed in the previous stage becomes the test user model at first. It is compared to all template user models which are stored in the database. The result of matching process is either identity of the test user or the user model becomes template in case no template model matches it and it is stored in the database. The distance between two user models (vectors) is based on t statistics of Welch's t-test.

We conducted a study on 17 users browsing a real running e-shop. An experiment was designed to fully encourage users to perform intended actions by gamification of the user experience. We collected over 90 strokes for each user consisting of at least 4 points. Minimum number of strokes made by one user is 42.

From the view of suitability for identification we examined how distinguishing the values of the characteristics are among the users. To quantify distinctiveness of the characteristic we perform comparison of values for each pair of users using t-test. The distinctiveness could be expressed then as a ratio of number of comparisons indicating difference to all comparisons. The most distinctive characteristic is duration of a single click with distinctiveness rate of 0.77.

Characteristics with higher distinctiveness rate have been selected into the user model. Using the user model containing 17 features we evaluated our identification method and achieved 63.1% success rate of identification.

In our future work we are going to examine changeability of the characteristics over the time and their dependency on hardware used. We are also going to test other classification methods to perform identification task.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 197-203.

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

References

[1] Pusara, M., Brodley, C.E.: User re-authentication via mouse movements. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security - VizSEC/DMSEC '04*, (2004), pp. 1–8.

[2] Gamboa, H., Fred, A.: A behavioral biometric system based on human-computer interaction. In: *Proceedings of SPIE*, (2004).

[3] Zheng, N., Paloski, A., Wang, H.: An efficient user verification system via mouse movements. In: *Proceedings of the 18th ACM conference on Computer and communications security - CCS '11,* (2011), pp. 139-150.

Activity-based Search Session Segmentation

Samuel MOLNÁR*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xmolnars1@fiit.stuba.sk

Automatic search goal identification is an important feature of personalized search engine. The knowledge of search goal and all queries supporting it helps the engine to understand our query and adjust sorting of relevant web pages or other documents according to our current information need. To improve the goal identification the engine uses other factors of user's search context and combine them together by different relevance weight. Although, most of factors utilized for goal identification involve only lexical analysis of user's queries and time windows represented as short periods of user's inactivity.

Recent works tackle the problem of search session segmentation by lexical, semantic and behaviour driven approaches [1, 2, 3]. Most of the approaches are based on utilization of time windows within defined interval. Jones and Klinkner [2] identified the task identification problem as a supervised classification problem, and tried four different timeouts (e.g. 5, 30, 60 and 120 minutes). Time-based approaches lack the precision in segmenting search session [2], since many search task are interleaving and user's search intent might span multiple days. Therefore, time-based or temporal features (e.g. inter query threshold or thresholds of user's inactivity) are usually used in conjunction with lexical or semantic features.

Most of the proposed approaches consider lexical features of queries for determining query similarity [1, 2, 4]. Jones and Klinkner [2] identified normalized levenshtein distance as the best lexical similarity measure for identifying goal boundaries. Semantic features of queries were utilized only in case of large corpora like Wikipedia[1] or Probase[2] [4, 5]. Proposed approaches propose 'wikification' for extending the meaning of a query in terms of concepts mined from Wikipedia or Probase.

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

[1] Wikipedia: https://www.wikipedia.org/

[2] Probase: http://research.microsoft.com/en-us/projects/probase/

Similar approaches for segmenting search sessions focus on utilization of user's behaviour described as Markov model [6]. Markov model proposed by authors in [6] outperforms other methods based solely on lexical or semantic features.

In our work, we focus on utilizing user activity during search for extending existing lexical and time factors. By analysing user search activity such as clicks and dwell time on search results, we better understand which search results are relevant for user's current information need. Thus, we utilize user's implicit feedback to determine relevance between queries by search results they share. Strong relationships between queries provide similarity measure between queries by the number of shared link adjusted by user's implicit feedback. Semantic analysis of queries and search results snippets is another factor we introduced for clustering queries into sessions. Utilization of encyclopaedias like Wikipedia and Freebase can provide a way of understanding concepts and user's intention behind the query and thus provide another clustering factor. We plan to integrate our model of weighted factors utilizing user activity and semantic analysis to existing search engines or servers like Elasticsearch.

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

References

[1] Ozertem, U., Chapelle, O. Learning to Suggest : A Machine Learning Framework for. In *SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 2012. pp. 25–34.

[2] Jones, R., Klinkner, K.L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008. pp. 699–708.

[3] Chen, E. Context-Aware Query Suggestion by Mining Click-Through and Session Data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08* (2008). 2008. pp. 875–883.

[4] Hua, W. et al. Identifying users' topical tasks in web search. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*. 2013. pp. 93.

[5] Lucchesse, C. et al. Identifying Task-based Sessions in Search Engine Query Logs Categories and Subject Descriptors. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11* (2011). pp. 277–286.

[6] Hassan, A. et al. Beyond DCG: User Behavior as a Predictor of a Successful Search. pp. 221–230.

Method for Novelty Recommendation Using Topic Modelling

Matúš TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matus@tomlein.org

There is a large number of news and other articles being published on the web by various large and small portals every day. However, the content in these articles is often repeated among them and most of them contain little novel information.

When a person reads an article about a specific topic, it is very likely that they might find dozens of similar articles on the Web, giving the same information in a different way. Such articles are not interesting to the reader. However, there might also be numerous related articles that contain significant novel and interesting information. The problem we deal with is to identify articles with novel and relevant information.

There were several attempts to create news recommender systems that applied novelty detection methods to provide an interface for users to find articles with novel information. They applied various difference metrics for novelty detection, like inverse cosine similarity, Kullback-Leibler divergence, density of previously unseen named entitites, quantifiers and quotes. Using similarity measures as the basis for novelty detection is also common in other works, which mostly focus on sentence-level novelty detection. Sentence-level novelty detection was the main topic of several TREC Novelty track workshops.

So far, the use of topic modelling in novelty detection has not been widely explored. However, there has been a research comparing topic modelling with cosine similarity in novelty detection and it showed promising results in favour of topic modelling [2].

Our method uses topic modelling in order to calculate the novelty and relevancy of articles. Topics are sets of relevant words with some probabilistic degree of distribution with them [2]. We use the Latent Dirichlet Allocation algorithm for topic modelling.

The reason why we think topic modelling can be useful in novelty recommendation is that it provides a way to work with the information in articles on a higher level of abstraction. It allows us to work with information using topics as

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

opposed to using keywords. This is particularly useful if we want to track similar information across articles and find novel groups of information in them.

It is important to ensure that the recommended articles are relevant to the interests of the readers, i.e. to what they previously read about. To achieve this, we cluster articles into groups based on their similarity and recommend novel articles from within the clusters that the user previously read about. We designed a simple method for clustering articles based on their topics. We decided to design and implement our own method because we wanted to make use of our topic model and for its simplicity.

Topics retrieved from LDA have various qualities. Some contain important information; some are just groups of words without significant importance or meaning. These less important topics can have an impact on the performance of our method and so it is useful to give them a lesser importance when considering their contribution.

We also want to give a lesser importance to topics that group information the user already read about. This is a crucial part of our method that ensures the novelty in our recommendation. To meet this goal, we employ topic ranking. We give each topic a numeric rank that represents its importance and novelty to the user. The rank of a topic is calculated separately for each user based on their user model.

We use an algorithm inspired by the method proposed in [1] that calculates the novelty of an article based on the Inverse Document Frequency (IDF) of its terms. We use the average IDF of the 100 best terms of a topic to calculate its rank. As the corpus of documents for calculating the IDF against, we use the articles the user read.

We evaluated our method in a preliminary experiment. The goal of the experiment was to find out the advantages and disadvantages of our method compared to two other commonly used methods for novelty detection.

The experiment went on for a day and a half and 5 subjects (university students) took part in it. They compared 152 pairs of articles. The articles being compared were retrieved from several well-known tech blogs.

We compared our method with two baseline methods used for novelty recommendation: inverse cosine similarity and IDF scored novelty detection [1].

The results were optimistic, giving better results in terms of relevancy of articles and also in recommending articles the participants chose to read next. The method for novelty detection using IDF scoring of terms gave better results in terms of novelty, however their relevancy was poor.

Amended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 185-190.

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

References

[1] Karkali, M., Rousseau, F., Ntoulas, A.: Efficient Online Novelty Detection in News Streams.

[2] Sendhilkumar, S., Nandhini, N., Mahalakshmi, G.: Novelty Detection via Toppic Modeling in Research Articles. airccj.org, 2013, pp. 401–410.

Evaluating Context-aware Recommendation Systems Using Supposed Situation

Juraj VIŠŇOVSKÝ[*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
visnovsky.j@gmail.com

Among many solutions coping with content adaptations to users' needs, recommendation systems seem to stand above all. The main purpose of recommendation system is to deliver relevant information to the user and thus simplify his navigation in data overflow.

User's actions and decisions, while using software system, are influenced by many factors. These factors also known as contexts describe the situation and the environment of the user. Personalized systems may collect a wide range of different types of context. Such may be user's current mood, health condition, activity, location, etc. Including contexts in the recommender system, makes it possible to generate more specific recommendations. This may lead to recommendation quality enhancement.

When evaluating or choosing one of several recommendation algorithms, we may employ one of three possible approaches: offline evaluation, online evaluation and user study [3]. Offline evaluation is easy and cheap to conduct, but its outputs are less reliable than outputs from online evaluation and user study [4]. Although, both user study and online evaluation are more accurate, they have one major drawback, which is their costly conduct [4].

In this paper, we focus on improving user study evaluation approach for context-aware recommendation systems.To reduce the drawbacks of the user study approach (setting up environment according to contexts and gathering satisfying set of users) we propose using supposed situations. Our goal is to maximize the amount of evaluated recommendations.

Proposed evaluation approach requires input data of context-aware recommendations and user study participants. Firstly, we analyze recommendations data and identify its frequent patterns. At the same time user study participants are being modelled. Participant's model reflects test his ability to evaluate recommendations using supposed situations. We identified and analyzed two approaches to measure participants' reliability. A general approach is based on measuring social perspective taking skills, which reflects the participant's tendency to

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

adopt psychological perspectives of another person. Another, more specific approach is based on experiences of experiment participant.

We conducted a pilot experiment to evaluate reliability of user study participants. In this experiment we let six participants fill out a form defining their short- and long-term contexts. Then we addressed them Interpersonal reactivity index inquiry. Finally, we let the participants to rate twenty recommended items.

In the following step, we let a subset of the original set of participants to evaluate a random set of supposed situations, based on explicit feedback from participants acquired in the previous phase. Experiment participants evaluated a total of 87 supposed situations. We observed that good results acquired by inquiry measuring social perspective taking skills did not affect the ability of participants to correctly evaluate supposed situations. Then we examined effects of the rate of experience with presented contexts on evaluation accuracy. We observed that there is no correlation between these two variables.

In our next experiment we will introduce two approaches to measure experiment participant's reliability. We will adopt Gehlbach's inquiry [2], which outputs should be more reliable than the outputs of Interpersonal reactivity index inquiry [1], used in pilot experiment, as it is harder for the participants to sanitize its results. We also assume that higher number of participants will be beneficial as it will bring diversity.

Then, we plan to carry second experiment to evaluate proposed user study evaluation approach. As our goal is to improve user study coverage, we will examine the attributes of common user study approach with proposed user study experiment using supposed situations. Our will concentrate mostly on the number of recommendations covered by each approach. We also expect our approach's accuracy to be slightly inferior to common user study experiment, but we believe in increase of F-measure.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 80-85.

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

References

[1] Davis, M.: A multidimensional approach to individual differences in empathy. In: *JSAS Catalog of Selected Documents in Psychology*, American Psychological Association, 1980, vol. 10, no. 4, pp .85.

[2] Gehlbach, H.: A New Perspective on Perspective Taking: A Multidimensional Approach to Conceptualizing an Aptitude. In: *Educational Psychology Review*, Kluwer Academic Publishers, 2004, vol. 16, no. 3, pp. 207–234.

[3] Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: *Recommender systems handbook*, 2011, pp. 257–297.

[4] Shani, G., Gunawardana, A.: A survey of accuracy evaluation metrics of recommendation tasks. In: *The Journal of Machine Learning Research*, 2009, vol. 10, pp. 2935–2962.

Traveling in Digital Space (PeWe.DS)

Automated Syntactic Analysis of Natural Language

Dominika ČERVEŇOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
cervenova.dominika@gmail.com

Natural language as one of the most common means of expression is also used for storing information on the web. To work them more effective and faster, we need to process text in natural language in a way that computers could understand.

Natural language processing is, however, a difficult and problematic process, because of the informality and not very good structuring of the natural language. The processing typically consists of the sequential application of different analysis components, which try to solve several problems such as phonological, syntactic, or context ambiguity, homonymy, polysemy, etc. Syntactic analysis, as a part of the natural language processing, discovers formal relations between syntagms in a sentence and assigns them syntactic roles. That can help make natural language and information stored using it more machine-processable.

Our goal is to analyze possibilities of maximizing the automation of this process and to minimize human manual work. We are working on a method that will be able to automate the syntactic text analysis process as much as possible. Currently, we focus on analyzing existing tools for various languages. There already are some parsers that can perform syntactic analysis in languages that are more simple and easier to formalize (like English, for instance), but we are also exploring options for Slavic languages (e.g., Russian, Czech or Slovak language) where automated syntagms recognition is a nontrivial problem.

In general, there are many approaches to automated syntactic analysis. Machine learning, for example, appears to be very useful in this domain. With enough training data – e.g., corpus of annotated sentences for specific language – it is possible to train a parser to recognize syntagms with state-of-the-art accuracy. One of the greatest advantages of this approach is that we can use one parser to analyze any language we have a corpus for [1]. Having enough pre-annotated data there is no need to have special linguistic skills to make parser work for any language. However, the accuracy

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 21–22.

of this type of parser varies depending on amount of specific features and rules of the language and it also depends on a quality of the annotations used for training.

Another successful approach is a rule-based parsing. This approach was also used for Slovak language by Čižmár et. al [2]. Their parser, based on rules created using *Pravidlá slovenského pravopisu* and Slovak national corpus, can recognize 98 % of predicates and subjects; objects, adverbials and attributes were recognized correctly in 72-85 % of all cases. These results are good but it is important to recognize not only syntagms alone, but also relations between them. Moreover, especially for adverbials and attributes, the recognition should be more accurate.

Creating automatic parser for Slovak language is a difficult task and there is currently no tool or approach that would provide such accuracy for Slovak as there are e.g., for English. Our aim is to create a method that will be able to recognise syntagms and relations between them automatically with as much accuracy as possible. We plan to use approach similar to machine learning, probably with some extensions. As a training data syntactic annotations made be people at the Slovak National Corpus at Ľudovít Štúr Institute of Linguistics will be used.

To create a method, applicable on a syntactic layer of a language, we have to analyse a morphological layer first. To syntagms recognition, we need to know for instance lemma of every word. Morphological information are also included in the annotations of Slovak National Corpus, however, they are not always complete. As a first step we need to fill the missing values. We plan to use some of the existing tools here. Unlike syntactic analysis, the morphological has been more successful in Slovak language. Even at our faculty there has been made a research in this field before.

Before we create our own parser, we will try to use data from corpus to train existing parsers, originally made for other, but similar languages, e.g., Czech, Slovenian or Russian. This could help us to identify many problems, connected with language differences, we should be aware of, by creating our own method.

We plan to evaluate our method using a software prototype and as a gold standard a part of syntactic annotations of the Slovak National Corpus will be used.

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

References

[1] Buchholz, S., Marsi, E. (2006). Conllx shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X),* ACL, pp. 149–164.

[2] Čižmár, A., Juhár, J., Ondáš, S. (2010): Extracting sentence elements for the natural language understanding based on slovak national corpus. In *Proceedings of the International Conference on Analysis of Verbal and Nonverbal Communication and Enactment, COST'10,* LNCS Vol. 6800, Springer, 2010 pp. 171–177.

Linking Slovak Entities from Educational Materials with English DBpedia

Ľuboš DEMOVIČ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
demovic@gmail.com

Currently, the Web provides a large amount of information, important facts and services to access them. It also has the potential to become the largest source of information and it is, therefore, desirable to be able to automatically gather information about entities and their relationships.

Data published on the Web is largely unstructured, without clear marking of entities and their relationships [1]. Linked Data describe a set of principles for publishing interlinked structured data.

Availability of vast amounts of interconnected data opens up great opportunities to create new generations of applications capable of using these Linked Data structures. With intelligent processing of the available data we can create useful methods aimed at: quick search, translation, personalization, recommendation, and context enrichment or user navigation [2].

In our research work we propose a method that allows the identification and extraction of entities and facts in the Slovak language using search queries. Subsequently, we use the acquired facts for correct and automated linking with the English DBpedia dataset. As DBpedia is the core of the Linked Open Data cloud, such interconnection of entities enables us to use additional information from various datasets. The results can be used for various tasks of personalized Web, e.g., for enriching the information presented to the user with additional facts.

For the purpose of linking Slovak entities, we use the Wikipedia API. One of its services is the OpenSearch service that returns the users' demand for the most popular Wikipedia articles. It is like the autocomplete for Wikipedia, thus for a specific term it returns several articles sorted by popularity. Next, we choose the best result from the offered choices. Now we have linked a Slovak entity with the Slovak Wikipedia.

We use another service from the Wikipedia API to link the Slovak Wikipedia article with all its available linguistic variations throughout Wikipedia. This is convenient, because most articles on Slovak Wikipedia have its English counterpart. At

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

this point, we have a Slovak entity link with the English Wikipedia, which means that we also have a link to the English DBpedia.

The algorithm for linking of entities in the Slovak language consists of five steps:

1. Search for entities via the Wikipedia OpenSearch API.
2. Find the English version of the entity for the selected result via Wikipedia API.
3. Disambiguation check of entities.
4. Choose the correct DBpedia URI
5. Link with the DBpedia dataset.

Figure 1 schematically shows how the algorithm works. We extract the entities using popular search queries from categorized study materials and subsequently we link these entities with DBpedia. As a result, we get DBpedia's entity together with all of its connections.

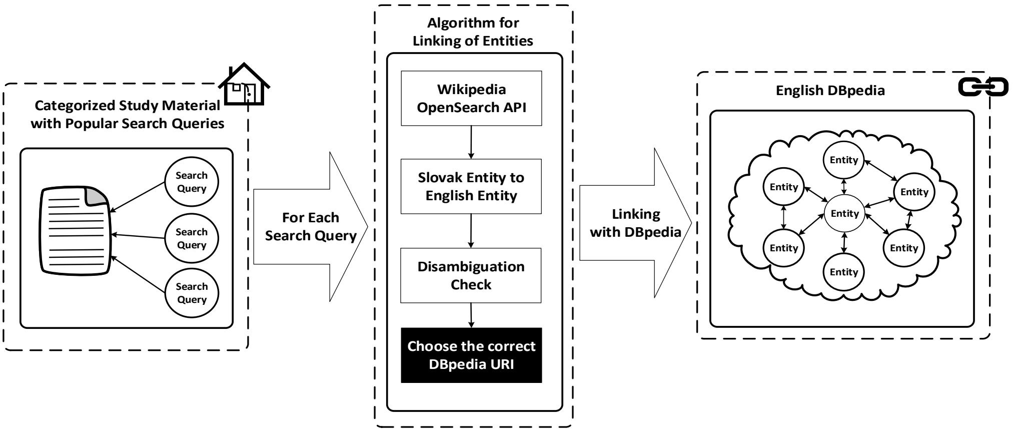


Figure 1. Diagram of the algorithm for linking of entities in Slovak language with DBpedia

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 137-142.

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

References

[1] DeRose, P., Shen, W., Chen, F., Doan, A., & Ramakrishnan, R. (2007). Building structured web community portals: A top-down, compositional, and incremental approach. In *Proceedings of the 33rd international conference on Very large data bases*, pp. 399-410. VLDB Endowment.

[2] Weikum, G., Theobald, M. (2010). From information to knowledge: harvesting entities and relationships from web sources. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 65-76. ACM.

Utilization of Linked Data in Domain Modeling Tasks

Michal HOLUB*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
michal.holub@stuba.sk

The idea of Semantic Web has found many followers among the web researchers. A lot of datasets publicly available use the Linked Data principles. These may be a perfect source for additional metadata utilizable in various tasks of web personalization, recommendation, information retrieval, and data processing. However, only few works actually pursue the idea of wider adoption of such datasets. The aim of our work is to use Linked Data for various tasks of web personalization.

We propose a method for relationship discovery among various concepts forming a concept map. The concept map is intended as a basis of domain models. For this purpose we use unstructured data from the Web, which we transform to concepts and discover links between them. Primarily, we aim at creating concept maps usable for recording the technical knowledge and skills of software engineers, and research fields of interest of scientists (especially in the domain of information technologies).

We examine the utilization of such concept maps and Linked Data utilization in order to 1) improve the navigation in a digital library based on what the user has already visited [7], 2) find similarities between scientists and authors of research papers and recommend them to the readers browsing a digital library [2], 3) analyze Linked Data graphs and find identities and similarities between various entities [5], 4) enhance the displayed articles based on linking entities to DBpedia [4] and recommend additional interesting information to the reader within a digital library, 5) enable users to search for information using natural language queries in English [3].

In all of the problems mentioned above we were able to improve the results of current research methods. We also proved that using Linked Data and concept maps in such problems has potential to improve results in various Web personalization and adaptation tasks.

Linked Data are being used in various datasets forming a Linked Data Cloud. In the center of this cloud there are two main datasets: DBpedia [1] and YAGO [6]. Both

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

use Wikipedia as their primary source of information. The goal of these datasets is to extract and define as many entities as possible, so that others can link to them.

For the purpose of describing the knowledge of software developers we propose the creation of a concept map. The map is composed of a set of concepts representing various technologies and principles the developers are familiar with.

We use a similar approach to describe the professional interests of IT researchers. However, in this case the concept map includes concepts representing various research areas, problems, principles, methods and models studied by the researchers.

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

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.

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.).

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, (2007), pp. 503-517.

[2] Chen H., Gou L., Zhang X., Giles C.L.: CollabSeer: A Search Engine for Collaboration Discovery. In: *Proc. of the 11th Annual Int. ACM/IEEE Joint Conf. on Digital Libraries*, ACM Press, (2011), pp. 231-240.

[3] Chong W., Xiong M., Zhou Q., Yu Y.: PANTO: A Portable Natural Language Interface to Ontologies. In: *The Semantic Web: Research and Applications*, LNCS Vol. 4519. Springer (2007), pp. 473-487.

[4] Exner P., Nugues P.: Entity Extraction: From Unstructured Text to DBpedia RDF Triples. In: *The Web of Linked Entities Workshop*, Boston, USA, (2012).

[5] Halpin H., Hayes P.J., McCusker J.P., McGuinness D.L., Thompson H.S.: When owl:sameAs isn’t the Same: An Analysis of Identity in Linked Data. In: *Proc. of the 9th Int. Semantic Web Conf. on The Semantic Web – Volume Part I*. Springer (2010), pp. 305-320.

[6] 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, (2007), pp. 697-706.

[7] Waitelonis J., Harald S.: Augmenting Video Search with Linked Open Data. In: *Proc. of Int. Conf. on Semantic Systems*, Verlag der TU Graz, Austria, (2009).

Query by Multiple Example Considering Pseudo-Relevant Feedback

Adam LIESKOVSKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
adam@lieskovsky.com

Digital libraries aggregate enormous amount of diverse publications and are therefore great centralized resources of information for scientists and researchers. With the increase of available information volume, which is prevalent in the space of digital libraries as well as in the other parts of Web, comes hand in hand the difficulty of searching and acquiring documents that are most relevant for the user's query. When using these systems, searching itself is not the primary activity of the user; on the contrary the user is trying to use and apply gained information later as a source for his work. Selection among search results requires user's attention, so it is a general effort and tendency that methods and interfaces used on the web are intuitive and effective.

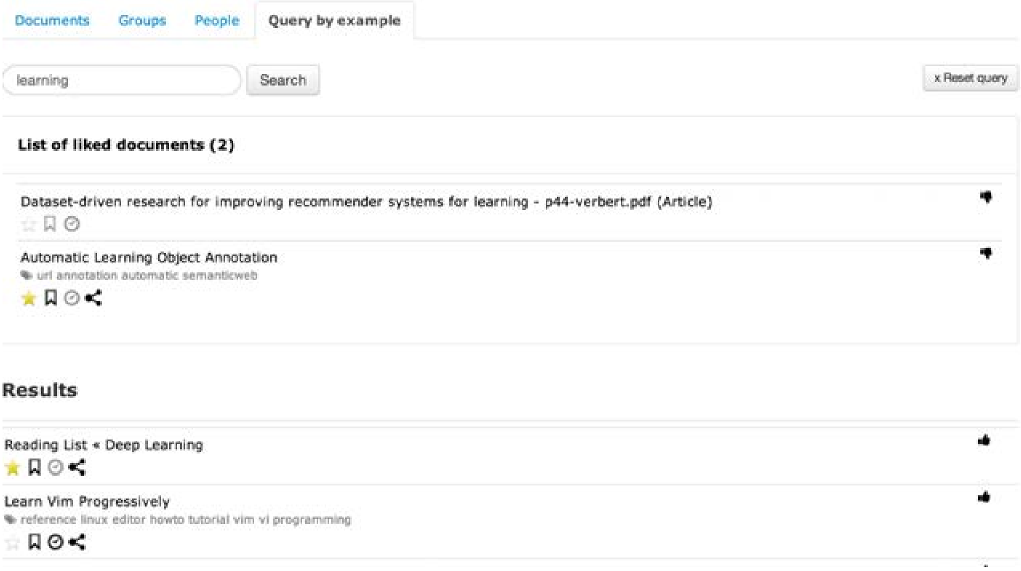
Nowadays the most popular approach of searching is searching by keywords. Novice users often cannot use it the proper way. Prerequisite for achieving satisfying results is accurate formulation of query by identifying correct keywords and certain level of knowledge of the searched domain. One of the alternative approaches that can overcome mentioned problems is *query by example* which has become fairly popular in the domain of multimedia [1, 2], because of the available metadata, that describes the media. We work with a query by multiple examples, where user selects documents as examples of relevant results which are further used to formulate the query.

In our work, we propose a method of iterative formulation of a query by multiple examples using user's explicit feedback for the query formulation. At the beginning, user starts with single keyword search. We take *n* most relevant (top ranked) documents for user's query and apply pseudo relevance feedback [3], which automates the user's evaluation of results and assumes that these documents might be a good source of metadata, for our further query.

Related documents and their similarities are being judged solely from the aspect of available metadata omitting any further full-text analysis. Beside author, author keywords, year, title and publication we have available metadata about references and citations between bookmarked documents. User can explicitly state and select some of

[*] Supervisor: Róbert Móro, Institute of Informatics and Software Engineering

the returned documents as positive examples and impact of these documents' metadata on the query will be increased.



List of liked documents (2)

Dataset-driven research for improving recommender systems for learning - p44-verbert.pdf (Article)

Automatic Learning Object Annotation
url annotation automatic semanticweb

Results

Reading List « Deep Learning

Learn Vim Progressively
reference linux editor howto tutorial vim vi programming

Figure 1. Query by example interface showing explicit feedback in Annota. Documents explicitly selected by a user (by a thumbs-up button) as positive examples are shown in the list of liked documents.

We implemented and evaluate our proposed method in Annota[1] (see Figure 1), which is a system for bookmarking and annotating digital documents. Firstly we would like to test our created interface, with the use of an eye tracker in UX lab with a follow up questionnaire. We also intend to evaluate the effect of pseudo relevance feedback on the results of a query.

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

References

[1] Helén, M., Lahti, T.: Query by example methods for audio signals. In: *Proc. of the 7th Nordic Signal Processing Symposium*, (2006), pp. 302–305.

[2] Rasiwasia, N., Vasconcelos, N.: Image retrieval using query by contextual example. In: *MIR '08: Proc. of the 1st ACM Int. Conf. on Multimedia Inf. Retrieval*, (2008), pp. 164–171.

[3] Wu, H., Fang, H.: An incremental approach to efficient pseudo-relevance feedback. In: *SIGIR '13: Proc. of the 36th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, (2013), pp. 553–562.

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

Researcher Modeling in Personalized Digital Library

Martin LIPTÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 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 rapidly. This has two serious consequences. First, many interesting works are unnoticed. Secondly, researchers spend too much time reading articles that turn out to be low-quality ones, 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 is used to personalize any feature of the 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, information on user knowledge is essential in educational domain [2], but in domain of digital libraries, other characteristics of the user like interests become important. In case of digital libraries, we deal with researcher models.

We propose a researcher model that leverages various user data available in digital libraries. We implemented the model in the Annota digital library[1], other digital libraries offer more or less user and domain data.

- papers the user has stored in her library
- papers the user has authored
- user’s activity on the ACM web pages of papers
- tags and folders the user has used for organizing her library
- tags and folders the user has used for navigation
- terms the user has searched for

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

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

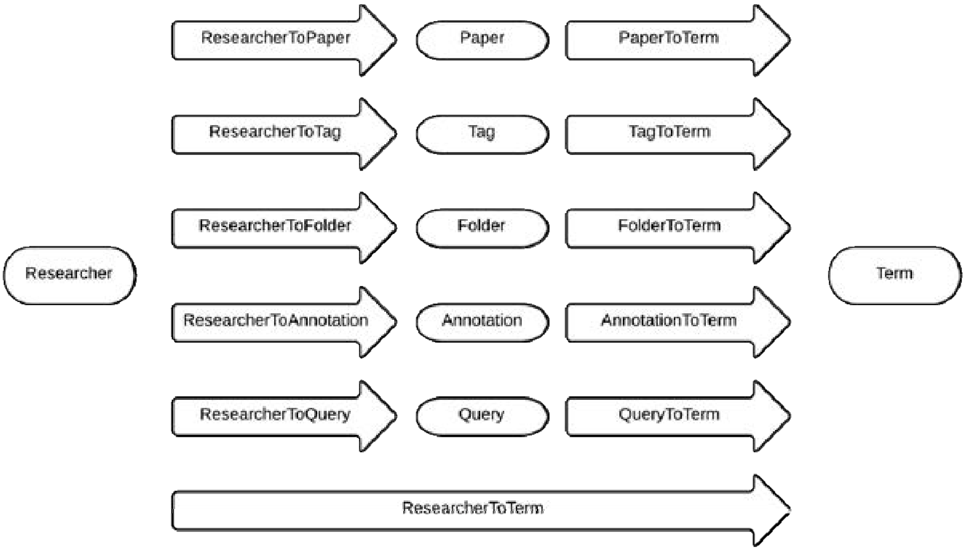


Figure 1. Researcher Model.

Digital libraries thus contain numerous relations between papers, authors, tags, etc. A graph is suitable to reflect these relations. We propose a researcher model (see Fig. 1), which combines the existing relations in the digital library to create new relations.

The model comprises multiple entities and relations. The final relation is *ResearcherToTerm*. Its weights represent a vector of terms relevant to the researcher. The researcher model is a vector of terms from outside, but a graph inside. Therefore the components of the model are reusable and the model is extensible.

We evaluate the model by investigating how the researchers perceive their computed interests. We plan to perform the experiment using a game with purpose.

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 W. on Int. and Knowledge oriented Tech., (2012), pp. 197-200. (in Slovak)

Using Navigation Leads for Exploratory Search in Digital Libraries

Róbert MÓRO[*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
robert.moro@stuba.sk

Researching a new domain in a digital library can be a challenging task even for a seasoned researcher and more so for novice ones, such as starting master or doctoral students. They can have a hard time formulating keyword queries, because they lack the needed domain overview and knowledge. The most natural way of navigation seems browsing, which does not force users to split their attention between the navigation interface and the search results. In addition, it supports the idea of navigation-aided retrieval as defined in [4] by understanding the search results as mere starting points for further exploration.

In order to emulate this behaviour and support exploration of the domain by the users, we provide them with navigation leads, i.e. links to relevant documents, with which we enrich the documents' summaries (or abstracts). In our work, we examine the influence of the navigation leads' different visualizations on the users' performance of exploratory search tasks in a digital libraries domain.

The eye-tracking technology has been increasingly utilized for evaluation of search systems. In [3] it was used to determine, at which elements of faceted search interface the users look the longest, finding out, that on the first results page they spend the most time looking at the search interface and facets and then on the consecutive pages they do not pay it much attention and examine the filtered results.

Other important metrics when evaluating exploratory search are learning and cognition which are closely linked to the concept of cognitive load. Measuring cognitive load can be a challenging task. The procedure usually consists of secondary task performance and subjective ratings [1], however other techniques can be used, such as concurrent and retrospective reporting, eye-tracking or concept-mapping [3].

In order to evaluate how the different visualizations of navigation leads affect the users' performance of exploratory search tasks, we have proposed three types of visualization, namely *visualization in text* of an abstract (or summary), *under the text* and *in a cloud of terms* next to the list of search results. Advantage of the first

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

approach is that the leads can be viewed in their context; on the other hand, only words or phrases already present in abstracts can be used as navigation leads. When visualizing the leads under the text, they lose they immediate context, but do not get in the way of reading. Lastly, visualization in a cloud tries to mimic the tag cloud with one exception – only leads for currently retrieved set of documents are selected into the cloud. For each visualization type, the same method of navigation leads' selection was used which computes relevancy of keywords extracted from abstracts by AlchemyAPI[1] by taking into account also tags added by users and keywords identified by the articles' authors.

We conducted a user study with five participants – bachelor and master students, whose task was to explore new domain using the provided navigation interface in a web-based bookmarking system Annota[2]. We hypothesized that visualizing leads in the text or under it will prove to be more immersive, thus resulting in more interaction with the search results (more time spent reading the abstracts as well as more read abstracts) in comparison with the visualization in a cloud and that consequently the users will acquire better understanding of the domain and the problem at hand with less (extraneous) cognitive load.

In order to evaluate our hypotheses we used different metrics, such as task success, time spent on task etc. We also collected gaze tracking data using the eye-tracker *Tobii X2* and used *Tobii Studio*[3] for the evaluation of the collected data. The audio and video of the experimental sessions was recorded as well.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 212-219.

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

References

[1] Cierniak, G., Scheiter, K., Gerjets, P.: Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, (2009), vol. 25, no. 2, pp.315–324.

[2] Gog, T. Van et al.: Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, (2009), vol. 25, no. 2, pp.325–331.

[3] Kules, B. et al.: What do exploratory searchers look at in a faceted search interface? In: *JCDL '09: Proc. of the 9th ACM/IEEE-CS Joint Int. Conf. on Digital libraries*, ACM Press, New York, (2009), pp. 313–322.

[4] Pandit, S., Olston, C.: Navigation-aided retrieval. In: *WWW '07: Proc. of the 16th Int. Conf. on World Wide Web*, ACM Press, (2007), pp. 391–400.

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

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

[3] http://www.tobii.com/en/eye-tracking-research/global/

Collocation Extraction on the Web

Martin PLANK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
plank.martin@gmail.com

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.

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).

Other approaches employ methods based on the linguistic properties of collocations. Manning and Schütze [2] describe these characteristic properties:

– Non-(or limited) compositionality. The meaning of a collocation is not a straightforward composition of the meanings of its parts.

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

- Non-(or limited) substitutability. The parts of a collocation cannot be substituted by semantically similar words (synonyms).
- Non-(or limited) modifiability. Many collocations cannot be supplemented by additional lexical material.

In our work, we propose a novel method based on the limited modifiability of collocations. The modifiability of a combination of words can be computed according to the frequencies of n-grams. The method can be explained on a simple example of collocation *to pull my leg* (to tell me something untrue):

1. The frequencies of collocation candidate and its components are computed. Also, the headword (important in the next steps) is identified (*leg*). Headword is one of the collocation components, which has a high semantic significance.
2. Frequent bigrams, which contain the headword, are identified (*right leg, long leg,* etc.). We call the certain number of the most frequent bigrams *headword supplements*. Their frequencies are computed, too.
3. The candidate is modified by the headword supplements. The result is a list of candidate modifications: *to pull my right leg, to pull my long leg,* etc.
4. The frequencies of headword supplements and candidate modifications are compared and the modifiability of a collocation candidate is computed.

The process of judging the collocation candidate is based on the following hypothesis: If the frequencies of candidate modifications are significantly lower than the frequencies of the original candidate, its components and the headword supplements, the candidate is probably a collocation. In other words, if the candidate can be supplied by other lexical information, it has high modifiability. Otherwise, if it cannot be supplied, it has low modifiability and is probably a collocation.

In the experiments, we compared our method with the state-of-the-art association measure *pointwise mutual information*. We compared precision and recall of these methods. Both methods achieved comparable results.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 161-166.

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

References

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

[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: *Proceedings of the ACL Student Research Workshop*. ACLstudent '05, Association for Computational Linguistics, 2005, pp. 13-18.

Discovering Identity Links between Entities on the Semantic Web

Ondrej PROKSA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
ondrej.proksa@gmail.com

In our times, few millions of specific pages increase on the World Wide Web daily, from which Linked Data can be obtained. Linked Data appear on the Web in different form and goal of gathering structured data is a better possibility of device processing.

Currently, for the purpose of connecting entities between different datasets, the relationship of identity (also known as owl:sameAs) is widely used. However, it turns out, that in some cases the claim, that two entities are identical, is incorrect [3]. The problem might be caused when we have people who create the relationship with their namesakes or there are different entities that are connected with relationship of identity, but their properties are not fully identical. In the case, when two entities are identical, it is possible to create new relationships, which enrich Linked Open Data cloud with relationships across the datasets.

Our main goal is to discover relationships of identity between entities in order to create new connections between existing data sources and datasets in the Linked Data Cloud. Our goal is to propose a method, that will automatically search for connections and owl:sameAs relationships using graph algorithms and specific rules. We use similarities from sub-graphs of properties and classes for determination of identity between two entities. In the case that entities are identical, it is possible to enrich new relationships, which are across the datasets in sub-graphs.

Our proposed method is universal to any particular domain, because it uses sub-graph of properties and classes for entities comparison. Method is applicable on any particular graph, in which similarity relationships between the entities exist. It is possible to use the method in the search for identical organizations among the public government data as well as in discovering duplicate authors with data from digital libraries and also for connecting a new dataset to the cloud of Linked Open Data.

Because the Linked Data datasets use various ontologies to describe their content, there is a problem of ontology diversity, which could cause that identical entities are not connected using owl:sameAs. The method performs ontology alignment [4] on

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

each of the found graph patterns. Finally, it aggregates similar ontology classes and properties. The method was evaluated on four datasets form the Linked Open Data cloud and using this method the authors were able to discover new, missing relationships between the datasets.

We propose a method for finding similarities between entities in a graph. We use this similarity to determine whether two entities are identical or not. This determines whether they should be linked using owl:sameAs relationship. We based our method on a hypothesis that the matching of entities is reflected in the similarity between the sub-graphs composed of classes and properties of the individual entities. This approach was also explored in the ontology matching problem [1].

The similarity between entities (sig. SGN) depends on the similarity of their properties, graph distance between entities and graph distance between neighbouring entities. We define the total similarity as a sum of similarities of its individual components:

- The similarity of properties between entities
- Distance between the entities
- Average distance between adjacent entities

The resulting values of the similarity are from the interval <0, 1>. SGN = 1.0 means that the two entities are 100 % similar (i.e., identical), where as SGN = 0.0 means that the two entities are not similar at all (their similarity is 0 %).

We have developed a prototype and evaluated it on a dataset from domain digital libraries. We analyzed the new dataset Annota [2] that was created using the principles of linked data. We tried to find duplicities authors comparing our method. We have proposed two possibilities find candidates for same authors.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 167-172.

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

References

[1] Aumueller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and Ontology Matching with COMA++. *Proc. of the 2005 ACM SIGMOD International Conf. on Management of Data*. SIGMOD ’05, New York, NY, USA, ACM, 2005.

[2] Bieliková, M., Ševcech, J., Holub, M., Móro, M.: Annota - poznámkovanie dokumentov v prostredí digitálnych knižníc. *Proc. of the Annual Database Conf.*

[3] Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameAs isn’t the same: an analysis of identity in linked data. *Proc. of the 9th international semantic web conference on The semantic web* - Volume Part I. ISWC’10, Berlin, Heidelberg, Springer-Verlag, 2010.

[4] Zhao, L., Ichise, R.: Graph-based Ontology Analysis in the Linked Open Data. *Proceedings of the 8th International Conference on Semantic Systems*. I-SEMANTICS ’12, New York, NY, USA, ACM, 2012.

Exploring Multidimensional Continuous Feature Space to Extract Relevant Words

Márius Šajgalík*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
marius.sajgalik@stuba.sk

With growing amounts of available text data the descriptive metadata become more crucial in efficient processing of it. One kind of such metadata are keywords, which we can encounter e.g., in everyday browsing of webpages. Such metadata can be of benefit in various scenarios, such as web search or content-based recommendation.

In our work we focus on vector representation of words to simulate the understanding of word semantics. Each word is thus represented as a vector in N-dimensional space, which brings the advantages of using various vector operations like easy similarity measuring between pairs of words, or vector addition and subtraction to compose meaning of longer phrases. We can also calculate what words are the most similar by finding the closest vectors, or a vector that encodes a relationship between a pair of words, e.g., a vector transforming singular form into plural one, etc. With word vectors, we can encode many semantic and also syntactic relations [3].

Such representation has a big potential in natural language processing (NLP) and we can only anticipate that in a few years it will completely supersede all those manually crafted taxonomies, ontologies, thesauri and various dictionaries, which are often rather imprecise and erroneous. Moreover, there is no means of measuring similarity directly between pairs of words in this hand-crafted data. Most relations are just qualitative and described by their type (e.g., an approach in [1] reveals only a relation type, but it cannot determine this relation quantitatively) and thus, all existing methods for measuring (semantic) similarity are limited to achieving only rather imprecise results.

We research the computation of keywords in vector space. This perspective on the keywords extraction problem also brings new interesting challenges and unsolved open problems. So far, we developed a method of extracting relevant words in clusters of words with similar meaning (see Fig. 1).

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 37–38.

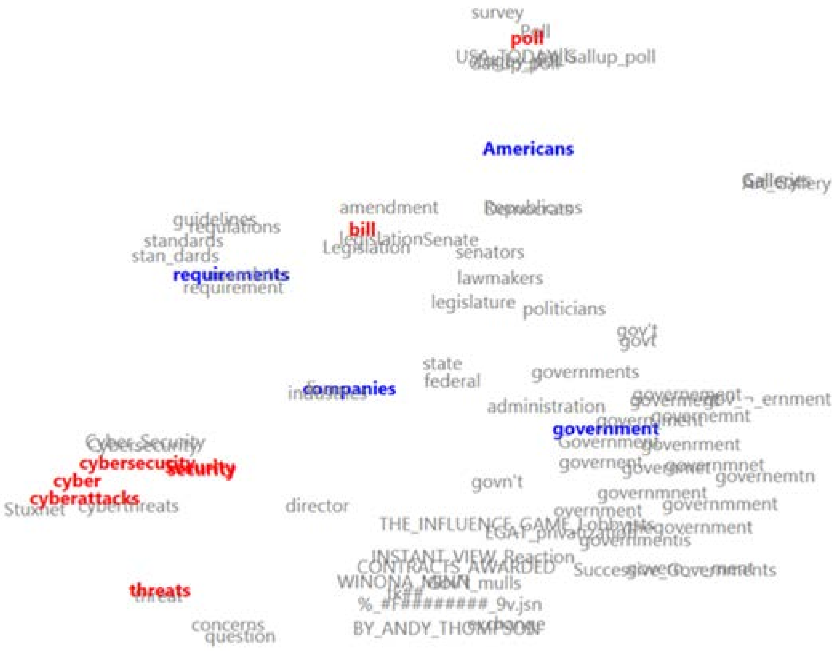


Figure 1: Top 100 most relevant words extracted from a website titled "Cybersecurity poll: Americans divided over government requirements on companies" and visualised by t-SNE [2]. Red words are keywords selected by website editors and blue words are keywords that we would also manually and subjectively choose as the most relevant for this article.

We can see that each cluster contains a keyword, but is cluttered with other similar words that often correspond to less common synonyms or their misspelled alternatives, so that computed data is still noisy and needs to be cleaned. This could be achieved by using frequency statistics, which is our next task to complete.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 93-100.

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

References

[1] Barla, M., Bieliková, M.: On Deriving Tagsonomies: Keyword Relations Coming from Crowd. In: Proc. of the 1st Int. Conf. on Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. pp. 309–320 Springer-Verlag (2009).

[2] Van der Maaten, L.J.P.: Barnes-Hut-SNE. In: Proceedings of the International Conference on Learning Representations, 2013.

[3] Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of NAACL HLT, 2013.

Anomaly Detection in Streaming Data

Jakub ŠEVCECH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
jakub.sevcech@stuba.sk

During the last few years we can hear all over us a buzzword *Big Data*. The definition of this term is rather fuzzy, but one of the most frequent one is that it is a common name for techniques used for processing data which are characteristic by its big amount, big velocity and/or big variability. The most common technique for dealing with such data is batch processing. However this type of processing is not viable in many applications mainly due to delays caused by the batch job processing time. When we require online or close to real time processing we have to consider the data as a stream and thus consider application of various stream processing methods.

The domain of streaming data processing caught increased attention in recent years, but the majority of works published so far focused on search and analysis of streaming data collected in time series databases, hence static collections of stream samples. They achieved substantial successes in search and pattern comparison, association mining, time series prediction and so forth [5]. The majority of these works focused on processing of static data collections and they did not focused on problems specific to processing of streaming data in real or near-real time. Until recently, only minor attention was given to stream processing. With the rising interest in Big Data, this domain is becoming the topic of interest for many researchers and practitioners.

Real time processing of streaming data introduces new restrictions and problems, methods for processing of static collection of time series samples are not designed for [3]. The main restriction is fast, potentially unbound stream of data, which is often accompanied with great variability of the data and the restriction of a single pass through the data during the processing. The main directions of current streaming data processing research are change detection, clustering [1], classification [4], pattern detection and association rules mining [2] and time series analysis.

In our work we approach processing of data streams with focus on methods for anomaly detection. The main challenge is to be able to process big number of various metrics running on the streamed data and to be able to do so in a single pass through the data. Our aim is to create and evaluate a method for anomaly detection in streaming

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

data based on pattern detection on the top of individual metrics running on the data and on classification of stream state using detected patterns and supervised learning.

The first restriction we have to face is the problem of frequent patterns counting. Comparison of patterns against current state of a stream is rather expensive operation. When we want to detect patterns, we have to select a set of patterns that have the biggest probability to be frequent and to limit the set of patterns we will look for in the data. This is thorough problem as we can only use single pass through the data and it is hard to predict future pattern frequency in time of their construction. A work described in [1] is elaborating this problem in further detail and it propose some viable methods for dealing with the problem. In our work, we build on these methods and we use them in anomaly detection in an evolving data stream. We propose a method for anomaly detection composed of two phases:

- *Patterns detection* in metrics running on the data stream. The result of this phase is a set of patterns matched on the stream in every time. The stream of source data is reduced into the stream of matched patterns.
- *Anomaly detection* in the stream of patterns phase uses common machine learning algorithms for anomaly detection and categorization of stream state in specified time window.

We will evaluate the proposed method on various sources of streaming data such as Twitter or Bit.ly and on various applications such as log analysis or standard datasets used in machine learning applications.

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

References

[1] Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S.: A framework for clustering evolving data streams. In: Proc. of the 29th international conference on Very large data bases, VLDB Endowment, (2003), vol. 29, pp. 81-92.

[2] Cheng, J., Ke, Y., Ng, W.: Maintaining frequent itemsets over high-speed data streams. In: Advances in Knowledge Discovery and Data Mining, Springer, (2006), pp. 462-467.

[3] Ling, C., Ling-Jun, Z., Li, T.: Stream Data Classification Using Improved Fisher Discriminate Analysis. In: Journal of Computers, (2009), vol. 4 no. 3.

[4] Wang, H., Fan, W., Yu, P. S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, (2003), pp. 226-235.

[5] Zhao, Q., Bhowmick, S. S.: Sequential pattern mining: A survey. In: ITechnical Report CAIS Nayang Technological University Singapore, (2003), pp. 1-26.

Processing and Comparing of Data Streams Using Machine Learning

Miroslav ŠIMEK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
mirzekdt@gmail.com

From many different approaches of machine learning, multilayered self-teaching neural networks (a.k.a. Deep Belief Networks) using the unsupervised learning approach are gaining popularity nowadays. They did not used to be accepted and were largely ignored by most of the experts in machine learning community for almost 40 years. One of the reasons was simply because of too little computational power of available technology at the time.

However, today it is already producing interesting results for example in computer vision. One of the successful projects using unsupervised learning of neural networks was recognition of human faces based on many snapshots from videos on youtube.com. This model was trained in 2012 for three days on 1000 computers with 16000 cores altogether [1].

There are several strategies for learning multilayer neural networks. Five of them are listed here [2]:

1. Denial
2. Analogy with evolution
3. Procrastination
4. Backpropagation
5. “Wake-sleep” algorithm

The most effective learning turned out to be the combination of strategies 3, 4 and 5. Strategy number 3, called procrastination is about usage of hidden layers. Neural networks with one hidden layer use this layer to find the patterns and features in the input layer. These features are much more useful for deciding on how the output will look like than just the raw input data.

Multilayered neural networks take this approach to higher levels of abstraction. First hidden layer finds the features in the input layer, the second hidden layer finds the

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

patterns and features of the features in the first hidden layer and so on. This approach is also a bit closer to how our brain works with multiple levels of abstraction.

The problem with multilayered neural networks is that the usually very powerful backpropagation algorithm (strategy number 4), does not work here as it is loosing power with every layer. This is where unsupervised learning using strategy number 3 comes useful to pre-train the hidden layers one by one separately to find the patterns and features in layer underneath [2]. After this stage of unsupervised learning the backpropagation algorithm is once again useful to fine-tune the model.

Fifth strategy, "Wake-sleep" algorithm is about usage of both passes through neural network for unsupervised learning, from bottom layer up to the top and from top layer down to the bottom. Bottom-up pass ("wake" phase) is about recognizing the output based on input. The binary states of units in adjacent layers will be used in the "sleep" phase to train generative phase with top-down pass, which is about generating the input based on output. The binary states can be now used in training recognition weights. In "Wake-sleep" algorithm we initialize the weights with small random values and then we alternate between these two passes of learning [2].

Our goal is to find methods and new ways of training to utilize the potential of multilayered neural networks and unsupervised learning to process and compare large streams of unlabeled data. We want to focus on data from eye tracker and simple sound recordings like long pronunciations of single letter for example letter "r".

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

References

[1] Quoc V. Le et al.: Building high-level features using large scale unsupervised learning. In *Proc. of the 29th International Conference on Machine Learning*, ICML 2012, Omnipress, pp. 81-88.

[2] Hinton, G. E.: To recognize shapes, first learn to generate images, In P. Cisek, T. Drew and J. Kalaska (Eds.), *Computational Neuroscience: Theoretical Insights into Brain Function*. Elsevier, 2007

Determining the Relevancy of Important Words in a Digital Library using the Citation Sentences

Máté VANGEL[*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
mate.vangel@gmail.com

The number of articles and publications in digital libraries is enormous and is persistently increasing. Therefore keeping everything organized in the world of digital libraries has become impossible without automation of some of the processes. The main concept is realized by methods of text mining, which is used to extract and link relevant information (objects) from texts. Text mining can be used to fulfil various kinds of tasks like domain modelling, automatic text summarization or navigation in a cloud of keywords.

Keyword extraction is usually done using the text of the specified document, but in digital libraries there are some other possible options for extracting relevant words. One of these possibilities is to use the available information related to the article, which is called metadata.

There are many kinds of metadata in digital libraries, for instance keywords provided by the authors, tags, year of publishing, category in which the article is located and tags associated by users. Citations can also be considered as an important source of keywords, because they can characterize the article. They can also highlight different, but relevant aspects of the analysed article, which can be relevant for other researchers.

There are various possible solutions for keyword extraction. However, they are mainly using only one source of available information – article's main text. We aim to extract keywords with the help of document metadata in our work, and we consider citations as the main source (input) for keyword extraction.

Citations can be used for keyword extraction in many ways. Using citations in keyword extraction can give distinct results compared to extraction from the abstract of the article [2], which is also a very popular and efficient input for extracting and evaluating keywords.

[*] Supervisor: Róbert Móro, Institute of Informatics and Software Engineering

We plan to analyse citations from different aspects. One of the important aspects is the citation context, which means the environment of the cited text [1]. Citation context can be important from two distinct views:

- Where is the cited text located in the article
- What is surrounding the cited text

By using the first view we plan to analyse the structure of the article, because articles in digital libraries share the same structure, i.e. at the beginning of the article there is always the title, then name of the author(s), publication year, abstract, text of the article and the conclusion at the end. Information located in the abstract and the conclusion usually reflect the main facts of the article, so citations from these parts are often more relevant than from other parts.

By using the second view we want to discover the surroundings of the cited text. In its surroundings we will search for relevant words and also for other citations. If we will find a keyword or other citations in its vicinity, then we will consider the analysed citation more relevant. An important factor will be the distance of the found keywords and other citations from the original [3].

Other aspects which can be analysed are the authors of citations, the article in which is the analysed article cited and the popularity of the citation (how many times was the given text cited).

In order to identify the most relevant keywords, we can use the spreading activation method on a citation network. In digital libraries most of the article references are directly connected to the referenced document, so we can recursively find and analyse all referenced articles and apply our method for keyword extraction from citations.

We plan to evaluate our method of keyword extraction in a web-based bookmarking system Annota[1] on the dataset of articles from ACM Digital Library[2].

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

References

[1] Qazvinian, V., Radev, D.R.: Identifying non-explicit citing sentences for citation-based summarization. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics - ACL '10*, (2010), pp. 555–564.

[2] Liu, S., Chen, C.: The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, (2013) pp.627–639.

[3] Elkiss, A. et al.: Blind men and elephants : What do citation summaries tell us about a research article. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, (2008), pp. 51–62.

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

[2] http://dl.acm.org

Using Tags for Query by Multiple Examples

Tomáš VESTENICKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
vestenicky@icloud.com

Nowadays, the most widely used approach for searching information on the web is keyword-based. The main disadvantage of this approach is that users are not always efficient in keyword choice. This is why we aim to help them with query formulation or offer them different, more natural approach.

Query by example paradigm is often utilized in image and music search [1], because we can easily determine the topic and assess similarity. In the digital libraries domain, documents can have multiple topics; one may be more important or more relevant than the other. This is why we decided to represent the document by tags and use these to aid searching process.

Our method is based on building the query by choosing positive examples of documents or typing keyword-based query as a starting point and then specifying information radius of results by selecting positive or negative examples utilizing explicit relevance feedback as a query refinement method. We use user-added tags and author keywords to determine document similarity. Users add tags to documents which aids searching process, because users choose what is relevant for them from the particular topic. Therefore, tags can be used for more fine-grained query refinement by enabling users to see and/or remove tags from documents selected as positive or negative examples. In cases, where documents have not yet been tagged, we use keywords provided by the document's author.

While searching, users can label documents (displayed as search results) as positive or negative, based on their preference. Labelling the document means labelling all its associated tags or keywords, which are then added to the corresponding set. Tags and keywords can be separately removed from these sets for further query refinement. Labelling as positive increases the document (or tag) relevance and vice versa. After labelling the first document we disregard the textual query and display results using only our method.

Keywords, which are used when user-added tags are not available, are extracted by AlchemyAPI[1] service. We display top five keywords according to the relevance

[*] Supervisor: Róbert Móro, Institute of Informatics and Software Engineering

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

score provided by AlchemyAPI. Tag relevance score for the purposes of our method is computed using TF-IDF scheme [2]. Each tag has its score based on relevance to the corresponding document:

$$TFIDF(t,d,D) = \frac{n(t,d)}{\sum_{w\epsilon d} n(w,d)} \times \frac{\log |D|}{|d \,:\, t \in d|}$$

where *t* represents tag, *d* document, *w* word and *D* stands for the set of all documents. Function *n(x, d)* counts every occurrence of *x* in document *d*.

Labelled tags will be shown in the form of tag cloud (where the word size represents its relevance score) in the left sidebar alongside the results.

In extreme situations, our method follows these rules:

1. In cases, when user labels the same tag (or keyword, hereinafter referred to as “tag”) differently during one session, this tag will be kept in both sets. Each tag will keep its score while reducing the other one’s effect (removing tags is left to the user)
2. If user labels the same tag as positive/negative more than once, it will be represented by the highest value

Search results are ordered by their score, which is recalculated after every change user makes in positive or negative set of tags. The score for each document is calculated by following formula:

$$relevance = \sum_{t\epsilon DPT} s(t)d(t) \quad - \sum_{t\epsilon DNT} s(t)d(t)$$

where *DPT* is document’s positive tags (document’s tags present in positive set), *s(t)* stands for tag’s search score and *d(t)* is tag’s document score (result from TF-IDF).

We focus on the domain of digital libraries of research articles and we evaluate our proposed method in bookmarking service Annota[2] by means of a user study.

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

References

[1] Rasiwasia, N., Vasconcelos, N.: Image retrieval using query by contextual example. In: *MIR ’08: Proc. of the 1st ACM Int. Conf. on Multimedia Information Retrieval*, ACM Press, (2008), pp. 164-171.

[2] Falessi, D., Cantone, G., Canfora, G.: A comprehensive characterization of NLP techniques for identifying equivalent requirements. In: *ESEM ’10: Proc. of the 2010 ACM-IEEE Int. Symposium on Empirical Software Engineering and Measurement*, ACM Press, (2010).

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

Web Search Employing Activity Context

Ľubomír VNENK *

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
lubomir.vnenk@zoho.com

At the age of information overload it is not easy to find exactly what we are searching for. Even if modern search engines try their best to choose the most valuable search results, they do not have enough information to show only desirable results. It is mostly because of simplicity of a search query. An average query consists of 2-3 keywords [1] that may exactly describe user's problem, but it may be not enough. The query may have many others meanings and search engine is not able to choose, which one you intended. To find out right meaning we need to know context behind the query, i.e. what led user to search. In our work we aim to find out what context is behind a query.

Many projects were trying to get user's context by evaluating his click-through. They find out what user is interested in and what is not interested in thanks to following links that user clicked and didn't clicked, for example in work of Leung and Lee [2]. However, there is no way to find out the right meaning at the beginning, when there is no click-through.

We hypothesise that searcher's search need comes in many cases from an application used recently. It means, the context of a query can be found inside an application and can be used to find out meaning of the query even if there is no click-through. To find out context of an application we developed an activity logger. It tries to catch each application actual context. Context of an application is represented as weighted keywords captured by the activity logger. To get larger application context, activity logger consist of three independent parts each aimed on specific type of activity:

- Tabber
- Annota
- Wordik

Tabber is desktop activity logger that captures user's activity and interaction between applications. It catches application switch event and is able to tell when and which application was active and how long. It also catches current window name and an

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

application type which may tell us something about current work with the application and what type of work is usually made with the application.

Second activity logger, Annota, is Mozzila Firefox extension that captures user's activity inside web browser. It captures URL address, title and time when a web-page is active, so we can get user's click-through. It also captures text that user selects or copy-pastes, because that text is important and can tell us something more about a website relations, because if user copies text at one website and pastes it to another, they must be somehow connected.

Last, but not least activity logger is Wordik. It is a Microsoft Word addin that captures content of a document user is actually writing. It captures names of topics the user writes and the text that is around currently edited section.

Logged information have various significance and must be processed to reflect this fact and to be prepared for future processing. Information is split into keywords and we assign weight to each keyword. Weight reflects how much is the keyword significant in regard to application itself.

Very important task is determining, which application has fired the search need, i.e. which application is related to the query. For this purpose we measure syntactic and semantic distance between query and each application's context that user used recently. An application with the lowest mix of distances is considered as the application that is connected to query. Connection between an application and query can be also revealed by analysing user's interaction between query and an application. If copy/paste event occurs between an application and one of the query results, we can assume that the application is connected to the query.

We verify success-rate of our methods that try to find connection between an application and query by comparing to explicit feedback methods. We modified Google search result page and allowed user to click on an application that is connected to the query from the list of applications that he used recently. Finally, the copy/paste interaction between search results and the applications resulted to have almost as high precision as explicit feedback.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava.

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

References

[1] Kamvar, M., Kellar, M., Patel, R., Xu, Y.: Computers and iphones and mobile phones, oh my!, In: *Proc. 18th Int. Conf. World wide we,* (2009), pp. 801

[2] Leung, K., Lee, D.: Deriving concept-based user profiles from search engine logs, In: *IEEE transactions of knowledge and data engineering journal,* (2010) , pp. 969-982

User Experience (PeWe.UX)

Methodics of Game Evaluation Based on Implicit Feedback

Peter DEMČÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
piers.demcak@gmail.com

Game software is a relatively new medium, but is has already proven its significance by occupying a major share in the profits of the entertainment industry. But aside from being a delight to play, games such as GWAPs and serious games, can be also used to substitute the conscious effort and attention of users with an activity that the users will do on their own. These games show in effect, one of the core characteristics of all games: they essentially sugar-coat a problem solving activity into an authentic game experience.

Game experience is the subjective manner, in which the player perceives and playfully interacts with a game. It is the goal of every game - whatever its purpose may be, to create a game experience, and to capture (and keep) the interest of its players.

Game experience makes use of basic cognitive skills such as modeling, focus, imagination and empathy. One of the aspects of game experience is immersion, which is the degree, to which the player is invested in playing of the game. As such, if a game aims to reach its goals, it is essential to achieve as high a level of immersion as possible. There are three increasing levels of immersion - engagement, engrossment, and total immersion[1]. The first level, engagement, highly depends on the players own game preferences, and also on the players ability to successfully learning the controls of the game.

It is apparent, that the concept or learnability from the area of evaluation of user interfaces applies to games, just like it does with any software. Games do communicate with players through a game interface. However, there are also some aspects to learning to use an application, which are characteristic to games. In games, players have to get to grips not only with the elements of the user interface itself, but also gradually learn the inner workings of the game mechanics behind it, without breaking the growing level of immersion. This is why a study of the learnability of a game interface in connection to the game immersion appears to be of importance.

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 51–52.

For the purpose of creation of the intended gaming experience, the ability to evaluate the interface learnability and the gaming immersion is highly valuable. However, because of the subjective character of user experience, both require feedback from the players to fully grasp.

The limitations of explicit feedback come from the difficulty of executing a detailed observation of the player's mental state without disturbing it. Hence, the importance of implicit feedback, which is based on the recognition of the user's mental state in regards to their natural behavior. Some of the information about the direct usage of the game application can be collected directly by monitoring the input devices, such as frequency of mouse clicks or the movements of the cursor. One of the other means of gathering interesting implicit feedback is through eye tracking.

Mapping of the eye movements to cognitive functions shows promise[2], even for the evaluation of games. Some of the possible approaches using eye tracking, are the identification of the fitting and disturbing elements of gameplay, the game passages which are the most and the least immersive, or recognition of the states of presence and gameflow with the player.

Our goal is to use the methodological approach in our research, to design a set of reusable principles which can be used for game evaluation based on eye tracking information. Then, we plan to apply these principles to several games with different user groups and different kinds of game play, to verify the results of our method.

Acknowledgement. This contribution was partially supported by the Research and Development Operational Programme for the project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the European Regional Development Fund.

References

[1] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *Int. J. Hum. Comput. Stud.*, vol. 66, no. 9, pp. 641–661, Sep. 2008.

[2] M. Bartels, "Eye Tracking Insights into Cognitive Modeling," *Proc. 2006 Symp. Eye Track. Res. Appl.*, vol. 1, no. March, pp. 27–29, 2006.

Analysis of User Behaviour on the Web

Patrik HLAVÁČ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
patrik@hlavac.sk

Development of web systems has changed considerably in recent years and it also changes the importance of the estimation of user attributes and further recommendations. Analyzing user behavior on the Web and interaction with a web browser via computer is a nontrivial issue, where new solutions have recently opened new research and application directions with better availability of technology and equipment. While existing solutions are based on monitoring the behavior directly in the browser by using only basic peripheral devices, now we have the possibility of directly monitoring user's gaze and focus blocks of content on the website. Devices referred to as eye trackers provide irreplaceable information about a user's gaze.

One objective of this work is to propose a user model suitable for collecting data from interactions in a Web environment. We consider the user model as a summary of information about a given user, which we can obtain by collecting implicit or explicit feedback directly from a system involved. With more frequent interaction, we gain more knowledge of his behavior and we can characterize him better.

The primary task will be to gather information through sensor device to identify the fields of view of the user's interest in the content on the screen along with the use of other devices (mouse, keyboard, microphone), which allow the acquisition of implicit feedback. These data from interaction will be processed in the user model.

Collecting accurate information about the content and the text with which the user came into contact will help us to determine the amount of information received and also help estimate the user's knowledge from the user model. These can be verified by explicit feedback, a test or questions.

Acknowledgement. This contribution was partially supported by the Research and Development Operational Programme for the project “University Science Park of STU Bratislava”, ITMS 26240220084, co-funded by the European Regional Development Fund.

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

References

[1] Labaj, M.: User Feedback for Personalized Recommendation. Dissertation Project Report. pp. 10–16 (2012)

[2] Joung, Y., El Zarki, M., & Jain, R. A user model for personalization services. In *Proc. of Fourth International Conference on Digital Information Management*, 2009. ICDIM 2009. IEEE, pp. 1–6 (2009)

Web Applications Usability Testing by means of Eyetracking

Martin JANÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xjanik@stuba.sk

In present a big part of software applications has transferred to web. When we speak about how natural is the use of a specific application (by a user) in a way of its purpose we refer to its' usability. Usability of an application has a great impact on application's success. Because of that, the question of applications usability is a subject of active research which is focused on how to effectively monitor user while interacting with application, analyze user's activity and identify certain patterns of behavior that could cause a problematic situation. Purpose of this research is not just to identify problematic situations, but also to categorize them and recommend actions to solve the problems [1].

In the process of evaluation of usability, user's feedback has a great importance. We distinguish between explicit feedback which involves ratings or votings and implicit feedback, including time spent on web page, mouse clicks, key typing or gaze tracking. While explicit feedback data are more accurate but are harder to obtain, implicit feedback data are easily acquired but because of difficulties to interpret them, they are often less accurate [2].

Gaze tracking studies discovered that the eye movement of the user, while reading, creates a characteristic pattern made from fixations and saccades. During a fixation, eyes are intently staring at one point. Saccades are fast movements between two fixations. Gaze tracking analysis is based on an important presumption that there is a relation between fixations, our gaze and our actual thoughts. This kind of analysis is difficult because of the need of special, eye tracking device. On the other hand, it has a great potential to achieve new results in evaluation of applications usability through evaluation of specific application or by discovering new relations between various signals of implicit feedback, where gaze tracking can serve as feedback type and validity confirmation [3].

In our study, we would like to focus on implicit feedback which is acquired through quantitative testing. While qualitative testing can point out the problems

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

relating to design of the application, we cannot find those problems accurately through quantitative testing. We would like to research this problem through usability testing including gaze tracking.

Gaze tracking brings new aspect for evaluating usability. It offers information of which objects attract attention and why. By following the object gaze order we can tell how users search through web applications, creating specific gaze pattern. Based on gaze patterns we can identify users' patterns of behavior. Specific unwanted patterns of behavior, which the owners of web applications would like to eliminate, are also present. Aimless movement across the web application, long time of users' inactivity, user repeatedly visiting the same web element, are some of the undesirable patterns of behavior.

We aim to create a method which will be able to identify unwanted patterns of behavior in domain of content management systems. Identification will be based on implicit feedback, particularly on feedback from gaze tracking. Our goal, next to identification of unwanted patterns of behavior, is to recommend a set of steps to resolve a problem, which is likely the cause of specific pattern of behavior. By improving our method we would like to be able to predict undesirable patterns of behavior to avoid their creation.

Acknowledgement. This contribution was partially supported by the Research and Development Operational Programme for the project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the European Regional Development Fund.

References

[1] Hema Banati, Punam Bedi and P. S. Grover, Evaluating Web Usability from the User's Perspective. *Journal of Computer Science*, Volume 2, Issue 4, 2006, pp. 314-317.

[2] Nathan N. Liu, Evan W. Xiang, Min Zhao, Qiang Yang, Unifying explicit and implicit feedback for collaborative filtering. *Proceedings of the 19th ACM international conference on Information and knowledge management*; 2010, pp. 1445-1448; ISBN: 978-1-4503-0099-5.

[3] Tobii Technology AB; Tobii Eye Tracking - An introduction to eye tracking and Tobii Eye Tracker. Tobii Technology AB, 2010, 12 pages.

Low-Cost Acquisition of 3D Interior Models for Online Browsing

Filip MIKLE, Matej MINÁRIK, Juraj SLAVÍČEK, Martin TAMAJKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
imaginecup2014@googlegroups.com

Building a dense 3D model in a usual 3D editing tool requires a lot of time and a significant skill in this field. Using our solution, everyone, even non-technically skilled person is able to model real places in a few minutes. Modeling interiors this way looks much more like filming than 3D modeling. It takes only a few more minutes to get a 3D model than to take a set of pictures of interior.

Naturally, there were always been efforts to make this process automatic. Laser scanners provide professional, accurate, but expensive way of modeling of interiors, which generally requires a lot of time and preparation. In comparison to laser scanners, our solution requires affordable and commonly used device - Kinect for Windows. It's also much faster to create 3D model using our product.

The first part is a scanning application, designed for real estate agents. The second part is a browsing web application, designed for ordinary people. Customers of real estate agency can easily walk through different interiors previously scanned. When the scanning is finished, scanning application processes the data and generates group of object files. Subsequently, agent uploads these files through our web application and the model is ready to be browsed.

We decided not to develop from absolute scratch. Scanning and pointcloud-to-mesh transformation and optimization parts are mainly based on features provided by open-source library PointCloudLibrary (PCL) and its dependent libraries. The Web-application is a part of our product and it’s based on ASP.NET.

We found out, that we need to remove redundant information from our 3D models. This redundancy, arising in scanning-phase, causes enormous increase of model's size (every point in space could be contained in model multiple times). That's why we developed *Voxel Tree*. It allows dynamic addition of environment features to model, guaranteeing, that every information will be in model contained only once. It also allows to choose density of model (the size of atomic part of model - voxel, in meters).

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

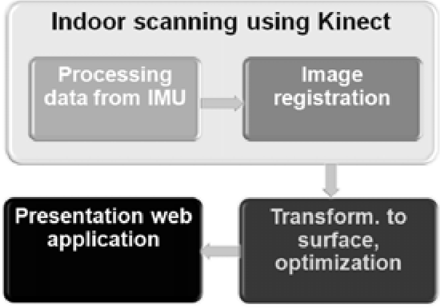

Indoor scanning using Kinect
Processing data from IMU
Image registration
Transform. to surface, optimization
Presentation web application

Figure 1. Architecture overview.

The alignment of neighbor frames is reached using the computer vision techniques. One of algorithms used for this purpose, is Iterative Closest Point (ICP) [1]. This algorithm would be costly and not affordable in real time, if it would work with all the points got from Kinect. This is, where keypoints come to use. We extract keypoints based on depth and also on color information from environment, ignoring ones that do not have known distance from camera (unknown depth). Everywhere, where appropriate, we use parallelism, recording significant increase on performance. Although the contribution of parallelism on CPU is large, our future plans lead to make as many as possible (and appropriate) computations on GPU (currently, we process to 3 frames per second).

One of the inputs to Iterative Closest Point algorithm is initial estimation of device´s orientation. This additional (and very important) information is taken from IMU (inertial measurement unit), electronic device that measures and reports on a craft's velocity, orientation, and gravitational forces, using a combination of accelerometers and gyroscope and magnetometer. Obtained measurements are converted to Euler angles, which represent current orientation of the Kinect.

Point clouds obtained in previous parts of scanning process are not the best for further manipulation, so we represent our models as a surface. We use a method that creates bounding box which is a 3D object containing all points of original cloud. Bounding box is then divided into uniform voxel grid which is essentially a little cube. Each point in the original point cloud is then assigned to the corresponding voxel according to its 3D coordinate. Then we make each wall of each voxel solid or transparent taking into account if wall can be seen from concrete position of Kinect.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava.

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

References

[1] Besl, Paul J. and McKay, Neil D. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* February 1992, Vol. 14, 2.

Implicit Feedback-based Discovery of Student Interests and Educational Object Properties

Veronika ŠTRBÁKOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
veronikastrbakova@gmail.com

In the present, searching and recommendation on the web are becoming more and more common. Whether it concerns search and recommendation of articles in news and digital libraries, of study materials, or different products in e-shops, it is essential to know the characteristics of the objects being recommended, and the characteristics of the person manipulating these web objects. These characteristics are collected via implicit feedback. Inaccurate information, collected from evaluation of implicit feedback from human behavior, has significant influence on the accuracy of recommendation.

With the increasing possibilities of monitoring users on the web, like signals from eye tracking camera, blood pressure, body temperature and pulse sensors, we gain the ability to evaluate implicit feedback with great accuracy, and with that, gain the related interpretation of various signals of activity in different domains.

Despite the existing implicit methods of evaluation for various signals of user activity, which aim to explore its characteristics, there is still room for improvement. Our research is aimed at the attributes of users and educational objects with the use of implicit feedback indicators, and their interpretation for use in the domain of education. By the research of chosen implicit feedback indicators, individually and with each other, we will explore their mutual relations. We plan to monitor the user activity, while the users are studying. To support our approach, we identified the following basic implicit feedback indicators, which can provide us the overall picture of the users behavior:

- Number of scrollings on the page, the overall time of scrolling on the page
- The number of moving of the page, the distance that the site was moved
- Different mouse movements
- The number of keystrokes
- The way of leaving the page

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

- Time spent on the page, time to first click
- Gaze of the user
- Number of highlighs / copies of individual pieces of text, size of highlighted / copied sections of text
- Annotation the text on the page
- The number of searches on the site

In this work, we plan an attempt to interpret the appropriate individual recorded signals of the user activity. Based on our findings we aim to improve the user model and to propose a method for the use of collected information in the domain of education on the web. Further, we plan to experimentally prove our findings in the context of educational objects and in the context of user modeling.

Acknowledgement. This contribution was partially supported by the Research and Development Operational Programme for the project “University Science Park of STU Bratislava”, ITMS 26240220084, co-funded by the European Regional Development Fund.

References

[1] Badi, R. et al.: Recognizing user interest and document value from reading and organizing activities in document triage. In: *Proceedings of the 11th international conference on Intelligent User Interfaces - IUI ’06*, ACM, 2006, p. 218-225.

[2] Barla M.: Towards social-based user modeling and personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1):52-60, 2011

[3] Fox, S., Karnawat, K., Mydland, M., Dumais, S. & White, T.: Evaluating implicit measures to improve web search. In: *ACM Transactions on Information Systems (TOIS)*, ACM, 2005, p. 147-168.

[4] Ricci, F., Rokach, L., Shapira, B., Kantor, P.: Recommender Systems Handbook, 1st edition. Springer-Verlag New York, Inc., October 2010, p. 10-23.

Software Development Webification (PeWe.PerConIK)

Keeping Information Tags Valid and Consistent

Karol BALKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
balko.karol@gmail.com

Metadata have become inseparable part of modern system development. Metadata as structured information describe the information resources. In our research, we are focused on the metadata within the project PerConIK, metadata used to describe features of the source code, such as the number of white characters or denomination of the copied code.

Within PerConIK project, metadata are being denominated by an expression *information tags*. Information tags are stored in a repository and are hold a reference to the source code they are describing. However, these information tags can be invalidated during modifications and refactoring of the source code either because their references to source code become outdated or simply because the feature they are expressing is no longer valid.

In our project we aim to create a method to control the consistency and validity of theses information tags. We use several ways in a combination in order to provide a solution.

Our method works with an abstraction of a source code, an abstract syntax tree, which makes the method language independent. One approach, which we are exploring nowadays is to examine similarities of different abstract syntax trees. To find this similarity we use a method called *robust tree edit distance*, which shows good performance in temporal and spatial complexity. On the basis of the resulting distances we compute clusters of these abstract syntax trees, which allows for a faster comparison of any new source code that come into the system.

If the evaluation proved our method, we would be able to transfer information tags to source code which is new in system, based on the abstract syntax trees similarity. Although we can't claim certainty that these information tags are absolutely valid, but we can determine their similarity to the nearest cluster of abstract syntax trees.

We were considering k-means as our clustering algorithm first, but its problem is in addition of new abstract syntax trees, which requires re-clustering of all known abstract syntax trees again. So in the next stage of our research, we want to use Birch

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

algorithm, which has a lot of advantages against k-means clustering algorithm, especially in incremental clustering of added abstract syntax trees.

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] R. Koschke, R. Falke, and P. Frenzel, "Clone Detection Using Abstract Syntax Suffix Trees," in *Reverse Engineering,* 2006. WCRE'06. 13th Working Conference on, 2006, pp. 253 – 262.

[2] M. Pawlik and N. Augsten, "RTED: A Robust Algorithm for the Tree Edit Distance," in *the 38th International Conference on Very Large Data Bases,* 2011, pp. 334–345.

[3] W. Yang, "Identifying Syntactic Differences Between Two Programs," vol. 21, no. JULY, pp. 739–755, 1991.

Source Code Review Recommendation

Matej CHLEBANA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xchlebana@stuba.sk

Software quality assessment is a complex, but substantial activity in terms of the project success. Early detection of errors can alert project managers on issues that may lead to prolonging or complete abolition of the project. Except for detecting errors it also helps maintaining consistency of source code and reducing the risks associated with an unexpected quit of a team member. Revision of developers' source code is time-consuming especially in larger companies where the new versions of the source code are generated in small intervals of time.

Software project management has a complex and extensively defined position. Project managers monitor and control the work of designers, developers and testers, but sometimes they also actively participate in these activities. While developers focus on the source code itself, architecture and performance, managers focus on a higher level, which includes: direction of the project, allocation of resources, functionality and user experience [1]. As the developers and managers focus on different activities in the project so they are interested in different data coming out of a project.

Quality assessment takes into account various metrics. One group of such metrics is measuring the activity of developers to analyze the mutual collaboration of developers in the team [2]. However, evaluation of data obtained from developer's activity monitoring is a challenging process. Every developer is different, has different experiences, skills, strengths and weaknesses. Some developers create better code during the day, another during the night when they are not influenced by the noise of the surrounding area. Circumstances related directly to the developer have an impact on the quality of the source code, such as illness or other life situations.

Identification of these special circumstances and characteristics of developers is difficult. Some environmental influences on the work of the programmer simply cannot be detected (e.g., problems they are currently dealing with in their private lives). Although we are not able to fully identify these special circumstances, we can work with the information we collect, for example, through a system being developed within the research project PerConIK (Personalized Conveying of Information and

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

Knowledge). PerConIK aims to support business applications in software house using empirical software metrics. This system collects empirical data through software tools and extensions for development environments (Microsoft Visual Studio 2012 and Eclipse) and web browser (Mozilla Firefox) that are installed on a workstation of a developer [3]. Next, it is necessary to analyze it and find metrics and features that are related to source code on top of which this activity was acquired. Our goal is to use process mining to create a model in which we can identify and select potential risk within source code. Source code can be then recommended for a detailed review if necessary.

Motivation for examining this area is mainly to increase the success rate of software products by trying to warn the code reviewer for potential problems within source code (whether due to a higher likelihood of errors or breaking of the coding conventions). The aim will be to separate "good" from potentially hazardous source code therefore to eliminate the necessity of reviewing all codes.

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] Buse, R. P. L., & Zimmermann, T. (2012). Information needs for software development analytics. In *2012 34th International Conference on Software Engineering (ICSE)* (pp. 987–996). Ieee. doi:10.1109/ICSE.2012.6227122

[2] Corral, L., Sillitti, A., Succi, G., Strumpflohner, J., & Vlasenko, J. (2012). DroidSense: A Mobile Tool to Analyze Software Development Processes by Measuring Team Proximity. In C. A. Furia & S. Nanz (Eds.), *Objects, Models, Components, Patterns* (Vol. 7304, pp. 17–33). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-30561-0_3

[3] Rástočný, K., & Bieliková, M. (2011). Information Tags as Information Space for Empirical and Code-based Software Metrics. In Proceedings of 17th International Conference on Fundamental Approaches to Software Engineering. (submitted)

Identifying Hidden Source Code Dependencies from Developer's Activity

Martin KONÔPKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xkonopkam@stuba.sk

Evaluation of software projects is important part of the process when reaching desired goals in planned time schedule is of high interest. Delivering quality software products in time requires to monitor them, evaluate their attributes in order to manage and support the development process. Traditionally, we evaluate software from the source code perspective as a result of its development process. With numerous metrics we are able to quantify the source code, evaluate its quality, complexity or maintainability [2].

Software source code is the result of development process, i.e., developer's activity. A developer performs various activities during the development which we can monitor [1]. Main focus is on developer's activities directly associated with development, i.e., programming, modelling or checking-in work results. Secondly, the work results, source code contents and time-based information about developer's work depends on activities associated with development indirectly, i.e., searching for solutions, studying existing source codebase and documentation. In the end, these activities are interweaved with activities which are not associated with development but still has high impact, i.e., emotional state of developer, environmental context, etc.

Traditional source code metrics do not rely on the mentioned types of developer's activities. However, we can search for connections of activities with resulting attributes of software to identify hidden patterns and relationships in the development process as well as to gather more information about it. One such application is to compare the metrics and techniques based solely on syntactic analysis of source code with approaches based on evaluation of monitored activities.

In our work we focus on identification of hidden source code dependencies from developer's activity to enhance existing source code dependency graph. Source code dependencies represents connections of software components, e.g., call references, inheritance or containment. A developer uses dependency graph to find out impact of a change in one software component on others. Based on the source for identification of dependencies, we distinguish between:

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

- Explicit dependencies – identified with syntactic analysis of source code,
- Implicit dependencies – identified from developer's activity.

Implicit dependencies describe which components are related to each other, even when no explicit connection exists between them at all. Implicit dependencies are identified from developer's activity of interactions with the source code during the development, e.g., when developer was studying existing solution in the source code, copied code fragment from one component into another. Interactions with source code also describe which components relate to each other in context of a particular task which developer was working on. We identify implicit dependencies from these activities with source code files:

- Time-based activities – open, close and switch to source code file.
- Copy-paste code fragment between two source code files.

Implicit dependencies overlap with explicit ones but also introduce new connections to the existing dependency graph. A developer searching for any dependent components may then find the components which are loosely coupled in the source code, but highly coupled during the runtime or development.

Our work is part of research project PerConIK – Personalized Conveying of Information and Knowledge (http://perconik.fiit.stuba.sk) [1]. We use services and logs provided by this project for evaluation of our method. We also see its possible application in code review and identification of problematic places when a code reviewer searches for places a developer worked on. If a reviewer identifies problem in one place a developer worked on, there may be problems in other places she touched as well. Surely, it is possible to extract such information from the list of checked-in files, however it is often important to prioritise this list as well.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 474-479.

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., Návrat, P., Chudá, D., et al.: Webification of Software Development: General Outline and the Case of Enterprise Application Development. In: *AWERProcedia Information Technology & Computer Science: 3rd World Conf. on Information Technology*, Barcelona, (2012), pp. 1157-1162.

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

Context-based Improvement of Search Results in Programming Domain

Jakub KŘÍŽ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
jacob.kriz@gmail.com

During programming, a programmer faces a great deal of many different problems related to his work. He often uses web search engines in order to try and solve them.

When searching the web, users usually enter only a few words as a search query, which may often lead to unsatisfactory results [1]. Programmers are expected to be usually better at putting together a search query, however, because they search much more often than regular users, it is likely that they get inattentive and get unsatisfactory search results as well.

Programmers do many things during a typical programming session apart from just writing the source code. They do all these things with a certain intention, in a certain context. By understanding this context we can make the results of the programmers' web searches more accurate and relevant. In this work we propose a method for extracting context metadata in the programming domain, from the sources specific to programming, in order to build a programmer's context model.

Based on our experience with the programming domain we group the problems programmers usually face into three categories:

- *Conceptual problem* – the programmer is trying to understand the idea of an algorithm or to come up with an algorithmic solution to a problem
- *Technical problem* – the programmer is solving a problem associated with using the methods of the language or library or API methods which he is currently using
- *Error* – the programmer is solving an error which occurred

The context model should represent the programmer's intentions in any given point in time. We base the structure of this model on the types of problems we described:

- Conceptual part
- Technical part
    - Language

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

  - Framework / libraries
  - Key identifiers
- Error part
- Programmer’s state

The *conceptual part* says about the conceptual intentions of the programmer. It is composed of a set of ranked keywords. These keywords are extracted from the active part of the source code, the part with which the programmer currently works, using a custom tf-idf algorithm, which is modified to extract conceptual keywords and only from the active part of the code. Similar extraction method for conceptual keywords has been successfully used in previous works [2].

The *technical part* includes information about the technologies the programmer is currently working with. The *language* part contains the identifier of the programming language; the *framework/libraries* part contains the identifiers of currently used frameworks and libraries. The *key identifiers* part is composed of a set of ranked key identifiers of core, library or API methods and classes. They are extracted using a similar, modified tf-idf algorithm as the conceptual keywords. The *error part* contains the identifier of the last error which occurred.

The *programmer's state* part says in which state the programmer is currently in. The state is detected by analysing the programmer’s activity in the IDE. The method considers several factors, like the time of his last typing or time since last compilation and uses supervised learning algorithm to classify his state.

We further use the context model to improve the results of programmer’s web searches. We do this by reranking the search results. Based on the detected programmer’s state the method picks the relevant part of the context model. When ranking the search results, the method boosts the rating for documents, which contain words or identifiers which were also found in the context model. We believe the context model, as proposed is very versatile and can possibly be used in many ways.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 492-497.

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] Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.,* 2000, vol. 36, pp. 207–227

[2] Ohba, M., Gondow, K.: Toward mining ”concept keywords” from identifiers in large software projects. In: *Proceedings of the 2005 international workshop on Mining software repositories.* MSR ’05, New York, NY, USA, ACM, 2005, pp. 1–5.

Modelling Developer's Expertise

Eduard KURIC*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
eduard.kuric@stuba.sk

Estimation of developer's expertise allows, e.g., developers to locate an expert in a particular library or a part of a software system (someone who knows a component or an application interface). In a software company estimation of developers' expertise allows, for example, managers and team leaders to look for specialists with desired abilities, form working teams or compare candidates for certain positions. It can be also used to support so-called "search-driven development". Relevance of software artifacts is of course paramount, however, trustability is just as important. When a developer reuses a software artifact from an external source he has to trust the work of an external developer who is unknown to him. If a target developer would easily see that a developer with a good level of expertise has participated in writing the software artifact, then the target developer will be more likely to think about reusing.

Our new idea is to automatically establish developer's karma based on monitoring his working activities during coding in an integrated development environment (IDE), analyzing and evaluating the (resultant) code he creates and commits to local repository. Primarily, we focus on answering these questions:

- What is the overall developer's karma and contribution for a software project, i.e., a degree (level) of his experience and familiarization with a functionality of a software system?
- Who has appropriate expertise for a particular part (component) of a software project?

In this area, several approaches have been proposed. The Expertise Browser [3] and Expertise Recommender [1] use a simple heuristic (Line 10 Rule) that a developer who commits to a file has expertise in the file. The Expertise Recommender establishes developer's expertise as a binary function, i.e., at a certain time only one developer can have an expertise in a file and it depends on who last changed it (last one wins). An argument is that a developer who was last to make a change has the most fresh mind about the code. It has limitations because it does not take into account a degree of real developer's source code contributions, i.e., it does not reflect his overall know-how

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

about the code as a whole. The Emergent Expertise Locator [2] and Expertise Recommender improve the Expertise Browser so that they consider a relationship among changes in a file when determining expertise. However, both approaches represent developer's expertise too straightforward, i.e., a developer who completely changed the implementation of a code fragment has no influence on the expertise of a developer who created the code.

In contrast with existing approaches trying to establish developer's expertise, we analyze his developing activities and the (resultant) code more deeply. To establish the overall developer's karma for a software project, we investigate software artifacts (components) which the developer creates. In other words, we take into account developer's “karma elements” as:

- *degree of authorship* – the developer's code contributions and the way how the contributions were contributed to the component;
- *authorship duration and stability* – the developer's know-how persistency about a component;
- *technological know-how* – the level of developer’s knowledge of used technologies (libraries), i.e., how broadly/effectively;
- *component importance*.

We have developed a tool for monitoring developers while creating code. It allows us to capture, track and evaluate different events, e.g., we log copy/paste, find and navigation actions in a web browser, Visual Studio, OneNote; biometric information; utilization of hardware resources; and states of running apps. We have inferred the particular developer's karma elements based on our observation of developers' 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] McDonald, D.W., Ackerman, M.S.: Expertise recommender: a flexible recommendation system and architecture. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (CSCW '00). ACM, New York, USA, 2000, pp. 231-240.

[2] Minto, S., Murphy G.C.: Recommending Emergent Teams. In *Proceedings of the Fourth International Workshop on Mining Software Repositories* (MSR '07). IEEE Computer Society, Washington, USA, 2007, pp. 5-.

[3] Mockus, A., Herbsleb, J., D.: Expertise browser: a quantitative approach to identifying expertise. In *Proceedings of the 24th International Conference on Software Engineering* (ICSE '02). ACM, New York, USA, 2002, pp. 503-512.

Employing Information Tags in Software Development

Karol RÁSTOČNÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
karol.rastocny@stuba.sk

1 Introduction

A management of software development process is a crucial part of the software engineering, from which the success of software projects is dependent. This management mostly relays upon quality and freshness of software metrics and analysis over these metrics. Software metrics can be based on source code or empirical data about software developers. Code-based metrics are well known and many approaches based on them have been proposed. But empirical software metrics are still uncovered part of software engineering even though they contain important information about software development process and they can be used e.g. for forecasting significant trends, similarly as empirical data (e.g., implicit user feedback) in the web engineering. Reasons lies in fact that empirical data collection is time expensive and erroneous [2].

2 Management and maintenance of information tags

We proposed a solution of these problems based on collecting, storing and maintaining developer-oriented empirical data abstracted to information tags and also empirical software metrics. However, empirical data stored in information tags can be invalidated by developers' activities, so they have to be managed and maintained. For this reason we proposed an approach for information tag management and maintenance based on event stream processing, in which we evaluate stream (C-SPARQL) queries and execute tagging actions, when queries are evaluated successfully.

We performed preliminary performance evaluation of the proposed approach, during which we have evaluated non-trivial C-SPARQL query with 2s window over stream of RDF triples and we have measured response time. In our testing environment (with 2x2.8GHz CPU), we reached satisfactory response (1s with 100ms tolerance), when we generated up to 300 triples per second (see Figure 1). It might look as an

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

unsatisfactory result, but a bottleneck was caused by overhead of our testing sandbox, that consumed more than 75% of resources. In addition, we receive 2 RDF triples per second from 10 developers.

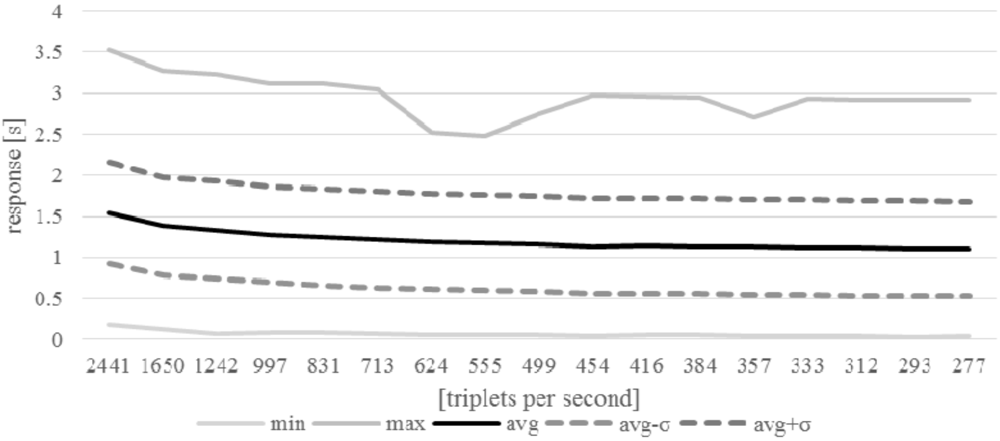


Figure 1. Response times of the Tagger by a number of RDF triplets posted to the Tagger, with C-SPARQL query set up to 2s sliding window.

3 Future work

Nowadays we are working on an approach for automatic learning of tagging rules based on analysis of developers' activity streams and changes in information tag space. We have two working versions that we plan to evaluate in parallel. The first one is based on clustering of similar developers' activities that similarly affected information tags. The second approach is based on process discovery techniques [1]. This solution does not analyse modifications of information tags but it gets new versions as combination of addition of new information tags and removal of old information tags.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 513-520.

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] Van der Aalst, W.: Process Mining: Overview and Opportunities. *ACM Trans. Manag. Inf. Syst.*, vol. 3, no. 2, (2012), pp. 7:1–7:17.

[2] Johnson, P.: You can't even ask them to push a button: Toward ubiquitous, developer-centric, empirical software engineering. In: *The NSF workshop for new visions for SW design and productivity*, Nashville, (2001), p. 5.

Assessing the Code Quality and Developer's Knowledge

Jana PODLUCKÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
j.podlucka@gmail.com

Quality or its absence is one of the main aspects for the evaluation of the source code. However, it is not only influenced by programmers' knowledge of the programming language but also by the actions that he faces during his work.

Mark et al. [1] discuss the influence of outer actions on the programmers' work. They have defined the value of diversion as the time needed by a person to return to the context of the work interrupted. In addition, several other aspects can be connected with these diversions, such as stress. The value of interruption is not only based on the type of a diversion, but it can be raised or lowered by the personal factors of the programmer himself.

Khan et al. [2], on the other hand, studies the mood and its impact on debugging performance of the programmer. In their experiment programmers were watching several mood-inducing movie clips and were tasked to make a debugging test afterwards and their productivity was compared.

Copy-paste is a frequently used trick that helps to spare time and effort during the programming. However, one of the main mistakes connected with it is, according to Li et al. [3], failure to remember the need of renaming identifiers after the paste action. This error can be easily discovered by compiler, if there is no identifier with the same name declared. If there is one, it could ultimately create an error which is extremely difficult to find.

Context of the source code creation is connected to the programmer and to the conditions of the code creation. As a context of the programmer we could consider following factors:

– Outer environment – represents the environment programmer works in and all the influences he faces there. It can be, for example, the office where he is disturbed by his colleagues.

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

- Inner environment of the programmer – this means his experiences in the field and his actual mood and state as well. Experience of the programmer can be described as the number of years he spent by programming.
- Time – explains, when the programmer is creating the code. It can be a part of the day, a day in the week, holidays or various unexpected situations. A code created in these situations could possibly contain more errors.

Our method to determine the bug probability in source code is based on evaluation of the programmers' activities in the work. Various combinations of these activities could lead to a rise or a drop in the probability rate.

At our disposal, there is a data set with the logs of several programmers, working on various software projects. Based on this, we are able to analyze the activity of programmers during the software development. We try to discover a connection among bug reports and contexts of programmers during the creation of particular part of code affected by the bug. Based on the influence of various conditions on the final code, we will evaluate these conditions and, in the same time, we will examine various combinations of the conditions and their impact. Final result of our work will be a numerical evaluation that will determine the probability of bug appearance.

Our project is focused on determination of code quality based on the context of programmer. We examine various factors that influence programmer during his work and the significance of their impact on the source code with the ultimate goal of preventing bug creation and failures in the process of programming.

Amended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 595-596.

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] Mark, G., Gudith, D., & Klocke, U. (2008, April). The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 107-110). ACM.

[2] Khan, I. A., Brinkman, W. P., & Hierons, R. M. (2011). Do moods affect programmers' debug performance? *Cognition, Technology & Work*, *13*(4), 245-258.

[3] Li, Z., Lu, S., Myagmar, S., & Zhou, Y. (2006). CP-Miner: Finding copy-paste and related bugs in large-scale software code. *Software Engineering, IEEE Transactions on*, *32*(3), 176-192.

Personalised Search in Source Code

Richard SÁMELA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
richard.samela@gmail.com

Web is a very huge information channel, which provide a lot of text based web pages, pictures, videos or sounds. For finding information about this types of content, users use web search engines. Part of these users is group of programmers. More often they are searching content based on software development. This content can be in the form of source code. Web contains a lot of free source code repositories. Also, programmers, which are working for IT companies have often access to their own source code repositories. Programmers have more choices where to find inspiration, advice or some other solution of their development problems. Therefore, when programmers are trying to help themselves, it may takes a lot of time. This spent time is influenced by quality of personalisation. More information about programmer, means better search results and less time spent by programmer for searching [1].

We would like to collect as much information about programmer as we can. For create quality user model, we need to resolve, which fact about programmers are relevant for us.

Every source code is referring to programmer, who wrote it down. We need them to build context profile, by retrieval names of source code authors (programmers). Except of getting source codes of some programmer, we can use technologies, which programmer learnt and has worked with them. Next we will evaluate knowledge score from specific domain model, all of the programmers [2].
Programmer's user model should contains these attributes:

- ranked experiences
- backward searching queries
- ranked searching results
- programmer's activity in development environment
- programming languages, to which, is programmer able to understand

In this thesis we will analyse various methods for creating a user model of programmer, which take care about knowledge, experiences and abilities to write a

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

code in programming languages, well-known by programmer. Next we will analyse various approaches for personalisation in source code, based on user model principles. We propose a method for creating programmer's user model automatically and usability in personalised searching in source code.

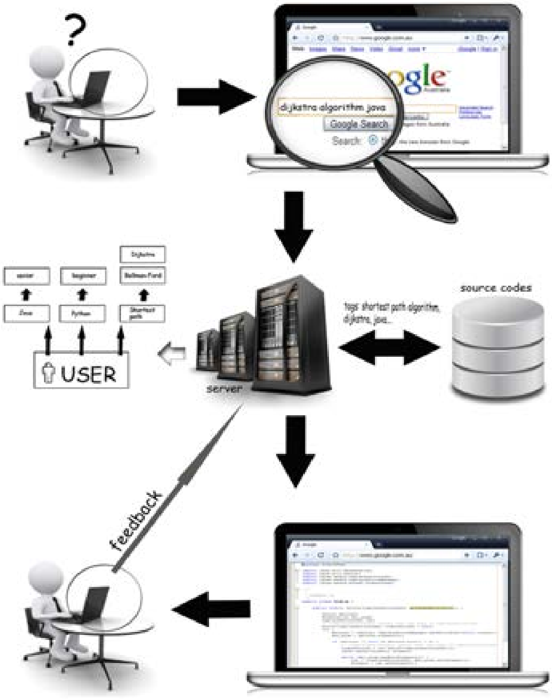


Figure 1. Process of obtaining user information.

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.

Browsing Information Tags Space

Andrea ŠTEŇOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
andrea.stenova@gmail.com

Software systems process information whose amount is growing exponentially thanks to the advent of the Internet. We have found out recently, that it is no longer enough for us to store the data, but we need something more. To be able to organize and effectively browse stored information, we use metadata (data about data), which can provide information describing what data are about, describe their structure, or give us additional information about the user who created them and how they were created. Information tags, an example of metadata types, contain structured information associated with a particular piece of content, such as the number of clicks on the link in a page, keyword characterizing paragraph in the article or the number of lines of method's source code.

Adding metadata to our content has various advantages. Their importance lies in ability to store, organize and provide information about the data whom they are assigned to. This gives us simpler and faster searches in metadata. Metadata can also allow a better understanding of the data and display the data change over the time.

Metadata are usually generated and processed by machine, and their amount makes it difficult for people to effectively read and understand them. There are various problems with users' comprehension of data. However, since the metadata can contain valuable information about the content, as well as the users working with it, it is important to enable users to understand and analyze this information. To ensure readability and understandability we can use navigation of user through the data with the help of their visualization. Visualization helps us to understand the huge amount of data and allows us to navigate within them more easily. Research methods are therefore dealing with how to handle metadata and how to navigate users through them.

Our work is focused on information tags connected to source code, which give us additional information about the code, for example rating of method, or bug in a class. These tags represent characteristics of source code, which they are assigned to. Main disadvantage is that characteristics are scattered around the project and so far there is no option of their easy browsing besides text search. For user, that can be for example

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

student of the subject Team project, these data are really valuable. They can be used to revise his source code, or to check the work of other team members.

Therefore we propose a method to support browsing information tags that are assigned to the source code. The proposed method combines the visualization of source code using fisheye view, visualization of information tags and filtering of shown space of tags using faceted browser. Through combination of these tools we are creating system for user navigation that enables easier searching of source code with selected characteristics.

Amount of information tags in system depends on the size of the project and the willingness of users to enter them manually. Growth of project increases the number of these tags and at some point, visualization of these tags become unusable. Faceted browser allows user to search in information tags that are connected to source code and filter space of these tags and therefore filter space of source code. Different values of facets represent possible values of information tags or ranges of these values. Using faceted browser, user can define search queries by selecting different values of facets. Facets and their possible values are defined in system fixedly, and they depend on type of tags that can be found in the system.

Information tags can be connected to a line or several lines of source code. Source code is divided into projects, packages, classes and methods that create tree of nodes, with relationships between them. We visualize this source tree using fisheye view, which scales the space and enlarges the nodes that are interesting for user. We enlarge nodes that meet the search query, defined by user in faceted browser. Other nodes of tree are shown to the user to preserve context and overview of project.

We visualize the information tags on multiple levels - not only at the place, they were created, but also at individual nodes, by aggregating values of tags of the same type in subnodes. This offers better overview not only about information tags, but also about their values throughout the project.

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

[3] 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, p. 24–43.

Modeling Programmer's Expertise Based on Software Metrics

Pavol ZBELL*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
pavol.zbell@gmail.com

Knowledge of programmers expertise in a software house environment is usually utilized to effective task resolving (by identifying experts suitable for specific tasks), better forming of teams, effective communication between programmers, personalized recommendation or search in source code, and thus indirectly improves overall software quality. The process of modelling programmer's expertise (building the knowledge base) usually expects on its input some information about programmer's activities during software development such as interactions with source code (typically fine grained actions performed in an IDE – integrated development environment), interactions with issue tracking systems and revision control systems, activities on the Web or any other interaction with external documents.

In our research, we focus on modelling programmer's expertise using software metrics such as source code complexity and authorship. We assume that programmer's expertise is related to complexity of the source code she is interacting with as well as to a degree of authorship of that code. By considering both metrics we intend to find answers for these questions:

– What is the programmer's familiarity of a software component? Is it affected more by programmer's degree of authorship or her read-only interactions (like trying to understand desired code, hence complexity) with it?

– Who works on common library APIs and who just uses them, i.e. are we able to distinguish software architects from business logic programmers?

There are several approaches to model programmer's expertise of which most are based on a simple heuristic – Line 10 Rule [1]. On the other hand, more sophisticated models exist such as Expertise profile [1]. Expertise profile is designed to distinguish between programmers who create methods and who call them, it is a composition of Implementation expertise (i.e. programmer's authorship degree) and Usage expertise (i.e. programmer knows which method to call and how to call it). Degree of knowledge

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

model [2] is a similar solution which takes degree of authorship and degree of interest into account. The degree of authorship is a combination of "first authorships" (first emergence of code by original author), deliveries (changes to the code by the original author) and acceptances (changes by others) and represents long term knowledge of the component. The degree of interest reflects component selections and edits, and represents short term knowledge. In case of source code complexity we have not fully explored the possibilities of its utilization and measurement. We believe that approaches based just on LOC (lines of code) metric or its variations can be further improved, e.g. by static analysis of the source code. Our idea is to explore alternative approaches like weighting AST (abstract syntax tree) nodes or call graph based metrics.

In comparison to the existing approaches we intend to focus more on modeling programmer's knowledge from her interactions in an IDE. An extended analysis [3] shows us that programmers interact with source code in many different ways and hence we may improve our expertise model by taking interaction types or patterns into account (e.g. who uses advanced refactoring tools on some code probably has significant knowledge of it). However, the foundation of our expertise model will still be a combination of source code complexity and authorship degree primarily derived from programmer's interactions in the IDE.

We are continually implementing our solution as an extension to Eclipse IDE since the Eclipse platform is easily extensible and has rich possibilities for user interaction tracking. We plan to evaluate our research on data from academic environment or (preferably) real software house environment.

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] Schuler, D., Zimmermann, T.: Mining usage expertise from version archives. In *Proceedings of the 2008 international workshop on Mining software repositories*. 2008. pp. 121.

[2] Fritz, T., Ou J., Murphy, G., C., Murphy-Hill, E.: A degree-of-knowledge model to capture source code familiarity. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1*. 2010. pp. 385-394.

[3] Murphy, G. C., Kersten, M., Findlater, L.: How Are Java Software Developers Using the Eclipse IDE? In *IEEE Software*. 2006. Vol. 23, no. 4, pp. 76-83.

Technology Enhanced Learning (PeWe.TEL)

Student Motivation in Interactive Online Learning

Tomáš BRZA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
`xbrzat@fiit.stuba.sk`

It is important for students to be motivated to learn and gain new knowledge in fields they are studying. For this reason many teachers try to make their classes as motivating as possible. Modern age enables teachers to use different ways of teaching, one of them being interactive online learning system. One of the advantages of online learning is that student can move at his own pace, repeat any lecture he did not fully understand.

Online learning systems also bring certain difficulties, one of them being less human contact with the teacher. This can cause loss of motivation for student as he is not physically present in class.

Our project is about increasing motivation of students using interactive online learning systems in order for them to learn more. Motivation, which refers to "the reasons underlying behaviour" [2] is divided into two categories: Intrinsic and extrinsic motivation. As Deci et al. [1] observe, "Intrinsic motivation energizes and sustains activities through the spontaneous satisfactions inherent in effective volitional action. It is manifest in behaviours such as play, exploration, and challenge seeking that people often do for external rewards". Extrinsic motivation, on the other hand, is influence by rewards, whether they are physical, for example money, or not, for example promotion or higher prestige and standing in community.

We are focusing on multiple elements which should increase both extrinsic and intrinsic motivation and therefore complement each other in motivating students. These elements are divided into three categories:

- Rankings
- Community
- Feedback

Rankings involve points for completing tasks, for helping others and badges for completing challenges inside the system. Badges will be awarded to students by

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 85–86.

proving themselves in completing challenges set by the teachers. For example completing certain amount of task will award student a badge.

We think that important part of motivating students is to enable them to be part of larger community. In this way, students can work together towards goal whether it is just helping each other with simple tips and advices, or by trying to solve tasks together or even compete in trying to solve them faster than others. To further increase community and ways to friendly compete with other students we decided to make their badges and results public. This means that students can compare each other's results and decide to push their limits to prove to others that they are better.

Feedback is category consisting of elements that are supposed to help struggling students. With each attempt to solve a task that fails because the solution proves to be inadequate, student is given reason for evaluation that returned information that task has not been completed successfully. With this reason provided student knows where he needs to improve and what he needs to focus on. He can complete even harder exercises which will provide more intrinsic motivation based on the success and also more extrinsic motivation that is accompanied by badges and points awarded for completing these harder tasks. Without such information, student would not know what is wrong and even despite his willingness to successfully complete the task, he might not be able to and that could cause frustration and anger and eventually loss of motivation.

We will compare these categories among each other and evaluate their effects on students. Since motivation can cause changes in behaviour of student in different manner we wish to evaluate all of the possible effects such as time spent in system, amount of tasks completed, time spent completing tasks, amount of attempts submitted, etc.

We expect that providing students with feedback will increase amount of times they submit tasks for evaluation, as opposed to leaving it alone and moving to next exercise. Badges will foster behaviour which is rewarded by them, in our case it is number of tasks completed, number of comments posted on forum and special badge for being first to solve any task.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 101-106.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Deci, E. L., Koestner, R., Ryan, R. M.: A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. In: Psychological Bulletin, (1999), pp. 627–668. (658)

[2] Guay, F., Chanal, J., Ratelle, C. F., Marsh, H. W., Larose, S., Boivin, M.: Intrinsic, identified, and controlled types of motivation for school subjects in young elementary school children. In: British Journal of Educational Psychology, (2010), pp. 712.

Finding and Harnessing Experts for Metadata Generation in GWAPs

Peter DULAČKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
dulacka@gmail.com

To get most out of the online multimedia, each object needs to be labelled with as exact metadata as possible. In our research we focus on harnessing crowdsourcing games to obtain such metadata. One of the biggest known problem of using the games is a need to validate all user-generated artefacts – which is being done by other users.

The artefact generation cycle in crowdsourcing games consists of: (1) *Artefact creation* (players mostly don't know about this process), (2) *Artefact validation* by multiple players and (3) *Artefact confirmation* after the given threshold is met.

Such cycle brings possibility of accidental removal of expert-generated artefact due to lack of knowledge by non-expert players who perform validation. Currently this process is being realized by: (1) multi-player game design providing real-time validation [1] or (2) single-player game design with a-posteriori artefact evaluation [2].

To speed up this cycle (and possibly skip second step of validation) we propose finding and harnessing experts among players – if we knew which players expertise in given domain is high, we could skip validation of their artefacts and give their voice bigger weight when they are used in validation process of other' players artefacts.

Expert finding is currently being performed in two major fields: (1) closed corporate environment and (2) public CQA portals (e.g. Stack Overflow). Current approaches successfulness varies between 40%-50%. Expert finding methods are mostly based on analysis of forum posts and connections between users:

- *Simple statistical metrics* (such as number of answers)
- *InDegree metric* (number of edges entering given node)
- *Z-Score* (number of edges entering and leaving node)
- *PageRank, Expertise rank, HITS algorithms*

In our work we focus on music and player's expertise in musical domains. We created online radio WoodstockFM[1] and incorporated a custom-made game module, which is

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

[1] http://woodstock.fm

directly connected to the radio. The game module consists of fact-based games which test player's knowledge in the domain of song, which is currently being aired by the radio. Radio displays no information about the song (not even the title) and player needs to answer to questions such as track title, artist, year the artist started performing etc. Apart of fact-based questions game module provides metadata generating games – the principle behind the method is based on our crowdsourcing game City Lights [2].

To analyze fact-based questions and recognize experts, we propose expert finding method based on HITS algorithm [3] altered to be used in crowdsourcing games. The alteration consists of these limitations:

- There are two types of nodes in graph: users and tasks.
- Only user-task relation is allowed.
- The relation can be created only when user answers the task correctly.

We proposed numerous extensions to match our needs in music domain. We conducted an experiment with 6 participants (one of them a-priori marked as a possible expert due to his education). By analysing their answers to fact-based questions, we were able to obviously recognize expert user among others.

We also performed a validation task to check whether their expertise score is related to their performance in task of music metadata validation. The final score didn't vary too much, however level of false positives and false negatives generated by participants in this task was lower for users with higher expertise.

Our preliminary results are promising. In our future work we would like to focus on enhancing the expert recognition and implementation of more artefact-generating games. We also need to implement game elements to enhance fun-effect of the game.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 50-55.

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

References

[1] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, p. 57, Aug. 2008.

[2] P. Dulačka, J. Šimko, and M. Bieliková, "Validation of music metadata via game with a purpose," in *Proceedings of the 8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012, p. 177.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

Dynamic Score as a Mean for Motivation of Students in an Educational System

Richard FILIPČÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
richard@filipcik.sk

There are many ways on how to motivate student during the use of web-based learning system. It is score and pointing systems which belong in most frequent and used ways to support motivation. Most educational systems use various methods of score computation – from the simplest which just count amount of performed activity to the advanced which take many factors into account.

We propose dynamic score computation method based on users activity, so system can modify „weights" of particular activities and induce students to perform all of them more or less equally. We are also making an effort to implement this method into learning system ALEF [1]. By this method we aim at increasing students' motivation thanks to smarter score computation.

This method relies on several factors important for score regulation. While some of them are constant and do not change over time, others are not and they can vary depending on the actions being performed:

– activity weight, which is value estimating effort necessary to perform this activity,
– activity preference set by the teacher, defining how important is the activity in the certain time period and
– activity priority, computed by the system itself, it is based on current status of activity of all students that is amount of activities performed and their relative ratio.

Proposed method calculates and regulates score for the student s at time t following expression below:

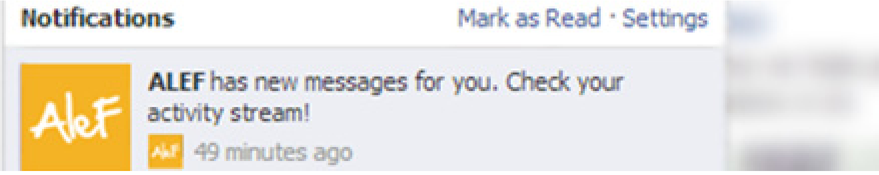
$$score(s,t) = score(s,t-1) + \sum_i partial\_score(C_i, s, t) \qquad (1)$$

where *score(s,t)* is score regulation function returning score for the student *s* at time *t*. We can see it sums up student's previous score and score additions for all activities performed at time *t*.

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

To give the student feedback from the system to know when and how factors affecting score regulation change, part of our method is also a feature called activity stream, which is similar to news streams popular on many social networks. In the stream student can see messages about preference or priority changes.

To bring messages from the stream right to the student, there is an option to connect his ALEF account with account on Facebook, so he/she can see what's new in the notifications bar. Thanks to rising popularity of various social networks it is likely that many students have their own accounts too. Bringing ALEF stream messages into Facebook notifications bar should motivate student to use ALEF more frequently. Notifications visualization is shown in Figure 1.



Notifications

Mark as Read · Settings

ALEF has new messages for you. Check your activity stream!

49 minutes ago

Figure 1. Visualization of ALEF message in Facebook notifications bar.

We plan an experiment consisting of two phases. In the first phase we will observe the behaviour of students when there will be no explicit feedback from the system about activity priority nor activity preference changes. The second phase will continue with new score regulation method, however, this time students will be explicitly informed about changes in preference or priority. We will compare results obtained from the both phases of experiment with each other and will determine how decisive was providing feedback by activity stream for students decisions on which activity to perform as next. Our evaluation will also be integrated with some more experiments based on ALEF which are aiming on external source adding and question-answer learning objects [2].

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 3-8.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Šimko, M., Barla, M., Bieliková, M.: ALEF: A framework for adaptive web-based learning 2.0. In: *Key Competencies in the Knowledge Society* [online]. Springer Berlin Heidelberg, Berlin, Germany, 2010. pp. 367-378.

[2] Šimko, J., Šimko, M., Bieliková, M., Ševcech, J., Burger, R.: Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects. In: *5th Int. Conf. of Computational Collective Intelligence Technologies and Applications*, ICCCI 2013, Springer LNAI. Vol. 8083, pp. 62-71, 2013

Facilitating Learning on the Web

Martin GREGOR*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
`greczsky@gmail.com`

Today, we have many opportunities to learn something new. One of the most preferred ways of learning is Web, where we gain information through reading articles on the Web. This type of learning is a part of life-long learning, which can be supported by computers and technology [6]. One of the analyzed problems is inaccuracy of user knowledge modeling, which can support learning on the Web. Web browsing requires user interaction (obviously). User interaction can be observed and measured as feedback from a user. This type of feedback, which does not require any additional effort from user, is called implicit feedback. Typical implicit interactions on the Web are for example clicks, text selections, moves, scrolls, time spent on page. From implicit feedback we can deduce what user reads and what he learns.

Feedback collection and evaluation is one of approaches used to user modeling. The basis of every personalized educational system is user model [5], specifically user knowledge model and collection of implicit and explicit feedback [1]. System stores information about knowledge of a user in his/her model. The most important thing for personalization of educational system is the precision of user knowledge model that is used for personalization. Inaccuracy of user knowledge model is a main problem of educational systems. Information in user knowledge model is deduced from interaction data originating in implicit and explicit feedback.

We can deduce information from implicit feedback on two levels [2]. The lower level is information deduction from one elemental interaction of user. The higher and better level is information deduction from combination of some interactions which user did. Implicit feedback is easily obtainable and trustworthy but it is sometimes really difficult to deduce information about user [2]. The second type of feedback is explicit feedback which requires additional effort from user, for example answering the questions, providing some ratings. On the other hand, we can consider untrustworthiness caused by wrong motivated user intent among disadvantages. Advantage of explicit feedback is possibility to check correctness of implicit feedback.

Our aim is to propose and evaluate a method of user knowledge modeling on the Web. In this work we focus on the area of term learning on the Web (this covers wide

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 91–92.

range of applications, e.g., technical vocabulary acquisition, foreign language learning, etc.). Our solution extends existing model and method of user behavior monitoring [4] and collects and evaluates implicit feedback from generic implicit indicator inspired by the work of [3]. Our main idea is to monitor a number of mouse entries ("mouseenters") to an area of interest of a web document. We use this number to predict what user reads and what user learns. In addition, we proposed a collection of specific implicit indicators for domain of term learning as well as method for their processing into the user model. These two implicit indicators are translation of term by user and term exploration by user.

The main contribution of our approach is term-based user knowledge modeling based on a number of mouseenters to an area of a document. The number of mouseenters to document areas is a novel implicit feedback indicator which we have proposed as a result of our research including several small experiments. The preliminary results show that this indicator outperforms state-of-the-art indicators utilized for predicting user's reading behavior.

To date, we have implemented our method as a JavaScript library. In addition, we have already done a preliminary experiment to evaluate partial hypotheses of our approach. Our future work covers conducting a more complex experiment to evaluate the stated high-level hypotheses.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 143-148.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Brusilovsky, P.: Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 1996, vol. 6, no. 2-3, pp. 87–129.

[2] Claypool, M., Le, P., Wased, M., Brown, D.: Implicit interest indicators. In: *Proc. of the 6th int. conf. on Intelligent user interfaces*. IUI '01, New York, NY, USA, ACM, 2001, pp. 33–40.

[3] Hauger, D., Paramythis, A., Weibelzahl, S.: Using Browser Interaction Data to Determine Page Reading Behavior. In: *User Modeling, Adaption and Personalization*. Volume 6787 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 147–158.

[4] Horváth, R., Simko, M.: Enriching the Web for Vocabulary Learning. In: *Proc. of 8thEuropean Conf. on Technology Enhanced Learning, EC-TEL,* 2013, pp. 609–610.

[5] Tozman, R.: Learning in the Semantic Web. *eLearn*, 2012, vol. 2012, no. 3.

[6] Trilling, B., Fadel, C., for 21st Century Skills., P.: *21st century skills: learning for life in our times*. Jossey-Bass, San Francisco, 2009.

Adaptive Collaboration Support in Community Question Answering

Marek GRZNÁR*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
marek.grznar@gmail.com

Nowadays, users are lost in a great amount of information available on the Internet. Many times they come into a situation when information, which they search, is not easily found anywhere on the Internet using traditional search engines. With the development of Web 2.0, there is an option to obtain such information by asking a community [1]. This kind of systems based on the sharing of knowledge to each other is being used lately. One type of these systems is Community Question Answering (CQA). Typical examples of such CQA systems are Yahoo! Answers and Stack Overflow.

The existing CQA systems, despite of their increasing popularity, fail to answer significant number of questions in required time. In some of current popular CQA systems, only 17,6% of the questions are answered sufficiently [2]. This critical problem was confirmed also in other studies, such as authors in [3] found out that only 11.95% from questions were answered in one day and in two days, there were just 19,95% of answered questions.

There are several options how to adaptively support users during community question answering process and thus how to achieve more successful results. One of these options for supporting cooperation in CQA systems is a recommendation of questions to a user which is a suitable candidate for providing the correct answer (Question Routing). Various methods have been proposed to help find answerers for a question in CQA systems, but almost all work studies heavily depends on previous users' activities in the particular system (QA-data). These methods use different aspect for question routing, such as:

1. users' knowledge;
2. users' activity;
3. users' motivation.

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

In our work, we focus on utilizing users' non-QA data as a way of personalized support during question routing. By analyses of users' non-QA activities, such as blogs, micro-blogs, friends etc., we can identify a suitable user for answering a specific question better. One of the examples could be to get social connections of answerer from social network (e.g. Facebook) and then we expect, that friends will more likely answer each other.

Using non-QA data is also helpful in solving of the cold start problem, because when there is a new user in system, it does not have information about this user and cannot recommend him or her any question. Additionally, many users in CQA systems are inactive. It means that they do not ask any question or they do not get answer on any question. This users are called lurkers [4]. We think that non-QA data also helps to recommend questions to this kind of users.

In the process, we also consider the traditional QA data such as votes, given answers, etc. Our method also includes existing semantic methods for question routing. We employ state-of-the-art methods for extracting semantics (LDA, LSA).

We want to evaluate our proposed method on a dataset from existing CQA systems (e.g. from Stack Overflow). After these experiments, we want to evaluate results from dataset obtained at our own faculty CQA system named Askalot[1]. Moreover, we will implement the proposed method as a part of Askalot and consequently, we can employ this system in a live experiment in which we will recommend questions to students at our faculty.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Q. Liu and E. Agichtein, "When Web Search Fails , Searchers Become Askers : Understanding the Transition," pp. 801–810, 2012.

[2] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10,* 2010, pp. 1585–1588.

[3] T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 2012, p. 783.

[4] M. Muller, "Lurking as personal trait or situational disposition," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12,* 2012, p. 253.

[1] askalot.fiit.stuba.sk

Natural Language Processing by Utilizing Crowds

Jozef HARINEK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
j.harinek@gmail.com

The amount of information stored in natural language on the web is huge and still growing. In order to make a better use of this information, we need to process the natural language and transform it to a form that machines are capable of understanding. This is performed by Natural Language Processing (NLP).

However, NLP is a difficult task. It has several levels in which it can be performed [1]. The first and most basic one is to recognize sounds. Then it goes to recognition of speech, in which the machine can divide sounds into words. Next step in NLP is morphological analysis. Here are the words analyzed and their grammatical categories are extracted. Next layer in analysis is syntactic layer. In this layer, the syntax of the text is recognized. It is represented by sentence components and relations between them. Another layer is morphemic layer which adds information about morphemic structure of the words. Last two layers are semantic and context layer, in which semantic structure is identified and put into context of the given text.

In our work we are focusing on syntactic analysis of Slovak language. This is a field that is still growing and much work is to be done, due to its specificity and grammatical structure [2]. To support NLP in Slovak language, we also plan to employ principles of crowdsourcing.

Crowdsourcing is a process, in which one uses the power of crowd to perform a task that is typically large and needs lots of experts to be involved [3]. To be successful, the crowd needs to be motivated and given tasks need to be small enough to be performed by non-professionals. There are several ways of motivating people. They can be motivated by a financial reward, by enjoyment they experience during fulfillment of the task (Games with a purpose), by added value of performing the task (they learn something), etc. [3].

In our work, we want to use the methods of crowdsourcing to help us annotate large scale texts, a task that normally needs lots of experts and time. We plan to verify our method in a software tool, which is specially designed to support education in

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

elementary schools. Children will perform sentence analysis assignments given by teacher. We also plan to verify data, that was manually annotated, obtained from Ľudovít Štúr Institute of Linguistics at the Slovak Academy of Sciences.

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

References

[1] Cimiano, P. *Ontology learning from text*. Springer, 2006.

[2] Čižmár, A., Juhár, J., Ondáš, S. Extracting sentence elements for the natural language understanding based on slovak national corpus. In *Proceedings of the International Conference on Analysis of Verbal and Nonverbal Communication and Enactment*, COST'10, LNCS Vol. 6800, Springer, pp. 171-177, 2010.

[3] Quinn, A.J., Bederson, B.B. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1403-1412, 2011.

Analysis of Interactive Problem Solving

Peter KIŠ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
koritnak@gmail.com

This study aims to analyze the way users solve interactive tasks in the world of information technology. The main objective of this work is to examine the relationship between personality traits of players with their gaming expressions and thus creating a characteristic model of gameplay behavior for each characteristic group of players. An additional aim is to explore the impact of gaming conditions on emotional responses of players.

For this purpose we designed a casual browser turn-based game Hexa with logical and strategic elements. The game contains 15 levels that are formed by different hexagonal game map with predefined shape. Each game-map cell has one random color out of several available colors and can create a union with its neighbor cells of the same color. The core principle of the game is to occupy as many cells as possible. In each turn, player picks one of the available colors, and the cells of the selected color which are neighbors to previously occupied cells are occupied by the player in the current turn. Every level can be played in 4 different game modes: player against none, player against computer, player against player on 1 PC, player against player on 2 PCs.

During playing of the game, we track player's in-game interaction like mouse moving and clicking, colors picked and game time. Our aim is to collect data that reflect player's gameplay in comfortable conditions. Playing the game in calm conditions is necessary to create precise model of gameplay that we want to link to player's personality. The personality is measured by Big Five personality test consisting of 60 questions. We adjusted the results for population of Slovakia. Logged raw data are used to compute multiple gameplay indicators as mouse movement speed and effectiveness, speed and the level of optimality of player's moves. All players will be divided into several groups based on their personality traits and for each group we would search for its characteristic gameplay indicators.

For next experiment we want to track player's gaze by eye-tracker. The experiment consists of multiple eye-tracked levels for each player. The first would be used to calibrate eye-tracker to fit conditions of current player (head-display distance, eyes positions…) and the others would be used to create characteristic model

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 97–98.

representing a way the player is analyzing the game map with his gaze. We assume that data from eye-tracker would be accurate enough to show us precise trajectory and speed of player's gaze, the time when eyes movement stops and game map regions having the greatest player's attention. We are interested in correlations between created model and player's personality.

The aim of the last experiment is to demonstrate the impact of gaming conditions on emotional responses of players. The different gaming conditions are represented by 4 game modes. We assume that playing the game against different opponents would create different emotional response – fun. In this experiment, players will be divided into pairs (buddies), where for each player we would measure his galvanic skin response during playing games against different opponents. The first measurements will be made while playing the game without opponent. The second measurement will be carried out while playing a game against AI-controlled opponent. The third series of measurements will be conducted as well as the previous, with the only difference that the opponent in the game players would be his buddy (second member of the pair). Last measurements would be again during playing the game between friends, but in this case will be played on one computer while players would alternate after each move. We assume that the physical presence of both players of a pair would be important in this case, because by this measurement we want prove that physical contact during playing multiplayer game increases emotional reactions compared to playing against each other "at a distance". Players in all series of measurements will play a few games with the measured values of skin impedance which would be then averaged to average skin impedance for every game mode. By projecting the averaged values in a chart we will be able to capture the differences of the entertainment value of the game depending on the selected game mode.

The results of this project imply to create a model that helps us better understand player's game play influenced by his personality traits. This model may be used by game developers to make their games or applications with interactive content better balanced for every targeted player or user.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Atkins, M.S. et al.: A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2006. s. 1027-1036.

[2] Lindley, C.A., Sennersten, C.C.: An Investigation of Visual Attention in FPS Computer Gameplay. In *VS-GAMES '09 Proceedings of the 2009 Conference in Games and Virtual Worlds for Serious Applications*. 2009. s. 68-75.

[3] Brown, R. et al.: The Gameplay Visualization Manifesto: A Framework for Logging and Visualization of Online Gameplay Data. In *Computers in Entertainment (CIE) – Theoretical and Practical Computer Applications in Entertainment*. 2007. Vol. 5, no. 3.

Keyword Map Visualisation

Matej KLOSKA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matej.kloska@gmail.com

Nowadays, people are creating more digital documents – data elements – than ever before. If we want to search and navigate in those documents, we need efficient way how to interpret relations between documents. Information visualisation has become a large field and various "subfields" are beginning to emerge. The question is, whether there is an inherent relation among the data elements to be visualised. Proper relations creation and information visualisation implies high success rate in looking for desired information in documents.

If we work with a large number of data elements, we often need efficient representation of content – e.g., to describe each document with an appropriate set of keywords. Keyword sets alone do not guarantee quality of maps, i.e., the ease of navigation in information space. Quality in this sense strongly depends on interconnections of keywords between documents. There are several techniques how to create keyword connections. Most of them are based on clustering techniques.

Another problem is, how to properly visualise keywords and relations between them. The importance to support users and provide adequate user experience is also crucial problem, which we have to keep in mind. The more fail proof interface, the more valuable maps.

We propose a method for keyword map visualisation for educational system ALEF with support of content managing system COME^{2}T (Collaboration- and Metadata-oriented Content Management EnvironmenT).Our method will be implemented and evaluated using existing system COME^{2}T [1]. The COME^{2}T allows easy administration of lightweight semantics for the provided content – digital documents and user-created annotations. It is being used as a content management system for educational system ALEF.

Graph structure highly affects visual quality of output map. Basically, there are two key issues targeted to a structure. First of them is *graph size*. The higher the count of vertices in map, the more difficult to work with such graph. To solve this issue, we introduce structural clustering in order to virtually reduce number of visible vertices on

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

screen. Structural clustering is more suitable in comparison to traditional semantics clustering that is based on clustering of keywords' semantics.

Second issue related to graph structure is *graph density*. Dense graph is a graph with the number of edges close to the maximum number of edges. On the other hand, sparse graph is the opposite – low number of edges. The higher the number of graph edges the lower the readability for the user.

We have identified three features that would improve user experience and make navigation easier – keyword map overview, navigation bar and coloring of nodes and edges. The keyword map overview will help user in every moment know where in map (s)he exactly is. It would be particularly useful in the case of zoom and pan where only a small part of map is shown on screen.

Currently, there is no other simple way how to provide user simple feedback on position in map in COME^{2}T, especially when clustering is applied. Navigation bar would enhance track of navigation in map when user wade in any cluster. It is similar to well-known address bar in web browsers or file explorers. In our case, bar provides information about users' map exploration path with respect to cluster hierarchy.

The third proposed feature is colouring of map elements. Visual feedback is a key feature for every user. We would like to provide interface for custom colouring of nodes and edges in addition to automatic one based on defined rules and analysis of graph. Automatic colouring is easy to implement and use because of predefined edges types.

Since we are still in a phase of implementing the proposed utilities to the system COME^{2}T, we here describe the evaluation plan. Our method would be evaluated in several phases. In first phase we will evaluate the impact of user interface and visualization library changes on time required for creation defined maps for provided courses. We expect, that time required for creation will be smaller than for the early method. In second we will evaluate the impact of supporting features like navigation bar and keyword map overview when a user comes back to previously created maps. We expect that supporting features will help a user orientate in map easier and recognize work already done. Final evaluation of method will be evaluated using the same tasks as in evaluation made before implementation of our method.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 9-14.

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

References

[1] Šimko, M., Franta, M., Habdák, M., & Vrablecová, P. (2013). Managing content, metadata and user-created annotations in web-based applications. In *Proc. of the 2013 ACM Symposium on Document Engineering, DocEng '13*, ACM, pp. 201–204. doi:10.1145/2494266.2494270

Observing and Utilizing Tabbed Browsing Behaviour

Martin LABAJ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
labaj@fiit.stuba.sk

In our work, we focus on observing, analysing, logging and utilizing the tabbed browsing behaviour both within adaptive web based systems and on the open Web. Adaptive systems in general make use of information about users (user model), content (domain model), etc. Apart from methods that analyse content, the user behaviour, both as explicit and implicit feedback, is often used to create such models. Tabbed browsing could express implicit user actions and we can possibly improve user models, domain models, or aid in recommendation.

The tabbing (also called parallel browsing) is a more accurate description of user activities during browsing than previous and simpler linear browsing models [1], which considered visits to resources in a linear fashion, where each “click” (a page load) replaces previously opened page. In tabbing models, in the same way as in the real browsing activity, the user can have multiple pages opened at once (opening them in separate tabs) and switch between them at any time during their existence. Users do use tabs in various situations – *reasons for using tabs* [2], such as:

– keeping bookmarks of pages to read by keeping them opened in tabs,

– comparing pages,

– looking for additional information in tabs about topic on a given page,

– opening multiple search results to tabs, and

– keeping to-do lists expressed as tabs.

The parallel browsing behaviour, however, cannot be reliably inferred from typical server-side logs. It can be observed with the aid of client-side scripts embedded within web pages (observing tabbing activities of all users of the application, but only for pages within such application, no tabbing is observed for other pages) or from a browser extension (observing tabbing on all web applications being visited in the

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

augmented browser, but only within a smaller group of users who have the extension installed).

We model the user tabbing behaviour from events sourced either from such a browser extension, where we realized prototype implementation in a Brumo extension framework, or from scripts included in pages served by Adaptive LEarning Framework system (ALEF). The events include traditional page loads, page unloads, but also focus and visibility tracking. These events are tracked per each tab during life-time of a page loaded in it, creating an activity vector. Such vectors from multiple pages are combined and user tabbing activity (e.g. open link in the same tab, open link to a new tab, switch between pages) is reconstructed.

Possibilities of using the tabbing data are for example in item (page) recommendation, where tab switches between items may represent relation between those two items. Another possible usage of user parallel browsing behaviour is in annotating the content in an adaptive system with additional resources from different web pages based on the user tab sessions from the system, across various domains and back to the system. Ultimately, our goal is to augment user and domain modelling in adaptive systems by taking parallel browsing into account. We therefore add another level of inference on top of the processed logs with recognized tabbing actions, where we try to recognize tabbing scenarios and use those to find user interest or content relations.

We are currently modifying the single-application logger (used within ALEF system) to consider tab switch delays and subsequent time spent on the page in a notion that one page rates another one when the user performs switch action and remains there (the time is a weight of the rating). We also propose a browser extension called Tabber, which allows users to view and analyse their usage of browsers tabs, while its data can serve as a dataset of browsing the open Web.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

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 International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, IEEE, (2006), pp. 262–269.

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

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

[4] Labaj, M., Bieliková, M.: Tabbed Browsing Behavior as a Source for User Modeling. In: *LNCS Vol. 7899 User Modeling, Adaptation, and Personalization: 21th International Conference (UMAP 2013)*, Springer, (2013), pp. 388-391.

Acquisition and Determination of Correctness of Answers in Educational System Using Crowdsourcing

Marek LÁNI*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 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 community question answering (CQA) portals and web sites, which are being used to gain knowledge and information. Therefore, the systems are beneficial to the users, however the users can be beneficial to the systems too. We can say, that it is a win-win relationship. The benefit for the systems comes from the content, which is often crowdsourced, i.e. generated by users themselves. Representatives of such systems are for instance portal stackoverflow.com or answers.yahoo.com. Thanks to these systems, it is possible to get an answer to almost any question. Nevertheless, it is important to filter and rate the answers, because the correctness of the user generated content cannot be guaranteed. Filtering and rating are activities, which users can perform in the majority of CQA systems. These activities consist, e.g. from assigning thumbs up/down or marking the answer as correct.

A basis of this CQA principle can also be used as an exercise in TEL systems. Teaching processes many times comprise scenarios, in which teachers ask questions and students (crowd) try to provide the correct answer. By this activity the students can clarify their understanding or gain new knowledge and the teacher can get an overview of the knowledge of the students. The whole process can be either part of a lessons or part of a TEL system as an exercise. In the exercise, it is possible to add additional features to this process and enhance it. Such a feature can be the evaluation of answer correctness. In case the students are involved in process of evaluation of correctness of answers, i.e. peer-grading principle is used, students can learn not only by answering, but also by evaluating and according to Sadler and Good [2] it brings different pedagogical advantages. This principle of interactive exercise based on answer correctness evaluation was used in the related work [3] on which our work is built. In our work, we took the exercise and enhanced it with a new features as

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

question answering, discussion etc., while we tried to provide the best possible contributions for improvement of the learning experience. We have also qualitatively verified this exercise with several users and they agreed, that the exercise is helpful in learning process.

Nevertheless aim and research contribution of our work is focused not only on providing the new interactive learning exercise, but also on analysis of collected data. We use different methods to interpret crowd evaluations of the answers correctness and to determine, if the crowd is capable to evaluate the answers similarly to expert. There are many works in context of CQA, which tries to determine answer correctness, but only few of them use crowd evaluations of answer correctness, however they are not giving much importance to these evaluations and they use them just as one factor in determination of answerer expertise level [1]1, 2]. As we think, that the crowd has big potential in answer correctness determination, we have created methods of interpretation of the crowd evaluations to determine answer correctness. These methods are:

- *Filtering of outliers* – in clear cases (cases with the small distribution of evaluations) we determine outlying evaluations resp. outliers and filter their evaluations out in unclear cases (cases with the big distribution of evaluations)
- *Evaluations weighting according to* their *distribution* – we have created more methods, which focuses on examination of distribution of evaluations and which try to prove different assumptions resulting from evaluations distribution.
- *Machine learning* – this method uses neural network, where the input is vector of evaluations and output is estimation of expert evaluation.

We plan to evaluate these interpretation methods by comparison with expert evaluation assigned to every answer and as a reference values, we plan to use the simple average of the evaluations of the answers.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Kao, W. Ch., Liu, D. R., Wang S. W.: Expert finding in question-answering websites: a novel hybrid approach. In: *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, ACM, 2010, s. 867-871.

[2] Sadler, P., Good, E.: The Impact of Self- and Peer-Grading on Student Learning. In: *Educational Assessment*, Volume 11, Routledge, 2006, *s.* 1–31.

[3] Šimko, J., Šimko, M., Bieliková, M., Ševcech, J., Burger, R.: Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects. In: *5th Int. Conf. of Computational Collective Intelligence Technologies and Applications*, ICCCI 2013, Springer LNAI. Vol. 8083, pp. 62-71, 2013.

Recommendation in Adaptive Learning System

Viktória LOVASOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
viktoria.lovasova@gmail.com

Recommender systems provide the user with recommended items he might prefer, or predict how much he might prefer each item. These systems help users to decide on appropriate items, and ease the task of finding preferred items in the collection [1].

It is very common that students learn via adaptive web based learning systems. Every user learns with a different speed, that is why it is appropriate to estimate a user's knowledge and recommend him the most useful learning object (LO). In modern browsers, the user can view different web pages in multiple windows or tabs, so he can focus on one page and switch between others. If we could capture such behaviour with the use of parallel browsing while he is learning, we could improve recommendations.

We propose and implement a recommendation system in ALEF based on users' parallel web browsing behaviour with the aim to improve recommendations. ALEF is an adaptive learning education system [2] being developed at FIIT that is used in several courses. We track user's behaviour in ALEF via client-side scripts. We cannot be sure what pages other than ALEF's pages are opened by the user, because the script scope is only within the system being tracked. Because we recommend only learning objects from ALEF, this is not a severe limitation. We consider the following actions important in our research: page *load*, page *unload*, *time spent* on the page, *switching* between tabs.

According to such browsing behaviour, one learning object could rate another one using the user's tab switches between them. We proposed the following formula where $R_{i,j}$ is the rating for LO_i from LO_j, v is a value which is added according to the time spent on the tab after the switch in seconds: $R_{i,j} = R_{i,j} + v$. Let LO_1 and LO_2 be learning objects in ALEF and suppose that the user switches from one tab with LO_1 to other tab with LO_2, he spends there 2 minutes, and then he switches back to LO_1 (Figure 1). From these actions, we can assume that there is a relation between LO_1 and LO_2, since the user spent some time on LO_2. When another user visits LO_1, we can recommend him the LO_2.

We evaluate the content-based similarity between the objects rated by user switch pairs by computing their cosine similarity on the concepts assigned to them in the

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

domain model. If the switch pairs correlate with the content similarity, the algorithm could replace content analysis in domains, where it is costly – e.g. video or image domain.

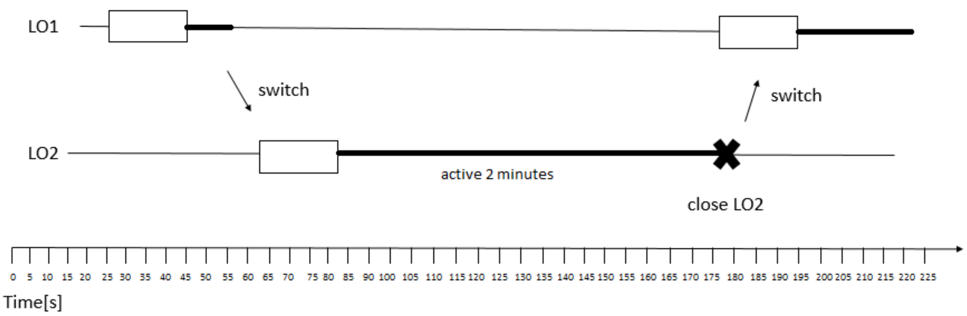


Figure 1. Switching from LO_1 to LO_2.

If the correlation is not observed, it can mean that the user browsing behaviour express different relations than those stemming from content. In that case, we will prepare a dataset by selecting pairs that were rated high according to browsing behaviour (our method) together with those rated by content similarity, and, as a control (for comparison), we mix in low-rated and random pairs. An expert on the course will evaluate the pairs to identify relations.

We aim for realizing the method in the live system that is used in the FLP course within ALEF where students prepare for seminars and exams over longer time periods. They will have a widget on the right side, where there will be recommendations for them. Half of the recommendations will be generated from the sequence recommender, the other half from parallel based recommender. They will be displayed intermixed together. There should be more clicks on the recommendations generated from the parallel recommender than from the sequence one. The recommended objects will be logged in the database. The recommendations generated from the two recommenders can be evaluated against each other

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 113-118.

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

References

[1] Ricci, F., Rokach, L., Shapira, B., 2010. Recommnender System Handbook, *Evaluating Recommendation Systems:*258.

[2] Šimko, M., Barla, M., Bieliková, M. IFIP Advances in Information and Communication Technology, *ALEF: A framework for adaptive web-based learning 2.0.* 2010:367-378.

Automatic Web Content Enrichment Using Parallel Web Browsing

Michal RAČKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xracko@stuba.sk

In the domain of technology enhanced learning, it is important to determine relevant information about learning objects and relationships in it. Adaptive learning systems are expanding the area of personalizing the learning needs of individual users. Assuming that users behave similarly when following the same goal e.g. searching for additional information for a given topic, we propose a method for automatic web content enrichment based on their actions in the domain of open Web.

Currently, there are a large number of web browsers allowing tabbed browsing. All of them in their latest versions support this kind of browsing [1]. Some of them allow persistently maintaining selected tabs, or renew tab state even after the user closes the application. Various researchers found that the users use tabs in a ways like temporary lists, parallel search results, etc. Majority of these ways of usage was not explicitly planned.

The aim of this project is the relationship discovery between sites frequently visited by users using multiple tabs. What are the relations between them, or whether they can be linked together based on the way the users have accessed them. These links may not depend on the hyperlinks between sites, but based on the user browsing behaviour. The proposed method is evaluated in ALEF adaptive learning system where the aim is to make easier or automatize adding external resources to learning objects.

We can model a user browsing session as a sequence of actions performed during browsing [2]. Processing the tabbing actions among browser tabs can pose a problem since each tab is seen as a separate dimension. Therefore we have to propose a method that will flatten those dimensions into one sequence of user actions. Then we identify sequences that led to adding external resources to adaptive learning system. When analysing parallel browsing behaviour data, we consider different ways of switching among tabs. Each of these actions is used irregularly. Research has shown [3] that re-visitation rate of tabs is much larger than the number of sites visited. Other important information when examining user browsing behaviour include tracking how much time

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

he/she spent browsing the content of the page [2] and thus we can reconstruct the actual page viewing.

When browsing the web, users also perform various irrelevant actions which we remove in pre-processing, such as fast switching between tabs. Then we classify pages into categories: adaptive system, digital library, search engine and other. Web browser usage data is a continuous record, so we have to be able to identify individual user sessions. A user session can be determined based on the number of existing tabs, and when the number reaches zero, it means the end of session.

In the last phase of pre-processing, we divide the continuous record into individual sessions and then divide these sessions into loops. A loop is defined as the smallest sequence of actions, which starts and ends in the same learning object. Between the initial and final action, there must be at least one other switch action between pages. The loops that are not closed by the end of a session are discarded. The resulting loops contain potentially relevant external resources to the learning object that began the loop.

We adjust the weighting of web pages with a function defining ratio between the active time spent on the site and its significance. The category represents 70% of final page weight, which is assigned from 0 to 1. Browsing time weight multiplier is logarithmically dependent on the time spent on the page, where the upper limit is empirically defined to be 20 minutes. Final formula for weight calculation is (1).

$$w = 0.7 * w_c + 0.3 * \log(t) \tag{1}$$

w_c is category importance and t is browsing time in seconds. Method output is a set of pairs learning object – external resource. Best method for discovery of loops potentially containing relevant external resource to a learning object was Naive Bayes that has 93.24% recall rate for chosen class. Trained classification method is used for further loop classification.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 173-178.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Dubroy P., Balakrishnan R. 2010. A study of tabbed browsing among mozilla firefox users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10),* ACM Press, pp. 673-682 (2010)

[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] Zhang H., Zhao S. Measuring web page revisitation in tabbed browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11).* ACM, pp. 1831-1834 (2011)

Adaptive Support for Collaborative Knowledge Sharing

Ivan SRBA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
srba@fiit.stuba.sk

Besides standard search engines, current possibilities of the Web allow us to employ many supplementary sources of information. These nontraditional sources of knowledge are often based on collective intelligence. Concept of collective intelligence refers to shared knowledge which emerges from common collaboration of community of users that share common practice, interests or goals. Collective intelligence is present in many popular web systems, such as forums, social networking sites or wikis. In recent years, the new forms of systems based on collective intelligence has appeared. One of them is Community Question Answering.

Community Question Answering (CQA) is a service where people can seek information by asking a question and share knowledge by providing an answer on the particular questions. One kind of CQA systems is providing users with a possibility to ask any general question without any topic restriction (e.g. Yahoo! Answers or Wiki Answers). On the other hand, there are also topic-focused CQA systems dedicated to specific domains (e.g. Stack Overflow where users concern with questions related to the programming).

A process of question answering in CQA systems consists of several steps. At first, an asker posts a question by formulating a title and a description of a problem which is the subject of the particular question. In addition, it is usually necessary to select an appropriate question's topic (a category or a set of related tags). Afterwards, other users can collaborate and provide their answer-candidates on the posted question. Answerers can vote for the most appropriate answer-candidate and thus help the asker, the CQA system and all users, who are involved in question answering process, to identify answers with the highest quality. The asker can finish the answering process by selecting the best answer, which satisfies his/her information needs best, and consequently the question is marked as answered and moved to the archive of solved questions.

The main goal of CQA systems is to harness the knowledge potential of the whole community to provide the most suitable answers on the recently posted questions in the

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

shortest possible time. We assume that searching for the answer to the question is actually also a specific way of learning. Moreover, providing an answer provides also a possibility to answerers to acquire a new knowledge. Therefore in our project, we present a novel perspective on CQA systems as collaborative learning environments.

In recent years, we witness increasing amount of research studies which concern with different aspects of CQA systems. The significant part of them focuses on providing adaptive support (e.g. [1]). We focus on *question routing* which is probably the most important part of each CQA system. It refers to a recommendation of potential answerers who are most likely to provide an appropriate answer on the newly posted question [2]. In our project, we propose a novel method for question routing on the basis of existing methods for question routing while promoting diversity of routed questions to maximize the learning potential of CQA process.

In addition, when we consider CQA systems as an innovative learning environments, we suppose that their potential for supporting of *organizational knowledge sharing and collaborative learning* is only to be discovered. In educational organizations, concept of CQA systems can be employed as a complement to formal learning in particular courses or even as an immediate component of learning process where community of students together with teachers can participate on solving questions related to students' learning. In business environment, CQA systems can be utilized to workplace learning while solving the questions about different problems employees run into during their work. And finally, in a research context, different research groups would be able to take advantage of asking questions about their research activities and receive knowledge from researchers who are experts in the particular domain.

We plan to evaluate the proposed method by employing a dataset from Stack Overflow which is considered as one of the most successful CQA system on the current Web. Alongside, we develop a CQA system named Askalot which is designed for organizations and more specifically for universities where students can take advantage of learning aspect in question answering process. Askalot will provide us also a possibility to study concepts of CQA in organization environment and consequently to apply the proposed method also in a live experiment.

The main contribution of our project is innovative perspective on CQA systems as environments which can support knowledge acquisition. We reflect this perspective in the proposal of question routing method as well as in organizational CQA system Askalot which is already in use at our faculty.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Li, B., King, I.: Routing questions to appropriate answerers in community question answering services. In: *Proc. of the 19th ACM international conf. on Information and knowledge management - CIKM '10,* ACM Press, (2010), pp. 1585–1588.

[2] Szpektor, I., Maarek, Y., Pelleg, D.: When Relevance is not Enough: Promoting Diversity and Freshness in Personalized Question Recommendation. In: *Proc. of the 22nd international conf. on World Wide Web,* (2013), pp. 1249–1259.

Collaborative Learning Content Enrichment

Martin SVRČEK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova2, 842 16 Bratislava, Slovakia
mato.svrcek@gmail.com

Situation, in which two or more people learn or attempt to learn something together, is called collaborative learning. Currently there are number of approaches to improve collaboration. In the context of collaborative learning, the Web has become a medium in which students ask for information, evaluate one another's ideas and monitor one another's work, regardless of their physical locations. These features are very important for the collaboration and make the web more interesting. When we use these features wisely, it can be very helpful for students.

Typical examples of the use of collaboration on the web are educational systems. There are several educational systems, such as ALEF [1], which allow the use of collaborative learning. ALEF is a system with many features that support collaboration. First is tagging. An important feature of tags is that tags in a certain way describe the document. They represent keywords and they are very similar to relevant domain terms (RDT) [3]. Students can add tags to the learning content and share it with other students [2]. In addition to tags, external sources are present in the ALEF, allowing to add interesting sources of information.

Within the educational system ALEF, our goal is to enrich learning content and support students in learning. For this purpose we use annotations. When learning, students often encounter many new terms. And they want to know what these terms mean. Therefore, we need explanations, which will be linked to these terms. This is why we decided to enrich learning content using a new type of annotations – *definitions*. Definitions constitute an explanation of terms. This is the main focus of definitions because we want to support students in learning. Students can easily add and find unknown terms and understand them.

The proposed tool allows students to add definitions in two ways:

1. ALEF definition (AD) - annotation, where the source of information or explanation is learning content available in educational system ALEF. Students will be able to select some text in the learning content and assign a definition to it.
2. Own definition (OD) - annotation, where the source of information or explanation is external source somewhere on the Web. Students will be able to add their own

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 111–112.

definitions, which are useful for them. Students will be able to add their own definitions that are useful to them and could also help others.

In the context of the student learning support, our goals are 1) facilitation of student's acquisition of relevant and correct definitions of key terms being part of the course learned, 2) convenient way of learning by reducing the number of actions necessary to search for term explanations, and 3) support navigation in the course through definitions of unknown terms in a document (utilizing the dedicated widget).

In addition to student learning support, we can enlarge and enrich conceptual metadata about the educational content, e.g., relevant domain terms can be compared lexically by using terms' definitions. As a result, we can improve metadata-based services in the system such as recommendation. For example, by comparing explanations of definitions we might be able to detect synonyms as a part of automated domain modeling. A more detailed description of this aspect is out of scope of our research.

In order to evaluate our approach, we will perform several experiments in real world setting of educational system ALEF. The evaluation is planned to be performed on the course *Functional and Logic programming*. We will conduct an uncontrolled long-term experiment motivating the students to use our type of annotation. After a defined period of time, we will analyze and evaluate this data. Evaluation of the experiment will follow the abovementioned goals of our method.

First, we will determine whether the definitions represent the key concepts in the document and we will also verify whether students are able to identify the correct explanation to definition. The results of students' actions we will validate by comparing them with opinions of experts. We also want to find out how many students use the definitions.

We see great potential that our definitions bring into the educational system ALEF. They can greatly help students in order to understand the subject-matter. However, in the future, the definition can be used in many other areas that we mentioned above (other option). We believe that we can get a lot of interesting findings and contribute to the improvement of collaboration and learning in general.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 125-130.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Bieliková, M., et al. ALEF: from Application to Platform for Adaptive Collaborative Learning. In *Recommender Systems for Technology Enhanced Learning*. Springer, 2014, pp. 195–225.

[2] Móro, R, et al. Towards collaborative metadata enrichment for adaptive web-based learning. In *Proc. of the 2011 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Int. Agent Technology-Vol. 03*. IEEE CS, 2011, pp. 106–109.

[3] Harinek J., Šimko, M. Improving term extraction by utilizing user annotations. In *Proc. of the ACM symp. on Document Engineering*. ACM, 2013, pp. 185–188.

Using Parallel Web Browsing Patterns on Adaptive Web

Martin TOMA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
martin.toma.svk@gmail.com

Web browsing became one of the most common activities in professional and also private life for many of us. The interface for web browsers was arguably not changed very much from the early days of Web. However, some new features, including the possibility of parallel browsing using tabs, were introduced. Tabs became very popular and widely used feature, which is present in every modern web browser.

This feature has an unquestionable impact on how people browse Web pages today. Parallel browsing in basic (see Figure 1) occurs when Web browser user is browsing multiple Web pages using multiple browser windows or by using tabs inside these windows. Users have many reasons why they use tabs and parallel browsing in general, such as those found in Mozilla browser user study [1]:

- short-term, visual bookmark,
- parallel searching (branching from search result page), and
- opening interesting link on background without interrupting the current process.

Users therefore certainly have motivation to use the tabs as the best tool to perform parallel browsing. According to the research [2], 57% of all tab sessions involve parallel browsing and users are multitasking by splitting their browsing activity into different tabs rather than viewing more pages overall. The same research also expressed the belief that this type of behavior has been growing recently and we foresee this continuing to gain popularity and that studies of this type of navigation behavior have been lacking in Web and hypertext communities, despite parallel browsing capturing fundamental interactions with hyper-links.

That's why we have decided to focus on analyzing the parallel Web browsing behavior and identifying patterns in it. Our main goal is to utilize these patterns via some kind of appropriate recommendation aimed for the Web browser user.

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

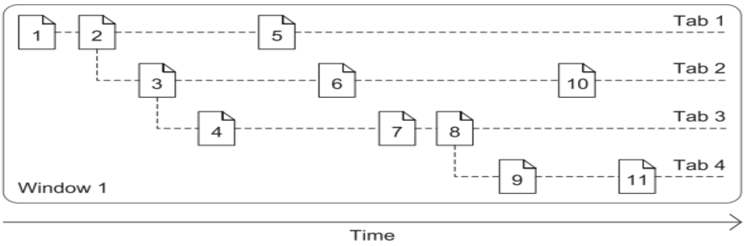


Figure 1. An example of Web browsing session with multiple tabs over time [3].

In order to detect repeating patterns in parallel Web browsing behavior we will, in the beginning, use data from Brumo project, which is aimed at browser based user modeling. The data have been already optimized for parallel browsing actions detection. After the pattern detection stage, we will analyze the recommendation possibilities. Currently we are thinking about two basic concepts of recommendation:

- Web content recommendation (within ALEF domain or on open Web)
- Web browser actions recommendation (suggesting that the user should bookmark this page, etc.)

There is also a possibility that we discover such a pattern, which will lead to other and possibly a very different kind of recommendation.

The main output of this work, web browser extension, will be capturing parallel browsing activity and also providing recommendation. We plan to implement it as a Google Chrome Web browser extension.

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

References

[1] Patrick Dubroy and Ravin Balakrishnan. 2010. A study of tabbed browsing among mozilla firefox users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 673-682.

[2] Jeff Huang and Ryen W. White. 2010. Parallel browsing behavior on the web. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia (HT '10)*. ACM, New York, NY, USA, 13-18.

[3] Jeff Huang, Thomas Lin, and Ryen W. White. 2012. No search result left behind: branching behavior with browser tabs. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*. ACM, New York, NY, USA, 203-212.

Accompanying Events

Experimentation Session Report

Róbert MÓRO, Ivan SRBA

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{robert.moro,ivan.srba}@stuba.sk

1 Motivation

Everyone, who has ever participated in research, knows, how hard it is sometimes to evaluate the proposed approach or a method, especially if the participation of human evaluators or in general participants is required. It is, therefore, one of the goals of the PeWe research group to not only provide its members with a constructive feedback during the phase of their method's or experiment's proposition, but also to help them with this (final) part of their projects.

Encouraged by the enthusiasm and activity during the experimentation session at the *Spring 2013 PeWe Workshop* [1], we decided to organize similar session also this year. The aim was to connect the experimenters with potential participants (other PeWe members) and create conditions for the experiments to be successfully carried out. This way, the experimenters could have gained preliminary results and useful feedback that will hopefully help them to prepare larger experiments and finish their theses with outstanding research results.

2 Experiments

Experimentation session during the *Spring 2014 PeWe Workshop* consisted of 8 experiments that together resulted in hours of active participation of many workshop attendees. The experiments were very diverse: the participants evaluated quality of their research interests model, learned new vocabulary, played a game, were recommended news or movies, searched for external sources enriching the learning content, linked Slovak entities with English DBpedia or solved tasks in a learning system while being monitored by an eye-tracker. The session lasted for several hours and officially ended at 2 a.m., which greatly exceeded our expectations.

Results of this experimentation session will contribute to the successful bachelor and master theses finalization. We believe that it will help the students not only finalize their work on theses, but also to prepare their main research contributions in form of research papers and submit them to the international conferences or peer reviewed scientific international journals.

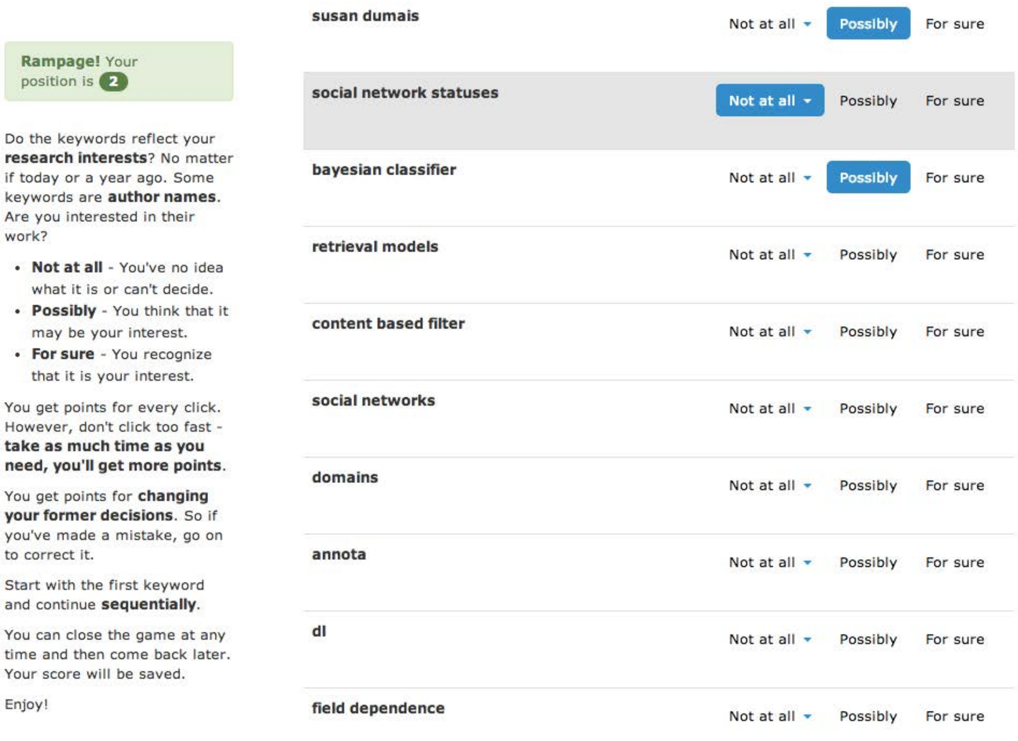
Spring 2014 PeWe Workshop, March 21, 2014, pp. 117–124.

2.1 Researcher model survey

Experimenter: Martin Lipták, Experiment supervisor: Mária Bieliková

The objective of the experiment was to evaluate a researcher model in the Annota (annota.fiit.stuba.sk) digital library by determining correlation between the researcher model terms and what the researchers themselves actually consider as their research interests. The experiment used a game with purpose to examine which terms were relevant. The game displayed 50, 100, 150, 200 and for some users even more terms most relevant to the user according to the researcher model. The task of the user was to decide if the terms were relevant. There were 3 options - Not at all, Possibly, For sure (see Figure 1). The users could come back and change their decision at any time.

The experiment had 15 participants who performed more than 4000 actions in the game (evaluations and corrections) and spent almost 3 hours playing. The correlation between the strength of the terms in the researcher model and the user evaluations was found. The terms that were strong in the model were mostly also evaluated by the users as important. The users were asked for the opinion on their researcher model after the experiment. All of them considered the first terms in the game, which were stronger in the model, also generally more relevant to their interests than the terms further in the game.



Rampage! Your position is 2

Do the keywords reflect your **research interests**? No matter if today or a year ago. Some keywords are **author names**. Are you interested in their work?

- **Not at all** - You've no idea what it is or can't decide.
- **Possibly** - You think that it may be your interest.
- **For sure** - You recognize that it is your interest.

You get points for every click. However, don't click too fast - **take as much time as you need, you'll get more points.**

You get points for **changing your former decisions**. So if you've made a mistake, go on to correct it.

Start with the first keyword and continue **sequentially**.

You can close the game at any time and then come back later. Your score will be saved.

Enjoy!

susan dumais Not at all ▾ [✓ Possibly] For sure

social network statuses [✓ Not at all ▾] Possibly For sure

bayesian classifier Not at all ▾ [✓ Possibly] For sure

retrieval models Not at all ▾ Possibly For sure

content based filter Not at all ▾ Possibly For sure

social networks Not at all ▾ Possibly For sure

domains Not at all ▾ Possibly For sure

annota Not at all ▾ Possibly For sure

dl Not at all ▾ Possibly For sure

field dependence Not at all ▾ Possibly For sure

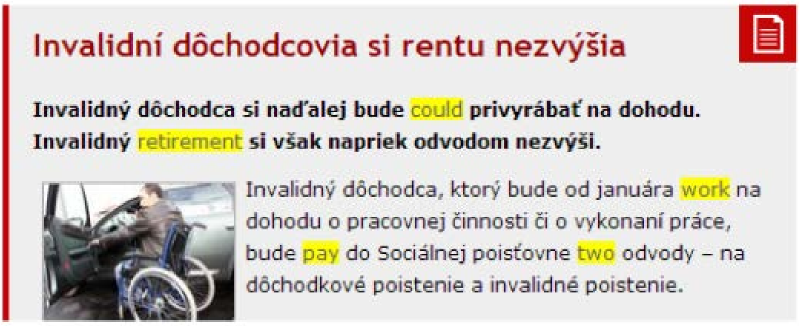
Figure 1. Researcher model evaluation interface in Annota.

2.2 User term knowledge modelling

Experimenter: Martin Gregor, Experiment supervisor: Marián Šimko

We have performed a qualitative experiment to compare two methods for user knowledge modelling using user created implicit feedback as user's read-wear indicators. The domain of an experiment was vocabulary learning on the Web. The experiment consisted of 6 participants. First half of participants used original method of user knowledge modelling and second half of participants used our proposed method of user knowledge modelling. Each participants read documents on the Web. The documents was enriched by terms in foreign language (see Figure 2). The terms could be translated to native language via click. After the experiment, we asked participants a few questions about a quality of knowledge modelling and a text enhancement.

Each user spent 20 minutes reading the documents and additional 10 minutes answering the questions. Results shows that original and proposed method have the same level of quality in text enhancement. Each half of participants read six documents on average. Each participant remembered two new words in foreign language and could translate five words on average. Each participant evaluated the quality of knowledge modelling of each method. Our method had 73% of words on average correctly modelled. Original method had 56% of words correctly modelled on average.



Invalidní dôchodcovia si rentu nezvýšia

Invalidný dôchodca si naďalej bude could privyrábať na dohodu. Invalidný retirement si však napriek odvodom nezvýši.

Invalidný dôchodca, ktorý bude od januára work na dohodu o pracovnej činnosti či o vykonaní práce, bude pay do Sociálnej poisťovne two odvody – na dôchodkové poistenie a invalidné poistenie.

Figure 2. Article at sme.sk (in Slovak) enriched by English terms.

2.3 Playing browser game Hexa

Experimenter: Peter Kiš, Experiment supervisor: Jozef Tvarožek

The main objective of our experiment was to collect enough gameplay data to examine the relationship between personality traits of players with their gaming expressions. The experiment was created to evaluate our research on bachelor project named Analysis of Interactive Problem Solving.

For this purpose we implemented a casual turn-based browser game Hexa (see Figure 3), used to track how the players perform mouse movement and clicking while playing the game and what turns they played. The game consisted of 4 different game modes, including the game against computer or multiplayer mode for two players. Every player was able to play the game anytime on their laptops during the experiment session. Each player spent playing the game between 15 to 45 minutes. After the experiment ended, we asked all participants to fill in Big Five personality questionnaire.

For evaluation process we used more than 500 logs from 21 players. Selected data was transformed into multiple gameplay indicators, for example: mouse movement speed and effectiveness, optimality of each played turn, etc. For each gameplay indicator we tried to find correlations with player personality traits. We created linear regression models that showed us for example that the indicators of mouse move speed and effectiveness are in high correlation with personality trait of conscientiousness.

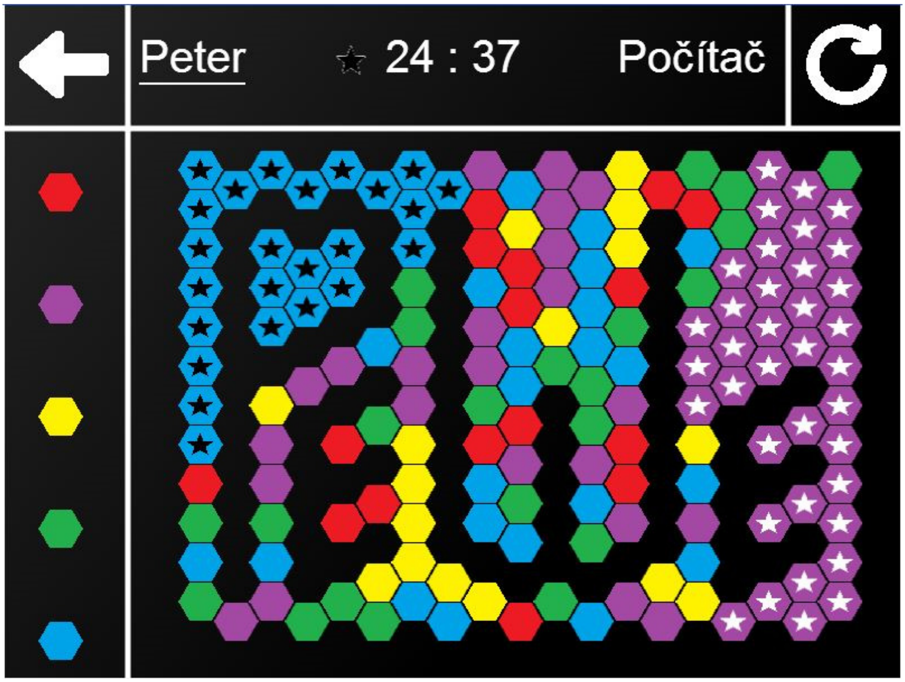


Figure 3. Level called "PeWe" played against computer.

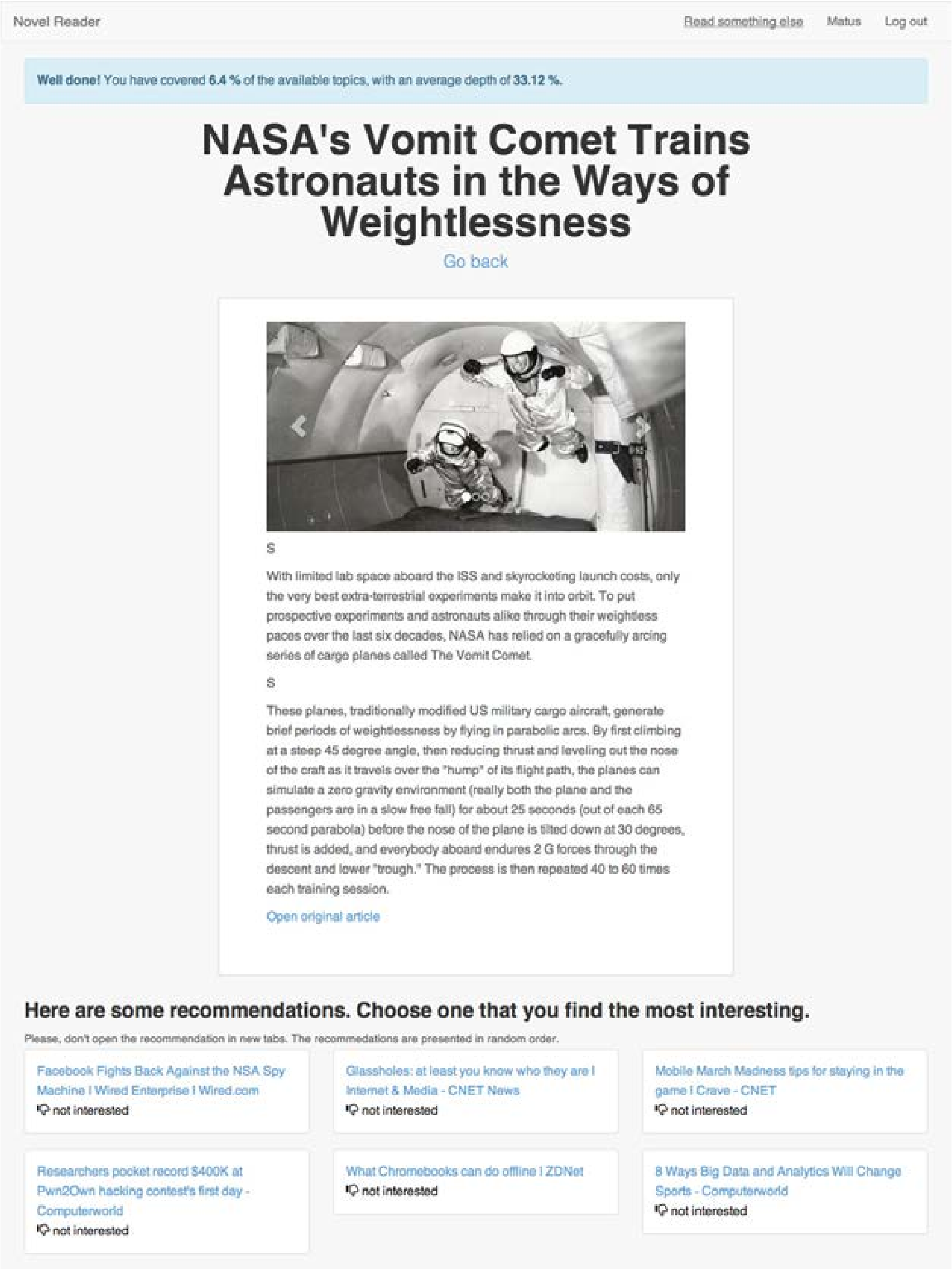
2.4 Reading and exploring news using a novelty-based recommendation

Experimenter: Matúš Tomlein, Experiment supervisor: Jozef Tvarožek

The goal of our experiment was to evaluate three content-based methods for recommendation. Two of the methods took novelty of the recommendations into account and one did not. One of the methods for novelty recommendation was ours. It used topic modelling to find out the novelty of articles based on their topics.

The articles being recommended were from several well-known tech blogs. The interface of the experiment simulated a news reading portal and showed an article that the user read with several recommendations that they could choose to read next. The participants were asked to choose one of the recommended articles that they found the most interesting.

Altogether, 23 participants took part in the experiment and they read a total of 310 articles. Each student read 13.5 articles on average with a standard deviation of 6. The data, we collected, were used to compare the methods for recommendation using statistics like precision, recall and the F1 measure and also using a ranking measure, R-score. Based on the results we concluded that recommendations based on novelty should be combined with recommendations that do not incorporate novelty to let the users choose based on their preferences.



Well done! You have covered **6.4 %** of the available topics, with an average depth of **33.12 %**.

NASA's Vomit Comet Trains Astronauts in the Ways of Weightlessness

S

With limited lab space aboard the ISS and skyrocketing launch costs, only the very best extra-terrestrial experiments make it into orbit. To put prospective experiments and astronauts alike through their weightless paces over the last six decades, NASA has relied on a gracefully arcing series of cargo planes called The Vomit Comet.

S

These planes, traditionally modified US military cargo aircraft, generate brief periods of weightlessness by flying in parabolic arcs. By first climbing at a steep 45 degree angle, then reducing thrust and leveling out the nose of the craft as it travels over the "hump" of its flight path, the planes can simulate a zero gravity environment (really both the plane and the passengers are in a slow free fall) for about 25 seconds (out of each 65 second parabola) before the nose of the plane is tilted down at 30 degrees, thrust is added, and everybody aboard endures 2 G forces through the descent and lower "trough." The process is then repeated 40 to 60 times each training session.

Here are some recommendations. Choose one that you find the most interesting.

Please, don't open the recommendation in new tabs. The recommedations are presented in random order.

Facebook Fights Back Against the NSA Spy Machine | Wired Enterprise | Wired.com
not interested

Glassholes: at least you know who they are | Internet & Media - CNET News
not interested

Mobile March Madness tips for staying in the game | Crave - CNET
not interested

Researchers pocket record $400K at Pwn2Own hacking contest's first day - Computerworld
not interested

What Chromebooks can do offline | ZDNet
not interested

8 Ways Big Data and Analytics Will Change Sports - Computerworld
not interested

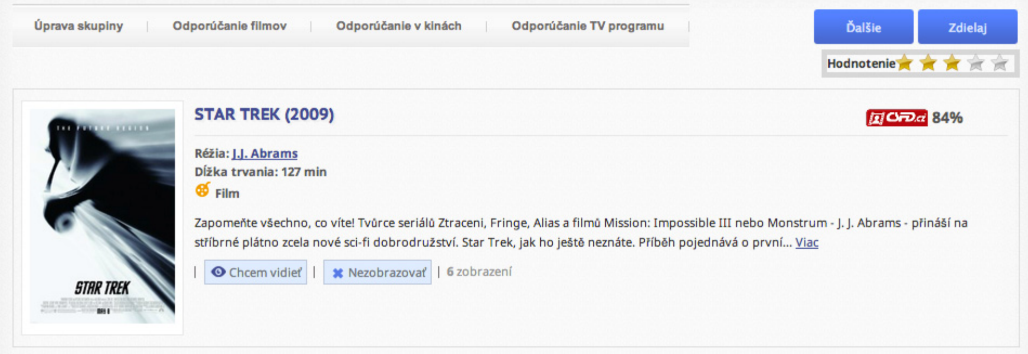
Figure 4. Experimental interface simulating a news reading portal.

2.5 Evaluating group recommendation of multimedia content

Experimenter: Eduard Fritscher, Experiment supervisor: Michal Kompan

The goal of our project was to propose a group recommendation method that will consider the personalities involved in the group. In our method we automatically generated the Big Five personality model by the use of data extracted from Facebook. Then we aggregated the personality models with the user preferences in our proposed aggregation strategy.

In our experiment, we tested our method against a widely used method in the domain of group recommendation. User formed groups and evaluated the generated recommendations. The recommendations, that they have evaluated, were generated by our method or the other in a random order. This way we obtained a control group to prove our hypothesis, which was that by the use of user personalities we can improve the group satisfaction. After we evaluated the collected data from the experiment, it showed that our proposed method performed better than the reference method.



STAR TREK (2009)

Réžia: J.J. Abrams
Dĺžka trvania: 127 min
Film

Zapomeňte všechno, co víte! Tvůrce seriálů Ztraceni, Fringe, Alias a filmů Mission: Impossible III nebo Monstrum - J. J. Abrams - přináší na stříbrné plátno zcela nové sci-fi dobrodružství. Star Trek, jak ho ještě neznáte. Příběh pojednává o první... Viac

Figure 5. Example of a group recommendation.

2.6 Monitoring web browsing behaviour in process of searching for external sources

Experimenter: Michal Račko, Experiment supervisor: Martin Labaj

We proposed a method for web content enrichment using parallel browsing. Method analyses web browser logs and extracts pairs of related web pages based on user behaviour that is the same when fulfilling same task. E.g. users follow same pattern when searching for information related to a particular web page. We can generalise this behaviour to patterns and use them for different tasks of web usage mining.

Main purpose of the experiment was to find out how users work with web browser in information retrieval tasks on the web while browsing in tabs – using parallel browsing. We found out very interesting user behaviour patterns while searching for external sources in ALEF (alef.fiit.stuba.sk) adaptive learning system in SeBe (Semantic Beer) course. All users used tabs and the majority of actions were tab switching actions. Thus, it is important to understand how users work with them.

We asked 14 participants to study information in SeBe course and find relevant external sources from open web domain. Every user screen was recorded for further analysis and all user actions were monitored using BRUMO (brumo.fiit.stuba.sk) browser extension. Average time spent fulfilling this task was 35 minutes and we gathered important data for partial method evaluation. Subsequently we applied the proposed method on gathered logs and extracted additional external sources. Results of this work were published at the student research conference IIT.SRC 2014.

2.7 Linking Slovak entities with English DBpedia

Experimenter: Ľuboš Demovič, Experiment supervisor: Michal Holub

The aim of this experiment was to verify the accuracy of linking entities in Slovak language from educational materials with English DBpedia. Every user in the experiment was able to see:

1. A random study material from any category.
2. Extracted entities and keywords that are linked to the English DBpedia.
3. To all linked entities, we showed the abstract of entities from Wikipedia to correctly identify the connections between them.

We have more than 7,000 educational materials divided in more than 40 categories. Based on the content of the study material, users determined with explicit feedback whether the connections were correct or incorrect. Our main hypothesis was that we would able to automatically link Slovak entities with DBpedia with the accuracy of at least 85%.

Experiment was attended by 11 people. Each participant performed the experiment on average 20 minutes. We tested connections to 272 different entities. Experiment results are very positive because we surpassed our hypothesised precision. The precision of the automated linking of popular search queries with DBpedia was 87.59%. Relevance of search queries to study materials was 98.14%. The proposed method appears to be promising and offers interesting results.

2.8 Implicit feedback in learning environments

Experimenter: Veronika Štrbáková, Experiment supervisor: Mária Bieliková

In our experiment, we focused on the observation of users studying in the adaptive educational system ALEF. Each participant was tasked with doing exercises from an ongoing Functional programming course. If the users did not know how to answer the questions from the exercises, they could study also the related study materials included in the course. We chose the questions according to the users' skills in the given domain, while trying to maximize the studying of the study materials. During the whole experiment, we were using eye-tracker technology. We observed the gaze of the users, the movements of the cursor, and the presses of the keys on the keyboard (see Figure 6). We recorded the behaviour of each user using the web camera with additional information recorded on paper - opening of new tabs in the browser, interesting cursor movements, and other behaviour characteristic for the given user.

The objective of our experiment was to collect the highest amount of data as possible, and afterwards, to learn to work with this data. We analysed the data using the Tobii Studio software (www.tobii.com/en/eye-tracking-research/global), with additional data coming from the ALEF system database. We connected the two sources of data and compared them with the objective of discovering interesting user behaviour during study. This will help us with our ultimate goal, i.e. to study the indicators of the implicit feedback of the users, trying to improve the existing user model inside the adaptive educational system ALEF.

Figure 6. Eye tracker experiment in progress.

3 Summary

The experimentation session turned out to be a huge success this year as well. We again witnessed enthusiasm of all the participants, many of whom attended not one, but three or four experiments. This was surely thanks to the experimenters, who came up with interesting experiments that were fun to attend, such as playing a game against friends or study beer science in a SeBe course in ALEF and search for external sources. We hope that the session helped all the experimenters to gain valuable feedback or preliminary data. We are looking forward to the next year.

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.

References

[1] Móro, R., Srba, I.: Experimentation session report. In: *Proc. of 13th Spring 2014 PeWe Workshop: Personalized Web – Science, Technologies and Engineering*. Gabčíkovo, Slovakia, 2013, pp. 113–119.

TV Log and Program Data Analysis Workshop

Jakub ŠIMKO, Michal KOMPAN

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
[jakub.simko,michal.kompan]@stuba.sk

1 Workshop aims

One of the activities undertaken at the spring 2014 edition of PeWe workshop was focused on the analysis of a data sample containing time frame of four months of TV "set-top-box" logs. The logs were acquired from one of the major cable TV providers in Slovakia. The sample was extensive: the logs themselves comprised over 10 mil. of rows (each representing a time period for which a particular user has watched a particular program), with over 10 thousands users and up to 200 channels. For more metadata (and prior to the workshop), we mapped these logs to program entries of electronic TV program guide (EPG), which we obtained from the provider as well.

The research aim of the workshop was the all-purpose analysis of the sample. The participants were asked to assess any interesting facts from the data. We have though, suggested the interest direction towards facts usable for user modelling with the potential for personalized TV program recommendation, for example analysis of viewer's interests. This was motivated by the aim of one of our applied research projects (through which we acquired the dataset), which focuses on personalized recommendation of multimedia resources from TV and archives.

As for pedagogical aims, we wanted to give the participants an opportunity to work with real (and appropriately extensive) data. We also aimed to strengthen the bonds and collaboration potential of PeWe group members, especially vertically (i.e. to mix junior and senior students) and enable them to share their experience. The participants were split into ten groups, each having at least four members of mixed seniority. The groups worked on the analysis during the morning and afternoon session of the workshop day.

2 Topics

The groups focused on a wide variety of topics and analysis types. One group commenced a basic analysis of distribution of programs and genres aired through the time period covered by the logs. Four groups analysed, how the broadcasted content affects the attraction of viewers and how the behaviour of viewers can be predicted.

Other two groups analyzed options of expanding the meta-information on the program entities by linking them to external databases. As for more distinct topics, one group analysed, whether it is effective to store programs in archive for premium users for more than few days, other designed architecture for efficient TV log data stream analysis. In the following sections, we describe the individual group contributions in detail.

2.1 Genres and their distribution across TV programs

Group leader: Karol Rástočný

Our group analyzed genres distribution in TV programs of the provider's offer during a day. The main goals have been to identify requirements of the provider's customers and to propose optimal distribution of genres. To fulfill these goals we utilized PostgreSQL as the data store, our tool written in Ruby for analyzing data and Microsoft Office Excel to evaluate and to visualize calculated results.

At the beginning we calculated percentage of genres in a real customers' viewership. This analysis gave us surprising results - 49.14% of time has been wasted by programs (genres) that had not been watched by customers. After that we sorted genres by percentage difference of watched and broadcasted time. This second analysis shows that the provider wanted to settle customers and he spend the most of time by broadcasting mostly desired genres (e.g., documents, sport or news). But the provider dedicated two times more broadcasting time to these genres than was necessary. These findings support our assumptions that it is necessary to update the provider's program portfolio to reach better coverage customers' expectations, which should decrease expenses and increase customers' satisfaction.

2.2 Relationship between number of watches and program properties

Group leader: Martin Labaj

The group focused on viewer count in relation to different programs and their properties. We used Unix text processing tools along with Ruby to pre-process the data and then followed with database tools and RapidMiner. The main challenge in tracking TV program audiences lies in the discontinuity of logs. Viewers often switch channels (e.g., due to commercials or another program starting on different channel) and the programs are seldom watched from start to end. In the pre-processing stage, we therefore aggregated log chunks for each user over programs and compared them with the program lengths.

We consider program watched by the viewer, if they watched for at least 50 % of its length, not necessarily in one continuous section. When looking at average viewer count by the release date of the program, we found that in the daytime, the most watched year is by far the 1952 – due to several classical fairy-tale movies from this year. In the prime time, the viewer count by release year was more even. The list of most watched single programs was dominated by episodes of evening news, but with watchers averaged over multiple episodes, they were outran by popular movies.

2.3 What attracts viewers to stay on a channel during switching streaks

Group leader: Marián Šimko

Digital television providers and television companies are highly interested in increasing their profit by selling more ads. In order to sell ads efficiently, it is necessary to understand interests and moods of potential viewers and offer them what they really want to view.

Among many challenges how to keep attention of a viewer belongs broadcasting a TV show that have a great potential to attract a viewer. Attraction can be viewed in two ways: 1) a viewer should not switch a TV channel he actually watches (a channel we want him to keep watching), 2) while switching, a viewer should switch to a particular TV channel (a channel we want him to switch to). Various approaches to deal with keeping attraction can be employed. In the research session, we focused on content-based approach and tried to answer a research question: "Who or what is the best channel switchers' attractor?"

By analyzing the provided dataset we identified best attractors in categories actor/actress, director, genre and year of the show. The winners were Joe Pesci, Milos Volny, parody and 1977, respectively. In addition, we also identified their counterparts - channel switchers' banishers. They were Jonathan Frakes, Cliff Bole, nature and 1987, respectively. Note that obtained results consider the specificities of Slovak TV viewers, whose viewing behavior was contained within data we used. Therefore, they cannot be generalized. Additional studies have to be conducted to reveal world-level attractors and banishers.

2.4 Efficiency of particular genre airing

Group leader: Jozef Tvarožek

In the analysis we explored the efficiency of airing a particular genre, show, movie and actors. The efficiency is determined by the total watch time of a show versus the air time of the show. As the least efficient genres during the data collection period (November 2013 – January 2014) we identified: Nature (efficiency 4.56), Parody (5.46), Sports (6.37), Biography (6.62), Music (6.98) a Documentary (9.24). On the other hand, the most effective genres were Puppet shows (124.55), Tragedy (39.29), first runs of movies (37.06), Action (36.68), Fairy tales (33.84), Comedy (32.03).

Analyzing the individual shows, we identified the most efficient shows had efficiency above 400, namely – 22 bullets (529.91), Profesionals (468.185), Little Fockers (436.04), 14th Opera Ball (430.486), New living (407.12), Olympus Has Fallen (395.61). For example, Športové noviny (daily sports news) have 264.43, same as one of the most acclaimed fairy tale in the Central European region Tri oříšky pro Popelku. In case of Slovak actors, we identified as most efficient the comedy actors Peter Marcin (324.57), Zuzana Tlučková (293.09), Marta Sládečková (283.90), while the trio Andy Kraus, Viki Ráková and Tomáš Jančo have efficiency of 189.75. In comparison, a highly popular international actor with comparable airtime (cca 10,000 minutes) was Brad Pitt with efficiency coefficient of only 11.95.

2.5 Importance of programs by modified TF-IDF

Group leader: Márius Šajgalík

In our group we developed a statistical method for calculating the importance of a program. Our method is conceptually based on TF-IDF method, which is commonly used in the field of natural language processing. It combines word frequency in a document and inverse document frequency of a word to calculate importance of a word in document. Our contribution lies in generalisation of TF-IDF method to calculate the importance of a program as a combination of frequency of watching a program and inverse frequency of broadcasting of the program.

However, as in the case of program broadcasting they are not just simple discrete events, but rather continuous broadcasting of a data stream, the audience may not watch the entire program from beginning to end. Therefore, we developed second variant of our method, which in contrast to the first one, combines durations, not frequencies. This way, we improved the accuracy of calculating the importance of a program, which eventually considered not just a frequency or duration of watching a program by audience, but also a program *broadcastedness*, which in the first place ever gives to the audience an opportunity of watching a program and thus, significantly influences the overall viewership.

2.6 Predicting best broadcasting time and best target audience

Group leader: Michal Holub

The group's goal in this workshop was to determine the target audience and the best broadcasting time for a given program. We developed a set of analytical tools using Ruby on Rails and PostgreSQL database. These analytical tools can be used to predict which family is the best target audience for a given program, i.e. which family has the highest probability of watching the entire program. Moreover, we can predict the best time of the day in which the program should be aired.

In order to predict the appropriate target audience we took all data from set-top-boxes for each family and compared them with the airing times of the shows. For each attribute of the shows (category, length, director, etc.) we can compute the list of families with the highest interest for this particular attribute. We also compared the interest of families in different parts of the day (morning, afternoon, evening, night) in various types of shows. From the results we can see that e.g. the programs about nature attract more attention in the evening than in the morning. Our analytical tools can help to predict the interest of families and target airing of the programs accordingly.

2.7 Genre similarity and ranking of TV channels based on IMDb ratings of broadcasting movies

Group leader: Eduard Kuric

The fact is - despite the boom in smart mobile devices and the web services for renting, buying and watching movies - the TV is not dead, the TV is here to stay. Its role in delivering compelling viewing experiences-collective and individual-continue. The TV will develop as an even more valuable vehicle for entertainment and, increasingly, for education and information. Instruments integrated in TV such as recommendation engines will control users‘ TV viewing behavior and their satisfaction. They will be instrumental in attracting new service subscribers.

In our group, we analyzed similarity of TV channels based on broadcasting genres. For each TV channel and day of week, we compiled a vector of terms which contains genres of broadcasted programs. Based on the cosine distance the result is a matrix of similarities for each channel. Next, we focused on ranking TV channels based on IMDb ratings of broadcasting movies. For each movie we retrieved its rating from IMDb movie database. The result is a list of TV channels ordered by the amount of the best broadcasted movies.

2.8 Mapping programs to external resources

Group leader: Ivan Srba

In our working group, we focused on two different aspects of the dataset: a metadata enrichment with publicly available information from the website csfd.cz; and a comparison of user behavior on different platforms (STB and mobile devices). At first, we prepared a scraper for a popular community portal csfd.cz which contains user reviews about all kind of programs. We decided to employ this source of programs' metadata because it covers especially national programs (e.g. local news) which represent a significant part of broadcast in the most watched channels covered by the dataset. The prepared scraper searched for a particular program, and consequently extracted its metadata, such as a genre, a country of production, a year of production, an aggregated user evaluation or even a position in the rank lists (there are several ranks, e.g. the best/worst movies). This extension of the dataset allowed us to analyze interesting aspects of the channels' program. To mention a few of them: correlation between quality of programs and their popularity among TV viewers, or distribution of programs according to the countries of their origin.

Secondly, we analyzed how users' behavior differs on various platforms, more particularly on STB and mobile devices. We found out that there are not significant differences in the popularity of TV channels. On the other side, we discovered promising patterns in the popularity of particular programs. We can utilize these results in personalized recommendations according to the type of platform used by a viewer.

2.9 Efficiency of program archiving

Group leader: Róbert Móro

Currently, the TV provider provides its customers with a free access to the TV archive storing the programming for several weeks. We were interested, how the service is actually used by the customers, i.e. how often they watch programs from the archive and how far back they go. Our goal was to find out, if it is necessary to store the whole month or if the storage space could be saved without limiting the customers.

In order to analyze the data, we used combination of various tools – R and SQL for pre-processing, RapidMiner and Excel for statistics and visualization. From all the records in the dataset (about 7 million), only small fraction (about 37,000) accounts for the archive usage data. We grouped the records representing the same programs watched by the same customers. Then we chose only the last record from each group in order to see time offsets between the original air time and the time when the customer finished watching the program. We found out that 70% of all the customers finish watching archived programs within a week after its original air time and 82% within two weeks. However, we discovered also some anomalies, such as programs that were watched over a period of two months, which (to our knowledge) should not be possible.

2.10 Architecture for TV log stream data processing

Group leader: Jakub Ševcech

The majority of data analysis tasks is performed using batch processing on predefined set of data. This approach, however, provides information from the data delayed at least by the time of processing of the batch of data. In our group, we focused not directly on the analysis of provided data, but on techniques to analyze and visualize the data in real time, in the time the data is produced. To transform provided data, we created simple Ruby script, and we pushed the data into Kafka message queue.

In the next step, we designed and implemented a topology of workers (bolts) in Storm framework for parallel stream processing. We designed the topology to continuously count number of spectators per television channel, but more complicated topologies can be implemented in like manner. The results of this aggregation were saved to a fast key-value store Redis. Lastly we used a simple Ruby on Rails application and D3 JavaScript library to read and visualize the aggregated data.

In the experiment, we developed a simple configuration for streamed data processing that can be easily extended for other, more complicated, analysis and can find its applications in various analytics tasks or big data processing.

SeBe 10.0: The Birth of the Academy

Marián ŠIMKO, Róbert MÓRO, Jakub ŠIMKO, Michal BARLA

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
`{name.surname}@stuba.sk`

For the past five years we have organized SeBe workshops to establish beer driven research (BDR) as a standalone area of study and to provide researchers as well as practitioners with a common ground for exchange of their knowledge and experience and for discussion of current advances and future trends in the field.

Since the first edition of SeBe workshop we aimed to cover wide variety of beer- and research-relevant topics. We stimulated the emergence of Beer Science to be the tool in our hands and minds, the tool of further research [1, 2]. We gradually encouraged Beernovation [3], we tracked the impact of beer on human senses (taste [4], visual perception [5]). We also tackled the industry-related issues of beer [6] and we covered biological [7], artistic [8] and philosophical [9] dimensions of beer.

We have observed a gradual shift from heavy-weight to more light-weight beer ontologies, rising popularity of BEL (Beer-enhanced learning) and MOOBs (Massively overdue open bars), or tackling of problems associated with beer-processing and sometimes even with data sparsity (i.e., empty tankards).

In order to answer demands of praxis and ever-growing BDR community we are happy to announce establishment of a new academic institution – Academia SeBeana, building on the tradition and strong standards set by the previous SeBe workshop editions. Our ambition is to teach our prospective students all the knowledge and skills necessary for them to be successful in this highly competitive beer market as well as prepare them for their future research careers.

Academia SeBeana finished its accreditation process at March 21st, 2014 in at accreditation ceremony in Gabčíkovo, Slovakia. The accreditation was granted by Professor Mária Bieliková, Accreditation Committee member (see Appendix A).

In the academic year 2013/2014 we open study programmes in all three levels of study in the field of Beergineering: undergraduate programme awarding the degree of Bachelor of Beergineering (BBeng.), and graduate programmes awarding the degree of Master of Beergineering (MBeng.) and Cerevisiae Doctor (CeD.).

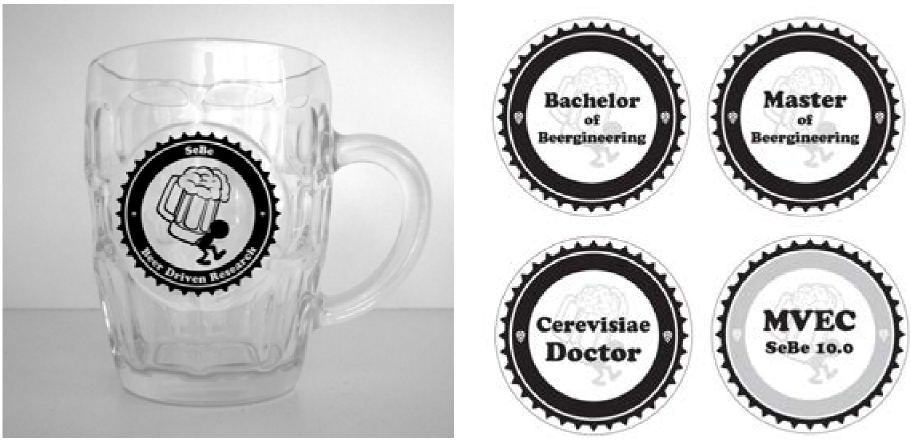
Our graduates:

- will gain a full university education in the given level of study focusing on beergineering processes, beer savouring, malt fermentation and beer-tapping techniques,

- will have knowledge in the field of beergineering in its broader social context and will understand interdisciplinary connections (Beer Social Science, Beer Psychology, etc.),
- will be capable of finding new high-quality beer brands and present their own solutions concerning the Beer driven research,
- will be aware of social, moral and economic aspects of Beer-Driven Research,
- will find employment as a member of creative team or its leader in different areas of industry and academia, where there are possibilities of development, deployment and maintenance of beer distribution (and consumption) systems.

33 students have enrolled in the first semester of Academia SeBeana. Each student obtained study index to have a vivid evidence of his/her study achievements. The students had the opportunity to access 10 educational activities aimed on practical aspects: experimentation and evaluation.

In order to earn a Bachelor of Beergineering degree, students had to collect 4 credits and pass the state exam, which was focused on examination of his/her knowledge of his study application bottle. To earn a Master of Beergineering degree, students had to collect 8 credits and pass the state exam, which – for this degree of study – covered broader discussion from the core of the field. Finally, to earn Cerevisiae Doctor degree, a student had to satisfy the strictest criteria: he/she had to have at least two publication in category ‘B’ and be awarded at least one prize for his submissions in scientific fora. He/she also needed to pass a state exam, which includes demonstration of exceptional and defense of the original contribution in the field.



(a) (b)

Figure 1. a) study index to keep evidence of study achievements with new SeBe logo, b) study achievement examples for SeBe 10.0 (MVEC stands for Most Valuable Experiment Contributor)

We are pleased to announce that already after the first beermester, 13 students graduated with degree Bachelor of Engineering and 4 students graduated with degree Master of Engineering (see Appendix B). We also paid tribute to our long-time brave supporter: Mária Bieliková earned Cerevisiae Doctor Honoris Causa degree in recognition of her outstanding pioneering and innovative academic work in Beergineering.

The Most Valuable Experiment Contributor prize was earned by MBeng. Ondrej Kaššák for his exemplary contribution to experiment evaluations. Ondrej managed to help the most experimentators and collected the most credits. Outstanding!

After the success of the first beermester, Academia SeBeana will continue to offer educational services in the precious and terrifically potential field of Beer-Driven Research.

Extended version of this paper is published in every bottle and every keg. It is born with each young barley sprout; it grows in every hop flower, within the each cell of yeast. It is present in the heart of each fellow SeBe researcher.

Acknowledgement. This work was partially supported by the SeBe secret sponsors and the SeBe coin initiative.

References

[1] Šimko, M. Barla, M. SeBe: Merging Research with Practice. November 2009, Modra, Harmónia, Slovakia (2009)

[2] Šimko, M. Barla, M. SeBe 2.0. The Emergence of Beer Science. April 2010, Smolenice, Slovakia (2010)

[3] Šimko, M. Barla, M., Šimko, J. Zeleník, D. SeBe 3.0: Means of Beernovation. November 2010, Modra, Harmónia, Slovakia (2010)

[4] 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, pp. 91–92 (2011)

[5] Šimko, M. Šimko, J., Barla, M. SeBe 5.0: Mug-Centered Design. October 2011, Modra, Slovakia (2011)

[6] 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, pp. 95–96 (2012)

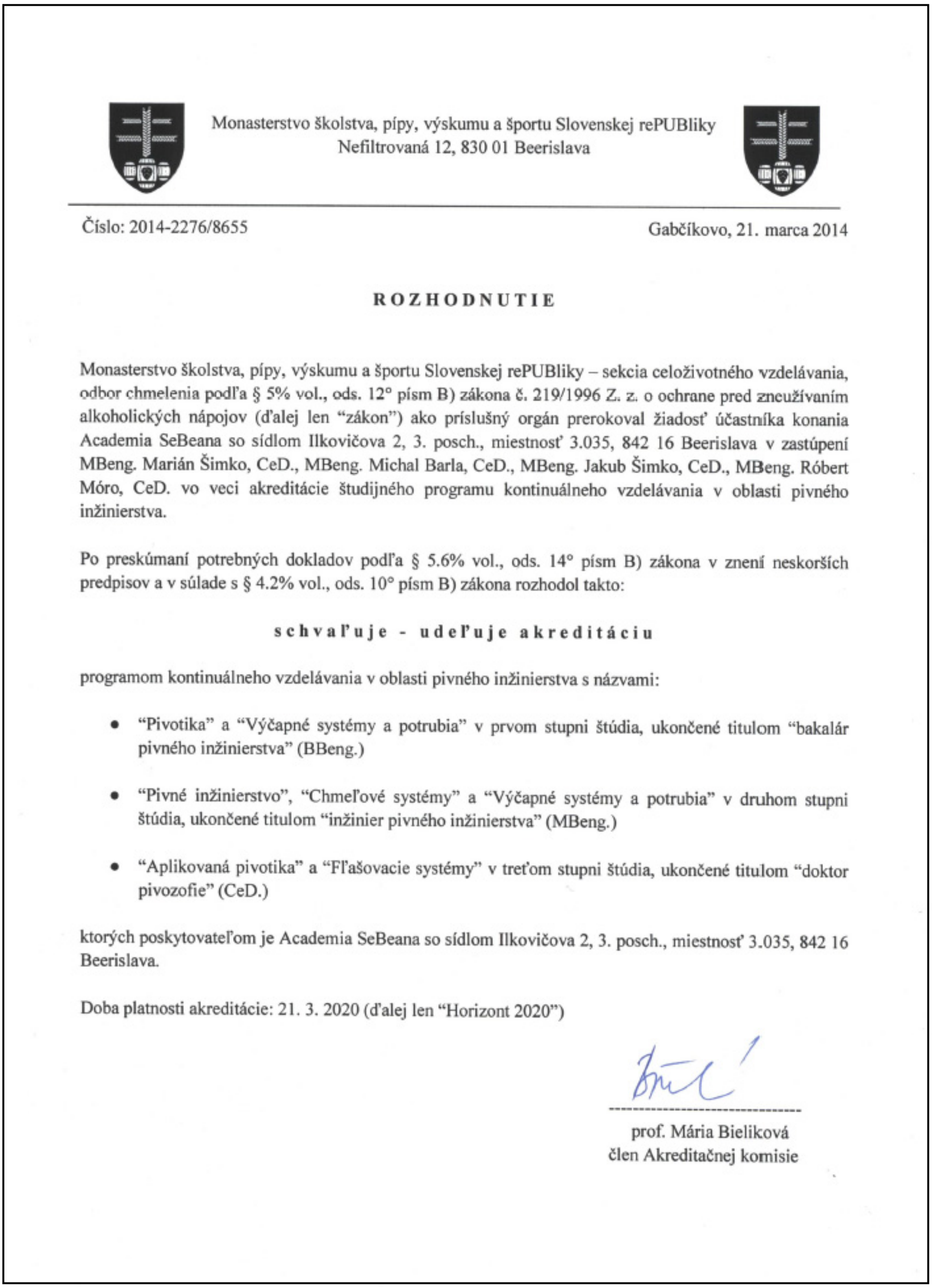
[7] Šimko, M. Šimko, J., Barla, M. SeBe 7.0: Healthy Body, Healthy Mind, Healthy Research. October 2012. Modra-Piesok, Slovakia (2012)

[8] Šimko, M. Šimko, J., Móro, R., Barla, M. SeBe 8.0: Beer – Unity of Science and Art. In Proc. of 13th Spring 2013 PeWe Workshop: Personalized Web – Science, Technologies and Engineering. Modra-Piesok, Slovakia, pp. 125–126 (2013)

[9] Šimko, M. Šimko, J. Móro, R., Barla, M. SeBe 9.0: Malting Pot of Knowledge. April 2013. Gabčíkovo, Slovakia (2013)

Appendix A: Accreditation file

Accreditation file grants accreditation for several study programmes within the study degree Beergineering (Accreditation file in Slovak).



Monasterstvo školstva, pípy, výskumu a športu Slovenskej rePUBliky
Nefiltrovaná 12, 830 01 Beerislava

Číslo: 2014-2276/8655

Gabčíkovo, 21. marca 2014

ROZHODNUTIE

Monasterstvo školstva, pípy, výskumu a športu Slovenskej rePUBliky – sekcia celoživotného vzdelávania, odbor chmelenia podľa § 5% vol., ods. 12° písm B) zákona č. 219/1996 Z. z. o ochrane pred zneužívaním alkoholických nápojov (ďalej len "zákon") ako príslušný orgán prerokoval žiadosť účastníka konania Academia SeBeana so sídlom Ilkovičova 2, 3. posch., miestnosť 3.035, 842 16 Beerislava v zastúpení MBeng. Marián Šimko, CeD., MBeng. Michal Barla, CeD., MBeng. Jakub Šimko, CeD., MBeng. Róbert Móro, CeD. vo veci akreditácie študijného programu kontinuálneho vzdelávania v oblasti pivného inžinierstva.

Po preskúmaní potrebných dokladov podľa § 5.6% vol., ods. 14° písm B) zákona v znení neskorších predpisov a v súlade s § 4.2% vol., ods. 10° písm B) zákona rozhodol takto:

schvaľuje - udeľuje akreditáciu

programom kontinuálneho vzdelávania v oblasti pivného inžinierstva s názvami:

- "Pivotika" a "Výčapné systémy a potrubia" v prvom stupni štúdia, ukončené titulom "bakalár pivného inžinierstva" (BBeng.)
- "Pivné inžinierstvo", "Chmeľové systémy" a "Výčapné systémy a potrubia" v druhom stupni štúdia, ukončené titulom "inžinier pivného inžinierstva" (MBeng.)
- "Aplikovaná pivotika" a "Fľašovacie systémy" v treťom stupni štúdia, ukončené titulom "doktor pivozofie" (CeD.)

ktorých poskytovateľom je Academia SeBeana so sídlom Ilkovičova 2, 3. posch., miestnosť 3.035, 842 16 Beerislava.

Doba platnosti akreditácie: 21. 3. 2020 (ďalej len "Horizont 2020")

prof. Mária Bieliková
člen Akreditačnej komisie

Appendix B: List of Graduates

STUDY PROGRAMME: BEERGENEERING

Master's degree
(Study programmes: Beergeneering, Hop systems, Draft systems and pipes)

MBeng. Márius Šajgalík
MBeng. Jakub Ševcech
MBeng. Miroslav Šimek
MBeng. Ondrej Kaššák

Bachelor's degree
(Study programmes: Beerformatics, Draft systems and pipes)

BBeng. Márius Šajgalík
BBeng. Jakub Ševcech
BBeng. Karol Rástočný
BBeng. Marek Grznár
BBeng. Miroslav Šimek
BBeng. Michal Računo
BBeng. Ondrej Kaššák
BBeng. Viktória Lovasová
BBeng. Jakub Mačina
BBeng. Dominika Červeňová
BBeng. Peter Dubec
BBeng. Peter Kiš
BBeng. Jozef Harinek

Index

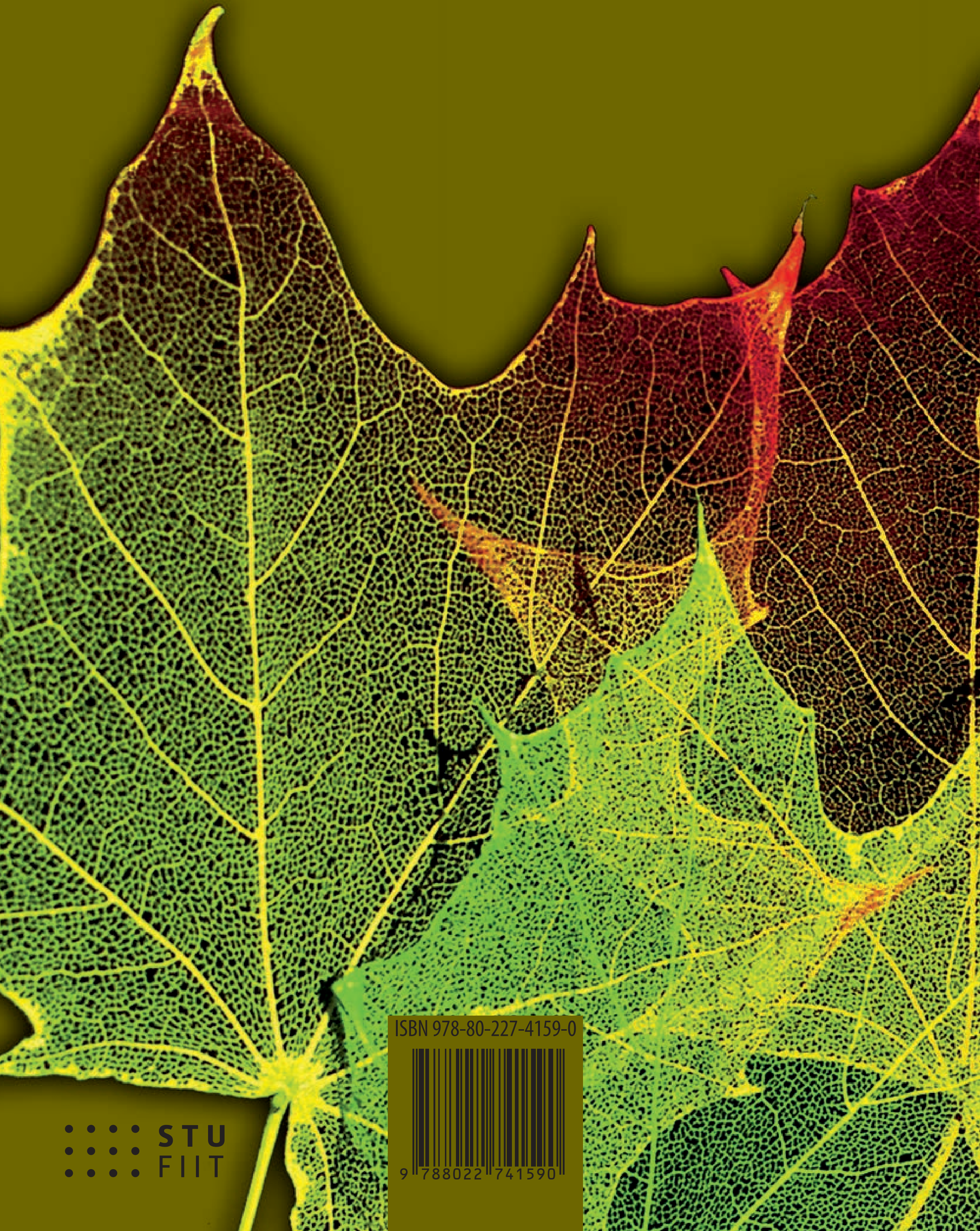
Mária Bieliková, Pavol Návrat, Michal Barla,
Michal Kompan, Jakub Šimko, Marián Šimko, Jozef Tvarožek (Eds.)

Personalized Web – Science, Technologies and Engineering
15th Spring 2014 PeWe Workshop

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

154 pages, Available at
http://pewe.fiit.stuba.sk/ontozur2014spring-proceedings/
Print Nakladateľstvo STU Bratislava
2014

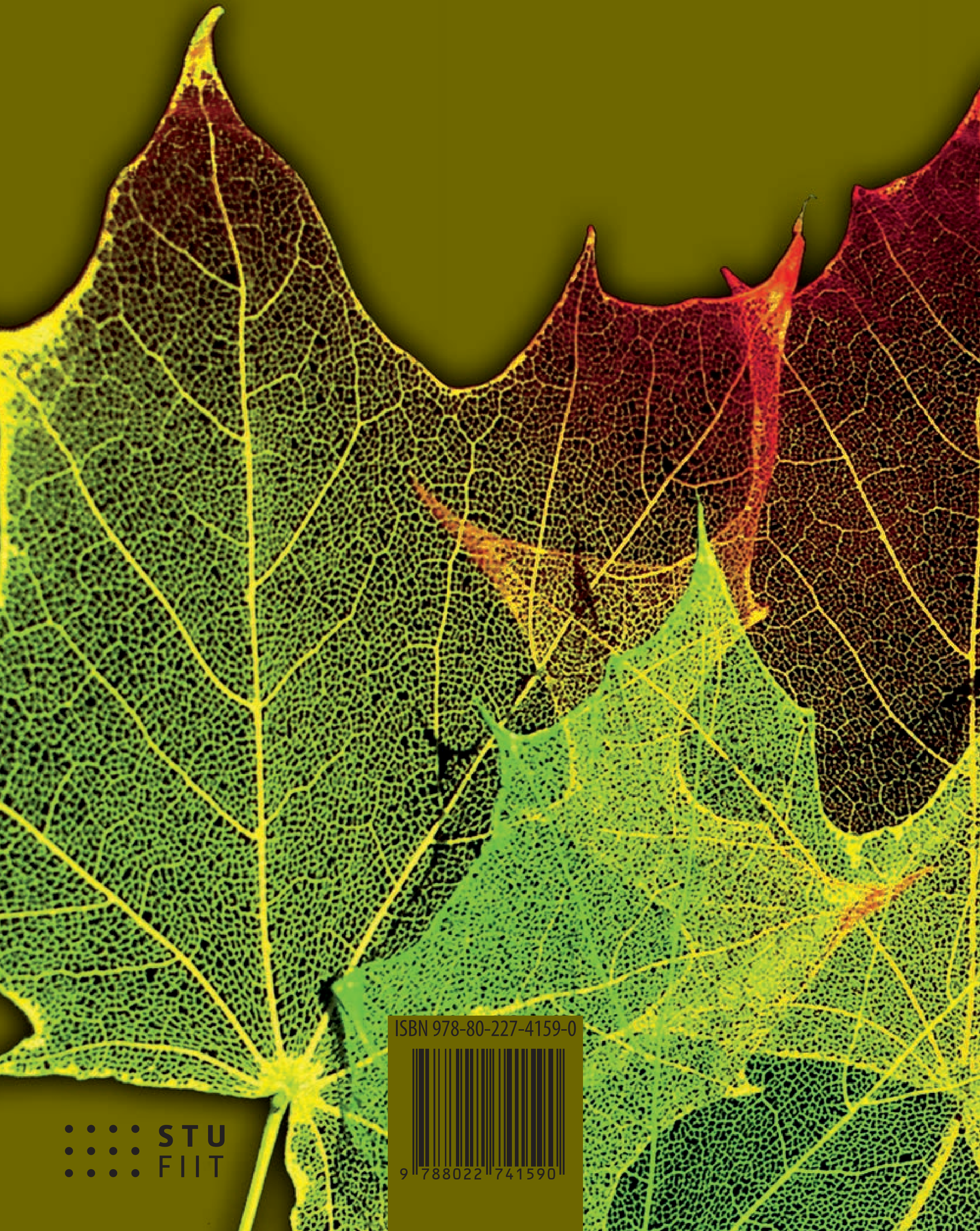
ISBN 978-80-227-4159-0



ISBN 978-80-227-4159-0

9 788022 741590

STU
FIIT



ISBN 978-80-227-4159-0

9 788022 741590

STU
FIIT