

Workshop on the Web - Science, Technologies and Engineering

7th Spring 2010 PeWe Ontožúr,
Smolenice Castle, Slovakia, April 2010
Proceedings

Mária Bieliková, Pavol Návrat (Eds.)

Keynote by Bebo White



Proceedings in
Informatics and Information Technologies

**Workshop on the Web – Science,
Technologies and Engineering**
7th Spring 2010 PeWe Ontožúr

Mária Bielíková, Pavol Návrát (Eds.)

Workshop on the Web – Science, Technologies and Engineering

7th Spring 2010 PeWe Ontožúr
Smolenice Castle, Slovakia
April 18, 2010
Proceedings



Slovakia Chapter



PeWe Group

S T U • •
• • • • •
F I I T •
• • • • •

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

Proceedings in
Informatics and Information Technologies

Workshop on the Web – Science, Technologies and Engineering
7th Spring 2010 PeWe Ontožúr

Editors

Mária Bieliková and Pavol Návrat

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

© 2010, The authors mentioned in the Table of Contents

Contributions are printed as delivered by authors
without substantial modifications

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

Executive Editor: Mária Bieliková

Copy Editor: Katarína Mršková

Cover Designer: Peter Kaminský

Published by:

Nakladateľstvo STU

Vazovova 5, Bratislava, Slovakia

ISBN 978-80-227-3274-1

Preface

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

Research in the field of the Web has more than 10 years of tradition at the Slovak University of Technology in Bratislava. Moreover, topics related to the Web attract many students. This volume is entirely devoted to students and their research. It contains extended abstracts of students' research projects presented at the 7th PeWe (Personalized Web Group) Workshop on the Web – Science, Technologies and Engineering, held on April 18, 2010 in beautiful Smolenice Castle, Slovakia. It was organized by the Slovak University of Technology (and, in particular, its Faculty of Informatics and Information Technologies, Institute of Informatics and Software Engineering) in Bratislava. Participants are students of all three levels of the study – bachelor (Bc.), master (Ing.) or doctoral (PhD.), and their professors.

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

- Search, Navigation and Visualization,
- Classification and Recommendation,
- User Modeling, Virtual Communities and Social Networks,
- Domain Modeling, Semantics Discovery and Annotations,
- Web Engineering, Semantic Web Services.

The workshop was launched by Bebo White's keynote on the hot topic of cloud computing. Bebo White is a Departmental Associate (Emeritus) at the SLAC National Accelerator Laboratory at Stanford University. Working as a computational physicist, he first became involved with the emerging World Wide Web technology while on sabbatical at CERN in 1989. Upon his return he was a part of the team that established the first American Web site at SLAC (the fifth site in the world). Ever since, his academic and research interests have evolved in parallel with Web technology.

The keynote was followed by student presentations of their research. The projects were at different levels according to the study level (bachelor, master or doctoral) and also according the progress stage achieved in each particular project. Moreover, we invited to take part also four of our bachelor study students who take an advantage of

our research track offered within their study programme and who just start their bachelor projects – *Pavol Bielik, Štefan Mitrik, Jakub Ševcech and Michal Tomlein.*

Bachelor projects:

- *Anton Benčíč, Roman Mészároš, Roman Panenka, Márius Šajgalík:* Geo-Based Social Network Navigation
- *Peter Bugáň:* Enhancing the Web Experience by Freely Available Metadata
- *Milan Lučanský:* Does HTML Tags Improve Results of ATR Algorithms?
- *Tomáš Michálek:* Collaborative Tagging for Word Relationships Mining
- *Jana Pazúriková:* Improving Social Skills Using the Social Exchange Framework
- *Ivan Srba:* Tracing Strength of Relationships in Social Networks
- *Maroš Unčík:* Collaborative Acquisition and Evaluation of Question by Learners

Master projects (started in the current academic year):

- *Marián Hönsch:* Virtual Community Detection in Vast Information Spaces
- *Martin Jačala:* Text Understanding and Analysis
- *Eduard Kuric:* Interactive Photo Retrieval Based on Semi-Automatic Annotation Using Visual Content and Folksonomies
- *Martin Labaj:* Recommendation and Collaboration through Implicit Identification in Social Context
- *Michal Lohnický:* Spatial and Time Navigation in Multimedia
- *Vladimír Mihál:* Exploring the Possibilities of Annotations in Learning Content
- *Karol Rástočný:* Browsing Similar or Related Data Entities by Breadth-First Search in the Semantic Web
- *Štefan Sabo:* Online Gathering of Information from Text Sources
- *Matej Valčuha:* Information Search Considering the User's Interest and Groups of Similar Users
- *Martin Virík:* Automated Recognition of Author's Writing Style in Blogs

Master projects (started in the previous academic year):

- *Michal Holub:* Website Navigation Adaptation Based on Behavior of Users
- *Tomáš Kramár:* Leveraging Social Networks in Navigation Recommendation
- *Michal Kompan:* Personalized Recommendation of Interesting Texts
- *Ladislav Martinský:* Improving Query Suggestion Capabilities Using Web Search Results
- *Pavel Michlík:* Personalized Exercises Recommending for Limited Time Learning
- *Jakub Šimko:* Enhancing Exploratory Search: Graphs, User Modeling and Search History
- *Dušan Zeleník:* Effective Representation for Content-Based News Recommendation

Doctoral projects

- *Michal Barla:* Towards Social-based User Modeling

- *Peter Bartalos*: QoS Aware Semantic Web Service Composition Approach Considering Pre/Postconditions
- *Pavol Mederly*: Towards Semi-automated Design of Enterprise Integration Solutions
- *Marián Šimko*: Lightweight Semantic Search Based on Heterogeneous Sources of Information
- *Ján Suchal*: Improving Search Using Graphs and Implicit Feedback
- *Jozef Tvarožek*: Cooking a Socially Intelligent Tutoring Platform
- *Michal Tvarožek*: Exploratory Search in the Adaptive Social Semantic Web

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

Finally we highly appreciate the financial support of our sponsor – The Foundation of Tatrabanka for support of publishing these proceedings and for support of the whole event.

April 2010

Mária Bielíková
Pavol Návrát

Table of Contents

Keynote

Clearing Away the Clouds: What is the Future of Cloud Computing? <i>Bebo White</i>	3
---	---

Students' Research Works

Search, Navigation and Visualization

Interactive Photo Retrieval Based on Semi-Automatic Annotation Using Visual Content and Folksonomies <i>Eduard Kuric</i>	7
Spatial and Time Navigation in Multimedia <i>Michal Lohnický</i>	9
Improving Query Suggestion Capabilities Using Web Search Results <i>Ladislav Martinský</i>	11
Geo-Based Social Network Navigation <i>Anton Benčíč, Roman Mészáros, Roman Panenka, Márius Šajgalík</i>	13
Browsing Similar or Related Data Entities by Breadth-First Search in the Semantic Web <i>Karol Rástočný</i>	15
Enhancing Exploratory Search: Graphs, User Modeling and Search History <i>Jakub Šimko</i>	17
Lightweight Semantic Search Based on Heterogeneous Sources of Information <i>Marián Šimko</i>	19
Exploratory Search in the Adaptive Social Semantic Web <i>Michal Tvarožek</i>	21
Information Search Considering the User's Interest and Groups of Similar Users <i>Matej Valčuha</i>	23

Classification and Recommendation

Website Navigation Adaptation Based on Behavior of Users <i>Michal Holub</i>	27
---	----

Personalized Recommendation of Interesting Texts <i>Michal Kompan</i>	29
Recommendation and Collaboration through Implicit Identification in Social Context <i>Martin Labaj</i>	31
Personalized Exercises Recommending for Limited Time Learning <i>Pavel Michlík</i>	33
Improving Search Using Graphs and Implicit Feedback <i>Ján Suchal</i>	35
Automated Recognition of Author's Writing Style in Blogs <i>Martin Virik</i>	37
Effective Representation for Content-Based News Recommendation <i>Dušan Zeleník</i>	39
User Modeling, Virtual Communities and Social Networks	
Towards Social-based User Modeling <i>Michal Barla</i>	43
Virtual Community Detection in Vast Information Spaces <i>Marián Hönsch</i>	45
Leveraging Social Networks in Navigation Recommendation <i>Tomáš Kramár</i>	47
Improving Social Skills Using the Social Exchange Framework <i>Jana Pazúriková</i>	49
Tracing Strength of Relationships in Social Networks <i>Ivan Srba</i>	51
Cooking a Socially Intelligent Tutoring Platform <i>Jozef Tvarožek</i>	53
Domain Modeling, Semantics Discovery and Annotations	
Enhancing the Web Experience by Freely Available Metadata <i>Peter Bugáň</i>	57
Text Understanding and Analysis <i>Martin Jačala</i>	59
Does HTML Tags Improve Results of ATR Algorithms? <i>Milan Lučanský</i>	61
Collaborative Tagging for Word Relationships Mining <i>Tomáš Michálek</i>	63
Exploring the Possibilities of Annotations in Learning Content <i>Vladimír Mihál</i>	65
Online Gathering of Information from Text Sources <i>Štefan Sabo</i>	67
Collaborative Acquisition and Evaluation of Question by Learners <i>Maroš Unčík</i>	69

Web Engineering, Semantic Web Services

QoS Aware Semantic Web Service Composition Approach Considering Pre/Postconditions <i>Peter Bartalos</i>	73
Towards Semi-automated Design of Enterprise Integration Solutions <i>Pavol Mederly</i>	75

Index	77
--------------------	-----------

Workshop participants Smolenice Castle

Thanks for a wonderful and stimulating time in Sydney

[illegible]



Clearing Away the Clouds: What is the Future of Cloud Computing?

Bebo WHITE

SLAC National Accelerator Laboratory
Stanford University
2575 Sand Hill Road, Menlo Park CA 94025 USA
bebo@slac.stanford.edu

In all likelihood, anyone who is a regular (or perhaps even casual) reader of the information technology press has experienced some of the hype surrounding *Cloud Computing*. It is sometimes difficult to distinguish whether the term is being used as simply a *buzzword* or *marketing term* or as a part of a genuine technical description.

Confusion or not, *Cloud Computing* is here and has become mainstream. In September 2008, the Pew Research Center reported, “69% of all Internet users have either stored data online or used a Web-based software application.” Gmail has millions of users and Flickr, with its millions of photographs, allows photo sharing at a scale never before possible or imagined. Whether they know it or not, users of these applications are taking advantage of *Cloud*-based computing and memory rather than resources available on their local devices. Such usage has been largely driven by *Web 2.0* and mobile-based applications.

In order to make sense of and participate in a meaningful discussion of *Cloud Computing*, a robust and unambiguous definition of the technology is needed. In general,

- *Cloud Computing is*
 - When computing services are provided over the Internet rather than locally on a user machine;
 - Computation is run on an supporting infrastructure which is independent of the applications themselves;
 - Cloud Computing infrastructure can take on many forms, but to the end user, the implementation is irrelevant, hence the “cloud” abstraction.
- *Cloud Computing is not*
 - Necessarily inclusive of Grid Computing, Utility Computing, or self-managing (local) computing;
 - Necessarily limited to Software as a Service (SaaS) or generalized network access to data.

- *Cloud Computing* environments can have any/all combinations of Distributed Computing elements.

There can be little doubt that *Cloud* technology will be an important part of computing at all levels in the future. Even so, the technology appears to be struggling to define its future. Does it really mean the end of large-scale industrial computer centers of the past? Is it really the first step towards a global computing infrastructure providing IT resources as a utility like electricity, water, telecom, etc.? How might individuals continue to benefit from *Cloud*-based systems?

The future of *Cloud Computing* does not depend solely on technological advancements. Its success will also depend on numerous social, political, and economic factors. For example, how can the Cloud remain open and non-proprietary but still allow monetization? What is the intersection between Cloud systems and social networks? How might *Cloud Computing* “show stoppers” such as data security, privacy, and intellectual property issues be addressed?

This talk is not going to provide answers to these complex questions. Instead, it will look beyond the hype and try to establish a baseline from which attendees might make intelligent decisions surrounding the technology and perhaps play a role in determining its future.

Search, Navigation and Visualization

Interactive Photo Retrieval Based on Semi-Automatic Annotation Using Visual Content and Folksonomies

Eduard KURIC*

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

Nowadays web photo management systems provide features for users, such as organizing, sharing and searching photos. With increasing popularization of digital and mobile phone cameras, there occurs a need of quick and exact searching. Content based indexing of photos is more difficult than text documents because the photos do not contain units like words. Searching is based on annotations and semantic keywords that are entered by a user and associated with photos. However, manual creating of annotations is very time-consuming and results are often subjective. Therefore, photo semi-automatic annotation is most challenging task.

Traditional approaches for semi-automatic annotation are based on combining keyword-based and content-based photo retrieval [3]. The user enters a query consisting of a target photo and keywords, typically only a caption. The aim is to find most similar photos and to extract related keywords. After a retrieval process, the user selects the best relevant keywords and associates them with the target photo. The process usually takes place in three steps. First, a keyword-based technique is used to obtain a list of candidate photos that are also associated with the input caption. Second, content-based photo retrieval technique is used to assemble a ranked list of visually related photos. Finally, a method is used to combine the ranked list into an annotation list which represents keyword proposals. These solutions employ global low-level features like color and texture for a content comparison of photos. However, the user query can include a full photo or a just part of the whole photo which we call object-of-interest.

In our work, we propose a novel method for annotating photos which extends existing solutions of searching similar photos primary according to objects-of-interest [2]. Often, those objects represent a foreground of a photo that is in comparison with a background less dominant. Therefore, in traditional approaches of content-based photo

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

retrieval foregrounds can be ignored (low-rated) despite of the fact that can represent most important elements of the photo. We use an interactive photo segmentation to determine objects-of-interest. To capture local photo information that is in object retrieval essential, we use scale invariant feature transform (SIFT) in combination with a hash-based method known as locality sensitive hashing [1].

A SIFT detector transforms a photo into a collection of feature points that are invariant to photo scaling, translation, rotation and partially to illumination changes. Such photo can be viewed as a bag-of-feature-points and any object in the photo is represented as a subset of the points (local descriptors). Unfortunately, with such representation, there arises a many-to-many matching problem in a high-dimensional space because the photo typically contains from hundreds to thousands of feature points. Therefore, our proposed solution provides the interactive photo segmentation whereby a user can select a subset of feature points of a target photo instead of a whole set. The query subset represents objects-of-interest and remaining points of the target photo are used to a refinement. A group of points of the subset can be associated with keywords and consequently each of database photos can contain exactly named objects-of-interest.

Our annotation process takes place in four steps. First, a system creates a candidate list consisting of database photos in which each one contains the same objects as the query. Second, the candidate list is refined by comparison with remaining points. Third, from the list there are gathered and ranked all named objects of which is created an annotation list (keyword proposals). Finally, other keywords associated with the candidates are ranked and combined with the annotation list.

By reason of using the local descriptors, the solution is high resistant to cropping and other common transforms against approaches based on global features. Our proposed solution does not require input caption but in the case of insufficient results allows extending input query of the keywords. Thus, our solution allows identifying objects in the photo and using relevance feedback user can improve performance of our content-based photo retrieval.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Ke, Y., Sukthankar, R., and Huston, L. 2004. An efficient parts-based near-duplicate and sub-image retrieval system. In *Proc. of the 12th Annual ACM Int. Conf. on Multimedia* (New York, NY, USA, 2004). New York, NY, pp. 869-876.
- [2] Kuo, Y., Chen, K., Chiang, C., and Hsu, W. H. 2009. Query expansion for hash-based image object retrieval. In *Proc. of the 17th ACM Int. Conf. on Multimedia* (Beijing, China, 2009). ACM, New York, NY, pp. 65-74.
- [3] Wang, X., Zhang, L., Jing, F., and Ma, W. 2006. AnnoSearch: Image Auto-Annotation by Search. In *Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition – Vol. 2* (2006). IEEE Computer Society, Washington, DC, pp. 1483-1490.

Spatial and Time Navigation in Multimedia

Michal LOHNICKÝ*

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

For most people a photo gallery is a way how to present experiences from their holidays, trips and events. Taking photos is a process of saving emotions via a camera and therefore the visualization of photo albums should communicate these emotions to make events unforgettable. Moreover, people spent their special events not alone, because as social human beings they share their events [1]. This is the reason why the process of creating photo albums should be a collaborative process. Last but not least, photo albums aggregate various experiences during holidays and a user's story is created by the chronological ordering of experiences.

Our work is aimed at the augmenting user experience while browsing photos and also other multimedia supporting the storytelling. This can be accomplished by various ways but to make the solution attractive for users the process of creating photo albums has to be as automated as possible. On the other hand, to take advantage of the collaboration we can ask the user to make more unordinary tasks to like write short descriptions of events etc. This is possible because when users create photo albums from mutual events there are overlap activities accomplished by both the users and the saved time can be used in a better way. Moreover, collaboration mostly means the motivation to create more attractive results, it is funnier, users write more comments and also more users remember more memories that have be archived.

However, we must not forget the automated part of creating photo albums that is essential, because we can ask the user to complete just limited amount of tasks. We have made an experiment consisting of five people. The result shows that people emphasize five elements in their storytelling (elements are ordered by relevance):

- *Events*: The users pick few most important events as a base of storytelling [1].
- *Order of events*: If the storytelling takes less than 5 minutes, the tellers order the events by relevance otherwise chronologically.
- *People*: The users clearly define the event attending persons.
- *Geographical localization*: People mostly link their stories to a locality (e.g. here was the hotel, the beach was approximately at 0,5 km distance etc.).

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

- *Other characteristics of the story:* Mostly pictured on photographs or varied facts connected to the place, weather etc.

We have also analyzed some portion of photo albums and we have found that there is a correlation between events and the amount of taken pictures during the hours of a day. So we can mark the local maximum of amounts and its surroundings as a single event and ask the user to name the event.

Moreover, a photography itself contains a lot of metadata in its EXIF, which can be analyzed and processed to augment the user experience while browsing the photo albums. For our work we identified two most important attributes in EXIF. The first one is geographical location where the photography was taken and the second one is timestamp.

Timestamp is important from two aspects – short time and long time aspects. Short time aspect is necessary to chronologically order events and also to order photos in these events to support the storytelling. The long time aspect is important because the digital photography has existed for 19 years and the timestamp is becoming more and more essential in archiving and browsing of photos. Users visit same places repeatedly and celebrate birthdays again and again etc. So it is important to provide visualization which can be used to navigate through time dimension.

The visualization of chronologically ordered photos should support the placement of photos in a space in a way that the storytelling would be easily recognized and understood. The solution of the placement is in the usage of maps as a background of visualization. The other advantage of the map placement is to offer another auxiliary element of storytelling. The information where the photo was taken says a lot about the photography even before a photo is viewed because the location is a kind of a connection between the photo and the events in the area. For example, we can easily gain the character of holiday (weather, beach, mountain, hiking etc.) from the location.

The main aim of our work is to propose an innovative navigation and browsing in photo albums according to the timestamps, geographical locations and collaborative storytelling. This kind of navigation in the combination with proper photo analysis and metadata discovering can create various views of a complex collection of photos in photo albums. This style of browsing photos can be used by the users for sharing photos in much higher quality, for finding photos which they miss in their photo collections, to view places where they intend to go in various time periods etc. It is also usable in the commercial sphere – in travel agencies, botanic monitoring etc. When we add the direction of taking photos to the location we can create an ideal presentation tool for real-estate companies.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Jomhari, N., Gonzalez, V. M., and Kurniawan, S. H.: See the apple of my eye: baby storytelling in social space. In *Proc. of the 2009 British Computer Society Conf. on Human-Computer interaction* (Cambridge, UK, 2009). British Computer Society, Swinton, UK, pp. 238-243.

Improving Query Suggestion Capabilities Using Web Search Results

Ladislav MARTINSKÝ *

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

Correct query formulation for web search is very important prerequisite for successful retrieval of desired results. Powerful tool for this purpose, used in most popular search engines, is query suggestion. The most important benefit is non-obtrusive motivation of user to further specify his needs, which is very helpful not only for him/her, but also for search engine. Capabilities of these existing solutions are however limited in scope only on the terms with high popularity. Approach presented in this diploma thesis provides alternative way to generate these suggested words, based on real time search results analysis. Main advantage is presence of suggestion for any meaningful query for which would user get one or more relevant search results, not only for popular ones.

Search results represent small units of wisely chosen information about particular query or context. These units consist of words with different levels of importance for user. Four main characteristics were chosen to distinguish which of them are applicable as suggested words to enrich query. Location of word, in which it appeared in result (title, description, url address, etc.) is measured by *Locality* characteristic. Each part of result has its own importance. Title is for example more important than description. *Uniqueness* represents number of different results in which particular word appeared. This characteristic can help to distinguish different contexts. When word appears only in one result, there is a chance, that it can represent unique context to the query comparing to other words. *Popularity* is percentage of occurrence of word to all the other words. Last characteristic *Distance* is count of words between examined word and query in one particular area of result. Words that occur close to query may be the correct ones which should user add to query to get desired results. Evaluation scale for every characteristic was set to 1-10. One potential suggested word can have maximum score of 40. Experimental results of this main part of application have shown good potential to choose relevant and helpful suggested words for user.

* Supervisor: Pavol Návrat, Institute of Informatics and Software Engineering

Very important solution has been published in article [1]. Authors are proposing a method to correct spelling errors in query with usage of actual search results. Main purpose why are authors using search results is unlimited amount of possible query forms and mistakes which cannot be efficiently stored in any database or data structure. Usage and advantage of search results is based on fact that they often contain misspelled occurrences, but also the correct ones. Similar problem of unlimited amount of possible query formulations is addressed also in my solution.

Important area to improve the effectiveness of presented solution is considering different needs and goals for every user. Suggesting words in general is helpful, but the amount of help depends on each user particular needs. This problem is addressed by second part of solution - personalization. Every action of user towards application (click on suggested word, click on result, etc.) represents his own needs and requirements and is stored in personal profile. This profile information is important for fifth added characteristic *Personalization*, representing the importance of particular word in according to his actions in past. Final result is ability to lever up important words for a particular user, which can potentially offer more relevant help to him/her.

To illustrate the flow of whole application, it is divided into four main parts/modules with following functions:

1. Data sources – gather and store results (Google, Yahoo, DMOZ)
2. Data processing – transformation of words to unique form (stemming, removing stop words, etc.)
3. Data evaluation – evaluate and sort (characteristics and personalization)
4. Presentation – provide suggested words to user (personalization)

Each module also represents a step in application flow. In brief the whole process starts when user finishes typing query. Application submits it to data sources and gathers data, which is processed and evaluated. Final chosen words, which are most relevant to query are presented via suggestion back to user. Clicking on any of suggested words restarts the whole process with modified query and optionally updates user's profile.

Approach presented in this work is not meant as a replacement for today popular solutions (Google, Yahoo, Ask.com), but rather as an extension. Since it needs little more time to generate suggestions, it can take place after regular solution fails and supplement it. Also the character of suggestion is little different with suggesting only whole words but not the possible endings of actual query word.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Chen, Q., Li, M., Zhou, M.: *Improving Query Spelling Correction Using Web Search Results*, [cited 19.2.2010], Accessible from <http://www.aclweb.org/anthology/D/D07/D07-1019.pdf>

Geo-Based Social Network Navigation

Anton BENČIČ, Roman MÉSZÁROŠ, Roman PANENKA, Márius ŠAJGALÍK*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
`icup_fiit2010@googlegroups.sk`

Social networks have been around for some time, but it was only recently that they have gained their popularity. These social networks are mostly accessed within a web browser, they usually operate on the well-known hypertext basis and the links between individual nodes in the network are represented by hyperlinks in the web site. Users always start at their own node and this allows them to quickly interact with people who are directly connected to them. On the other side if we need the users to often create temporal connections, then the mentioned tree navigation is not the best choice.

Because our system requires such interaction we decided to provide a map interface, as support [1] for social networks visualization. In this system we are providing users the opportunity to give away or lend some of their items, or get something from others if they need it. We decided to use existing social networks instead of creating a new one for obvious reasons, but user interactions build a new network of temporal connections. These connections as well as the permanent ones are later used in our evaluation algorithms that decide what users see in the map, based on region-specific characteristics and their own preferences. An example of items visualization is in the Figure 1.

Our system implements two types of evaluation algorithms. The first one is for static view, while the second one deals with the time component as well. The algorithm for static evaluation calculates rating coefficient for people, offers, requests and collections. While the map shows either individual users, or offers and requests together, it also shows collections organized by charities in both of those views. This means that the evaluation algorithm has to decide whether to promote an offer or a collection if a collision occurs. This is certainly not as straightforward as deciding among a group of colliding offers.

The dynamic evaluation algorithm is used when the user is viewing activities from the past or watching them in real-time. If the user views activities in real-time, the algorithm maintains the frequency at which these actions occur by accommodating the frequency of shown actions to the real action flow. This means that with the rise of user-relevant actions over certain period of time, the percent of shown actions remains

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

the same, thus increasing the number of shown actions over that time period. The percent coefficient is calculated from the frequency of actions over longer period of time, so it is adjusted to the current trend.

The aim of evaluation algorithms is to provide a view, where both the region and the user characteristics are relevant. Regional relevancy is achieved by analyzing trends of activities and preferences in the region, while user relevancy is evaluated using the active user's activities and preferences.



Figure 1. Visualizing of users and items on map.

The system as a whole uses service-oriented architecture [2] using Windows Communication Foundation. We have chosen this architecture and technologies, because this way the client can abstract from the functionality that is provided by server and it allows us to implement both web and mobile clients using a single paradigm.

Present aims to address as many people as possible and to empower this we use existing social networks, where the information about Present and information from Present are spread. Social networks are also used for people to promote positive competition and to motivate them to help others and our planet.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] MacEachren, Alan M.: How maps work: representation, visualization, and design. The Guilford Press, (2004).
- [2] Datz, T.: What you need to know about service-oriented architecture. CIO, (2004).

Browsing Similar or Related Data Entities by Breadth-First Search in the Semantic Web

Karol RÁSTOČNÝ *

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

The difficulty of finding relevant data on the Web is increasing as web repositories grow. Therefore, we propose an approach for browsing the Semantic Web, which can help users find relevant results, i.e. how to find data in the Semantic Web, and how to browse similar and/or related data entities.

We extend the faceted browser Factic¹, which displays results in thumbnail matrix/list view with an additional view where results are categorized in hierarchical clusters. This view helps users to browse large numbers of results in a more organized way. We propose an approach to hierarchical cluster creation and labelling using semantic similarity computed from metadata.

Our clustering approach is based on Bordogna's and Pasi's hierarchical-hyperspherical divisive fuzzy c-means clustering method [1]. This approach has good results, but there are performance issues. We see a problem in the process of identification of optimal count of clusters in the first level of clustering. We propose an approach in which we make an interval approximation of the optimal number of clusters. As optimal number of clusters we select the best based on the quality function defined by Mecca et al. [3]. Because this quality function is based on the removal of edges with maximal length from a minimum spanning tree, where vertices represent clustered items and length of edges represents similarity, it may not return the optimal number of clusters for fuzzy clustering. To address cluster label creation, we propose a novel method based on common facets in the similarity vector, where we proceed from the lowest level of clusters to the topmost level.

While faceted browsers help users to find information they offer less support for the exploration of resources similar to an already found result. We address this via view-based search within the Semantic Web using navigation in a 2D graph. As with other tools for navigation (e.g. Paged Graph Visualization [2]), the process starts with one central node, which represents the initial result and some nodes around it

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

¹ <http://mirai.fiit.stuba.sk/meia>

representing its facets. After that, users can browse the graph by expanding nodes representing facets. To prevent the graph from becoming unclear due to node expansion, which can result in many (irrelevant) nodes, we propose tools based on result clustering, facet marking and the hiding of nodes and graph components.

Clusters are created from results that have the same facets displayed in the graph and behave in the same way as results (see Figure 1). This means that clusters are connected to facets and users can display their facets. We give users the ability to filter new results after expanding nodes by marking facets as wanted, unwanted and visible. This marking sets if new results new results should, may or are not allowed to have direct connections to marked facet. The hiding of nodes gives users the opportunity to choose which nodes they do not want to see anymore. The hiding of some nodes in the graph can segment the graph into several components, thus enabling the hiding of whole graph components instead of sequentially hiding individual nodes.

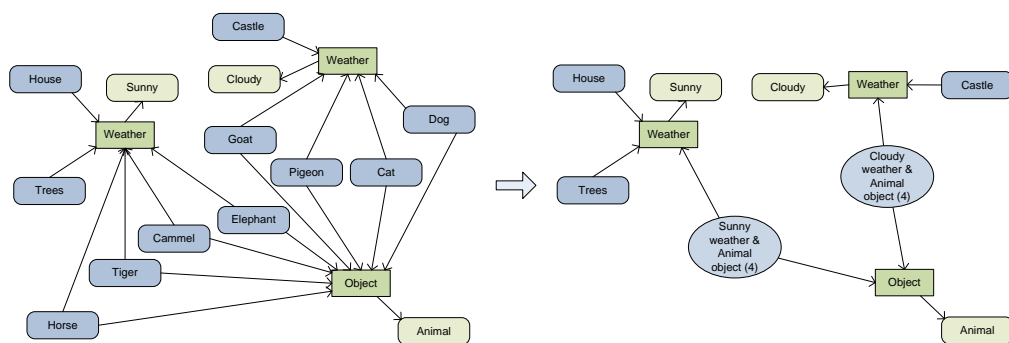


Figure 1. Example of cluster creation in a graph.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010.

References

- [1] Bordogna, G., Pasi G.: Hierarchical-Hyperspherical Divisive Fuzzy C-Means (H2D-FCM) Clustering for Information Retrieval. In: *Proc. of the 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology*, IEEE Computer Society, Washington (2009), pp. 614-621.
- [2] Deligiannidis, L., Kochut, K.J., Sheth, A.P.: RDF data exploration and visualization. In: *Proc. of the ACM first workshop on CyberInfrastructure: information management in eScience*, ACM, New York (2007), pp. 39-46.
- [3] Mecca, G., Raunich, S., Pappalardo, A.: A new algorithm for clustering search results. *Data & Knowledge Engineering*, Vol. 62, No. 3 (2007), pp. 504-522.

Enhancing Exploratory Search: Graphs, User Modeling and Search History

Jakub ŠIMKO*

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

We focus on exploratory search (ES), which is oriented toward helping users during learning and investigative search tasks where the goal is not a single web document or fact [1]. ES fights the “information space invisibility problem”, best described with typical example of web search, where a user has to guess search keywords rather than picking a query from a list. ES gives users clues at each step of the search session which are produced based on extraction and analysis of information space semantics and presented with respect to the user’s personal interests.

We facilitate exploratory search in several ways. First, we develop a novel approach intended to reduce user effort required to retrieve and/or revisit previously discovered information exploiting web search and navigation history (its state of art was well described by M. Mayer [2]). We collect streams of user search actions and identify user agendas (i.e. groups of actions that form sessions with a common user goal, the principle was implemented in various projects [3]). The semantics of each action (text query, facet restriction, visited document) is represented by a term vector, constructed with the help of lemmatization and term extraction API’s (WordNet, OpenCalais). Actions are grouped by cosine similarity of their vectors and the time gaps between them using fuzzy rules. Our early experiments of session identification over the AOL corpus show promising results.

Based on the identified sessions, we construct and persistently store visual trees representing session history (see Figure 1). Trees, visible to user, provide an overview of the current (complex) session and improve orientation among visited results. We also provide users with a History Map – a scrutable graph of semantic terms and web resources, constructed by merging individual session history trees, using the Delicious Taxonomy, and the associated web documents (see Figure 2). The History map has full-text search capability over individual history entries and enables navigation throughout the visualized history graph. We evaluate our approach on the Web via supervised and unsupervised live user experiments.

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

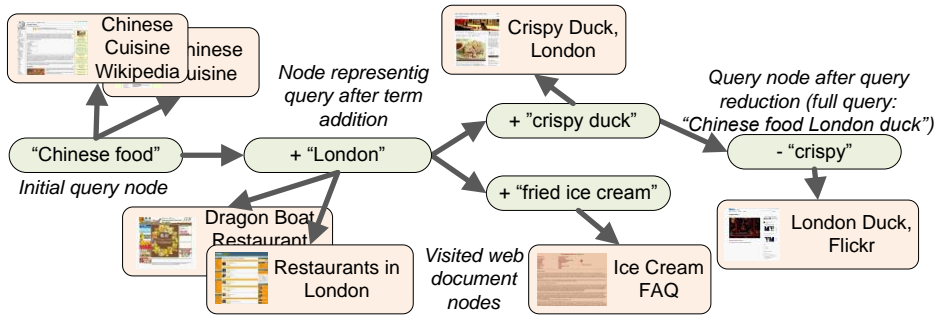


Figure 1. Search History Tree example.

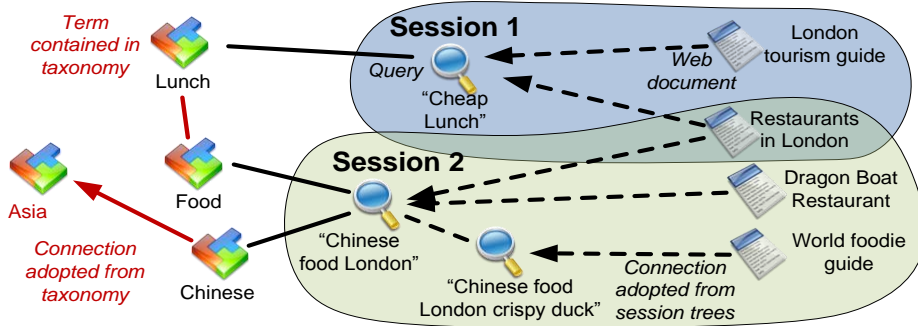


Figure 2. History Map example.

Also, our interest lies in discovering relationships among web documents and terms, that are useful for providing navigation clues. Based on identified search sessions, we map their initial queries to end results and refine suggestions for future occurrences of such queries. We focus on discovering relationships not detected by search engines.

We also discover relations between the terms themselves via a special web search game in which users formulate queries in a specific format to minimize the number of returned results. The query format forces players to use terms that are related together, where multiple occurrence of the same term combinations results in a relationship.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM 49 (2006), pp. 41-46.
- [2] Mayer, M.: Web history tools and revisitation support: A survey of existing approaches and directions. Foundations and Trends in HCI, vol. 2, no. 3, 2009, pp. 173-278.
- [3] Jansen, B. J. et.al.: Defining a session on web search engines: Research articles. J. Am. Soc. Inf. Sci. Technol., 58(6), 2007, pp. 862-871.

Lightweight Semantic Search Based on Heterogeneous Sources of Information

Marián ŠIMKO*

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

To satisfy user's information needs, the most accurate results for entered search query need to be returned. Traditional approaches based on query and resource Bag-Of-Words model comparison are overcome. In order to yield better search results, the role of semantic search is increasing. However, the presence of semantic data is not common as much as it is needed for search improvement [1]. Although there are initiatives to make resources on the Web semantically richer, it is demanding to appropriately describe (annotate) each single piece of resource manually. Furthermore, it is almost impossible to make it coherently. The current major problem of the semantic search is the lack of available semantics for the resources, especially when considering the search on the Web [2].

To overcome this drawback, we propose an approach leveraging lightweight semantics of resources. It relies on resource metadata model representing resource content. It consists of interlinked concepts and relationships connecting concepts to resources (subjects of the search) or concepts themselves. Concepts feature domain knowledge elements (e.g. keywords or tags) related to the resource content (e.g. web pages or documents). Both resource-to-concept and concept-to-concept relationship types are weighted. Weights determine the degree of concept relatedness to resource or other concept, respectively. The domain representation we obtain is straightforward and adopted to its goal – it is designed to address the specifics of the Web environment and enables to improve personalized search. Furthermore, it resembles lightweight ontology thus allowing automated generation.

When acquiring metadata, we process heterogeneous sources of information: the content and social data. We consider:

- keywords supplied by the author himself,
- keywords (concepts) generated automatically by content processing,

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

- tags supplied by both web-application users or users of some social tagging service.

We assume that by using heterogeneous sources of information the acquired domain model will be more accurate and thus feasible to enable advanced behavior such as recommendation or personalized search in web-based applications.

Having domain model as described above, we examine the possibilities of search improvement. We propose two variants of so called *concept scoring computation* taking place online during searching (see Figure 1). With concept scoring we extend the baseline state-of-the-art approaches to query scoring computation expecting an improvement of the search. For the computation we consider two approaches: statistical and topological. First approach takes into account statistical aspect of available metadata, while second one analyses a subset of metadata topology. Utilizing metadata we are able to assign the query to particular topic (set of concepts) and yield more accurate search results with respect to related resources.

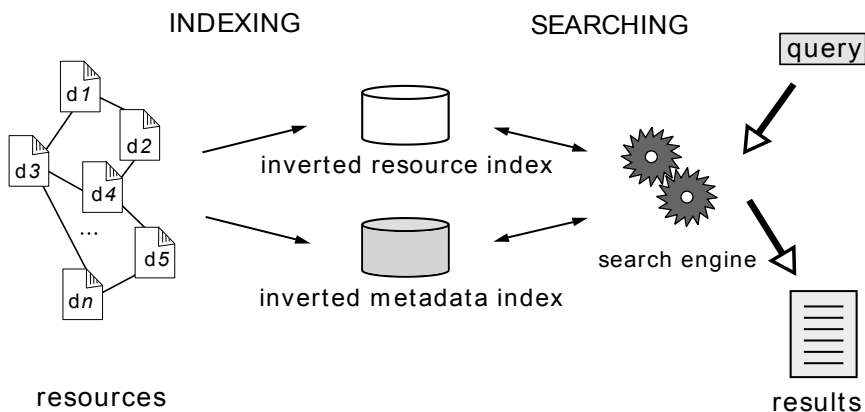


Figure 1. Combined scoring computation overview.

In the current stage of the research we are working on the evaluation of the proposed approach by building on the Lucene information retrieval library.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., Castells, P.: Semantic Search Meets the Web. In *Proc. of the 2008 IEEE International Conference on Semantic Computing*, pp. 253-260, (2008).
- [2] Sabou, M., Gracia, J., Angeletou, S., d'Aquin, M., Motta, E.: Evaluating the Semantic Web: A Task-based Approach. In *LNCS 4825: The Semantic Web. Proc. of the 6th International Semantic Web Conference, ISWC 2007, Busan, Korea*, pp. 423-437, (2007).

Exploratory Search in the Adaptive Social Semantic Web

Michal TVAROŽEK*

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

The evolution of the Web as a dynamic global virtual socioeconomic space resulted in many issues that affect both individual users and the entire human society. We need not only to address information overload and the navigation problem but also accommodate for novel trends in web use such as the push towards exploratory search [2], more interactivity, involvement, personalization. To achieve these objectives, we need to take advantage of principles and approaches outlined in present web initiatives – the Adaptive Web, the Semantic Web and the Social Web.

In our work, we combine and extend several existing approaches in order to create an advanced exploratory search browser for both the semantic and legacy web taking advantage of personalization, social wisdom and semantics. Ultimately, our goal is to provide users with a seamless exploration experience within a common seamlessly integrated Adaptive Social Semantic Web environment.

We build upon existing approaches to faceted browsers and advanced visualization such as VisGets [1], and propose an enhanced faceted browser extended with support for semantic information spaces to facilitate:

- exploratory search in terms of investigative and learning tasks,
- automated user interface generation to accommodate for web dynamics,
- user modeling and personalization to address information overload,
- collaborative content/meta-data creation to harness the power of social wisdom.

To evaluate our approach we developed Factic – a personalized faceted browser based on the aforementioned principles (see Figure 1). Factic is also integrated with additional support approaches for exploration such as history tracking and tree visualization, graph visualization and incremental navigation in the information space, and custom content rendering tools to facilitate content exploration [3].

We evaluate our approach both via synthetic experiments and user studies in several application domains – digital images, scientific publications, job offers. Our

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

Search

Browse

Explore

Logged in as michael

Logout

Aspect ratio

Landscape 16:10 (1)

Landscape 16:9 (2)

Landscape 3:2 (5996)

Landscape 4:3 (65)

Landscape 5:4 (5)

Portrait 16:10 (1)

Portrait 3:2 (432)

Portrait 4:3 (28)

Portrait 5:4 (1)

Author

Mária Bieliková (1492)

Camera

NIKON CORPORATION NIKON D300

NIKON CORPORATION NIKON D70s

FUJIFILM FinePix S5Pro (4)

NIKON CORPORATION NIKON D70 (1)

Panasonic DMC-FX07 (2)

Canon EOS-1D Mark II (6)

FUJIFILM FinePix Z2 (1)

SONY DSC-N1 (85)

Category

Resource > Image

ListView

MatrixView


MoreResults

Slideshow

View

BatchEdit

Train



Type:

Photo Resource Image

Functional properties:

Created with:

NIKON CORPORATION NIKON D300

Lighting:

Half dark

Thumbnail:

http://mirai.fit.stuba.sk/photos/tb/tb_6624753e841aded0eca3fcb918e2134.jpg

Created on:

2008/12/01

Created at:

05:29:14

Image owl:

Photo.owl:1

Search

Browse

Explore

URI:

http://mirai.fit.stuba.sk/ontologies/photo.owl

Load

Help

Save All

Load Previous

Clear Cache

22-rdf-syntax-ns:type

photo

rdf-schema:Resource

Image

Format

photo.owl#format_9c8113aa5789fb

Shows

elections/2008

Cats

Created with

NIKON CORPORATION NIKON D300

Author


Camera > NIKON CORPORATION NIKON D300

ListView


MatrixView

MoreResults

Slideshow




Vyber08_230_misc08_07




Vyber08_228_priroda08_0




Vyber08_227_priroda08_0




Vyber08_229_priroda08_0




Vyber08_225




Vyber08_233_priroda08_0




Vyber08_226_priroda08_0




Vyber08_224_priroda08_0




Vyber08_232_priroda08_0




Vyber08_234




Vyber08_231_priroda08_0




Vyber08_219_misc08_04



Vyber08_218_priroda08_0



Vyber08_217_priroda08_0



Vyber08_219

Figure 1. (A) Generated facets in Factic with a list-based result overview showing all result properties (top left). (B) A matrix result overview with image thumbnails and the correspondingly generated annotation pane for collaborative content creation (bottom right).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

- [1] Dörk, M., Carpendale, S., Collins, C., Williamson, C.: VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Trans. on Visualization and Computer Graphics*, 2008, vol. 14, pp. 1205-1212.
- [2] G. Marchionini: Exploratory search: from finding to understanding. *Communications of the ACM* 49 (2006), pp. 41-46.
- [3] Tvarožek, M., Bieliková, M.: Reinventing the Web Browser for the Semantic Web. In: *Proc. of WI/WIRSS '09*, IEEE Computer Society, 2009, pp. 113-116.

Information Search Considering the User's Interest and Groups of Similar Users

Matej VALČUHA*

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

Due to constantly expanding amount of information on the internet it has become increasingly difficult to find the information we need. Therefore we use search engine in which we enter our request. Usually we provide one or two words to specify the item or the term. For the same requirement, the standard search engines give us the same result. Different users might be looking for completely different things even though they use the same requirements. Therefore, the result of the standard search engine might contain a lot of irrelevant links. Such problems can be solved by personalized search, which takes into account the particular interest of the user. Based on the user's profile, the personalized search can override the requirement and give it to the standard search engine or to change the links order that has returned as search results of the standard search engine. It can also combine both of these opportunities to achieve more relevant results to the user's interest.

To start such a search it is necessary to find a large source of information about the user from which one can read the area of interests. Social networks have a great potential for this tool. People disclose their hobbies, update their status, join different groups and become fans of their favorite movies, musicians and athletes. Groups have the advantage that even when the user is inactive, other users can contribute with relevant facts. All such information, like membership in a group or messages, can be used to create a profile of interests of the user. The users of social networks can also express their views on any topic or a question. This interest could be used in determining the relevancy of the results. The users would be given an opportunity to express their views on the results. These results would be shown to the others with similar interests and they would be able to increase the ranking of a relevant result or to decrease the order of an irrelevant result.

The user of the personalized search should be allowed to choose which of his groups would be added to the search profile. Membership in a group does not have to reflect the user's current or permanent interest.

* Supervisor: Pavol Návrat, Institute of Informatics and Software Engineering

The principle of the search is shown on the Figure 1. The user enters on Facebook or another social network his request to search through text. Then he sends this request along with selected groups which should be taken into account. All of the text is divided into words. Words are then adjusted to the basic form to make it easier to compare with. Consequently, the required terms are compared with the group titles. If a match is found, the group is scanned in more details and the most used terms are used to rewrite the requirement (query rewriting) to achieve better results. Otherwise, it searches the groups in details and it looks for the required term. If a match is found, the words from the group titles are used to complete the query. This rewritten requirement is then given to the standard search engine.

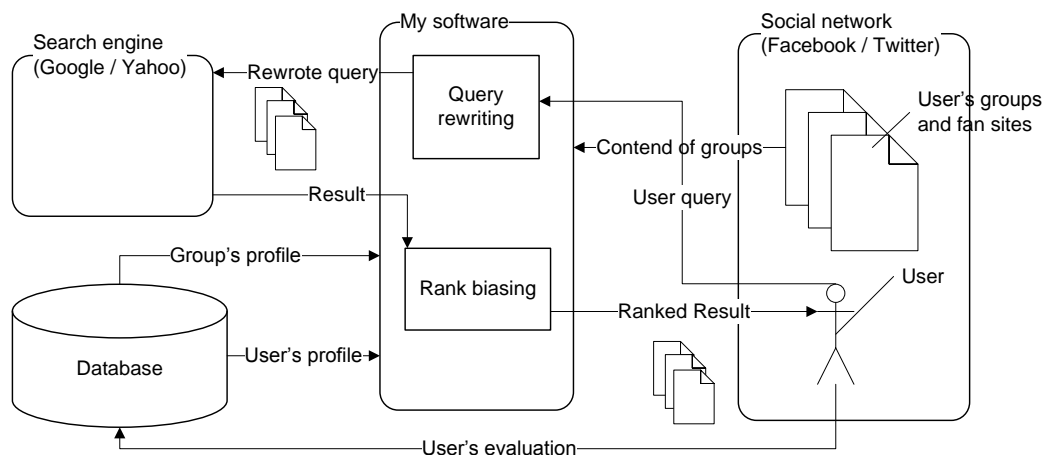


Figure 1. Architecture of the system.

In the next step the results of the standard search engine are processed, along with the profiles from the database. Each group and each user have their own profile. These profiles are compared with the results found by the search engine. Based on the weight of every word from the profile and on its frequency in the results every reference is assigned with a weight, which determines its order.

The actual evaluation from the database is given to the final references. However, only the evaluations that belong to the user or one of his groups are chosen. The weight of the evaluation is added to the weight for each link. These evaluations also appear under the link in the results as an advice from the users who have used them.

The final order of links is the sum of all of their weights. The more relevant link for the user the more weight should it have. After clicking on one of the results of the search, the selected page is displayed and also the option to add the evaluation of its relevance. Information about the click on the reference is recorded into the user's profile and those of his groups that match the chosen reference. After choosing the evaluation this is added to the other evaluations.

Acknowledgements: This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant KEGA 345-032STU-4/2010.

Classification and Recommendation

Website Navigation Adaptation Based on Behavior of Users

Michal HOLUB*

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

Web portals contain large amount of information from which various groups of visitors could benefit. Unfortunately, the website does not “know” its users and so the presented content is not personalized. Visitors often see the content in which they have no interest [2]. Web portals also contain large amount of links and it is difficult to choose which link to follow. Users surfing the web leave digital footprints that reflect their interests and preferences. We can analyze this information and use it for content personalization. Moreover, some users can discover an interesting page hidden deeper in the web portal’s hierarchy. Users who behave similarly do not have to find it again, they can benefit from mutual recommendation of pages.

We propose a method for adaptive navigation support and link recommendation within a web portal. It is based on the analysis of users’ navigational patterns and their behavior on the web pages. We also mine the portal to extract interesting information which is presented in a new way. Web pages of the portal are enriched with new sections with links that might interest the user.

Each user selects different approach while browsing through a web portal. One user can follow links to certain depth and then backtrack if he has not found the desired information. Other user can use more the breadth first approach. In this case he tries basically to visit all links from the menu and returns straight to the main page. Based on users’ activity we discover four basic navigational patterns [3] in their clickstreams and group the users according to the prevailing patterns. Within each group we compare users using cosine similarity method on their clickstreams. For each user u in every group we get a list of other users from his group sorted by their similarity to u . We select top N similar users and recommend links to u which they found interesting.

The way a user behaves on a web page reflects his interest in this page. We monitor actions he conducts which include *time spent*, occurrence of *scrolling events* and *copying text into clipboard*. Comparing these actions with actions of other visitors to the same page indicates the degree of the user’s interest. We do the comparison

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

using collaborative filtering method, as we describe in [4]. This enables us not only to compare users but also to predict user's interest in a web page he has not visited yet. For the prediction we take computed interest of his top N similar (according to clickstream comparison) users who already visited the web page.

We apply proposed method of adaptive navigation and link recommendation to the web portal of our faculty (www.fiit.stuba.sk). First we analyze the portal's web pages and extract useful information. Many pages inform about an upcoming event. Therefore we create a personalized calendar of events for each user and insert it to the web page. The calendar contains two kinds of events:

- Event which interests the user (according to our interest estimation method). We add reminder about this kind of event to user's calendar.
- Event which the user does not know about yet, but which similar users found interesting. We recommend link to this event and add it to user's calendar.

To evaluate proposed method we use a solution based on adaptive proxy server [1]. Adaptive proxy server is a platform that enables implementation of various methods and techniques of content and navigation adaptation. It tracks user's actions and modifies HTTP requests and responses. We implemented a plug-in which enhances the web pages with personalized sections. These include calendar with events personalized to every user and links to pages which we recommend him to visit. Recommendations are periodically computed by an independent tool for every user who visits the web portal. Our plug-in also inserts controls for explicit feedback. Users can express their interest (positive or negative) in the visited web page. They can also state if they would or would not recommend the page. Explicit feedback is used for evaluation of the contributions of proposed method to browsing experiences of visitors. Using proposed method we can personalize other sections of web page that contain links as well. We can sort these links according to expected interest of user. By comparing clickstreams we get communities of similar users which can be used also in other applications.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Barla, M., Bieliková, M.: „Wild web“ personalization: adaptive proxy server. In *Proc. of the 4th Workshop on Intelligent and Knowledge oriented Technologies* (Herľany, Slovakia, 2009), F. Babič and J. Paralič (Eds.), pp. 48-51 (in Slovak).
- [2] Barla, M., Tvarožek, M., Bieliková, M.: Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems. *Computing and Informatics*, Vol. 28, No. 4 (2009), pp. 399-427.
- [3] Canter, D., Rivers, R., Storrs, G.: Characterizing User Navigation through Complex Data Structures. *Behaviour and Information Technology*, Vol. 4, No. 2 (1985), pp. 93-102.
- [4] Holub, M., Bieliková, M.: Estimation of User Interest in Visited Web Page. In *Proc. of the 19th Int. Conf. on WWW* (Raleigh, NC, USA, 2010), ACM Press.

Personalized Recommendation of Interesting Texts

Michal KOMPAN*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
`kompan05@student.fiit.stuba.sk`

The data amount on the web is serious problem for the common user. One of the most relevant sources of information over the web are news portals (nytimes.com, reuters.com, etc.). Most of users prefer large renowned news metaportals. They include thousands of daily added news from the whole world and there is no chance to access them in a fast and comfortable way for every user. The only way to help the user is to filter large amount of information and reduce it to an acceptable amount. There are several filtering systems in this domain nowadays [3, 4].

The main problem in the content-based filtering is effective and enough expressive representation of items (or articles). This is often done by means of text summarization [1] or keywords extraction [2]. These techniques are commonly used in English based systems and cannot be easily applied to other languages. Keywords extraction and summarization brings better results as the other methods but are more time consuming. These methods cannot represent non-text documents without modification.

Proposed representation compresses article information value to short vectors, which are used for fast similarity computation over the specific articles time-window. This vector represents article in an effective way, so there is no need to store whole articles. Proposed method expects pre-processed article as an input and produces vector representation usually no longer than 30 words. Then these vectors can be easily used for similarity computations or we can use them in special structures for recommendation e.g. binary trees [5].

Our method for content-based news recommendation uses this effective article representation. We use similar articles to create recommended content based on implicit user model. The method for recommendation is based on three basic steps – computing article similarity, creating user model and recommendation based on first two steps (Fig. 1).

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

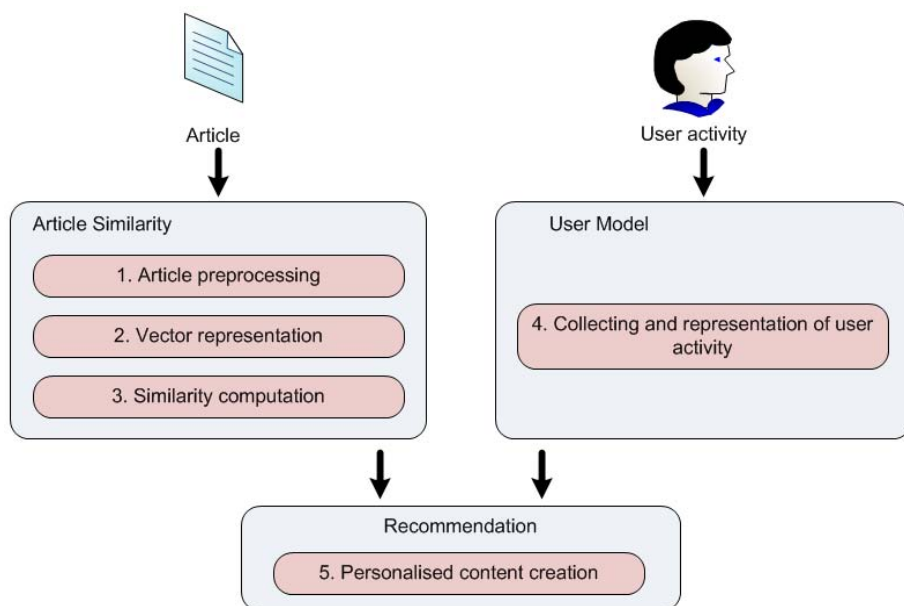


Fig. 1. Proposed news recommendation method.

In the article similarity step it is necessary to preprocess every article to reduce word space. Then is article represented in an effective vector representation, which is used in cosine similarity computation. As a result of article similarity step we obtain a list of similar articles for every article in the dataset. User model is created implicitly based on server logs by identification of visited and recommended article for unique cookie. Finally is the recommended content from both similar articles and user model created.

Acknowledgement. This work was partially supported by by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010.

References

- [1] Dakka, W., Gravano, L.: Efficient summarization-aware search for online news articles. In *Proc. of the 7th ACM/IEEE-CS Joint Conf. on Digital Libraries, JCDL '07*, ACM, New York, NY, 2007, pp. 63-72.
- [2] Kou, H., Gardarin, G.: Keywords Extraction, Document Similarity and Categorization, Tech.rep. No.2002/22, PRiSM Lab. of Versailles Univ., 2009.
- [3] Wongchokprasitti, C., Brusilovsky, P.: Newsme: A case study for adaptive news systems with open user model. In: *Proc. of Int. Conf. on Auton. and Auton. Systems*, 2007. ICAS07. pp. 69.
- [4] Yoneya, T., Mamitsuka, H.: 2007. Pure: a pubmed article recom. system based on content-based filtering. In *Proc. of Int. Conf. on Genome Inf.* 18, pp. 267-276.
- [5] Zelenik, D., Bielikova, M.: Dynamics in hierarchical classification of news. In *Proc. of the 4th Work. on Intelligent and Knowledge oriented Technologies (WIKT 2009)*, 2009, pp. 83-87.

Recommendation and Collaboration through Implicit Identification in Social Context

Martin LABAJ*

*Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
martin.labaj@computer.org*

In the field of e-learning, the identification of difficult and/or interesting parts of learning text can be useful feature for various tasks like rewriting the text, showing the student where to focus or offering help. However, methods, that acquire this information from inputs obtained by directly interacting with the user (explicit user feedback), for example by asking him to subjectively rate his comprehension, can lead to distraction in the learning process and require that users participate voluntarily.

In our work, we track implicit feedback/implicit interest identifiers including user scrolling, e.g. to which portion of text has user scrolled and what time he spent there (read wear [1]). Using statistical approach and taking intersections and overlays of timed viewports collected from many users over many page views into account, we can determine which part is the most time-consuming and therefore interesting or difficult. With enough users, details about various parts of scrollable content can be obtained very precisely, even with a precision of single words. Another basic important data consist of mouse clicks (click heatmaps) and mouse movement (flowmaps).

As in any method dealing with time based user action tracking, there is a possibility that user is pursuing different activities during evaluated time periods. While with mouse interaction it is evident that while those actions occurred, user has been truly working with content, in passive actions like scrolling to a part of content and viewing the displayed content for a period of time, we cannot determine directly from action (e.g. viewing/reading) itself whether user is working with tracked content nor even whether he is present at the computer. We try to resolve this by using camera pointed towards the user and employing two-level physical user tracking: (a) face tracking, where physical presence of user at the computer is detected and (b) eye tracking, where user gaze is evaluated. Both methods allow leaving out time periods when user is not directly using computer or in the case of eye tracking (where possible by the quality of used camera) even when he is using the computer, but he is working with different parts of screen and not with the displayed content. Readily available

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

solutions allow gaze tracking with enough accuracy for content (even text) evaluation even with low-cost commercially available webcams [3], therefore apart from enhancement of the spent time evaluation we also account the gaze location as an interest indicator.

Together with scrolling and mouse interaction, we assign to each word, picture and similar atomic fragments of content:

- Gaze samples – detected periods of gaze falling onto this fragment.
- On-screen time – how long has been user viewing screen with this fragment.
- Mouse interaction – clicks, selections for copying or annotating, continuous selecting (typically used as reading position aid), mouse-over events, etc. related to this fragment.

Based on these data, we calculate an attention index for each fragment similarly to attention time in [3]. Attention index of larger blocks is sum of attention indexes of its child fragments. These data are then used for document summarization and recommendation, document review recommendation and possibly even translated back to scrolling via content-adapted assistance to scrolling [2].

Subsequently, readily available data of user's active fragments of content can not only be used for content fragments identification and recommendation, but also in a social context. By augmenting the displayed content with indication of active fragments of other users, we provide users with information how are they doing in comparison with others. We are also hoping to increase user collaboration by providing ordinary user messaging augmented with indication of where each user (friend) currently works (reads) in the same content. This gives the user an option to contact those friends who are currently thinking about the same portion as he has problem with. As the user asks friends learning the same part, he is not distracting them away from their current study and he also obtains better advice.

While the main evaluation platform is ALEF (Adaptive LEarning Framework) system, one of the possibilities we consider for implementation is through Adaptive Proxy which would also readily bring this concept to open space (Web).

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T.: Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, ACM Press. 1992, pp. 3-9.
- [2] Ishak, E. W., & Feiner, S. K.: Content-aware scrolling. In *Proceedings of the 19th annual ACM symposium on User interface software and technology - UIST '06*, ACM Press. 2006, p. 155.
- [3] Xu, S., Jiang, H., & Lau, F.: User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 13th international conference on Intelligent user interfaces*, ACM, 2009, pp. 7-16.

Personalized Exercises Recommending for Limited Time Learning

Pavel MICHLÍK*

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

The objective of adaptive navigation is to help the student choose the best topics or learning objects to focus on, in order to maximize the learning efficiency. In this project we focus on exercises as an important part of preparing for an exam. Since the time for preparation is limited by the date of the exam or midterm, it might be not sufficient for learning all required concepts perfectly, especially for students who started preparing late. Our goal is to help the student to achieve as good exam result as possible. A strategy used by many students is going through all topics in the course very quickly and learning every topic at least to some extent, rather than learning few topics in detail. Our recommending method is designed to help the students to prepare for the exam using the former strategy.

To achieve proper learning time distribution between all required concepts, we attempt to determine optimal knowledge levels of all concepts at the end of learning time, which are achievable at the current learning speed. Using the overall knowledge level increase from the learning start (the sum of knowledge level increases through all concepts), we estimate the knowledge level increase from present time to the end of learning. The overall increase is then divided between all concepts in such way, that the final estimated knowledge levels of concepts correspond with the concepts importance given by the teacher. In an extreme case, where the student's knowledge was very low and there was little time left, the estimated knowledge level increase would be almost zero for every concept. Such learning strategy cannot be successful because the student cannot pass the test with almost no knowledge of all concepts. To prevent this condition, we set a minimal concept knowledge level – the estimated knowledge level for every concept is never lower than this limit.

To make a recommendation for a student, we compute an appropriateness value for each exercise in the course. Then, a predefined number of exercises with largest appropriateness values are recommended. Three criteria are used for each exercise evaluation: concept appropriateness, exercise difficulty appropriateness and time

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

period since the student's last attempt to solve the exercise. These criteria are orthogonal and for an ideal recommendation, all of them are supposed to be met. Therefore, the final appropriateness of the exercise for the student is computed as the minimum of the three partial results.

The purpose of concept appropriateness evaluation is to decide, whether the student should learn the concepts covered by the particular exercise. A vector of concept appropriateness values for the student is constructed and compared to the exercise's related concepts vector. The appropriateness value of a concept depends on:

- the teacher-given concept importance,
- the student's current knowledge level of the concept (the concept appropriateness falls rapidly after the student reaches the estimated optimal level – over-learning a concept would result in less achievable knowledge of other concepts),
- concept prerequisites (other concepts' knowledge that should be achieved before learning the particular concept – a concept is not appropriate for learning before its prerequisites are met).

The second criterion – exercise difficulty appropriateness – ensures that the difficulty of the recommended exercise matches the student's knowledge level. This prevents the student from being uninterested or discouraged.

The final criterion suppresses recommending of recently viewed exercises. After visiting an exercise, its appropriateness value produced by this criterion drops to zero and gradually returns to 1 over time.

The student knowledge model is represented by a vector of concept knowledge levels. Updating of the knowledge model is based on explicit user feedback – the student chooses one from a set of replies (from *solved* to *not understood*) and the knowledge levels are updated using the Computer-Adaptive Testing method [1].

Our solution is currently evaluated in controlled experiments within the Functional and logic programming course, using the ALEF framework for adaptive web-based learning [2]. Experiments consist of a pre-test, a learning session and a post-test to verify the adaptive navigation impact on students' learning performance.

Since our method uses weighted relations and fuzzy logic rather than strict rejection rules, we expect it to be able to deal with imperfections in the domain model. This can make the method usable with automatic generated domain models.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Linacre, J. M.: Computer-Adaptive Testing: A Methodology Whose Time Has Come. In: *Development of Computerized Middle School Achievement Test* (in Korean). Seoul, (2000). [Online; accessed May 5th, 2009]. Available at: <http://www.rasch.org/memo69.pdf>
- [2] Šimko, M., Barla, M., Bielíková, M.: ALEF: A Framework for Adaptive Web-based Learning 2.0. *World Computing Congress 2010, Key Competencies in the Knowledge Society (KCKS 2010)*, IFIP, Brisbane, Australia, Submitted.

Improving Search Using Graphs and Implicit Feedback

Ján SUCHAL*

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

With the coming era of semantic web, large, structured and linked datasets are becoming common. Unfortunately, current search engines mostly see web only as a graph of pages linked together by hyperlinks, thus becoming insufficient for searching in such new, structured and multidimensional data. When dealing with multidimensional data, identifying relations and attributes that are important for users to achieve their searching goals becomes crucial. Furthermore every user, can have different priorities, different goals which can even change in time.

One of the goals of this work is the extension of existing graph algorithms for multidimensional data, where the usage of tensor algebra and multigraphs can be useful, in contrast with currently preferred matrix algebra. Such extension of graph algorithms would be able to increase relevance and quality of search, and even enable new quality of query formulations.

Evaluation of relevance and quality of search can be done gathering implicit feedback (e.g. quality can be measured just by monitoring user interactions with the system). Another goal of this work is the exploitation of gathered (implicit or explicit) feedback from users to not only evaluates the underlying system, but also to analyse users' behaviour thus opening possibilities for adaptation and personalization.

The main goal of this work is the usage of implicit feedback in search and recommendation engines dealing with large multidimensional data, to improve search and recommendation result quality, speed and scalability.

In particular we focus our work on four major topics:

- Advanced techniques for mining knowledge and feedback from standard server logs, such as viral recommendation detection and probabilistic source identification, negative feedback from positive access logs and time-based trend characteristics.
- Performance and scalability issues of recommendation algorithms for real world large applications, such as collaborative news article recommendations and graph-based ranking algorithms.

* Supervisor: Pavol Návrát, Institute of Informatics and Software Engineering

- Dealing with uncertainty and unknown data especially in large sparse matrices and tensors with power-law distributions typical for real-world applications to search and recommendation engines.
- Comparing synthetic evaluation methods for search and recommendation algorithms with real implicit and explicit feedback-based methods.

Theoretical and practical results of our work in these areas include:

- Nearest neighbourhood collaborative filtering algorithm based on generic full text engine exploiting power-law distributions and yielding recommendations comparable to spreading activation graph-based model. Furthermore having linear scalability characteristics with respect to dataset size and easy parallelization to multiple machines/cores.
- Viral recommendation detection and probabilistic source identification from standard server access logs using time-based referer analysis and exploiting power-law distribution of recommendations [1].
- Negative feedback mining from standard server access logs applicable to news-based portals based on probabilistic „seen-but-not-clicked“ heuristic.
- Recommendation and evaluation framework for major news portal www.sme.sk.
- Application of spreading activation based recommendations on social network of slovak companies register (www.foaf.sk) [2].

Our future work is focused on more comprehensive evaluation of quality and performance characteristics for various datasets and parameter sensitivity.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Suchal, J.: Reconstructing Viral Recommendation Networks from Server Logs. In: *Student Research Conference 2009. 5th Student Research Conference in Informatics and Information Technologies Bratislava*, April 29, 2009: Proceedings in Informatics and Information Technologies, STU v Bratislave FIIT, 2009. ISBN 978-80-227-3052-5, p. 218-223.
- [2] Suchal, J., Vojtek, P.: Navigácia v sociálnej sieti obchodného registra SR. In: *Datakon 2009*, Proceedings of the Annual Database Conference. Srní, Czech Republic, October 10-13. 2009, Prague: University of Economics, 2009, ISBN 978-80-245-1568-7. p. 145-151.

Automated Recognition of Author's Writing Style in Blogs

Martin VIRIK*

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

In the past decade it has become much easier to create web content even for users with no experience with web technologies. Weblogs are the most typical and the most growing example of this trend. Thousands of bloggers use this hybrid genre to express their ideas, opinions and emotions, making blogs a rich space of topics and writing styles. In proportion to increasing number of blogs, the number of efforts to improve blog-search and recommendation algorithms has also grown. New requirements are aware of blog articles text quality and consider individual writing style an important blog characteristic.

In our research we focus on linguistic characteristics of blog articles in order to recognize and classify writing style of articles, blogs or even authors. We study the grammar and morphology of selected language and possibilities of computational linguistics to extract the features of document model necessary for further classification. In the first phase of our research we have been studying basic text mining and classification methods [1] and works related to the analysis of blog articles linguistic quality. Apart from user profiling, such as gender or personality profiling [2], a great effort has been on differentiating between informative and affective articles [3]. This and other genre based research has proven a large overreach of affective blogs, especially diaries. Methods analyzing reading difficulty are much related to the weblog classification [4]. We discovered a space for building models for multiple factors such as measures of syntactic complexity or prior knowledge of the reader.

We plan to build on system architecture for an advanced text mining system with background knowledge base as described by Feldman & Sanger [1] (see Figure 1). In our research we aim to accomplish preprocessing tasks and to create the processed article collection. This collection will be used by text mining algorithms, which discover patterns and trends and respond to user requests by considering his background knowledge and preferences.

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

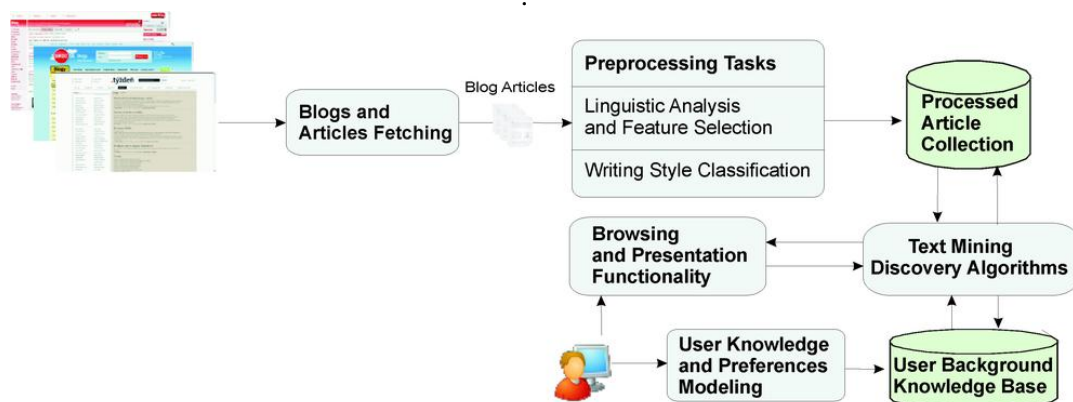


Figure 1. Proposal of system architecture for blog writing style classification (based on [1]).

We have considered many different types of features as potential markers of writing style, including lexical and syntactic complexity-based features or features mapping vocabulary difficulty level.

Average word length and syllables count are standard features measured in most readability indexes, such as ARI, Gunning-Fog Index or Flesch Reading Ease. Algorithms behind these indexes are relatively fast and easy to optimize to any alphabetic language. Together with grammatical and orthographic error count in unedited articles, these features could create a good picture of blogs lexical quality.

Capturing syntactic complexity of informal blog articles could bring more sophisticated view on text structure. Using advanced algorithms, such as Linked Grammar, can create space for discovering syntactic patterns, which could be used to characterize reading difficulty of writing style.

We plan to evaluate the results by applying our method on the collection of articles from live blogs and gathering implicit and explicit feedback from users.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010.

References

- [1] Feldman, R., Sanger, J.: Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York. 2007.
- [2] Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J.: Automatically profiling the author of an anonymous text, In *Communications of the ACM*, v.52, n.2, February 2009, pp.119-123.
- [3] Ni, X., Xue, G., Ling, X., Yu, Y. and Yang, Q.: Exploring in the Weblog Space by Detecting Informative and Affective Articles. *Comparative and General Pharmacology*, 2007, pp. 281-290.
- [4] Miltsakaki, E., Truitt, A.: Real-Time Web Text Classification and Analysis of Reading Difficulty. *Computational Linguistics*, June 2008, pp. 89-97.

Effective Representation for Content-Based News Recommendation

Dušan ZELENÍK*

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

Our work is based on advantages which could be achieved by the hierarchical representation of similarity between entities. As long as we are working with news we focused on representing similarities among text documents. Our method for similarity search composes a tree of news articles based on content similarity following the related work made by Sahoo [2]. In contrast to mentioned work, we preserve hierarchy especially for the lowest level of the tree where entities are not clustered explicitly, but considered as single entities. This way is our representation ready to produce clusters of similar entities on every level of cluster density. This representation grows incrementally and produces hierarchy, what is effective for growing datasets and dynamically changing domains, because of its logarithmic complexity of storing and retrieving similar articles [4].

In a connection to tree composing we had to solve problem of deep tree form. This form emerges when articles submitted to hierarchy are hardly similar. We use tree balancing to preserve homogeneity of the tree. Advantages of tree balancing are useful in domains where features describing entities are very rare, simple and intersection of features is small. Image tags or keywords extracted from text are then sufficient to compose a tree.

We apply our solution on the domain of news as a part of SMEFIIT project [1], where is the complexity of the method important. Since news are continuously published on mentioned news portal, articles became significantly time sensitive. Therefore, should be service for similar news providing up-to-date. We keep similarity of articles in the hierarchy, so the set of the most similar articles is retrieved very fast. Furthermore are all features of processed articles considered (including newly added).

To prove that our representation is not only fast by also reliable, we evaluated it in comparison with brute force similarity search. We achieved relatively high precision for top similar articles [3]. Articles with lower, but still considerable similarity are

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

omitted, what is consequence of non-existing transitive relations among extracted feature sets.

Finally, we utilized mentioned representation to generate recommendations for users - readers of the news website. It is important to provide recommendations real-time in such a domain. Otherwise, the user loses his patience very easily. The user demands correct and relevant results but keeps waiting very shortly. Complications occur especially in domains where the subject of recommendation is time sensitive and the dataset grows. Our representation utilized for the news recommending solves these issues, because of its low complexity.

One of the advantages is ability to generate recommendations according to content of articles. Another advantage is that recommendations are personalized. The content of an article is mapped on user's interests, which means that articles similar to the articles interesting for user are recommended. We solved complex problem of processing amount of articles to generate recommendations using our effective representation of similarities.

We use incrementally composed hierarchy of similar articles also as a hierarchy of user's interest stereotypes. Each stereotype is a tree node with set of ancestors - similar articles. Since the user reads specific types of articles, we presume that his interest stereotypes could be located in our representation and ordered by its relevance. The ratio of articles read and articles not displayed is a criterion for such a sorting. In a result, the recommendation consists of articles from more relevant stereotypes to cover all of the user's interests. Recommending such a mixture is better, especially because of multivariate nature of single reader and his interest. The content similarity is then effectively used to recommend newly added articles if relevant for specified interests of reader, even with mentioned drawback of omitting less similar articles.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Barla, M., Kompan, M., Suchal, J., Vojtek, P., Zeleník, D., Bielíková, M.: News recommendation. In *Proc. of the 9th Znalosti*, pp. 171-174, 2010.
- [2] Sahoo N., Callan J., Krishnan R., Duncan G., and Padman R.: Incremental hierarchical clustering of text documents. In *CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management*, NY, USA, 2006.
- [3] Zeleník, D.: News Recommending Based on Similarity Relations. In *Student Research Conference 2010*. 6th Student Research Conference in Informatics and Information Technologies Bratislava, April, 2010, Proceedings in Informatics and Information Technologies, STU v Bratislave FIIT, 2009.
- [4] Zeleník, D. and Bielíková, M.: Dynamics in hierarchical classification of news. In *WIKT '09: Proc. of the 4th Workshop on Intelligent and Knowledge oriented Tech.* (WIKT 2009), pp. 83-87, Košice, Slovakia, 2009.

User Modeling, Virtual Communities and Social Networks

Towards Social-based User Modeling

Michal BARLA*

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

Our work deals with enhancing the individual-based user modeling and personalization of adaptive web-based systems with knowledge encompassed within social networks. One of the problems we are aware of in the traditional user modeling is a cold-start problem, when adaptive system cannot provide any meaningful personalization to a new user, for who it does not have any information stored in his or her user model yet. However, such a new user is probably the one who deserves the most some help and guidance provided by the system, in order to get more familiar with its interface, provided functionality and the presented information space itself.

Another issue comes from the closed nature of prevailing user modeling approaches, resulting in only a few "personalized islands" within the whole Web, where majority of content and information services are provided using a failing "one-size-fits-all" paradigm.. The reason why the majority of adaptive approaches is built on the top of a closed corpus domain is that every adaptive system must track user's attitudes (such as knowledge or interest) towards domain elements, often realized in the form of overlayed user model. Closed corpus domain can provide a detailed, often manually prepared and non-changing conceptualization, which is easily used for user modeling purposes. In the case of an open corpus or vast and dynamic domain, we can not track user's relations to all documents or other pieces of information which exist within the domain. The solution is to provide a metadata model and its mapping to domain items, which serves also as a bottom layer for overlayed user model.

Our goal is to contribute to the cold-start problem in an open corpus domain by leveraging social information (such as relationships between a new user and other, already present users or membership of a user in a virtual community). The approach is motivated by social behaviour, which is inherent to the most of human beings. More precisely, the initial estimate of user characteristics is acquired as a weighted combination of characteristics other users interconnected with various types of relationships, acquired from various sources as well as based on common navigational patterns of users. The advantage of such approach is that it produces the standard user

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

model, which can be maintained by well-established approaches to the user modeling and which can be easily used by classical personalization and adaptation techniques.

We evaluate our method in a domain of information research, such as searching for documents in the open information spaces as the Web is or in closed but vast information spaces like digital libraries or electronic newspaper. We use a rather simple keyword-based (tag-based) user model representation coming from various text analysis techniques applied on web-pages visited by the user (see Fig. 1). More, we acquire various relationships between tags by analyzing folksonomies [1], considering co-occurrence and relatedness of tags within web pages, by employing linguistic knowledge from Wordnet, all in order to compare particular user characteristics or even whole user models. Our evaluation platform is an enhanced proxy server capable (apart from logging the information gained by analyzing the traffic) to personalize either user requests (e.g., disambiguate the search keywords [2]) or responses sent from particular web server (e.g., annotate or re-rank search results).

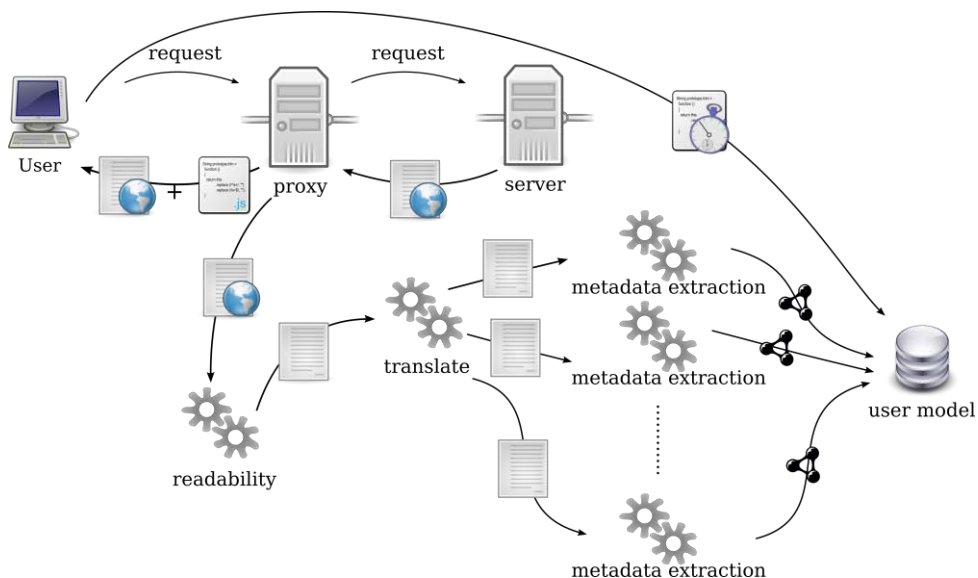


Figure 1. Keyword-based user modeling (evidence layer) in an open corpus domain (Web) through an enhanced proxy server.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Barla, M., Bielíková, M.: On Deriving Tagsonomies: Keyword Relations Coming from Crowd. In *International Conference on Computational Collective Intelligence – ICCCI 2009*, LNCS 5796, pp. 309-320, 2009.
- [2] Kramár, T., Barla, M., Bielíková, M.: Disambiguating Search by Leveraging the Social Network Context based on the stream of User's Activity. In *User Modeling, Adaptation and Personalization – UMAP 2010*, to appear.

Virtual Community Detection in Vast Information Spaces

Marián HÖNSCH*

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

Collaborative filtering is present within current web-based systems in many forms. At the beginning they were mostly either item based or user based, but as the time passed, many hybrid approaches combining several techniques from multiple disciplines emerged. However, the basic idea remained always the same: use past experiences of users to get benefits for an individual.

We analyzed research works related to collaborative filtering in domains of research papers recommenders, personalized learning systems, news recommenders etc.. Used forms of collaborative filtering differ in ways how they model the user, gather users opinions, compute similarity between users or items, define groups of users, present recommendations and many others. We will consider these observations and also other features in our work. As collaborative filtering is always tailored to the specific domain, we chose to focus on recommending news articles. Our main aim is to develop a method for detecting virtual communities among users in order to improve recommendations on a news web portal.

Virtual community detection is an advanced feature of recommender systems, built on the top of basic recommenders. We need to ensure at least a stable method of user modelling and similarity counting. In our case, user model is based on keywords, as we assume that the interests of a user can be projected into keywords extracted from content he reads [2]. To handle synonyms and ambiguous words present in a user model we need to employ additional knowledge about keywords. One approach is to build our own semantic model by processing all the available articles and tracking what words do appear together, another one is to use a semantic service such as WordNet. To handle ambiguous words we use other words from the same article, which help us to reveal the right sense. A much simpler alternative to semantically-based approaches is to do a word count.

Similarly to the representation of user interests, an article is also represented by keywords extracted from it (keyword-based domain model). When a user reads an

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

article the domain model (keywords of the article) is interlinked with the user profile. Keyword model of an article is constructed mainly from headlines and words from the first few sentences. We assume that these parts contain most relevant content.

In a user profile we gather all information that the system knows about the user. There can be many sorts of attributes and inputs. We model these attributes in layers (i.e., synonyms, accepted recommendations, negative feedback, long term information). A crucial task of every recommender system is to gather user opinions. In our system we want to consider feedback in a form that user has read an article, did or did not accept recommendation. To handle a “cold start” problem and the overload of a user profile we fade and forget items. Then two users would have profiles of comparable size, independently of how long they have been actually using the system.

A community is formed by people with similar user models. We cluster user profile so that every cluster represents an interest [3]. This approach is based on the assumption that different categories of articles are represented by different clusters of keywords. The categories correspond to the users interests. Then we group users based on these partitions. Community is represented by an aggregated keyword model extracted from user models of its members. Communities tend to change radically over time and appropriate reaction of the system to these changes is an actual research field. In [1] they propose methods how to count virtual communities with respect to changes over a short time. A newspaper domain is a good example. Only rarely people read the newspaper from yesterday. People also tend to read different categories of articles depending on the time of the day or day of the week (i.e., on weekends we read the Sunday part). Detected communities are volatile so we compute new compound of communities on a daily basis.

To test our approach we recommend articles. We compare the quality of our recommendations against existing collaborative filtering approaches. There are two types of recommendations, one is *novel items recommendation* (i.e., what others read before) and *suggestions to the article the user is actually reading* (i.e., other that read this have also read this). Our main contribution is detecting virtual communities based on user profile clustering, layered user model and community graphs. We deal with fluctuating and time-dependent changes in community detection.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Falkowski, T. and Spiliopoulou, M.: Users in Volatile Communities: Studying Active Participation and Community Evolution. In *User Modeling 2007*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 47-56.
- [2] Lops, O., Degemmis, M. and Semeraro, G.: Improving Social Filtering Techniques Through WordNet-Based User Profiles. In *User Modeling 2007*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 268-277.
- [3] Zhang, M. and Hurley, N.: Novel Item Recommendation by User Profile Partitioning. In *Proc. of the 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology*, Vol. 01, 2009.

Leveraging Social Networks in Navigation Recommendation

Tomáš KRAMÁR*

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

Finding a relevant document based on few keywords is often difficult. Many keywords are ambiguous, their meaning varies from context to context and from person to person. Some words are ambiguous by nature, e.g., a coach might be a bus or a person, other words became ambiguous only after being adopted for a particular purpose. There are also words whose meaning depends on the person who is using them; clearly, architecture means different things to a processor designer than to an architect. Based on the previous observations, we might conclude that using short queries is not a good idea. Unfortunately, this is how we search.

The order of documents provided by the search engine depends on the adopted relevance function; the most widely used search engine today – Google – uses a PageRank relevance function: the more links to a document, the more likely it is to appear at the top positions. This ordering is however not always compatible with user's information needs. We tackle the problem by implicitly inferring the context and modifying the user's query to include it [1].

The overview of the process is depicted in Fig. 1. The user requests a page via proxy (step 1) configured in her browser. Proxy requests the page from the target server (step 2) and extracts the characteristic document features (step 4) – a vector of document keywords, tags from delicious.com and ODP category. Based on user's activity and the extracted features a *social network* is built (step 5), where a weight of an edge denotes a similarity of two users connected by this edge. The stronger is their relationship, the more similar interests they have and the higher is the weight of the edge connecting them. Next, a *community detection algorithm* is run (step 6), to partition the network into clusters of similar users (based on the stream of their activity). The algorithm is designed to take advantage of the weighted relations in the graph and produces overlapping communities, i.e., a user may belong to multiple communities at one time.

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

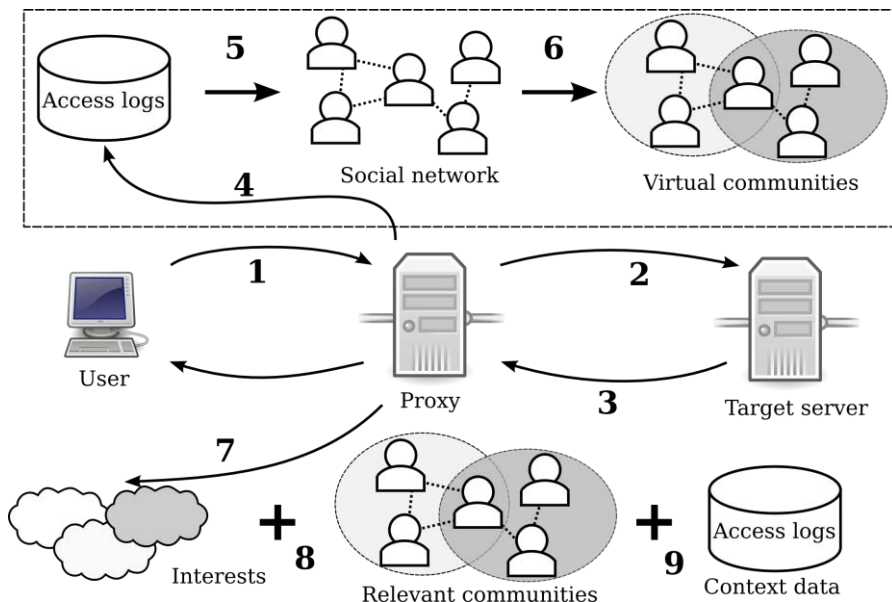


Figure 1. Overview of the query expansion process.

In order to identify the *search context*, we need to capture user's current interest (step 7), which in our case is a set of documents features the user is currently interested in. When we detect that a search has been initiated, the current interest helps us to determine all relevant (i.e., sharing at least one feature) communities (step 8). The top n matching communities are then considered as the *search context* and passed to the final stage of query expansion.

We use two approaches to infer new keywords, each using the data provided by the members of the communities (step 9). *Query stream analysis* follows a simple observation of how we do our searches. When a search query does not return relevant documents, it is redefined. The redefinition continues unless the user finds the information or gives up. We take all queries issued by users from the *search context* and search for query streams where at least one query matches the user's query. The last query is extracted from each successful query stream and used to enrich the original query. A *keyword co-occurrence analysis* is based on analyzing which additional keywords frequently occur with the words from the query in the documents viewed by the users from the current *search context*. The original query is enriched with the top n co-occurring keywords.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Kramár, T., Barla, M., Bielíková, M.: Disambiguating search by leveraging the social network context based on the stream of user's activity. In *User Modeling, Adaptation and Personalization - UMAP 2010*, to appear.

Improving Social Skills Using the Social Exchange Framework

Jana PAZÚRIKOVÁ*

*Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
pazurikova@yahoo.com*

More and more children and youngsters have difficulties with socializing – meeting new people, interacting with them smoothly, making friends and nurturing relationships they have. These children are often uncomfortable to converse with peers; they do not know well how to behave in a relationship, what is desirable and what they should avoid to do. One of the ways how to encourage and help young people to improve their social behaviour is to offer them an opportunity to gain experience without the fear of failure. Therefore, the aim of the project is to develop the method for improving social skills of young people and for learning the appropriate social behaviour within the group.

One mainstream approach in social psychology – the social exchange framework [1] proposes that actors exchange resources between themselves, bringing benefits to one side, while incurring costs on the other. Difference from economic exchange model is “its emphasis on the social structures within which exchange takes place”

The prototype of the method is devised as a strategic computer game. The user reveals a few personality characteristics, then they describe their dream friend, somebody they would like to meet and relate to. The goal of the game is to befriend this person. That can be achieved by virtual socializing, meeting new people, spending time together doing activities, communicating. In the social exchange framework, the user and their virtual friends represent actors, activities and dialogs are resources. Benefits they share are increase in values of needs and relationship characteristics. The cost is simply the possibility of only one transaction at time. There are two kinds of transactions – negotiated and reciprocal.

Negotiated transactions are discussed before they are made. The user starts this transaction by choosing what activity (e.g. going to the pub, camping, playing sports) and with which people. The system activates one transaction over specified amount of time, the activity and people are chosen randomly. Then it is required a short dialog – an offer and acceptance or decline. After agreeing to take part in the activity, the

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

process (Figure 1) emerges. The effect on every need and relationship property is calculated, influenced by activity parameter (e_n – the effect on needs, e_r – effect on relationships), how much the person like the activity (h - hobbies) and random number $\langle -1, 1 \rangle$ representing how successful the activity was.

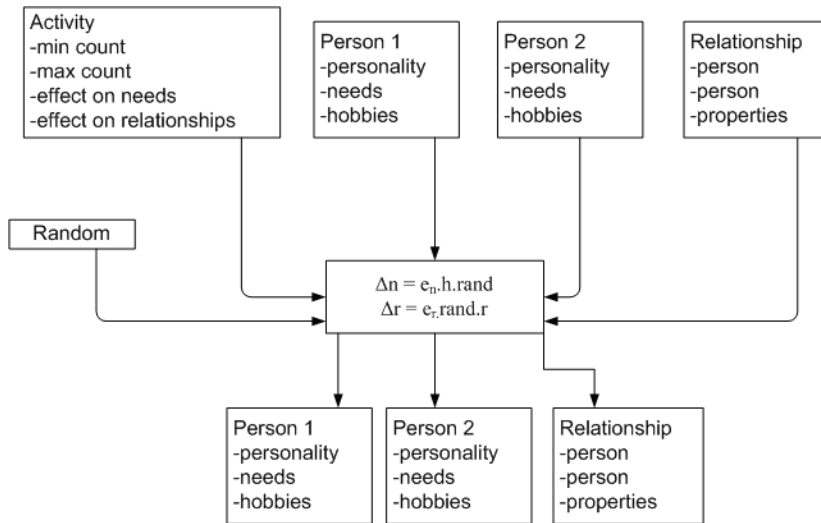


Figure 1. Activity Manager.

In reciprocal transactions, the actor initiates exchanges by performing a beneficial act for another (e.g. having a conversation) without negotiation and without knowing whether the other will reciprocate [1]. Dialogs are text-based, both the user and system choose questions and answers from the oriented weighted graph. They can progress in several directions in order to maximize the positive influence on the relationship.

The preliminary experiment is currently performed, focusing on of how accurately the system resembles the real-world social behaviour. In the main experiment, around 15 young people (aged 13-17) is asked to play the game for few hours and they are encouraged to befriend somebody they do not know well. After two weeks, the changes in their confidence and skills will be collected by a questionnaire. We anticipate the users will improve their behaviour in the first days or weeks after meeting a new person and their self-confidence will rise. The game scenarios are targeted at interpersonal interactions leading to gradually developing the relationship – from unknown youngster or an acquaintance to the best friend. We also plan to evaluate the method within an intelligent tutoring system [2].

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Burke, P.: Contemporary Social Psychological Theories. Stanford University Press, Stanford, 2006.
- [2] Tvarožek, J., Bieliková, M.: Feasibility of a Socially Intelligent Tutor. In: *Intelligent Tutoring Systems 2010*, (accepted), Pittsburgh, Springer, 2010.

Tracing Strength of Relationships in Social Networks

Ivan SRBA*

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

Nowadays we witness rapid expansion of new generation of services known as the Web 2.0. One kind of these services are social networks where people can arrange and express many types of relationships. The strength of these relationships among users can be really different and can rapidly change in the time. If we know about strong or weak relationship intensity between two users we can provide actions which depend on actual context of deployment. For example we can provide adaptive recommendation. Also the user can control evolution of his or her relationships to friends.

We have proposed a method for analysis of the evolution of user's relationships and its evaluation by means of developed web-based application, which approximates the user's relationships with other users in the time. This approximation is based on varied user's activities performed in social networks. Example of this activity is sending a message or uploading common photography. Such activity we denote as a rate factor. The rate factor can be shared among several *sources* of user's activities (social networks). Meanwhile for each source it can have different importance, which is represented numerically by a *weight*. The weight of rate factor expresses relative influence of the rate factor to the final relationship intensity. Weight for each source and rate factor is assigned experimentally. Examples of rate factors with weights determined according analysis and monitoring of users' behaviors are:

- common photography (positive influence, weight 0.13),
- boyfriend/girlfriend relationship (positive influence, weight 0.95).

Not only weight but also the count of all appearances of the rate factor (not only in relationship of two users who are traced) influences final relationships strength. This fact assures that also the frequency of using the social network has effect to result. Partial relationship intensity depends on time and the duration of influence too. To include these effects we differentiate rate factors of single activity, interval activity or unbounded activity. All mentioned effects are included in sequence of calculation.

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

We can use many sources of user's activities to evaluate proposed method. We chose for experiment well known and popular social portal Facebook. We developed web system Intensity Relationship Analyzer & Presenter (see Figure 1) to realize the proposed method. This application uses wrapper to connect to social network Facebook and to data mine rate factors via Facebook API.

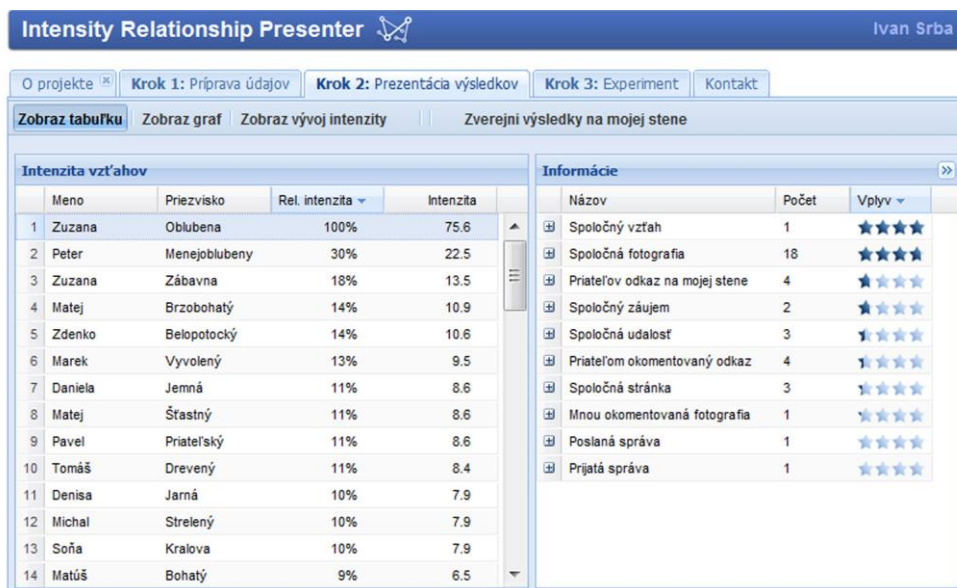


Figure 1. Presenting calculated results in Relationship Analyzer & Presenter.

In the experiment we hypothesize three results: first one is that the calculated intensity will describe similar distribution of user interaction among friends as in [1]. Second one is that the calculated intensity will represent similar evolution of relationships in time as it was described in [2]. Finally, third expected result is that the method will calculate relationship intensity for first ten best friends with 80% reliability.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Wilson, C., Boe, B., Sala, A.: User interactions in social networks and their implications. In: *Proc. of the 4th ACM European conf. on Computer systems*, Nuremberg (Germany): ACM New York, 2009, pp. 205-218.
- [2] Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the evolution of user interaction in Facebook, In: *Proc. of the 2nd ACM workshop on Online social networks*, Barcelona (Spain): ACM New York, 2009, pp. 37-42.

Cooking a Socially Intelligent Tutoring Platform

Jozef TVAROŽEK*

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

Educational technology has moved from tools that automate repetitive tasks such as grading tests to intelligent tools that provide personalized instruction. Intelligent tutoring systems give direct one-on-one instruction and feedback to students during problem solving. Students, however, often engage in off-task behaviors that diminish learning gains. Maintaining sustained student motivation is therefore important for effective learning, yet providing motivational feedback is often at odds with cognitive scaffolding. Various approaches for improving motivation have been proposed. The affective support for learning seems difficult to realize and therefore remains limited, while the narrative-centered story-based approaches are not directly applicable to traditional domains such as mathematics and computer science.

In this work, we propose to enhance computer-supported learning systems with a virtual conversational agent that employs socially intelligent dialog strategy to increase student motivation and guide students to instructional activities appropriate for their current context [1]. The activities (problem solving, course notes) are augmented by social features (synchronous group work, annotations, asynchronous discussions, etc) which are subsequently used by the tutoring agent to facilitate a socially encouraging learning path for individual students (Figure 1). The dialog strategy is induced by reinforcement learning method on Wizard-of-Oz natural language data collected online with the help of domain experts.

The tutoring agent does not directly participate in learning activities with students and its dialog capability can be developed separately from the domain content. In the process, we redesigned numerous techniques used in pseudo-tutor learning environments and tailored them to the socially intelligent tutoring context. The problems for students to solve are scripted in a template language that generates complex problems with hints, while optionally being semantically adapted to student's individual preferences. Decisions are made on the server, natural language dialogs and collaborative features within the client's interface are synchronized near real time.

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

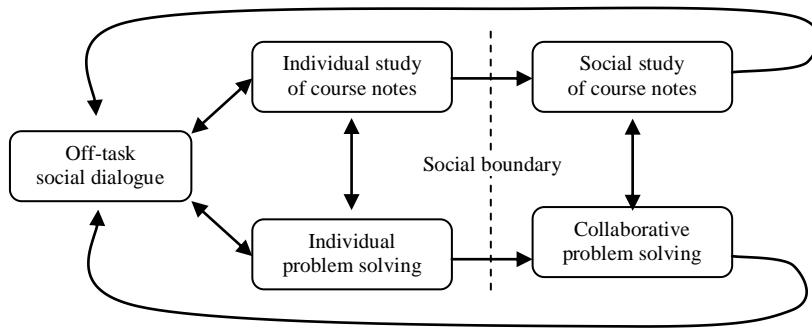


Fig. 1. Types of learning opportunities with admissible transitions (arrows), which are facilitated by the tutoring agent.

In present state of the platform, the tutoring agent influences the transitions between different learning activities by a set of rules that can recommend a good course of action for the student at any given moment. For example, when an examination is imminent, the student is advised to work on exercises from a similar problem set. When a student (or a collaborative group of students) does not seem to understand the most basic of facts during problem solving, she (or the whole group) is redirected to the corresponding course material. Rules for facilitating the transitions get more involved when the social boundary is crossed (Figure 1), as other people are a valuable resource with which the tutoring agent can “negotiate”. It is not possible for a human student to cross this boundary at will, and the transition *must* be facilitated by the tutoring agent. For example, when a student repeatedly demonstrates incompetent behavior (in terms of social/task abilities) the tutoring agent can refuse to put him in a group that would probably only impair the work of others due to his unfit behavior.

We apply these methods to increase motivation and learning gains in a learning system for middle school mathematics. Some 54% of students engage with the tutor quite naturally, while the others seem to be require more tangible benefits. Students in the socially engaged group liked the system and the tutor more, and they were also more successful in solving problems within the tutoring environment. The reinforcement learning strategy lets us create a working dialogue capability rapidly, without tedious dialogue scripting. We envision that advanced users (students and teachers) can put expertise in their own virtual presence, adding new virtual tutors capable to directly help others in learning.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Tvarožek, J., Bielíková, M.: The Friend: Socially-Intelligent Tutoring and Collaboration. In: *AIED2009*, IOS Press, Brighton, UK, 2009, pp.763-764.

Domain Modeling, Semantics Discovery and Annotations

Enhancing the Web Experience by Freely Available Metadata

Peter BUGÁŇ*

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

Much of information available on the Internet can be easily understood by people, but not at all by computers, which present the actual content of the web pages. The problem is often solved by adding machine-understandable metadata. In addition to “lightweight” metadata in the form of ‘meta’ html tag, which is devoted mainly to the web search engines, we know also more formal metadata defined within Semantic Web initiative. The vision is a Web, in which the meaning (semantics) of information and services is shared across the web applications. Metadata contained within the Semantic web create so called Linking Open Data cloud, that is growing every year, although this growth is slow. Only interesting Semantic Web applications can persuade users to participate more actively in the Semantic Web initiative, which would hopefully increase the growth of the available metadata until a self-strengthening threshold is reached.

When we are reading articles on the web, e.g., online newspapers, we are often unable to understand it quickly (especially longer sections) and we need to re-read it several times. Therefore, we underline the most relevant keywords on the web pages and annotate them with automatically generated content. Underlined keywords should help readers to quickly recognize the main words of an article while the actual content of annotations should allow for better understanding of the underlined word.

Process of annotation consists of these parts:

1. Web document processing – as we annotate only the content part of a web document, we can omit unnecessary parts such as menus, advertisements and focus only on segments which are of user's interests.
2. Lookup of words, that should be annotated
3. Annotations filtering
4. Creation of annotations content
5. Visualisation of annotations

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

For keyword search we use OpenCalais web service, which extract instances like well-known people, companies, organizations, geographical indications (states, cities, rivers ...) from the given text or URL. The important feature of OpenCalais is that it provides also metadata for the selected keywords such as additional facts about retrieved keywords (i.e., that Robert Hughes is a reporter from BBC) or relevance of the retrieved keyword according to the rest of the document. The metadata help us to reduce irrelevant results in search for content of annotation.

After keyword extraction, we decide which words are to be annotated – which are in our case the most relevant words found in the article. As we already mentioned, the relevance is acquired from OpenCalais.

Content of annotations is tailored according to type of a particular article. For instance, if an article is about sport, the annotation will contain information relevant to the sport topic. This allows us to provide different annotations of the same word used in different context. A word ‘Lisbon’ within a political article would be annotated with information about mayor of the Lisbon, in case of a sport article, the annotations will contain information about sport events which were or will be hosted in Lisbon or about athletes coming from this city.

Annotation could contain:

- Textual information (e.g. population of a country),
- Hyperlinks to the external resources (e.g., reference to a photography of a monument in the city, or reference to the article on Wikipedia),
- Reference to other entities (e.g., when information in the note is that, Bratislava is the capital of Slovakia, by clicking on Slovakia, we will view information about that instance).

As we have indicated, the information source of our annotations is Semantic web. For geographical instances we use specific resources like Dbpedia.org, Wikitravel.org and Factbook, which we evaluated to be the most relevant resources for this kind of information. When we search for metadata about people, we use semantic search engine Sindice.org.

The last part of the annotation process is actual visualization of annotations. We decided to visualize annotated words by underlining them. There is a possibility to setup different colours for different types of instances. Clearly, it is not necessary to underline and attach an annotation to each and every occurrence of an instance within a document, which would for sure overload the readers and would not be very comfortable. Therefore, we underline only first occurrence of the word. The contents of the notes are shown in tool-tips, which appear after hovering the mouse over the underlined word.

From the technical point of view, our solution is based on enhanced proxy server being developed at Institute of Informatics and Software Engineering (peweproxy.fiit.stuba.sk).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

Text Understanding and Analysis

Martin JAČALA*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
`jacala06@student.fiit.stuba.sk`

With increasing amount of online content available we need tools and methods for efficient visualization, exploration and navigation in this data. There are many sources of textual information providing us with high-quality, human written text on various subjects, such as events, celebrities or politics. Most of such published content comes from newswire portals around the world, i.e. The New York Times, BBC News, Reuters, etc. Another source of interesting information are personal blogs, where many people express their opinions on variety of topics, online discussions or various comments left by random web page visitors.

If we closely study the daily published information, we notice the individual textual artefacts are often related, even if the connection between them is not immediately obvious. The news captures the same person in various situations, interacting with others or attending certain events. We can gather a lot of interesting knowledge by identifying and tracking the person across various textual resources. Such knowledge is not easily obtained just by few texts a day as we probably do – automatic analysis using dedicated methods will fit in better.

The main goal of our work lies in entity recognition and identification in open web content. One of the challenges we face is the open, constantly changing world that needs an adaptive method able to continuously improve itself and learn to identify new, previously unknown entities without the need to define them first by hand. The enormous amount of input data also requires feasible method in terms of computational and storage complexity.

Currently, we study some Named Entity Recognition algorithms based on various principles, such as Support Vector Machines or Conditional Random Fields [1, 3]. In our study, we evaluate the current state-of-the-art implementations [2] and compare them to plain, approximate text matching algorithms with list of gazetteers.

Additionally, we are exploring the possibilities of using freely accessible content, such as Wikipedia in the process of training the evaluated algorithms. Since the English Wikipedia project contains more than 3.2 million articles, it becomes a viable source of training and testing data. The wiki pages define by itself a metadata for each

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

article up to some degree. We can easily extract set of entities of given category (e.g. persons) from Wikipedia's disambiguation or list pages (e.g. composers of 20th century).

The Wikipedia has been already provided with semantics; DBPedia project provides freely accessible, semantically annotated datasets. We are evaluating feasibility of the DBPedia's "instancetype_en" dataset providing more than 1.1M objects, where each object has one or more assigned categories from the ontology. There are total of 200 object types defined by the ontology (such as Thing, Place, Person, Organisation, etc.). However, the provided datasets does provide ontology and hierarchy of concept instances, but such data are unusable for training of machine learning or stochastic based algorithms.

Another interesting project is the "Yahoo! Semantically Annotated Snapshot of English Wikipedia (SW1)" [4]. As the name suggests, this corpus contains semantically annotated snapshot of English Wikipedia from 2006. This dataset contains more than 1.5M entries and 20.3 unique named entities in multi-tag format. This makes the data feasible for training, evaluation and comparison of various algorithms.

We would like to build a web based presentation layer, where users can look up discovered relationships between various entities. The interface will involve social feedback, where users can participate, correct mistakes, and further improve the extracted data and the process itself. Additionally, we can gather statistics and perform additional analysis on the input data, such as type and category of the article, polarity, subjectivity, writer's opinion, etc. We hope that by providing users with different views on the same information we will attract user interest and increase the chance of receiving human-generated feedback.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Finkel, J.R., Grenager, T., Manning, Ch.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.
- [2] Marrero, M., Sánchez-Cuadrado, S., Lara, J. and Andreadakis, G: Evaluation of Named Entity Extraction Systems. In *Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*, pp. 47-58.
- [3] Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of the 13th Conference on Computational Natural Language Learning (CoNLL 2009)*, pp. 147-155.
- [4] Zaragoza, H., Atserias, J., Ciaramita, M., Attardi, G.: Semantically Annotated Snapshot of the English Wikipedia v.1 (SW1) In *Proc. of the 6th International Language Resources and Evaluation (LREC 2008)*, pp. 2313-2316.

Does HTML Tags Improve Results of ATR Algorithms?

Milan LUČANSKÝ*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
lucansky06@student.fiit.stuba.sk

Our work focuses on mining relevant information from websites. We propose a method combining two different approaches to keyword extraction. The first one is used in a field of text mining. Automatic Term Recognition (ATR) algorithms are used to retrieve relevant keywords from plain text. Measures they use are based on different characteristics of text, e.g. statistical, probabilistic or both. The second approach is used to retrieve information from World Wide Web. It is important to have relevant keywords in HTML title tag, or to have external links to our page, because all these factors increase our page rank, which ensures better place in search results when searching for such keyword. We want to make the best of this information and support the extraction of keywords by analyzing semantics of HTML tags.

To achieve better results in keyword extraction we combine aforementioned approaches. Our research aims to retrieve descriptive words from different websites. Website usually consist of many (from tens to hundreds) web pages. First, we download the pages. Then we extract plain text from them. Plain text file is ready to be processed with ATR algorithm, which returns keywords. Each keyword has its own weight which denotes its relevance within the corpus (in our case within the website). Besides text processing we also consider HTML tags used to represent web content. We believe different tags flag different semantics. Similarly to SEO (Search Engine Optimization) approaches, we investigate semantic potential of the following tags[2]:

- Title tag,
- Heading tag,
- Anchor tag.

Title concludes whole webpage content – therefore we assume it contains exactly those words we are looking for. In an ideal case, *heading* resumes certain parts of webpage, so it also contains descriptive words. *Anchor* text from the rest of pages (but within the website) should contain keywords as well. There is a different chance to find keywords

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

in a title tag and a heading tag. To address this issue, words present in title tag have greater weight than words present in heading tag. A HTML tag weight we refer to as *TagIndex*. It is a number from interval $(0,2)$. Our aim is to find appropriate *TagIndex* values for different tags experimentally in order to extract the most relevant keyword from a web page.

To combine two different weights of words extracted with ATR algorithms and acquired from HTML tags, we multiply the weights as follows:

$$w_{i,j}' = w_{i,j} \cdot \text{TagIndex} \quad (2)$$

where $w_{i,j}$ is a weight obtained from selected ATR algorithm. As an example of ATR algorithm, we utilize well-known *tf-idf* measure:

$$w_{i,j} = \text{tf}_{i,j} \times \text{idf}_i \quad (1)$$

where $\text{tf}_{i,j}$ represents frequency of a term t_i in document d_j (term frequency) and idf_i relates to the number of documents within a corpus containing the term t_i (inverse document frequency).

We consider several possibilities how to evaluate our method. In the first experiment we use wikipedia corpus [1] containing articles about animals and wikipedia page-to-page link database¹. We acquire keywords using *tf-idf*, *Weirdness*, *C-value*, *Glossex* and *Termex*. Evaluation of the results will be done a posteriori, when a group of judges decide, whether the top N th proposed term is a part of the domain or not. After that, the precision will be computed. Subsequently we will repeat the experiment using the same five ATR algorithms, but with a modification in the form of multiplying the final weight by *TagIndex* for every candidate term. This modification should improve weight for those candidate terms that are present within one of three mentioned HTML tags. Evaluation will be done by the same group of judges at the same conditions. Finally, precision will be computed and the results will be compared to the results of ATR algorithms without modification.

While previous evaluation involves only wikipedia corpus, it is interesting to test the method on several different domains. Input will be a website containing pages about specific topic (e.g. cars, programming, baking ...). Corpus from different websites will be evaluated the same way as in the first option.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Zhang, Z., Iria, J., Brewster, Ch., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC08*, (2008).
- [2] SEOMoz. 2009. *Searching Engine ranking factors* (online), (visited 9.4.2010). <http://www.seomoz.org/article/search-ranking-factors#ranking-factors>

¹ <http://users.on.net/~henry/home/wikipedia.htm>

Collaborative Tagging for Word Relationships Mining

Tomáš MICHÁLEK^{*}

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
`michalek07@student.fiit.stuba.sk`

As amount of data on the internet is growing faster and faster, information and data are becoming hard to discover and explore. It is nearly impossible or possible in a very small scale to process unstructured text, photo or video content. Nowadays searching is heavily based on looking up a key words. It's like playing darts and hoping to hit the right words. Even then we won't get the information we are looking for, instead we get an amount of links referring to resources (articles, pages, etc.) where the desired information is wrapped up with a lot of data, which are at the moment useless for us. So even answering simple question or looking up for single fact can be a very long process.

We can split this content into several basic categories by its format. Structured or mostly structured text, which can be relatively easily processed automatically in a very large scale like RSS feeds.

Second category are resources with unstructured text data. This resources often contains also a lot of text vapid to user which can distort a search results and become very misleading. Delicious.com is online bookmarking web service. Users are using this system to bookmark interesting content they found on internet or to discover another based on their interests. Users can add to every bookmarked link tags to make it more discoverable and searchable. As many people mark resources we can extract more information telling us something about relations between this marks and resources.

Third category is content that is very hard to process automatically, like photo and video content. Pages like youtube.com or Flickr.com have to deal with this problem. Without people marking this content we won't be able to search in it.

In my bachelor project I am focusing on these systems, describing methods of tag relations mining and behaviour and nature of these systems. Despite of the content we are tagging all these systems have common structure creating a tripartite graphs. One

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

tagging instance is an edge connecting user, resource and tag. Another common feature of these systems is that tag distribution between resources follows a power law.

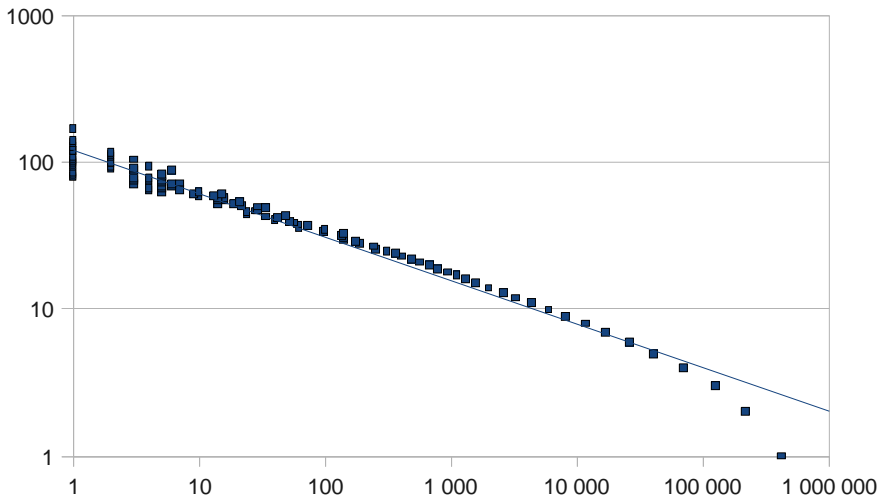


Figure 1. Tag distribution on resources. Dataset from delicious.com.

On Figure 1 is displayed tag distribution of tags. On the x axis number of resources; on y is marked how many tags it contains. Both axes are in logarithmic scale. This means that 90% of resources have between one and five tags. If we are able to discover relations between tags, we can extend these resources by new words making content more accessible to users.

But there can also be another motivation for word relationship mining. If we know a relations between words we can better understand content, meaning that we can process unstructured text nearly same way as structured. This allows extracting from text not just key words but also facts and information. This makes a huge difference in creating a search query. Instead of trying to hit a key words and scrolling through a great amount of links we can ask a question same way we are asking another person in real life and just get an short and factual answer.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Halpin, H., Robu, V., Shepherd, H. The Complex Dynamics of Collaborative Tagging. WWW 2007 / Track: E*-Applications.

Exploring the Possibilities of Annotations in Learning Content

Vladimír MIHÁL*

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

Modern web based educational systems employ principles of Web 2.0, utilizing tools allowing students to actively contribute to the learning content and to personalize the content and presentation to them. One of the techniques allowing contribution to the learning content is user annotation. Besides the contribution and active participation of the student, annotations provide other several possibilities of their utilization within the educational domain.

Annotations inserted into the learning document primarily serve as learning content based feedback or a contribution to the content. They can be short comments, personal notes of a student, reports of errors or mistakes in the learning text or questions related to the document, increasing overall information value. Obvious benefit for the students is that a greater amount of the relevant information is available to them and the whole information is organized in similar manner as the original learning content (since annotations are tightly bound to the content).

Content annotations also introduce the possible interactions between students. If the students are not limited just to commenting the learning content, they would also respond to the annotations of other students. It will eventually lead to the conversations and discussions, which will be embedded in the annotations and spread through the whole learning content. Discussions will provide to students actual information about upcoming events or current activities. The concept of short discussion is similar to forums or microblogs, which are already well known and used amongst students, what will partly serve as motivation. We presume that more discussed parts of the document are more relevant and interesting to students, what we can possibly use to improve recommendation of the content [1].

By creating annotations, the student provides us information about his current activities from which we can track his progress within the course. We can compare the activities and the progress of students with each other and recommend the learning content using this data along with the content based recommendation. It would be also

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

interesting to combine this information with the results of mid-term tests to discover the parts of the document, which were studied by students, who were more successful on the test, and which parts by less successful ones. We can potentially find most helpful parts of the content and recommend them to students, helping them to achieve better results.

Another possible use of user annotations, similar to previously discussed idea, is a bookmarking within the learning content. Students using annotations as bookmarks will select the parts of the textbook important to their current activities or future tasks. Such annotations can be used for browsing the content [2]. Discovering relations between bookmarks and actual tasks of a student is another possible subject of research. The association of tasks with annotations explicitly by students is unlikely to be used, since it requires additional effort from students. Using this data we can help weaker students with tasks, recommending them content, which helped other students.

Besides the comments written by students, annotations also provide another important data – selected fragments of the text, which are associated with the annotations. Since students select certain fragments to annotate them, these fragments of the learning document are presumably significant for them. When numerous students select the same fragment, it indicates that the selected fragment can be generally important within the context of the document. Such metadata can be used for searching in the course or to improve results of methods for discovering concepts from the text. Selected words can contain concepts or even be concepts themselves.

Using annotations to discover concepts and relations (and consequently creating a domain model of the course) is not limited just to the students' annotations. It can be also utilized as an interface used by an expert to select occurrences of concepts in the text and then generate the domain model. Selecting keywords within the text requires far less effort than manual specifying of the concepts and relations thus it can be useful for convenient creation of the domain model.

Primary goal of our current work is to explore possible uses of content annotations within the learning content and determine the most promising ideas of their utilization. Our aim is to reduce effort for the students and teachers, provide additional value to the students and to make learning content presentation more interactive and attractive to students.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z.: Optimizing web search using social annotations. In *Proceedings of the 16th international Conference on World Wide Web* (Banff, Alberta, Canada, May 08-12, 2007). WWW '07. ACM, New York, NY, 2007, pp. 501-510.
- [2] Yudelson, M. and Goreva, N.: Providing social navigation within annotated examples. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia* (Pittsburgh, PA, USA, June 19-21, 2008). HT '08. ACM, New York, NY, 2008, pp. 255-256.

Online Gathering of Information from Text Sources

Štefan SABO*

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

The project I am working on is dealing with gathering of information from text sources available on the web. The information wouldn't be gathered from databases, or already processed sources, but from random sources presented in natural language. These sources may come in many forms, like consumer forums, articles, user comments or various blogs. The aim of this project would be to create a system, able of locating, harvesting and subsequently analyzing such sources. This analysis would be able to provide us with information about various products, events, etc. The first idea would be to implement opinion mining about consumer products as cell phones, notebooks, or other electronics. Although this is just a first notion and could be subject to a change further down the road.

The rough schema of the system is presented on Fig.1. The system receives queries from users, processes them and produces results, which are returned back to the user. The data is gathered by the means of web crawling bots and stored in a database. After being gathered, the data is analyzed and the results are produced. The individual steps presented in this diagram don't need to be performed in this order. The returning of a result will always be preceded by issuing a query. However, the system-internet and the system-database interaction need not occur only after issuing a query. If we can specify, what general kind of data needs to be gathered before the query is issued, the data may be gathered on-the-fly all the time. Also certain parts of analysis can be done independently, without the knowledge of the query. Examples of such steps would be the author analysis, parsing, or syntactic preprocessing.

Similar systems dealing with opinion mining are no new idea [1, 2]. However, this system would differ in the means of information locating and gathering. To achieve reasonable reliability we need to analyze a rather wide database. In order to do so, I would like to utilize the social insect model, namely the metaphor of bees gathering food for their hive. The bees would be implemented as web-crawlers.

* Supervisor: Anna Bou Ezzeddine, Institute of Informatics and Software Engineering

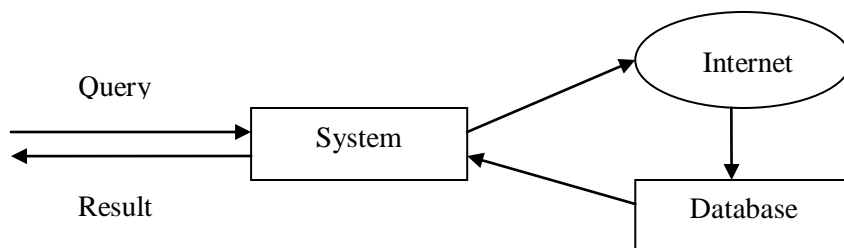


Fig. 1: Schema of the system interaction.

These crawlers would be able to search the web and download pieces of information for further processing. The advantages of this approach consist in the ability of web crawlers to make decisions and prefer the more suitable sources over the ones that are less interesting. The new information tends to be linked to more often than the older one. So the web crawlers searching the web are more likely to localize such a source, just like a bee stumbles upon a quality food source. These sources would subsequently be harvested, meaning that the text information would be downloaded to a database and ready to be analyzed.

The analysis would provide us with opinions about chosen products. After the implementation, series of tests would be performed. Aim of these tests would be to compare different strategies, or parameters and find out, how good would be produced.

Acknowledgements. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010.

References

- [1] Spangler, S. et.al.: COBRA - mining web for COrporate Brand and Reputation Analysis, In *Web Intelligence and Agent Systems*, Vol. 7, IOS Press, Amsterdam, 2009, pp. 243-254.
- [2] Cai, R. et.al.: iRobot: an intelligent crawler for web forums. In *Proceeding of the 17th international conference on World Wide Web*, ACM, New York, 2008, pp. 447-456.

Collaborative Acquisition and Evaluation of Question by Learners

Maroš UNČÍK*

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

Web evolution allows us enormous access to information of all kinds and applies in all spheres. But this evolution also evoked a flood of information. Finding and classifying certain information is becoming increasingly difficult. In relation with using the Web as fast and flexible tool to support education it arouses a need of approaches for simplify searching and presentation of information.

Currently many educational web-based systems publish not just a static text, they do much more. With adaptation to the individual needs of students they support learning, communication and advice. The common way to keep content up-to-date is content enrichment such as adding annotations [1]. We present a method for content enrichment. We design an approach for adding new quality and interactive content to learning materials and integrate students as active parts of learning process. This potentiates to get the quality content, which is useful for students themselves and their peers, as reviewed in [2].

The very important parts of any learning material are questions, which summarize the keys facts of the education materials. Even though there are some approaches to automatic extract relevant questions from the educational texts, the quality of extract questions is still low. Creating questions by an expert is extremely time-consuming and from the expert view it is often difficult to specify the difficulty of questions, to choose relevant questions as well as the right wording of the questions.

We add questions with collaborative aspect. Firstly, the questions are added by students, who also participate on reviewing and authoring. Secondly, their peers can answer the questions and thus is the interactivity ensured. We believe that proposed approach leads to a new and quality content and improves learning process.

As in our method the questions are added by students who are just learning particular topics, we should pay attention to their quality. It is necessary that these questions have a similar level of quality as the expert's questions. Our idea is to evaluate the quality of questions based on the explicit feedback of students in

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

conjunction with actions that students do in the educational system and also on the evaluation of expert (a teacher). For this purpose we designed the model for the question rating and the model for the rating of the user ability to create questions.

The entire method is divided into four steps, which are not necessary followed in the order as shown bellow, but are closely related:

1. Adding a question.
2. Answering a question.
3. Rating the ability of the student to create questions.
4. Rating the quality of the question.

The first and second steps are short-term from user's view, the user adds or answers questions. The third and fourth steps are a long term processes and it takes time to be able to calculate final values as we need some minimal amount of questions and answers inserted for proper rating. Rating of questions derives from the explicit rating of questions by students and implicit rating of questions, based on the actions of students in the system (the user rating model).

We include also a competitive element of motivation in form of gaining points in overall assessment. Students play a simple game based on receiving reward. Students do not know the exact procedure for the allocation of points, their job is to find a tactic that brings them the greatest number of points (in principle, adding quality questions and rate questions like others).

To evaluate our approach, we have designed and implemented a software component for adding questions, which is part of educational web-based framework ALEF [3]. We provide experiments in domain of functional and logic programming. The framework is based on the concept of system FLIP [4], but adds openness, flexibility and modularity. We plan to provide experiments at this adaptive system in the real study process. The experiments will be run for one week and the students will create the questions related to educational materials for programming language Prolog.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

- [1] Agosti, M. Ferro, N.: A formal model of annotations of digital content. In *ACM Transactions on Information Systems*, 2007.
- [2] Hsiao, I.-H., Brusilovsky, P.: Modeling Peer Review in Example Annotation. In: *Proc. of 16th International Conference on Computers in Education*, Taipei, Taiwan, 2008.
- [3] Šimko, M., Barla, M. and Bielíková, M.: ALEF: A Framework for Adaptive Web-based Learning 2.0. In: *Key Competencies in the Knowledge Society at World Computer Congress 2010 [submitted]*. Brisbane, Australia 2010.
- [4] Vozár, O., Bielíková, M.: Adaptive Test Question Selection for Web-based Educational System. In: *Proc. of SMAP 2008 - 3rd Int. Workshop on Semantic Media Adaptation and Personalization*. Prague, CR, 2008.

Web Engineering, Semantic Web Services

QoS Aware Semantic Web Service Composition Approach Considering Pre/Postconditions

Peter BARTALOS*

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

Web services present a topical research area with lot of attention. One part of this research aims to propose solutions to automatic composition of several web services into workflows bringing a utility which cannot be provided by single service. The desired goal of such composition is described in the user query. The automatic web service composition showed to be a challenging task [5]. The research of service composition in last years tends to focus on issues related to QoS [1, 3, 4], pre-/post-conditions [2, 6], user preferences (e.g. soft constraints), service selection considering complex dependencies between services [7]. Our work deals with the effectiveness and scalability of service composition aware of QoS, and pre-/post-conditions.

Our approach is based on a lot of preprocessing done before we are responding to user queries. During it we create data structures which are used to quickly answer the query. The most important is that we evaluate which services can be chained, i.e. which services produce data and have a post-condition required by the other services. This can be done without knowledge of any query which will be processed. The next important issue is that we precalculate different characteristics of the post-conditions to make fast evaluation whether the service produces condition satisfying the goal condition. The problem still remaining is to i) select the services producing the required outputs, and state (services appearing as final in the workflow), and ii) evaluate which services can be used, since they have provided inputs and how they interconnect (design of the data-/control-flow). The latter is significantly affected also by the selection of the service combination with the best aggregated QoS.

To find the services directly producing the required goal (final services of the workflow) a two step process is performed. First, we find services producing the required outputs. This is done in constant time. Second, we filter these services based on post-conditions. Here we use the precalculated characteristics of web services' post-

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

conditions. We compare them with the characteristics of the condition defined in the user query. The precalculated characteristics are stored in data structures supporting fast evaluation if the post-condition of some service implies the condition defined in the user query.

The design of the data-/control-flow is based on two processes. The first selects services which can be used because they have provided inputs. The second selects services which cannot be used because they do not have provided all inputs. The second process is not necessary to find a composition. It is used only to faster the *select usable services* process, which is necessary. The improvement in terms of lower composition time caused by application of *select unusable services* is in more than one order of magnitude. Our experiments showed that the combination of these processes saves a lot of computation time when looking for a suitable composite service.

After we have found the services directly producing the required goal and designed the data-/control-flow of the composite service, we get a prescription based on which we execute services to produce the user defined goal. The data structures used during the service composition are designed to be easily updateable in the case that new service becomes available or some service is removed. This is important to support fast reaction to the dynamic changes of the web services environment.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

- [1] Alrifai, M., Risse, T., Dolog, P. and Nejdl, W.: A scalable approach for qos-based web service selection. In *Service-Oriented Computing ICSOC 2008, Workshops*, Berlin, Heidelberg, Springer-Verlag, 2009, pp. 190-199.
- [2] Bartalos, P. and Bielikova, M.: Fast and scalable semantic web service composition approach considering complex pre/postconditions. In *WSCA '09: Proc. of the 2009 IEEE Congress on Services*, Int. Workshop on Web Service Composition and Adaptation, IEEE CS, 2009, pp. 414-421.
- [3] Bartalos, P. and Bielikova, M.: Semantic web service composition framework based on parallel processing. In *IEEE Int. Conf. on E-Commerce Technology*, 2009, pp. 495-498.
- [4] Huang, Z., Jiang, W., Hu, S. and Liu, Z.: Effective pruning algorithm for qos-aware service composition. In *IEEE Int. Conf. on E-Commerce Technology*, 2009, pp. 519-522.
- [5] Kona, S., Bansal, A., Blake, B., Bleul, S. and Weise, T.: A quality of service-oriented web services challenge. In *IEEE Int. Conf. on E-Commerce Technology*, 2009, pp. 487-490.
- [6] Kona, S., Bansal, A., Blake, B. and Gupta, G.: Generalized semantics-based service composition. In *ICWS '08: Proc. of the 2008 IEEE Int. Conf. on Web Services*, IEEE CS, 2008, pp. 219-227.
- [7] Yu, H.Q. and Reiff-Marganiec, S.: A backwards composition context based service selection approach for service composition. In *IEEE Int. Conf. on Services Computing*, 2009, pp. 419-426.

Towards Semi-automated Design of Enterprise Integration Solutions

Pavol MEDERLY*

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

Our aim is to reduce the effort needed to create integration solutions, i.e. specialized software systems that enable cooperation of disparate business applications (services), either within an enterprise or across enterprises. We are trying to achieve this goal by fully or partially automating the process of *technical design* of such solutions.

More specifically, the technical design of integration solutions includes choosing the overall architecture style (e.g. the one based on Process Manager or Pipes and Filters pattern), selecting an appropriate integration platform (e.g. a concrete Enterprise Service Bus, or ESB, product), and designing the solution components so that all the functional and non-functional requirements are met. Functional requirements are typically given by business analysts in the form of an abstract description of the workflow and data transformations the solution has to implement. On the other hand, typical non-functional requirements cover areas of availability, reliability, performance, security, logging and auditing, and maintainability of the solution, as well as ensuring compatibility with message formats, protocols and application programming interfaces (APIs) used by individual systems being integrated.

We concentrate primarily on *messaging-based integration solutions*, i.e. those that utilize a standardized message-oriented middleware infrastructure for communication between solution components. So far we have developed two methods that create designs of such integration solutions, taking into account a subset of the non-functional requirements categories described above – namely throughput, availability, logging, message ordering, choosing correct message format and content, and duplicate message handling. The methods are largely platform-independent and describe the solutions they create using enterprise integration patterns [1], the de-facto standard language in the area of messaging-based integration.

The first method [2] uses an *action-based planning* approach, representing properties of message flows present in the integration solution as the planner's states of the world, and potential solution components (i.e. business services as well as

* Supervisor: Pavol Návrát, Institute of Informatics and Software Engineering

integration services ensuring e.g. message format conversions, load balancing, fault tolerance, or message logging) as planning operators. Each operator has a set of preconditions (checking presence of specific literals in the state of the world) and effects (removing and adding some literals from/to the state of the world). These preconditions and effects correspond to laws inherent in the domain of messaging-based integration solutions. Our method encodes an integration problem as an action-based planning problem, executes a planner, and then interprets the plan found by the planner as a description of the integration solution.

The second method achieves similar goals using *constraint programming*. It encodes an integration problem as a Constraint Satisfaction Problem (CSP), representing properties of business and integration services and message flows between them as CSP variables, and the laws of messaging-based integration as constraints over these variables. Then it executes a CSP solver and interprets the solution found as a description of the integration solution (see Fig. 1).

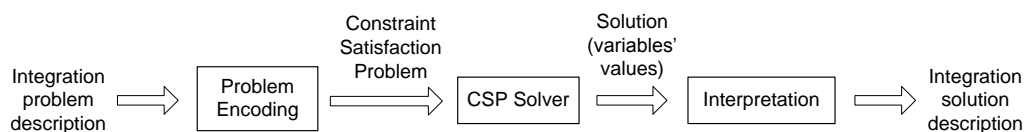


Fig. 1. A schema of the second method.

Current results show that these methods, especially the second one, are able to find solutions for practically-sized integration problems within reasonable time (seconds). Yet what is needed is to broaden the set of design issues tackled by the methods: In the future we would like to deal with the problem how to place and manage logical data elements in messages, how to resolve security issues, how to support diverse message transport protocols, and how to deploy solution components into ESB containers. We also plan to more precisely define the notion of solution optimality. We expect that in order to achieve these goals we would have to involve the developer in the solution-finding process. Finally, we plan to provide a code-generation module that would generate partially or fully executable code for selected integration platforms.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency (contract No. APVV-0391-06) and by the Scientific Grant Agency of Republic of Slovakia (grant No. VEGA 1/0508/09).

References

- [1] Hohpe, G., Woolf, B.: *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Pearson Education, Inc., Boston, MA (2004).
- [2] Mederly, P., Lekavý, M., Závodský, M., Návrát, P. Construction of Messaging-Based Enterprise Integration Solutions Using AI Planning. In: *Preprint of the Proc. of the 4th IFIP TC2 Central and East European Conf. on Software Engineering Techniques*, CEE-SET 2009, (2009), pp. 37-50.

Index

Barla, Michal, 43
Bartalos, Peter, 73
Benčíč, Anton, 13
Bugáň, Peter, 57
Holub, Michal, 27
Hönsch, Marián, 45
Jačala, Martin, 59
Kompan, Michal, 29
Kramár, Tomáš, 47
Kuric, Eduard, 7
Labaj, Martin, 31
Lohnický, Michal, 9
Lučanský, Milan, 61
Martinský, Ladislav, 11
Mederly, Pavol, 75
Mészáros, Roman, 13
Michálek, Tomáš, 63
Michlík, Pavel, 33

Mihál, Vladimír, 65
Panenka, Roman, 13
Pazáriková, Jana, 49
Rástočný, Karol, 15
Sabo, Štefan, 67
Srba, Ivan, 51
Suchal, Ján, 35
Šajgalík, Márius, 13
Šimko, Jakub, 17
Šimko, Marián, 19
Tvarožek, Jozef, 53
Tvarožek, Michal, 21
Unčík, Maroš, 69
Valčuha, Matej, 23
Virik, Martin, 37
White, Bebo, 3
Zeleník, Dušan, 39

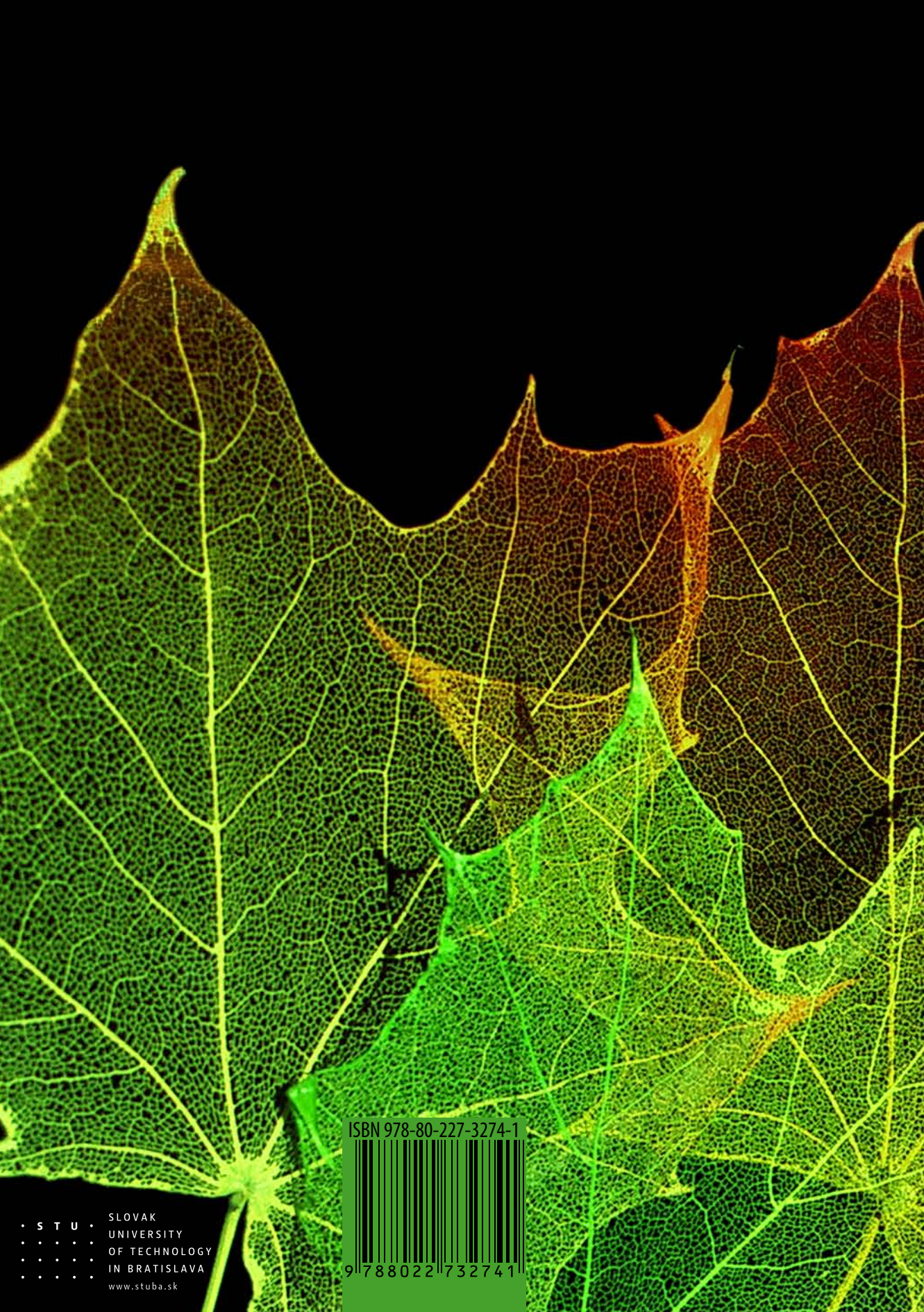
Mária Bieliková, Pavol Návrat (Eds.)

Workshop on the Web – Science, Technologies and Engineering
7th Spring 2010 PeWe Ontožúr

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

88 pages, 50 copies
Print Nakladateľstvo STU Bratislava
2010

ISBN 978-80-227-3274-1



• S T U •
• • • • •
• • • • •
• • • • •

SLOVAK
UNIVERSITY
OF TECHNOLOGY
IN BRATISLAVA
www.stuba.sk

ISBN 978-80-227-3274-1



9 788022 732741