

5th Workshop on Intelligent and Knowledge oriented Technologies

WIKT 2010 Proceedings

Michal Laclavík
Ladislav Hluchý (Eds.)



November 11 - 12, 2010
Bratislava, Slovakia



S T U . . .
.
F I I T .
.



The workshop was organized by

Institute of Informatics, Slovak Academy of Sciences, Bratislava
Faculty of Informatics and Information Technologies, STU in Bratislava
Faculty of Electrical Engineering and Informatics, Technical University of Košice

The workshop was supported by

VEGA 2/0184/10, AIIA APVV-0216-07

Program Committee

Hluchý, Ladislav – ÚI SAV Bratislava - chair
Bieliková, Mária – FIIT, STU Bratislava
Krajčí, Stanislav – PrF, UPJŠ v Košiciach
Laclavík, Michal – ÚI SAV Bratislava
Mach, Marián – FEI, TU v Košiciach
Machová, Kristína – FEI, TU v Košiciach
Návrat, Pavol – FIIT, STU Bratislava
Paralič, Ján – FEI, TU v Košiciach
Rozinajová, Viera – FIIT, STU Bratislava
Šaloun, Petr – PrF, OS Ostrava
Vojtáš, Peter – MFF, UK Praha
Zendulka, Jaroslav – FIT, VUT Brno

Organizing Committee

Ladislav Hluchý
Michal Laclavík
Oľga Schusterová
Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
E-mail: wikt.ui@sav.sk

Proceeding Editors

Michal Laclavík
Ladislav Hluchý
Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
E-mail: {laclavik.ui, hluchy.ui}@savba.sk

ISBN 978-80-970145-2-0

© Institute of Informatics SAS and the authors of respective articles, 2010

Predhovor

Teší nás, že vám môžeme predstaviť už v poradí 5. zborník z workshopu zameraného na inteligentné a znalostne orientované technológie - WIKT 2010, ktorý sa uskutočnil 11. – 12. novembra 2006 v Bratislave.

Po 4 rokoch sa workshop opäť koná na Ústave informatiky SAV v Bratislave, kde bol organizovaný aj prvý zo série WIKT workshopov. Tento workshop sa snaží podporiť výskum, vývoj a výmenu poznatkov v oblasti inteligentných a znalostne orientovaných technológií. Hlavným cieľom workshopu je vytvoriť podmienky pre stretnutie a výmenu informácií o bežiacom výskume, diskusiu o aktuálnych problémoch v predmetnej oblasti a možných spôsoboch ich riešenia, ale aj výmenu skúseností s použitím relevantných pokročilých technológií a softvérových nástrojov, ako aj spôsobov ich využitia a nasadenia pre riešenie úloh praxe.

Hlavné témy workshopu boli:

- znalostné technológie a ich aplikácie
- modelovanie znalostí a ontológie
- sémantické spracovanie informačných zdrojov
- spracovanie informačných zdrojov v slovenskom jazyku
- sémanticky a servisne orientované architektúry
- znalostné bázy a organizačné pamäte
- usudzovanie a odvodzovanie

Na WIKT 2010 bolo podaných 31 príspevkov z ktorých bolo 19 prijatých ako riadnych príspevkov a 9 ako postrov. Všetky príspevky prebehli riadnym recenzným konaním s kvalitnými recenziami, ktoré majú snahu vylepšiť výsledné príspevky a následnú diskusiu o danej téme na workshope, pretože cieľom nie sú len kvalitné vedecké články ale aj diskusia o záujímavých témach a teda sú na workshope vítané príspevky nasledovných typov:

- výskumný príspevok
- work-in-progress
- vizionársky príspevok
- znalostné praktiky
- ponaučenia a skúsenosti
- aplikačný príspevok

Chceli by sme poďakovať všetkým, ktorí prispeli k úspešnému uskutočneniu workshopu. Chceme poďakovať programovému a organizačnému výboru a hlavne všetkým autorom, za ich príspevky a prezentácie na workshope.

Michal Laclavík, Ladislav Hluchý
Október 2010
Bratislava

Preface

We are pleased to introduce the 5th proceedings of the Workshop on Intelligent and Knowledge-oriented Technologies - WIKT 2010, held on 11 – 12 November 2010 in Bratislava.

After 4 years, the workshop is again organized at Institute of Informatics SAS, where the first of WIKT workshops was held. The main goals of this workshop is to create conditions for meeting and intensive exchange of information about running research, discussions about current problems in the areas in question and about possible ways how to solve them, but also exchange of experiences with the use of relevant advanced technologies and software tools and about the ways to deploy them and to use them for solving real world problems.

The topics of the workshop include:

- Knowledgemodeling, ontologies
- Semantic Web
- Semantic processing of information resources
- Processing of information resources in Slovak language
- Semantic and service-oriented architectures
- Knowledge bases and organizational memories
- Reasoning and inference.

We have recieved 31 submissions for WIKT 2010. Programme Committee has accepted 19 submissions for regular presentations and 9 submissions as posters. All submissions were peer reviewed with aim to give important feedback for the final papers and workshop discussion, since the goal of the workshop is not only excelent research papers, but also other submission types to make discussion more fruitful. So following types of submissions are welcomed:

- research paper
- work in progress
- visionary paper
- best practice guidelines
- lesson learned
- industrial & applications papers

Many people have assisted in the success of this workshop. We would like to thank all the members of the Programme and Organizing Committees for their work and assistance for the workshop. We would also like to express our gratitude to all authors for contributing their research papers as well as for their participation in the workshop that made the event fruitful and successful.

Michal Laclavík, Ladislav Hluchý
October 2010
Bratislava, Slovakia

Table of contents

POZVANÉ PREDNÁŠKY	1
SLOVAK NATIONAL CORPUS TOOLS AND RESOURCES	2
<i>Radovan Garabík</i>	
GRAF LINIEK WIKIPÉDIE AKO ZNALOSTNÁ BÁZA PRE IDENTIFIKÁCIU RELÁCIÍ MEDZI KONCEPTMI	8
<i>Marek Ciglan</i>	
ODPORÚČANIE A PRISPÔSOBOVANIE	10
VPLYV VZOROV V SPRÁVANÍ NÁVŠTEVNÍKOV WEBOVÉHO PORTÁLU NA ODPORÚČANIA	11
<i>Michal Holub, Mária Bieliková</i>	
HYBRIDNÉ ODPORÚČANIE VO VÝUČBOVÝCH SYSTÉMOCH	15
<i>Pavel Michlík, Mária Bieliková</i>	
TRENDY A BUDÚCNOSŤ ODPORÚČANIA ONLINE	19
<i>Mária Bieliková, Michal Kompan, Dušan Zeleník</i>	
UČENÍ PŘÍKLADY – PERSONALIZOVANÝ ADAPTIVNÍ WEB	23
<i>Jan Nekula, Petr Šaloun, Zdeněk Velart, Petr Klimánek</i>	
INTERNET A TECHNOLOGIE	27
ADAPTIVNÝ PROXY SERVER: PREVÁDZKA A SKÚSENOSTI PO ROKU	28
<i>Tomáš Kramár, Michal Barla, Mária Bieliková</i>	
HRY S ÚČELOM OBJAVOVANIA SÉMANTIKY NA WEBE	32
<i>Jakub Šimko, Michal Tvarožek a Mária Bieliková</i>	
PŘÍKLAD VYUŽITIA WEBOVÝCH TECHNOLOGIÍ PRE INTERNETOVÝ MARKETING	36
<i>Adela Tušanová, Ján Paralič</i>	
POUŽITIE SOLR NA INDEXOVANIE A VYHLADÁVANIE DÁT	41
<i>Zoltan Balogh a Emil Gatial</i>	
DISTRIBUOVANÉ SPRACOVANIE DÁT NAD MAPREDUCE ARCHITEKTÚROU (HADOOP A HIVE)	49
<i>Martin Šeleng</i>	
AUTOMATIZOVANÉ VYTVÁRANIE POUŽÍVATELSKÝCH FORMULÁROV	55
<i>Emil Gatial a Zoltán Balogh</i>	
SOCIÁLNE SIETE A GRAFY	59
MODELOVANIE A ANALÝZA MALEJ KOMUNITNEJ SOCIÁLNEJ SIETE	60
<i>Gabriel Tutoky, Ján Paralič</i>	
GRAPH TRANSFORMATIONS FOR SEMANTIC EMAIL SEARCH	64
<i>Marcel Kvassay, Michal Laclavík, Štefan Dlugolinský, Ladislav Hluchý</i>	
VYUŽITIE SOCIÁLNYCH SIETÍ PRI VYHLADÁVANÍ V EMAILOCH	68
<i>Michal Laclavík a Ladislav Hluchý</i>	

SÉMANTIKA A ONTOLOGIE	73
SÉMANTICKÁ SIEŤ AKO SPOJITÝ SYSTÉM	74
<i>Stanislav Dvorščák, Kristína Machová</i>	
APPLICATION ONTOLOGY MANAGER FOR HYDRA.....	78
<i>Ján Hreňo, Peter Kostelník, Martin Sarnovský</i>	
AN ONTOLOGY DRIVEN APPROACH TO SOFTWARE PROCESS ENGINEERING	82
<i>Miroslav Líška, Pavol Návrat</i>	
DOLOVANIE INFORMÁCIÍ A ZNALOSTÍ	86
VYUŽITIE JBOWL KNIŽNICE PRI RIEŠENÍ ÚLOH DOLOVANIA ZNALOSTÍ Z TEXTOV	87
<i>František Babič, Štefan Bašista, Roman Dudek, Roman Mihaľ, Peter Savčák</i>	
DATA MINING FOR FOG PREDICTION	91
<i>Peter Bednár, František Albert</i>	
DISCOVERING OCCURRENCES OF USER-DEFINED PATTERNS IN HISTORICAL DATA REPRESENTING COLLABORATIVE ACTIVITIES IN VIRTUAL USER ENVIRONMENT	95
<i>Jozef Wagner, Ján Paralič, František Babič</i>	
POSTRE	101
WEB INFORMATION INTEGRATION IN KNOWLEDGE DISCOVERY	102
<i>Kristína Machová, Dominika Fodorová</i>	
POUŽITIE ALTERNATÍVNYCH PRÍSTUPOV PRE PLÁNOVANIE VÝROBNÉHO PROCESU .	106
<i>Tomáš Kasanický, Ján Zeleneka</i>	
DOLOVANIE UDAJOV V HYDROMETEOROLOGICKÝCH APLIKACIACH.....	111
<i>Martin Šeleng, Peter Krammer, Ondrej Habala a Ladislav Hluchý</i>	
TEXT DOCUMENT RETRIEVAL BY DOCUMENT SPACE DIMENSION REDUCTION WITH FEED-FORWARD NEURAL NETWORKS	116
<i>Lenka Skovajsová, Igor Mokriš</i>	
ICT-BASED TOOLBOX IN OCOPOMO PROJECT AND POTENTIAL METHODS FOR INTEGRATION	121
<i>Peter Butka, Marián Mach, Tomáš Sabol, Karol Furdík</i>	
MULTI-AGENT-BASED CONCEPTION OF MODERN AIRCRAFT DESIGN	125
PODNIKOVÁ INTELIGENCIA, ANALYTIKA A PROCES OBJAVOVANIA ZNALOSTÍ V DATABÁZACH	129
<i>Jozef Kovač</i>	
 AUTHORS INDEX.....	 133

Pozvané přednášky

Slovak National Corpus tools and resources

Radovan Garabík

L. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Abstract. The article presents current state of affairs in several projects conducted by the Slovak National Corpus department of the L. Štúr Institute of Linguistics, Slovak Academy of Sciences. We describe the Slovak National Corpus, Corpus of Spoken Slovak, tools used for linguistics analysis and an ongoing effort to create Slovak WordNet.

1 Slovak National Corpus

The Slovak National Corpus is a huge, representative corpus of modern written Slovak (since the 1953 orthography reform). Currently, the whole corpus contains over 700 million tokens. There are several specialised subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus, texts written until 1989). The corpus is automatically lemmatised and morphologically annotated and is indexed using the *Manatee* software [Ryc00]. To query the corpus, there are two possibilities – first, the users can use multiplatform (Tcl/Tk) *Bonito* client to access the *Manatee* server, using its own protocol. This approach provides the users with complete access to all the advanced querying, sorting and statistical features of the server, however requires installation of a specialized software. The other possibility is to use web based access, where only basic features are present. In both cases, the search interface provides CQL compatible query syntax.

However, in the last few years the ability of an average user to install arbitrary software (and use anything that is not web-based) declined considerably, and new corpus users often face an insurmountable obstacle in downloading, unpacking and running the *Bonito* client. Because of this, we are considering transfer of the corpus to *Manatee-2*, which provides complete web-based interface as a replacement of the Tcl/Tk client.

A separate corpus (although part of the whole Slovak National Corpus project) is a manually morphologically annotated corpus, whose main purpose is to be a source of train data for Slovak language tagger (and, to a lesser extent, for morphology annotation tools).

The size of the Slovak National Corpus source archives is 46 GB, however, a substantial percentage of this are original scan images (when converted into raw XML text, the size is about 6 GB uncompressed).

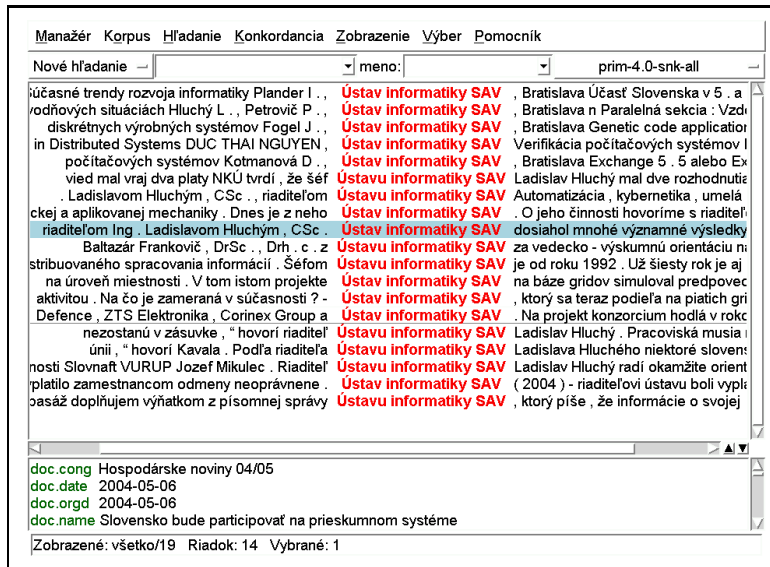


Fig. 1. Screenshot of *Bonito* client

2 Corpus of Spoken Slovak

Corpus of Spoken Slovak is a project to record reasonable amount of sound samples of contemporary Slovak, together with their manual phonemic transcription, automatic lemmatisation and morphosyntactic analysis. At the time of writing, the corpus contains about 160 hours of sound recordings, corresponding to 1.2 million tokens. Since the transcription is done manually (no reasonably accurate transcription software exists), the remaining task of morphosyntactic analysis is exactly the same as with the Slovak National Corpus texts.

The archive is kept in FLAC format, and we convert the whole recordings into Ogg/Vorbis and Ogg/Speex formats (for easier handling and transcription) and for the final linking through the corpus web interface we split the files into small chunks corresponding to dialogue turns. The source archive size is currently over 200 GB.

One of our primary goals was to make this corpus unencumbered by usual copyright and privacy concerns that plague similar projects. We have to take care not only of copyright law, but also the law on protection of personal data [Ná05]. We do this by removing any sensitive information (e.g. personal names) before including the recordings in the archive, and by including only those recordings where we have explicit expression of consent by all the relevant participants to include the recordings in our archive.

For transcription, we are using the *transcriber* software [BGWL01], with a detailed set of tags to annotate both internal speech features and external sound events influencing the recorded discourse.

Access to the corpus can be performed in two (almost independent) ways. One of them uses standard *Bonito* client, in the same way as the preferred access to the main

Slovak National Corpus. Each token provides following attributes: *pron*, *lemma*, *tag*, *dcount*. *pron* is the transcribed pronunciation, *lemma* and *tag* come from the standard automatic morphosyntactic annotation, *dcount* is the possible number of lemma-tag pairs.

The other way to access the corpus is to use specialized web interface, offering additional visual representation of transcription and annotation, as well as links to sound recordings themselves.

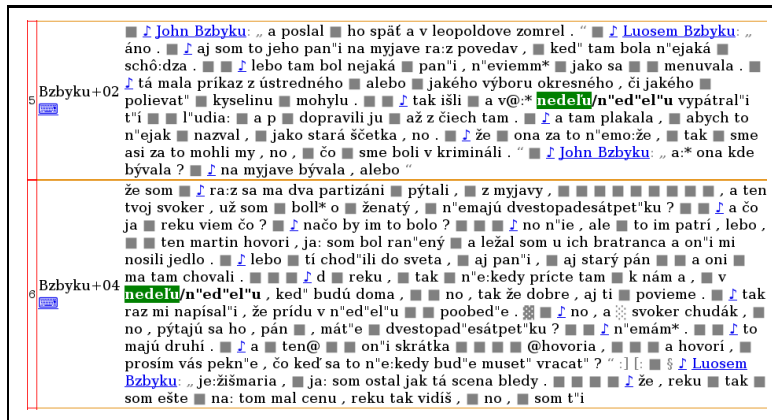


Fig. 2. Screenshot of Corpus of Spoken Slovak web interface

<pre> <Event desc="poz" type="noise" extent="instantaneous"/> niekedy/n"e:kedy prídŕe/pričte tam <Event desc="pi" type="pronounce" extent="instantaneous"/> k nám a, <Event desc="poz" type="noise" extent="instantaneous"/> v nedeľu/n"ed"el"u, keď/ked" budú doma, <Event desc="poz" type="noise" extent="instantaneous"/> <Event desc="mm" type="noise" extent="instantaneous"/> no, tak že dobre, aj ti <Event desc="pi" type="pronounce" extent="instantaneous"/> povieme. <Sync time="755.906"/> </pre>

Table 1. Example of annotation of Corpus of Spoken Slovak transcriptions

3 Linguistic analysis

The foundation of all subsequent analysis is assignment of unique lemma and tag combination to all the words in the analysed text (e.g. in our corpus). This is realised as a two stage process, first stage is morphosyntactic analysis, i.e assignment of all the possible lemma-tag pairs to a given token. Second stage is disambiguation – selection of one (correct) lemma-tag pair for a given word. We collected semi-automatically complete paradigms for 74 000 lemmata[Gar06] and stored manually verified and into a wiki-based database[Gar08]. The database contains complete paradigms, with an exception for third person plural of L-participle, where we keep only tag for general gender (*všeobecný rod*, tag ‘h’), since the forms of all the other genders are identical, and the paradigm is then automatically expanded to cover all the existing genders with corresponding tags. The morphological analysis then consists from looking up all the possible tags and lemmata for a given word form, and from guessing possible lemmata and tags for words not present in the database.

<pre>== Lema == mať == Paradigma == V1e+: mať VKes+: mám VKesb+: máš VKesc+: má VKepa+: máme VKepb+: máte VKepc+: majú VWesb+: maj VWepa+: majme VWepb+: majte VHe+: majúc VLesam+: mal VLesaf+: mala VLesan+: malo VLepah+: mali == Homonymia == [[mať]] ----- KategoriaVerbá</pre>

Table 2. Paradigm of the verb *mať*

3.1 Guessing

Quite an important part of the analysis is assigning a lemma-tag pair to words that are not present in the morphological database. While a reliable determining of lemma, part of speech and morphological tag when given an unknown word is impossible, it is nevertheless desirable to obtain at least some information about those words. E.g. even if we guess lemma incorrectly, getting at least correct part of speech will help in eventual subsequent syntactic annotation. Our guessing is based on suffix similarity – first, during the training phase, we build an array of suffices of existing wordforms. We use fixed length of 3 characters (determined empirically). During the guessing phase,

if the unknown word starts with a capital letter and is not situated at the beginning of a sentence, it is assumed to be a noun or a adjective (most common parts of speech for proper names), otherwise it could be also a verb, participle, adverb or a numeral. Special provision is implemented for potential adjectives beginning with the prefix *naj-* and verbs beginning with the prefix *ne-* (for superlatives and negated verbs).

3.2 Disambiguation

The second step is disambiguation, where each word is assigned a unique lemma and a morphosyntactic tag out of the possibilities assigned in the first step. For disambiguation, we use *morče*, an averaged perceptron model originally used for the Czech language tagging [SHRS09], re-trained on the Slovak manually annotated corpus.

<s>			
Po	po	Eu6	04
chvíli	chvíľa	SSfs6	02
ste	byť	VKepb+	02
zistili	zistiť	VLdpbh+	07
,	,	Z	01
že	že	0	02
to	to	PFns1	05
nejde	nejst'	VKesc-	01
.	.	Z	01
</s>			

Table 3. Example of an automatically morphosyntactically tagged sentence from the Slovak National Corpus

4 WordNet

There is currently an ongoing effort in collaboration with Technical University of Košice in building a basic Slovak WordNet database. We plan to use the database as a skeleton of a basic English-Slovak-German-Polish-Lithuanian dictionary¹. The building process consists of mapping automatically generated Slovak synsets to English synsets from WordNet v.3.0. The synset generation has been described in [Gen09]; the synsets are manually corrected before being added to the database. We use special annotation to mark synsets that do not have clear English equivalent. Our goal is to build synsets containing ten thousand most frequent words from the Slovak National Corpus (nouns, adjectives, verbs and adverbs), together with a complete set of their hypernyms (i.e. each Slovak synset will have a hypernym, unless mapped to those few English synsets that do not have a hypernym).

¹ As part of the Slovak Online (Lifelong Learning Programme DG EAC/31/08) project.

POS	synsets	%
noun	4669	51.6
verb ^a	1895	21.0
adjective	2265	25.0
adverbs	214	2.4

^a Negated verbs are not in the database.

Table 4. POS Composition of Slovak Wordnet Database

References

- [BGWL01] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5–22, 2001.
- [Gar06] Radovan Garabík. Slovak morphology analyzer based on Levenshtein edit operations. In Michal Laclavík, Ivana Budinská, and Ladislav Hluchý, editors, *1st Workshop on Intelligent and Knowledge oriented Technologies*, pages 2 – 5, Bratislava, 2006. Institute of Informatics, Slovak Academy of Sciences.
- [Gar08] Radovan Garabík. Storing morphology information in a wiki. In Olga Shemanayeva, editor, *Lexicographic tools and techniques*, pages 55 – 59, Moscow, 2008. IITP RAS.
- [Gen09] Ján Genči. Synset Building Based on Online Resources. In Jana Levická and Radovan Garabík, editors, *NLP, Corpus Linguistics, Corpus Based Grammar Research*, Brno, 2009. Tribun.
- [Ná05] Národná rada Slovenskej republiky. Zákon č. 428/2002 Z. z. o ochrane osobných údajov Z. z. v znení zákona č. 602/2003 Z. z., zákona č. 576/2004 Z. z. a zákona č. 90/2005 Z. z. *Zbierka zákonov Slovenskej republiky*, Bratislava, Slovakia, 2002, 2004, 2005.
- [Ryc00] Pavel Rychlý. *Korpusové manažery a jejich efektivní implementace*. PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2000.
- [SHRS09] Drahomíra Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. Semi-supervised training for the averaged perceptron POS tagger. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 763–771, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Graf liniek Wikipédie ako znalostná báza pre identifikáciu relácií medzi konceptmi

Marek Ciglan

Norwegian University of Science and Technology, Trondheim, Norway
ciglan@gmail.com

Abstrakt. Wikipédia je slobodná encyklopédia vytvorená masívnou kolaboráciou dobrovoľníkov z celého sveta. Kvôli jej rozsahu, bohatstvu informácií a štruktúre je mnohými považovaná za hodnotný zdroj sémantických dát, použiteľný v mnohých oblastiach informatiky. Wikipédia bola úspešne použitá v oblasti spracovania prirodzeného jazyka, pre obohacovanie systémov vyhľadávania informácií ako aj pre budovanie ontológií. Výnimočná je aj štruktúra Wikipédie, kde je každý článok venovaný jednej téme a články sú husto previazané hyperlinkami. V tejto práci využívame graf liniek Wikipédie, kde každý vrchol reprezentuje jednu tému a hrany reprezentujú relácie medzi témami, zodpovedajúce hyperlinkám medzi jednotlivými článkami. Našou snahou je identifikácia dôležitých relácií medzi zadanými vstupnými konceptmi. Priamočiary prístup, hľadanie najkratších ciest v grafe liniek, často nevedie k dobrým výsledkom. Je to spôsobené veľkým počtom ciest s najkratšou dĺžkou z ktorých je väčšina sémanticky nie príliš zaujímavých. Príčinou sú matematické vlastnosti grafu liniek Wikipédie; táto sieť vykazuje vlastnosti malého sveta - mocninová distribúcia stupňov uzlov, malá priemerná vzdialenosť vrcholov grafu, malý priemer grafu a vysoký lokálny zhlukovací koeficient grafu. Naš prístup spočíva v použití metódy šírenia aktivácie na grafe liniek Wikipédie. Šírenie aktivácie je algoritmus navrhnutý pre asociatívne prehľadávanie grafovej dátovej štruktúry. Hlavnou výzvou zvoleného prístupu je spôsob váhovania hrán grafu liniek, kde váha linky má reflektovať sémantickú blízkosť spojených konceptov. Predstavíme viacero prístupov k váhovaniu hrán založených na topologických vlastnostiach siete, čiastočnej sémantiky definovanej taxonómiou kategórií Wikipédie a korelácie počtu návštev stránok zodpovedajúcich jednotlivým konceptom. Následne predstavíme aplikáciu WikiPop, využívajúcu váhovaný graf liniek a algoritmus šírenia aktivácie na personalizovanú detekciu udalostí z štatistik návštevnosti stránok Wikipédie. Graf liniek Wikipédie je vo svojej podstate sieťou konceptov, kde hrany predstavujú reláciu medzi nimi. V záverečnej časti príspevku sa budeme venovať možnosti konštrukcie siete konceptov zo zbierky dokumentov, pomocou techník dolovania dát z textov.

Odporúčanie a prispôsobovanie

Vplyv vzorov v správaní návštevníkov webového portálu na odporúčania

Michal Holub, Mária Bieliková

Ústav informatiky a softvérového inžinierstva
Fakulta informatiky a informačných technológií, Slovenská technická univerzita
Ilkovičova 3, 842 16 Bratislava, Slovensko
{holub,bielik}@fiit.stuba.sk

Abstrakt. Na webových portáloch môžeme nájsť množstvo užitočných informácií. Prácu nám uľahčia odporúčania, ktoré berú do úvahy našu predchádzajúcu činnosť a činnosť nám podobných používateľov. V príspevku predstavujeme spôsoby, ako využiť sledovanie správania sa používateľov pri odporúčaní webových stránok. Správanie na jednotlivých stránkach používame na určovanie záujmu o ne. V postupnostiach navštívených stránok hľadáme vzory, vďaka ktorým vieme zistiť viac o cieľoch návštevníka. Niektoré z navrhnutých myšlienok overujeme prototypom adaptívneho systému, ktorý odporúča zaujímavé udalosti návštevníkom webového portálu našej fakulty.

1 Úvod

Veľké webové portály dnes obsahujú množstvo informácií. Každý z návštevníkov webového portálu má iné potreby a z týchto informácií ho zaujíma iba vybraná časť. Problémom pre portál je správne identifikovať potrebu používateľa a zobrazíť mu informácie, o ktoré má v danej chvíli záujem. O záujmoch používateľa nám môže viac prezradiť jeho správanie sa pri opakovaných návštevách webového portálu. Ak v správaní identifikujeme opakujúce sa vzory, môžeme podľa nich zoskupiť podobných používateľov a vytvárať pre nich odporúčania.

Návštevníci k informáciám na webe pristupujú rozličnými spôsobmi. Najbežnejším spôsobom je využitie hypertextových odkazov, ktorý sa používa v takmer polovici všetkých prípadov [4]. Ďalším často používaným spôsobom je využitie tlačidla *Späť* vo webovom prehliadači [6]. Podiel ostatných prípadov (ručné zadanie URL, výber odkazu z histórie, výber zo zoznamu obľúbených položiek, atď.) je zanedbateľný, v jednotkách percent. Z tohto dôvodu má zmysel optimalizovať najmä zobrazenie odkazov na webových stránkach.

Záujem o prezentované informácie môžeme zistiť porovnaním kľúčových slov s modelom používateľa [1]. Musíme však predpokladať, že všetky zobrazené stránky používateľa zaujali. Keď vieme, aké témy používateľa zaujímajú, môžeme mu odporučiť webové stránky, ktoré obsahujú príslušné kľúčové slová. Kľúčové slová môžeme tiež získať z anotácií odkazov, na ktoré používateľ klikol. Kliknutie na odkaz však vo všeobecnosti nemusí vyjadrovať používateľov záujem o stránku, na ktorú smeruje. Druhým spôsobom zistenia záujmu je sledovanie akcií.

2 Analýza správania pri navigácii

Návštevníci webového portálu za sebou zanechávajú digitálnu stopu v podobe odkazov, ktoré použili pri navigácii medzi stránkami webového sídla. Z každého sedenia vieme získať vektor, ktorého zložkami sú jednotlivé stránky. Usporiadané sú podľa poradia, v akom boli navštívené. Takto vytvorené postupnosti nám môžu odhaliť aktuálny záujem používateľa, ako aj jeho dlhodobé návyky pri návštevách daného webového sídla. Uvedené informácie vieme využiť pri odporúčaní odkazov a následne dokumentov, na ktoré vedú. Po analýze dokumentov vieme odporúčať konkrétne objekty (informácie), ktoré tieto dokumenty obsahujú.

Webový portál si môžeme predstaviť ako orientovaný graf. Jednotlivé stránky sú uzly grafu. Medzi stránkami sa pohybujeme prostredníctvom hypertextových odkazov, ktoré tvoria orientované hrany grafu. Pri používaní portálu tak vznikajú rôzne cesty medzi jednotlivými stránkami. O celi používateľa v rámci konkrétneho sedenia nám môže mnoho napovedať už prvá navštívená stránka. Na nej si používateľ vyberie niektorý z odkazov, čím určí cestu v grafe. Každý ďalší použitý odkaz konkretizuje zámer používateľa. V takomto prípade mu vo vhodnej chvíli môžeme odporučiť dokument na konci cesty, ktorý ostatní používatelia považovali za zaujímavý. Znížime mu tým počet odkazov, ktoré musí použiť.

Z dlhodobého hľadiska nám analýza postupností navštívených stránok môže prezradiť určité zaradenie používateľa do jednej z cieľových skupín webového portálu. Väčšie portály často obsahujú informácie určené rôznorodým skupinám používateľov, čomu je prispôbená aj navigácia. Prikladom môže byť univerzitný webový portál, ktorý obsahuje informácie pre pedagógov, študentov a verejnosť. Používateľ si ako prvé v menu vyberie, do ktorej z týchto skupín patrí. Keď si pri každej návšteve vyberie tú istú skupinu, môžeme túto informáciu využiť a následne mu priamo odporučiť sekcie s informáciami pre jeho skupinu.

V postupnostiach navštívených stránok tiež môžeme hľadať opakujúce sa vzory. Tie nám povedia viac o zvyklostiach používateľa pri práci s daným webovým portálom. Niektorí používatelia viac využívajú navigáciu poskytnutú portálom a pohybujú sa v kruhoch, iní častejšie využívajú možnosť vrátiť sa späť, čím ich postupnosti pripomínajú schody.

V postupnostiach navštívených stránok identifikujeme tieto základné vzory [3]:

1. Cesta – postupnosť, v ktorej sa žiadna stránka neopakuje.
2. Kruh – postupnosť začínajúca aj končiacia na tej istej stránke.
3. Slučka – postupnosť prechádzajúca už raz navštívenou stránkou.
4. Hrot – postupnosť, v ktorej sa vraciame späť po tej istej trase.

Použitý vzor vypovedá o návykoch používateľa, ako aj o jeho aktuálnych zámeroch. Napr. väčší výskyt hrotov signalizuje, že používateľ hľadá konkrétnu informáciu. Naproti tomu, výskyt slučiek a kruhov napovedá, že používateľ sa snaží objaviť, čo sa na portáli nachádza. V prípade, že u každého používateľa bude z dlhodobého hľadiska na konkrétnom portáli prevažovať jeden zo vzorov, môžeme túto skutočnosť využiť na rozdelenie používateľov do skupín. V rámci skupín môžeme následne odporučiť zaujímavé dokumenty a prispôbovať navigáciu [5].

3 Správanie na jednotlivých stránkach

Na konkrétnych stránkach webového portálu návštevníci vykonávajú rozličné akcie. Tieto vieme využiť na zistenie používateľovho záujmu o prezentované informácie. Podľa typu záujmu sme identifikovali tri kategórie akcií. Prvými dvomi typmi sú akcie vyjadrujúce čisto kladný (tlač stránky, pridanie do obľúbených položiek, skopírovanie textu do schránky) alebo záporný (zatvorenie stránky po veľmi krátkom čase od zobrazenia, zastavenie načítavania stránky) záujem. Posledný typ predstavujú akcie, ktoré môžu vyjadrovať oba druhy záujmu v závislosti od kontextu, v akom boli vykonané (čas strávený na stránke, miera pohybu myšou, miera rolovania stránky).

Akcie, ktorých význam závisí od kontextu ich vykonania, porovnávame s akciami vykonanými na danej stránke ostatnými používateľmi v minulosti. Podľa toho určujeme, či akcia vyjadruje kladný alebo záporný záujem. Ak napr. používateľ strávil na stránke nadpriemerne veľa času oproti ostatným používateľom, usudzujeme z toho, že stránka používateľa zaujala. Takto určíme záujem o každú videnú stránku.

Záujem môžeme vyjadriť rôznymi formami, napr. dvomi hodnotami (stránka používateľa zaujala/nezaujala) alebo spojitou (reálne číslo na zvolenej stupnici odrážajúce mieru záujmu). Takto môžeme určiť záujem nielen o samotný dokument (webovú stránku), ale aj o objekty na vyššej významovej úrovni, ktoré získame predspracovaním dokumentu. Ak napr. webová stránka informuje o nejakej krajine, priradíme záujem používateľa priamo k danej krajine. Dokumenty (objekty), ktoré zaujali viacero ľudí, môžeme odporúčať ďalším používateľom. Podľa záujmu o zhladnuté webové stránky môžeme určiť aj záujem o stránky, ktoré doposiaľ používateľ nevidel s využitím kolaboratívneho filtrovania.

Ako veľmi užitočné sa javí určovanie záujmu o zobrazenú stránku pre potreby odporúčania spojiť so vzormi nájdenými v postupnostiach odkazov. Ako sme už uviedli, používateľ sa pri navigácii na portáli vydá niektorou z ciest. Pre všetky stránky na tejto ceste vieme vypočítať predpokladanú mieru jeho záujmu a odporučiť mu tie najzaujímavejšie. Skrátime mu tým cestu k cieľovému dokumentu. Takto tiež môžeme odhaliť dokument, ktorý by používateľ mohol prehliadnuť.

4 Využitie vzorov v správaní pri odporúčaní udalostí

Vyberané časti konceptu prezentovaného v statiach 2 a 3 sme overili na webovom sídle našej fakulty (www.fiit.stuba.sk). Množstvo stránok univerzitného portálu obsahuje informácie o nadchádzajúcej udalosti. Práve sledovanie udalostí je dôležité pri návšteve takéhoto sídla. Tieto udalosti automaticky nachádzame a zostavujeme z nich osobný kalendár každého návštevníka. Navrhli sme pre tento účel metódu určovania záujmu o zobrazenú stránku. Takto určený záujem spájame s udalosťou, o ktorej stránka informuje. Udalosti s najvyšším vypočítaným záujmom umiestňujeme do kalendára ako pripomienku. Ďalej predpovedáme záujem o udalosti, o ktorých používateľ ešte nevie, a pridávame ich do kalendára ako odporúčania. Každý návštevník má svoj vlastný kalendár s udalosťami (pozri obr. 1).

Na zaznamenanie správania sa používateľov a modifikáciu webových stránok pridaním kalendára používame adaptívny proxy server [2]. Ten nám umožňuje

pridávať odporúčania do stránok ľubovoľného webového portálu. Takisto vieme odlišiť jednotlivých používateľov pomocou jedinečného ID (nevieme však povedať nič o osobe používateľa, súkromie tak ostáva zachované). Zaznamenávame tri akcie: čas aktívne strávený na stránke (t.j. vtedy, keď používateľ hýbal myšou), počet rolovaní stránky a výskyt *skopírovania textu do schránky*.



Obrázok 1. Osobný kalendár so zobrazenou odporúčanou udalosťou 10.5.2010.

Analyzovali sme tiež postupnosti navštívených odkazov a hľadali v nich opakujúce sa vzory. Našli sme všetky typy vzorov, pričom sme návštevníkov rozdelili do skupín podľa prevažujúceho vzoru. V ďalšej práci sa chceme zamerať na overenie, či používatelia takto vytvorených skupín budú mať spoločné záujmy, a či pre nich budú zaujímavé rovnaké odporúčania.

PodĎakovanie. Tento príspevok vznikol vďaka čiastočnej podpore grantov VEGA VG1/0508/09, KEGA 028-025STU-4/2010 a v rámci OP Výskum a vývoj pre projekt: Podpora dobudovania Centra excelentnosti pre Smart technológie, systémy a služby II, ITMS: 26240120029, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Barla, M., Bieliková, M.: On Deriving Tagsonomies: Keyword Relations coming from the Crowd. In LNAI 5796, Proc. of ICCCI 2009, Springer, pp. 309--320 (2009)
2. Barla, M., Bieliková, M.: Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In IADIS Int. Conf. WWW/Internet (2010)
3. Canter, D., Rivers, R., Storrs, G.: Characterizing user navigation through complex data structures. Behaviour & Information Technology, vol. 4, no. 2, pp. 93--102 (1985)
4. Cockburn, A., McKenzie, B.: What do web users do? An empirical analysis of web use. Int. Journal of Human-Computer Studies, vol. 54, no. 6, pp. 903--922 (2001)
5. Holub, M., Bieliková, M.: Estimation of User Interest in Visited Web Page. In: Proc. of the 19th Int. Conf. on World Wide Web, Raleigh, USA, ACM Press, pp. 1111--1112 (2010)
6. Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., Sommerer, R.: Smartback: supporting users in back navigation. In: Proc. of the 13th Int. Conf. on World Wide Web, New York, USA, ACM Press, pp. 63--71 (2004)

Hybridné odporúčanie vo výučbových systémoch

Pavel Michlík, Mária Bieliková

Ústav informatiky a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita, Ilkovičova 3, 842 16 Bratislava
{michlik,bielik}@fiit.stuba.sk

Abstrakt. V tomto príspevku opisujeme možnosti, ako kombinovať jednoduché metódy odporúčania do hybridných metód v kontexte výučbového systému. Navrhujeme princíp automatickej optimalizácie metódy váhovania. Ukazujeme, ako je možné využiť váhovanie a iné hybridné metódy na overovanie a porovnávanie nových metód odporúčania.

Kľúčové slová: hybridné odporúčanie, váhovanie, optimalizácia

1 Odporúčanie v adaptívnych výučbových systémoch

Adaptívne odporúčanie obsahu je jedným z veľmi vhodných spôsobov pre zefektívnenie učenia [2] [7] a tým aj uľahčenie získavania vedomostí študentovi. Existuje viacero prístupov k adaptívnemu odporúčaniam obsahu. Klasifikácia metód odporúčaní podľa [3] rozlišuje nasledujúce kategórie:

- *kolaboratívne metódy* – odporúčajú obsah podľa spätnej odozvy od ostatných používateľov a hľadajú podobnosti medzi používateľmi,
- *metódy založené na obsahu* – používajú atribúty obsahu (odporúčaných objektov) a model používateľa, ktorý je vytvorený na základe atribútov tých objektov, ktoré daný používateľ v minulosti videl a hodnotil,
- *demografické metódy* – rozdeľujú používateľov do skupín (stereotypov),
- *metódy založené na užitočnosti* – na rozdiel od metód založených na obsahu nevytvárajú dlhodobý model používateľa, ale určujú jeho momentálne potreby a vyhodnocujú užitočnosť odporúčaných objektov vzhľadom na tieto potreby,
- *metódy založené na znalostiach* – vyberajú vhodný obsah na základe používateľových preferencií a sady pravidiel.

Všetky druhy odporúčacích metód dosahujú (ako vo výučbových systémoch, tak aj v iných aplikáciách) pozitívne výsledky, aj keď na rôznej úrovni.

Študenti – používatelia výučbových systémov – sa okrem okamžitých preferencií obsahu môžu líšiť aj štýlom učenia, ktorý im vyhovuje. Výučbový systém ALEA [4] sa prispôbuje študentom podľa toho, či preferujú učenie od všeobecného konceptu ku konkrétnemu, alebo naopak. Rôzne metódy odporúčaní obsahu môžu vyhovovať rôznym štýlom učenia. V [5] opisujeme metódu odporúčaní príkladov, ktorá je

určená pre prípravu na test v obmedzenom čase. Cieľom je zabezpečiť, aby študent prešiel za daný čas čo najviac potrebných tém aspoň do takej miery, aby potom jeho výsledky v teste boli (vzhľadom na čas, ktorý venoval učeniu) čo najlepšie.

Iné výučbové systémy, napríklad [7], ponúkajú viacero alternatív toho istého obsahu. Študent má možnosť vybrať si napríklad medzi kratším, náročnejším a dlhším, ale jednoduchším textom. Podobne, jeden koncept môže byť vysvetlený textom, obrázkom, príkladom, alebo inými typmi obsahu, a každý študent si môže vybrať ten typ, ktorý mu najviac vyhovuje. Pritom preferencia študenta nemusí platiť globálne. Môže sa ukázať, že pre jedného študenta sú pri rôznych témach vhodné rôzne prístupy. Je tu teda priestor pre adaptáciu, kde by výučbový systém napríklad na základe zhľukovania študentov automaticky vybral vhodný spôsob prezentácie.

Na každý z týchto pohľadov môže byť vhodná iná metóda odporúčania. Hybridné metódy – kombinácie viacerých prístupov pri tvorbe odporúčaní – predstavujú vhodný mechanizmus, ako rôzne kritériá a aspekty odporúčania skombinovať so súčasným zachovaním jednoduchosti modulov systému. Hybridné metódy ponúkajú možnosť, ako využiť užitočné vlastnosti základných metód a zároveň vykompenzovať ich nedostatky – napríklad problém nového objektu pri kolaboratívnom odporúčaní, alebo absencia hodnotenia kvality objektu pri odporúčaní na základe obsahu [1]. V tomto príspevku sa zameriame na možnosti kombinovania viacerých základných prístupov a ich využitie v doméne adaptívnych výučbových systémov.

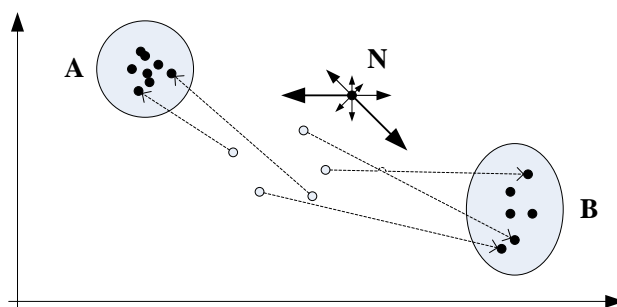
2 Hybridné metódy odporúčania

Medzi množstvo spôsobov hybridného odporúčania – kombinovania viacerých odporúčacích metód patria napríklad [3]:

- *váhovanie* – výsledným hodnotením každého objektu je lineárna kombinácia čiastkových hodnotení,
- *prepínanie* – automatické prepínanie medzi dvoma alebo viacerými metódami,
- *kaskádové odporúčanie* – jedna metóda odporúčania je použitá na zjemnenie výsledkov inej metódy,
- *zmiešané odporúčanie* – výsledky viacerých metód sú zobrazené súčasne,
- *pridávanie atribútov* – výsledkom jednej metódy sú atribúty obsahu, ktoré využíva iná metóda odporúčania.

V adaptívnom výučbovom rámci ALEF [6], ktorý v súčasnosti s väčším kolektívom vyvíjame, sme navrhli a implementovali podporu kombinovania odporúčaní váhovaním s pevne nastavenými váhami. Otvára sa možnosť robiť adaptívnu optimalizáciu týchto váh a vytvoriť tak vhodnú kombináciu odporúčaní individuálne pre každého študenta. Pri optimalizácii sa môže vyhodnocovať napríklad prírastok vedomostí za časové okno, explicitná spätná odozva (vyhovujúce alebo nevyhovujúce odporúčania), alebo používanie, resp. nepoužívanie odporúčaní študentom. Nástroje pre odporúčanie, ktoré vstupujú do kombinovania metód, môžu byť aj rozdielne nakonfigurovanými inštanciami jedného nástroja – napríklad odporúčanie len určitých typov objektov alebo nastavenie konkrétneho štýlu učenia.

Podľa takto získaných váh možno študentov zhlukovať, pretože váhy jednotlivých odporúčacích metód pre jedného študenta sú vlastne reprezentáciou jeho štýlu učenia. Zhluky sa následne môžu využiť pre ďalšiu metódu prispôbovania, prípadne pre urýchlenie spomínanej optimalizácie váh odporúčacích nástrojov pre nových študentov – ak zmeny váh pri optimalizácii smerujú k niektorému zhluku, môžeme tieto zmeny zosilniť, pretože predpokladáme, že nový študent bude podobný študentom v danom zhluku. Situáciu ilustruje obrázok 1.



Obr. 1. Príklad zhlukovania používateľov podľa preferencií metód odporúčania. Na osiach grafu sú váhy odporúčacích nástrojov vo váhovanom hybridnom odporúčaní. Ak sa z náhodnej počiatkovej konfigurácie používateľov (svetlé body uprostred) váhy optimalizáciou sústredia do zhlukov A a B, potom nový používateľ N, začínajúci takisto na náhodnej konfigurácii, bude v prípade posunu váh smerom k niektorému zo zhlukov urýchlený (šípky z bodu N).

V prostredí, kde naopak neočakávame rozdielne preferencie metód odporúčania, môžeme použiť kolaboratívnu optimalizáciu váh odporúčacích nástrojov, kde všetci používatelia vytvárajú jedinú spoločnú konfiguráciu odporúčania.

Výučbový rámec ALEF ponúka tiež možnosť kaskádového odporúčania, takisto s pevnou sekvenciou odporúčacích nástrojov. Podobne ako v predchádzajúcom prípade, vzniká možnosť prispôbovania sekvencie metód podľa dosahovaných výsledkov. Môže ísť len o jednoduché aktivovanie a deaktivovanie článkov postupnosti odporúčaní, alebo v zložitejšom prípade o zmenu poradia odporúčacích nástrojov, t.j. adaptívny výber hlavnej a zjemňujúcich metód odporúčania. Rovnako ako v prípade váh je možné využiť zhluky študentov s podobnou konfiguráciou odporúčania pre ďalšiu adaptáciu výučbového systému.

3 Overovanie metód odporúčania a zhodnotenie

Možnosťou kombinovania rôznych metód odporúčania v jednom systéme (nielen výučbovom) vzniká platforma pre overovanie a porovnávanie nových metód odporúčania. Najjednoduchší spôsob je nastavovanie váh odporúčacích nástrojov napevno pre každú testovaciu skupinu používateľov. Týmto spôsobom sme overovali metódu [5], kde boli dve rôzne konfigurácie rovnakej metódy odporúčania a jedna kontrolná skupina študentov s náhodným odporúčaním. Ukázali sme tak aj možnosť

jednoducho vytvárať kontrolné skupiny – náhodné odporúčanie, prípadne pevná referenčná sekvencia objektov, sa implementuje so stanoveným rozhraním ako odporúčací nástroj a nastaví sa príslušné váhy pre kontrolnú skupinu používateľov.

Iný spôsob porovnávania metód odporúčania, ktorý je použiteľný aj v nekontrolovanom prostredí, kde nemáme možnosť vytvárať ekvivalentné skupiny používateľov na korektné porovnanie, je použiť kombinovanie odporúčaní a optimalizáciu váh podľa toho kritéria, ktoré potrebujeme vyhodnocovať. Podľa výsledných optimalizovaných váh alebo ich priebehov v čase môžeme usúdiť, ktoré odporúčacie nástroje boli preferované a za akých podmienok. Ak sú vo výslednom kombinovanom odporúčaní čiastkové odporúčania viacerých metód, môžeme tiež sledovať, ktoré odporúčania používatelia nasledujú, t.j. ktoré ich zaujmú.

Problémom optimalizácie váh je, že pre viac optimalizovaných parametrov trvá dlho (potrebuje veľký počet interakcií so spätnou odozvou). V prípade výučbového systému bude možné použiť a kombinovať len málo rôznych metód odporúčania, zvlášť pri optimalizácii bez identifikovaných zhlukov. Na druhú stranu, väčší počet súčasne pracujúcich odporúčacích nástrojov je problematický aj z dôvodu veľkého zaťaženia systému. V prípade kolaboratívnej optimalizácie jedinej spoločnej konfigurácie odporúčacích nástrojov máme k dispozícii viac interakcií, a teda je možné použiť aj viac odporúčacích nástrojov (optimalizovať viac parametrov).

PodĎakovanie. Tento príspevok vznikol vďaka čiastočnej podpore grantu KEGA 028-025STU-4/2010 a v rámci OP Výskum a vývoj pre projekt: Podpora dobudovania Centra excelentnosti pre Smart technológie, systémy a služby II, ITMS: 26240120029, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Referencie

1. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In: IEEE Trans. on Knowledge and Data Engineering, vol. 17, pp. 734-749. IEEE Computer Society (2005)
2. Brusilovsky, P., Ahn, J., Dumitriu, T., Yudelson, M.: Adaptive Knowledge-Based Visualization for Accessing Educational Examples. In: Tenth International Conference on Information Visualisation, pp. 142-150. IEEE Computer Society (2006)
3. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. In: User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370. Hingham (2002)
4. Kostelník, R., Bieliková, M.: Web-based environment using adapted sequences of programming exercises. In: Proc. of Information Systems Implementation and Modelling – ISIM 2003, pp. 33–40. MARQ Ostrava, Brno (2003)
5. Michlík, P., Bieliková, M.: Exercises Recommending for Limited Time Learning. In: Proc. of the 1st Workshop on Recommender Systems for Technology Enhanced Learning, Procedia Computer Science, vol. 1, no. 2, pp. 2821-2828. Elsevier (2010)
6. Bieliková, M. et al.: ALEF: Web 2.0 Principles in Learning and Collaboration. In: Int. Conf. on e-Learning and the Knowledge Society, pp. 54-59. Riga Technical University (2010)
7. Yu, Z. et al.: Content Provisioning for Ubiquitous Learning. In: IEEE Pervasive Computing, vol. 7, no. 4, pp. 62-70. IEEE Computer Society (2008)

Trendy a budúcnosť odporúčania online správ

Mária Bieliková, Michal Kompan, Dušan Zeleník

Ústav informatiky a softvérového inžinierstva, Fakulta Informatiky
a informačných technológií, Slovenská technická univerzita,
Ilkovičova 3, 842 16 Bratislava, Slovensko
{bielik,kompan,zelenik}@fiit.sk

Abstrakt. Príspevok sa zameriava na oblasť personalizovaného odporúčania správ na internete, špeciálne na webe. Venujeme sa hlavným problémom, ktoré sa odporúčacie systémy snažia riešiť. Načrtávame riešenia, spolu s konkrétnymi príkladmi. Na základe existujúcich metód odporúčania navrhujeme vylepšenia. Ponúkame návrhy riešení a podmienky, ktoré by mali spĺňať. Tiež uvádzame spôsoby ako metódy odporúčania v doméne online správ overovať a vyhodnocovať. Vyhodnotenie je pritom orientované na metódy, ktoré pracujú s čitateľmi a samotnými článkami, ktoré sú predmetom odporúčania. Článok tak prispieva ako prehľad, ale najmä ako inšpirácia pre nové odporúčacie systémy.

Kľúčové slová: odporúčanie, personalizácia, články, správy, čitatelia

1 Úvod

Rozmach internetu, ktorému čelíme v posledných rokoch, prináša viaceré problémy. Pozorujeme enormný nárast informácií v každej oblasti ľudskej činnosti. Výnimku netvorí ani spravodajské portály, poskytujúce v snahe prilákať čo najviac používateľov stovky nových článkov denne. Priemerný používateľ strávi čítaním online správ denne približne 16 minút pri dvoch návštevách¹. Vzhľadom na množstvo pridávaných článkov tak používateľ nie je schopný pristupovať k takým článkom, ktoré by si za iných okolností prezrel. V priebehu rokov sa postupne mení aj samotný obsah online správ. Nejedná sa už len o textové správy, ale v čoraz väčšej miere správy obsahujú multimedialný obsah. Výraznou črtou online správ a správ ako takých je rýchla degradácia ich informačnej hodnoty.

Metódy pre odporúčanie na webe sú predmetom aktívneho výskumu od polovice 90 rokov. Vzhľadom na povahu najčastejšie odporúčaného obsahu (text) sú základné myšlienky prevzaté z oblastí vyhľadávania informácií, dolovania v dátach prípadne umelej inteligencie. Všeobecne môžeme úlohu odporúčania charakterizovať ako:

$$\forall c \in C, s_c = \arg \max_{s \in S} u(c, s)$$

kde C je množina všetkých používateľov, S všetkých prvkov (správ), ktoré sa môžu odporučiť a u je funkcia vyjadrujúca užitočnosť správy pre daného používateľa [1].

Trendy v metódach odporúčania sú v súčasnosti najmä v kolaboratívnych modeloch [8, 9]. Tieto metódy využívajú masu používateľov. V prostredí

¹ Aimmonitor.sk

internetových správ ide o monitorovanie čitateľa a článkov, ktoré prečítal [4], prípadne ohodnotil ako zaujímavé. Princíp vychádza z podobnosti čitateľov na základe časti rovnakých prečítaných článkov. Vďaka povahe odporúčaní založených na monitorovaní používateľov vznikajú problémy s ochranou *súkromia jednotlivca* alebo zneužívanie známych princípov na *manipulovanie odporúčania*. Okrem všeobecného problému nového *neznámeho používateľa*, nastáva aj problém s novou, *neznámou položkou* – spoločne označované ako *studený štart*. Nové, ešte neznáme články nie sú čítané a v rýdzo kolaboratívnych modeloch ich nemôžeme odporúčať inak ako náhodne. Ďalším bežným problémom sú *šedé ovce*, a teda používatelia, ktorý nesympatizujú s väčšinou ale ani nie sú na opačnom póle oblasti záujmov. Odporúčania pre týchto čitateľov nie je ľahké generovať tak, ako pre väčšinu.

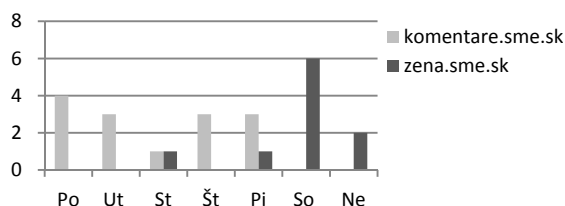
Metódy odporúčania založené na obsahu analyzujú jednotlivé položky a výsledky analýzy sa využijú na hľadanie obsahového súvisu. Odporúčania sú tak orientované na jednotlivca, bez vplyvu okolia [7]. S novými možnosťami práce s textom, obrazom či zvukom sa tieto metódy postupne znovu dostávajú do popredia [5]. Navrhli sme dve metódy odporúčania na základe obsahu a tieto overovali na dátach z portálu SME.sk [7]. Pri vyhodnocovaní metód sa ukázali aj známe problémy. V prípade článkov a ich významu je najčastejší problém *riedkosť matice* opisujúcej príslušný článok, čo však významne súvisí s reprezentáciou článku – navrhli sme úspornú vektorovú reprezentáciu významných slov článku doplnenú o ďalšie položky ako index čitateľnosti či kategórie článku. Spracovanie textu a odhaľovanie sémantickej podobnosti veľmi ovplyvňuje jazyk a gramatika [6]. Problémy vznikajú pri *synonymách* alebo *polysémach*. Ďalším problémom je *nadmerná špecializácia* odporúčania, ktorá nastáva ak odporúčania v čitateľovi vzbudzujú dojem opakovania sa správ. Tento stav môže privodiť aj odporúčanie samotné, keď návrhy článkov postupne zužujú priestor odporúčaní.

Kombináciou obsahových a kolaboratívnych metód vznikajú hybridy [1]. Keďže tieto metódy získavajú výhody oboch princípov, môžeme sledovať trend ich použitia. Popri získaných výhodách a čiastočne vyriešených problémoch však ostávajú problémy, ktoré sú spoločným menovateľom všetkých princípov. Často sa stretávame s používateľmi negatívne prístupujúcich ku odporúčaniam. Od obáv o stratu súkromia, cez nedôveru v metódu odporúčania, až po netransparentnosť riešenia tak používateľ prejavuje nespokojnosť [2].

2 Vízia odporúčania online správ

Okrem spomenutých problémov môžeme počas používania existujúcich systémov pozorovať aj ďalšie nedostatky. Zaujímavé je riešiť problematiku zmeny záujmov čitateľa. Ten sa mení nielen dlhodobo, ale i krátkodobo, dokonca periodicky vzhľadom na čas, ale aj miesto. Príkladom môže byť iný záujem o články ráno v práci ako po nedeľnom obede. Ak sa metóda odporúčania nie je schopná prispôbovať, alebo dokonca, ak časté zmeny záujmov používateľa negatívne ovplyvňujú kvalitu odporúčaní, potom je nutné objavovať znalosť vykresľujúcu meniace sa návyky. Následným využitím pravidiel alebo pravdepodobnostných modelov vieme začleniť do personalizovaného odporúčania i čas, prípadne priestor – lokalitu, v ktorej sa

používateľ nachádza. Pritom informácia o čase je základným atribútom, ktorý vieme pozorovať. A spomínaná lokalita sa v dnešnej a blízkej dobe mobilných zariadení stáva stále dostupnejšou. Na obrázku 1 môžeme pozorovať rozdielny záujem čitateľa počas týždňa vytvorený na základe analýzy záznamov čitateľov denníka sme.sk. Vybrali sme dve sekcie a priemerné čítanie počas mesiaca.



Obrázok 1. Porovnanie priemerného čítania dvoch sekcií na SME.sk pre vybraného čitateľa.

Zaujímavé je aj odporúčanie pre konkrétneho používateľa rozšírené na problém odporúčania pre skupiny. Jedná sa o rozšírenie priestoru „preferencií“ kedy sa okrem zohľadnenia aktivity konkrétneho používateľa, prihliada aj na jemu podobných používateľov, resp. všetkých členov danej skupiny, ktorá môže byť priamo definovaná, alebo získaná prostredníctvom odhaľovania intenzity vzťahov v prostredí sociálnych sietí.

V prípade veľkých spravodajských portálov, môže byť odporúčanie generované pre konkrétneho používateľa neefektívne. Preto sme navrhli generovanie odporúčaní pre „meta-používateľov“, ktorí reprezentujú širšie záujmové skupiny, či už prístupom založenom na odporúčanom obsahu, prípadne hybridnom modeli.

3 Spôsoby overenia metód odporúčania a záver

Overenie navrhnutých metód odporúčania je netriviálna úloha. V prvom rade je treba charakterizovať „úspešnú“ odporúčaciu metódu. Na jednej strane môžeme za úspešnú metódu považovať takú, kedy používatelia prečítajú viac správ ako pred jej zavedením. Na strane druhej to môže byť metóda, vďaka ktorej používateľ prečíta správ menej, avšak prinesú mu vyššiu informačnú hodnotu. Iným spôsobom merania kvality odporúčacích metód býva pomer prečítaných článkov ku počtu odporúčaných. Všetky tieto prístupy sa však snažia detegovať používateľov postoj k odporúčanému obsahu – či bol pre neho zaujímavý a užitočný. To nás vedie k priamočiaremu riešeniu, spýtať sa používateľov, či sú s daným odporúčaním spokojní. Postoj čitateľa však môže byť výrazne ovplyvnený chápaním a očakávaním používateľa. Je preto dôležité zvoliť správny spôsob prezentácie odporúčaného obsahu.

Pri vyhodnocovaní úspešnosti metód je však nevyhnuté zohľadniť grafickú reprezentáciu daného zdroja online správ, kedy môže aktuálne usporiadanie prvkov na internetovej stránke výrazne ovplyvniť správanie používateľa (titulná stránka, poradie v zozname odporúčaných správ a pod.). Tento problém môžeme pozorovať pri využití syntetických testov, kedy sa snažíme predpovedať správanie používateľa bez reálneho zohľadnenia aktuálnej grafickej reprezentácie spravodajského

portálu [3]. Takýto experiment využije údaje zozbierané počas reálneho využívania riešenia. Pomocou testovacej a trérovacej množiny tak zisťujeme úspešnosť metódy. Pri tomto teste treba zohľadniť problém vyhodnocovania úspešnosti predikcie správania čitateľa a nie úspešnosti odporúčania. Porovnania na testovacej množine tak musia zohľadňovať nie konkrétne články ale fakt, či je sémantika článkov pokrytá.

Metódy odporúčania sú aj po niekoľkých rokoch skúmania aktuálnou témou, kedy sa snažia reagovať na nové problémy a etablovať do viacerých domén. Súčasný prístup musí vyriešiť nie len enormným množstvom používateľov, odporúčaného obsahu, ale musia pracovať s viacerými typmi odporúčaného obsahu. Nezanedbateľné sú otázky týkajúce sa bezpečnosti a dôveryhodnosti takýchto systémov.

Dáta získané používaním internetových novín SME.sk sme použili na vytvorenie viac alternatív odporúčania [3,7,9]. Vytvorili sme obsahové a kolaboratívne odporúčania, ktoré overovali hlavne synteticky. Pozorovania týchto odporúčačích metód a ich nedostatkov sme využili pre víziu odporúčačieho systému budúcnosti.

PodĎakovanie. Táto Tento príspevok vznikol vďaka čiastočnej podpore grantu VEGA VG1/0508/09 a v rámci OP Výskum a vývoj pre projekt: Podpora dobudovania Centra excelentnosti pre Smart technológie, systémy a služby II, ITMS: 26240120029, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Adomavicius, G. and Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17, 6 (Jun. 2005), 734-749.
2. Ahn, J., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. 2007. Open user profiles for adaptive news systems: help or harm?. In *Proc. of the 16th int. Conf. on World Wide Web. WWW '07*. ACM, New York, NY, 11-20.
3. Barla, M., Kompan, M., Suchal, J., Vojtek, P., Zeleník, D., Bieliková, M., 2010. News recommendation. In *Proc. of the 9th Znalosti*, pp. 171-174.
4. Carvalho, C., Jorge, A. M., and Soares, C. 2006. Personalization of E-newsletters Based on Web Log Analysis and Clustering. In *Proc. of the IEEE/WIC/ACM int. Conf. on Web intelligence*. IEEE Computer Society, WDC, 724-727.
5. Das, A. S., Datar, M., Garg, A., and Rajaram, S. 2007. Google news personalization: scalable online collaborative filtering. In *Proc. of the 16th int. Conf. on World Wide Web. WWW '07*. ACM, NY, 271-280.
6. Kasper, W., Steffen, J., Zhang, Y. 2008. Using Semantics for News Navigation. *2008 IEEE International Conference on Semantic Computing*, 261-267.
7. Kompan, M., Bieliková, M. 2010. Content-Based News Recommendation. In *Proc. of the 11th Conf. EC-WEB*, 61-72.
8. Su, X., Khoshgoftaar, T. M. 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence, 2009*(Section 3), 1-20.
9. Suchal, J., Návrát, P. 2010. Full text search engine as scalable k-nearest neighbor recommendation system. *AI 2010, IFIP AICT 331*, Springer, 165-173.

Učení příklady – personalizovaný adaptivní Web

Jan Nekula¹, Petr Šaloun¹, Zdeněk Velart², Petr Klimánek¹

¹ Přírodovědecká fakulta, Ostravská univerzita v Ostravě 30. dubna 22,
701 03 Ostrava, Česká republika
{jan.nekula, petr.saloun, r09462}@osu.cz

² Fakulta elektrotechniky a informatiky, VŠB-TU Ostrava, 17. listopadu 15/2172,
708 33 Ostrava-Poruba, Česká republika
zdenek.velart@gmail.com

Abstrakt. Navigace nad příklady je možná pomocí technik adaptace a personalizace s využitím ohodnocení znalostí a vazeb v prostoru konceptů. Použití příkladů vede k přesnějšimu mapování výukových objektů na koncepty, což umožňuje přesnější navigaci. Student je navigován na příklady, které rozšíří jeho aktuální znalosti v aplikační doméně, nebo s ohledem na zvolené dílčí cíle. Navigace využívá jazykově nezávislé ohodnocení konceptů díky technologiím WordNet či Google Translate, a je připravena pro vícejazyčný provoz (např. v češtině, angličtině a turečtině s možností libovolného přepínání). V mezinárodní spolupráci připravujeme experimentální ověření a vyhodnocení námětů prezentovaných v tomto textu.

Klíčová slova: personalizace, adaptace, webový systém, výuka příklady, koncept, navigace,

1 Úvod

Existuje mnoho možností jak studenta navigovat a jak mu předložit výukové objekty (learning objects – LO). Pomocí technik adaptace a personalizace s využitím ohodnocení znalostí a vazeb v prostoru konceptů (concept space – CS) lze navigovat studenty v rámci kurzu, což prezentujeme v [1]. Tento přístup rozšiřujeme o možnost předložit vzorový příklad se zdrojovými texty – nabízíme tak zpřesnění možnosti průchodu kurzem. Případová studie bude realizována v námi vyvinutém systému XAPOS [2]. V rámci systému XAPOS je s příklady zacházeno stejně, jako se standardními výukovými objekty (webové stránky). Díky tomu můžeme využít systémovou vícejazyčnost a možnost personalizované a adaptivní navigace. Využití příkladů zpřesňuje mapování LO na koncepty, což vyplývá z jejich struktury – příklady obsahují pojmy z prostoru konceptů přímo, například názvy použitých funkcí. K mapování lze využít i informaci strukturovaný –text LO označovaný tagy dovoluje rozpoznat slova z názvu příkladu, kapitoly, nadpisu, popisu obrázku či zdrojového textu, případně i z komentářů ve zdrojovém kódu.

2 Související práce

Možnostmi výuky pomoci příkladů se zabýval například Brusilovsky, jeho systémy WebEx, později pak NavEx [3] využívají pro navigaci mezi příklady systém prerekvizit a výstupních konceptů (odpovídající požadovaným vstupním a naučeným výstupním znalostem). Dalšími možnostmi navigovat uživatele nad obsahem jsou:

- využití metod pro automatické odhady zájmu uživatele podle dosud navštívených webových stránek, viz [4], nebo - pomoci vytváření mapy témat (topic maps), jako v [5]. Při navigaci je také možno využít přístupů vyhledávající tzv. „long tail“ témata, která jsou v uživatelově zájmu, ale zároveň nejsou „přepopularizována“ [6].

3 Adaptivní a personalizovaný systém XAPOS

Systém XAPOS naviguje studenta nad LO reprezentující obsah – studijní materiál, ukázkový příklad, případně jiný multimediální objekt, jejichž koncepty jsou popsány v CS. Každý LO je provázán s jedním nebo více koncepty z CS. Koncepty v CS reprezentují základní pojmy dané problémové domény. Po jejich naučení se studenty se stávají znalostmi, o které v konečném důsledku jde. Vedle konceptů samotných patří do CS i vzájemné vztahy mezi koncepty. Vztahy mohou být strukturální, reprezentující strukturu CS (typickým příkladem je vztah subClassOf), nebo vztah vyjadřující závislost mezi koncepty (např. prerequisite, definedBy, requires).

Navigace studenta nad LO probíhá v závislosti na aktuálních znalostech studenta, nebo s ohledem na zvolené cíle. Základním principem navigace v XAPOSu je, že existující vztah v prostoru LO koresponduje s příslušným vztahem mezi koncepty v CS. Správné ohodnocení vztahů v CS umožňuje XAPOSu studenta navigovat s ohledem na jeho aktuální znalosti a pozici v obsahu, který LO představují.

K ohodnocení vztahů v CS využíváme upravený PageRank algoritmus, který byl navržen pro hodnocení vztahů mezi webovými stránkami a Google jej využívá pro jejich vyhledávání. Původní PageRank rozšiřujeme o započtení strukturálních a závislostních vztahů. XAPOS studentovi nabízí vhodné LO uspořádané dle jejich ohodnocení a dle aktuálních znalostí studenta. Vedle toho může student přejít na LO, který je množinou svých konceptů nejvíce podobný aktuálně zobrazenému LO. XAPOS dovoluje i přímý přechod na libovolný LO obsažený v kurzu.

Jazyková nezávislost XAPOSu při navigaci nad obsahově shodnými LO v různých přirozených jazycích je postavena nad hotovým kurzem s LO v jednotlivých jazycích. Pojmy v konceptech v CS jsou vytvořeny ve zvoleném jazyce (angličtině, tedy jednojazyčně), vztahy mezi LO a koncepty v CS jsou stejné, bez ohledu na jazyk LO. Navigace nad LO je potom transparentní vzhledem k jazyku zvolenému pro prezentaci obsahu LO, vychází jen z ohodnocení vztahů v CS a ze znalostí studenta. Volba jiného jazyka pro zobrazení obsahu i navigaci jsou pak jen záležitostí GUI.

4 Příklady jsou specializovaný obsah

XAPOS zpracovává LO obsahující příklad se zdrojovými texty stejně, jako každý jiný LO. Pojmy z příkladu se i díky tagům snadněji extrahují, jejich umístění v CS a vzájemné vazby mezi koncepty s pojmy jsou pak stejné, jako u běžných LO. U příkladů student zřejmě využije nabídku XAPOSu odkazující na příklady podobné.

4.1 Import nových příkladů do systému

V XAPOSu jsou LO označovány HTML tagy a jsou uloženy jako soubory v souborovém systému. Koncepty, jejich vztahy a přídavné informace (jméno, anotace, jazyk apod.) jsou uloženy v databázi. Při přidávání nových LO do XAPOSu se vkládané LO mapují na koncepty v CS. Mapování je v současnosti řešeno ručně. Vyvíjí se však podsystém pro automatické získávání pojmů z vkládaných HTML LO za využití technologií WordNet případně Google translate.

Zpracování příkladů probíhá semiautomaticky. Příklad s HTML značkami je do systému zařazen automaticky, manuálně se však musí zapsat seznam konceptů, na které je příklad v systému vázán. V této oblasti také vyvíjíme plně automatický systém, viz dále.

Na vstupu je příklad zapsán ve formě XML dle daného DTD – XHTML a našich tagů. LO příkladu obsahuje základní informace -- název, anotace, zadání (popis) a řešení v daném programovacím jazyce. Jazyk textu může být atributem každého tagu. Součástí LO jsou i vstupní a výstupní data pro otestování příkladu studentem.

4.2 Automatické zpracování příkladů a průchod kurzem

Pro mapování vzorových příkladů se zdrojovými texty na koncepty je výhodou, že přímo obsahují pojmy z CS – klíčová slova, vybrané knihovní funkce, názvy konstant..., a k tomu mají i přesnou syntaxi. Pro mapování je možno využít i dodatkové informace obsažené v názvu, anotaci a komentářů v rámci samotného kódu příkladu. Automatické zpracování příkladu je nyní v systému XAPOS ve fázi vývoje. Jakmile bude import zautomatizován, práci autora příkladů bude pouze jejich příprava. O samotné vložení, provázání a navigaci se postará XAPOS.

Při vstupu do kurzu má student nulové znalosti v problémové doméně. Jakmile student začne procházet jednotlivé LO, začíná se množina jeho znalostní rozšiřovat – personalizovaná navigace XAPOSu toho využívá, viz předchozí popis XAPOSu.

XAPOS může obsahovat i zkušební testy, kterými si student ověří stupeň zvládnutí obsahu. Úspěch i neúspěch v položkách testu aktualizuje množinu znalostí studenta – znalosti jsou přidávány i odebírány. Personalizovaná navigace nabízí další průchod kurzem dle aktuální množiny znalostí, při zapomenutí či chybě tak nabídne příslušné LO studentovi i opakovaně.

5 Případová studie

XAPOS byl využit pro experimentální ověření navigace modifikovaným PageRank algoritmem. Předkládaný obsah byla část kurzu programování ve funkcionálním jazyce Lisp. Důvodem je relativní neznalost jazyka Lisp mezi studenty – díky ní můžeme předpokládat stejné vstupní znalosti mezi studenty. Jazykem pro tvorbu konceptů a tvorbu vztahů mezi nimi je angličtina, LO kurzu se připravují souběžně v angličtině, češtině a turečtině. Připravovaný experiment bude zaměřen na řešení podobnosti obsahu LO při výuce pomocí příkladů.

6 Závěr a budoucí práce

V příspěvku jsme popsali rozšíření adaptivní a personalizované navigace nad LO v systému XAPOS o vzorové příklady. Hlavním přínosem práce je doplnění XAPOSu o využití znalosti struktury příkladů (tagy) i o snadnější extrakci pojmů pro tvorbu konceptů a jejich vztahů. XAPOS přitom zůstává beze změny personalizované navigace dosud využívané jen pro LO bez zdrojových textů.

V budoucnu plánujeme rozvoj automatického zpracování a zařazení nových řešených i neřešených příkladů. Zejména budeme řešit problém podobnosti příkladů se zdrojovými texty tak, abychom uživateli nabídli správný příklad z více podobných. Další oblastí rozvoje bude zpřesnění zpracování průchodu kurzem s využitím grafových i statistických nástrojů a vizualizace výsledků. V příštím semestru také uskutečneme praktický experiment navigující studenty, nad řešenými příklady se zdrojovými texty.

Výzkum byl částečně podpořen projekty SGS21/PřF/2010 a FRVŠ ČR 24/2010.

Literatura

1. Šaloun P., Velart Z., Concept Space Rating for Personalization of Learning Materials Based on Relations. 4th International Workshop on Semantic Media Adaptation and Personalization, SMAP 2009, pp 67-72.
2. Šaloun P., Velart Z., Nekula J. Navigation over multilingual content using one concept space: controlled experiment. 5th International Workshop on Semantic Media Adaptation and Personalization, SMAP 2010, accepted.
3. Yudelson M., Brusilovsky P., Sosnovsky S. Accessing Interactive Example with Adaptive Navigation Support. Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04)
4. Holub M., Bieliková M., Estimation of User Interest in Visited Web Page. Proceedings of the 19th international conference on World wide web. WWW '10, Raleigh, USA, 2010, pp 1111-1112
5. Dicheva D., Dichev C. Helping Courseware Authors to Build Ontologies: the Case of TM4L. In 13th Int. Conf. on AI in Educ., AI-ED 2007, July 9-13, LA, CA, pp. 77-84.
6. Mi Zhang, Hurley N., Niche Product Retrieval in Top-N Recommendation, the 2010 IEEE/WIC/ACM International Conferences on Web Intelligence (WI/IAT 2010), Toronto, Canada, 2010

Internet a Technológia

Adaptívny proxy server: prevádzka a skúsenosti po roku

Tomáš Kramár, Michal Barla, Mária Bieliková

Ústav informatiky a softvérového inžinierstva,
Fakulta informatiky a informačných technológií,
Slovenská technická univerzita v Bratislave
{Meno.Priezvisko}@fiit.stuba.sk

Abstrakt Pred rokom sme na dielni WIKT 2009 predstavili prototyp adaptívneho proxy servera, ktorý umožňuje prostredníctvom série zásuvných modulov spracovávať a modifikovať obsah správ posielených medzi používateľovým prehliadačom a cieľovým serverom. Vďaka tomu je možno sledovať a zachytávať aktivitu používateľa na celom webe a na jej základe vykonávať také prispôsobovanie, resp. personalizovanie webu, aké sme doteraz poznali iba z izolovaných adaptívnych webových systémov. V tomto príspevku opisujeme aktuálny stav projektu; sumarizujeme obdobie, počas ktorého bol proxy server používaný na FIIT STU; opisujeme experimenty, ktoré sme s využitím tohto proxy servera vykonali a zamýšľame sa nad budúcim smerovaním projektu.

1 Adaptívny proxy server

V súčasnosti má väčšina pokusov o personalizáciu, či iné obohatenie webu [2] jeden hlavný nedostatok: sú veľmi úzko špecializované na jeden konkrétny portál, či službu, čo zahŕňa aj proprietárne spôsoby tvorby modelu používateľa. Konkrétne metódy tým pádom nemajú prístup k iným údajom o používateľovi, ako k tým, ktoré používateľ poskytne v rámci služby, na ktorej je metóda nasadená a naopak, neexistuje možnosť personalizovať, či inak prispôbovať ostatné portály či služby, kde by to mohlo byť pre používateľa výhodné.

Nosnou myšlienkou adaptívneho proxy servera [1] je preklenúť túto bariéru, posunúť metódy modelovania používateľa a personalizácie zo servera smerom ku klientovi a vytvoriť tak platformu pre rozličné metódy s cieľom obohatovať a personalizovať celý “divoký” web. Kým klasický proxy server slúži iba ako preposielač správ medzi klientom a serverom, náš adaptívny proxy server môže tieto správy modifikovať. To umožňuje jednotlivým metódam zasahovať do štruktúry odpovede servera, upravovať HTML kód posielený klientovi a tým celé surfovanie personalizovať. Jednoduchým spôsobom sa tak používateľovi môžu ponúknuť nové služby, resp. vylepšenia existujúcich, poskytnúť mu tak známe rozhranie a ťažiť z jeho existujúcich návykov.

Adaptívny proxy server je založený na existujúcom proxy serveri Rabbit¹. Ten sme rozšírili o systém zásuvných modulov, ktoré umožňujú flexibilne odchy-

¹ Proxy server Rabbit, <http://www.khelekore.org/rabbit/>

távať a modifikovať HTTP požiadavku klienta, či odpoveď servera. Tieto zásuvné moduly sú rozdelené do dvoch kategórií – služby a spracovávajúce moduly. Služby poskytujú všeobecnú funkcionálnu nad správy, od základných, ako napríklad načítanie, či úprava tela správy, až po pokročilé, ako prístup k DOM modelu odpovede, či preusporiadanie alebo doplnenie výsledkov vo vyhľadávачi. Spracovávajúce moduly kompozíciou vhodných služieb tvoria telo samotnej metódy.

Za účelom tvorby modelu používateľa a neskoršej analýzy dát sme vytvorili niekoľko služieb a spracovávajúcich modulov, ktoré zabezpečujú zapísanie metadát o každej návšteve ľubovľnej stránky do databázy. Spolu s navštívenou adresou, identifikátorom používateľa a časom návštevy zapisujeme aj kľúčové slová a pojmy extrahované z každej stránky [5]. Na extrakciu pojmov sme vytvorili webovú službu Metall², ktorá z poslaného HTML kódu extrahuje hlavnú informačnú časť (napr. telo článku), tzn. odstráni nepotrebné časti ako sú menu, hlavičky a pätičky, reklamy a pod., odstráni HTML značky, získaný čistý text potom preloží do angličtiny (služba google translate) a agregáciou výsledkov rôznych webových služieb (OpenCalais, tagthe.net, Alchemy) a knižníc pre spracovanie textu (jkey-extractor) získa ohodnotený zoznam pojmov.

Pre potreby jednoznačného priradenia každého dopytu správne použiteľovi sme používateľom priradili unikátny, náhodne vygenerovaný 32 miestny reťazec, ktorý sa doplní do *User-Agent* hlavičky používateľovho prehliadača, ktorá sa posiela pri každej požiadavke a predstavuje teda vhodné miesto pre umiestnenie identifikátora. Na stránke projektu peweproxy.fiit.stuba.sk sme pripravili nástroj pre automatické nastavenie všetkých nainštalovaných prehliadačov. Poskytujeme tiež informácie pre pokročilejšie nastavenie proxy servera ako aj prístup k samotným záznamom, ktoré proxy server k danému identifikátoru uchováva (ľubovľný z nich má používateľ možnosť odstrániť).

Proxy server bol do krátkej testovacej prevádzky uvedený 25. marca 2010 a následne bol predstavený študentom našej fakulty. Ku dňu vyhodnocovania (4. mája 2010) ho používalo 40 používateľov, ktorí si cez proxy zobrazili takmer 650 tisíc webových stránok (z toho zhruba 111 tisíc bolo unikátnych).

2 Projekty na platforme proxy

Na platforme adaptívneho proxy sme vyvinuli a nasadili niekoľko projektov. Prvý z nich sa zaoberá zjednotnotením vyhľadávania. Cieľom tohto projektu [4] je pomôcť používateľom pri formulácii a vyhľadávaní nejednoznačných dopytov doplnením kľúčových slov, ktoré spresnia dopyt podľa záujmov používateľa. Ak napr. používateľ vyhľadáva výraz **jaguar** a zaujíma sa o autá, jeho dopyt sa preformuluje na **jaguar car**. Pre iného používateľa, ktorý sa zaujíma o zvieratá sa však tento dopyt preformuluje na **jaguar animal**.

V tomto projekte sme proxy platformu využili na vytvorenie modelu používateľa aj na integráciu do vyhľadávača. S použitím štandardných záznamov proxy (pojmy a kľúčové slová každej navštívenej stránky) sme pre každého používateľa vytvorili jeho *bag-of-words* model, zložený z charakteristických pojmov

² Metall, <http://peweproxy.fiit.stuba.sk/metall/>

a kľúčových slov získaných z ním prezeraných stránok. Následne sme prekryvom týchto modelov vytvorili váhovanú sociálnu sieť (váha hrany je určená veľkosťou prekryvu) a jednoduchým šírením aktivácie sme získali komunity podobných používateľov. Tieto komunity tvoria kontext používateľovho vyhľadávania: pri vyhľadávaní sa doplnia také kľúčové slová, ktoré sa v metadátach používateľov z komunity vyskytujú často spolu s vyhľadávaným výrazom.

Vďaka proxy serveru sme mohli metódu overiť priamo nad vyhľadávačom Google. Nami vytvorené moduly do proxy zachytávali každé vyhľadanie v tomto vyhľadávači, ďalšie moduly (služby) zabezpečili doplnenie výrazu o kľúčové slová a s použitím Google API vyhľadanie novej sady výsledkov. Nové výsledky sme potom doplnili do HTML odpovede vyhľadávača, takže pre používateľa sa výsledky zobrazovali úplne nerušivo, akoby ich vygeneroval samotný vyhľadávač.

Cielom druhého z projektov [3], vyvinutých na platforme adaptívneho proxy servera, je zjednodušenie navigácie po portáloch. Rôzni, ale podobní používatelia majú na portáli často podobné informačné záujmy; hľadaná informácia však často nie je dostupná priamo z hlavnej stránky, ale je potrebné sa k nej preklikať cez systém menu a odkazov. V tomto projekte sme zaviedli zdieľanie úsilia medzi návštevníkmi webového sídla – teda, ak jeden používateľ takúto dôležitú informáciu našiel, ostatní používatelia, ktorí ju tiež potrebujú, ju už nemusia náročne hľadať. Vhodným príkladom takéhoto portálu je sídlo našej fakulty `fiit.stuba.sk`, na ktorom hľadajú podobní používatelia podobné informácie v rovnakom čase: napr. bakalári informácie o odovzdávaní bakalárskeho projektu. Aj v tomto prípade sme využili proxy platformu na získanie dát o používateľoch (*clickstreams*) a na integráciu personalizovaného menu do zvoleného portálu – v tomto prípade už spomínané sídlo FIIT. Na základe podobnosti *clickstreamov* sme určili podobných používateľov a na základe návštevnosti jednotlivých stránok portálu ich dôležitosť. Získané informácie sme používateľom, s použitím proxy servera, prezentovali na portáli vo forme nerušivého rozšírenia samotného portálu – kalendár akcií (napríklad s termínmi skúšok, či odovzdávaniami projektov) a osobné novinky s informáciami neviazanými na dátum.

Oba projekty boli charakteristické tým, že nám umožnili zintegrovat naše metódy do existujúcich služieb a portálov a overiť ich tak v reálnom prostredí, kde si používatelia niekedy možno ani nevšimli, že pracujú s obohatenými verziami služieb. Bez použitia proxy servera by sme museli vytvárať náš vlastný vyhľadávač alebo portál, ktorý by sme museli používateľov “donútiť” používať. Navyše, najmä v prvom projekte sme dobre využili skutočnosť, že máme k dispozícii kompletnú aktivitu používateľa na webe, čo nám umožnilo komplexnejšie zachytiť jeho záujmy a pretransformovať ich do pomoci pri vyhľadávaní.

3 Budúcnosť projektu

Pri doterajšej prevádzke sa ako najväčší problém ukázala neochota používateľov surfovať cez proxy. Najčastejšie uvádzaný dôvod je obava o stratu súkromia, napriek tomu, že celé riešenie má otvorené zdrojové kódy, je anonymizované a navyše záznamy možno aj zmazať. Predpokladáme, že tento problém zmier-

nime zlepšením komunikácie, vytvorením vhodnejších nástrojov pre správu prístupových záznamov a najmä ponúknutím užitočných služieb. Ďalším často uvádzaným dôvodom nepoužívania je stabilita – počas prevádzky sa vyskytli občasné výpadky spôsobené *softlockmi* vo virtualizovanom jadre operačného systému na ktorom bol proxy server nasadený. Tento problém sme adresovali presunom na inú virtualizačnú platformu a pracujeme aj na monitorovaní jednotlivých služieb proxy platformy. Je to veľmi dôležité, keďže každý výpadok znamená stratu používateľov, ktorí okamžite nastavlia svoj prehliadač, aby proxy nepoužíval.

V súčasnosti tiež pracujeme na ďalšej verzii proxy, kde sa zameriavame na zjednodušenie API používaného na tvorbu modulov, tak, aby bolo ich vytváranie čo najjednoduchšie. Vylepšujeme tiež proces extrakcie metadát zo stránok, s cieľom zvýšiť ich kvalitu. Za týmto účelom sme vytvorili už spomínanú webovú službu *Metal*. Zamýšľame sa aj nad variantom, kedy by sa proxy server, resp. služby, ktoré vykonáva presunul úplne na stranu klienta v podobe rozšírenia webového prehliadača, ktoré by využívalo istú formu pamäte zdieľanej s ostatnými inštanciami u iných používateľov, resp. s centralizovanými službami.

V budúcnosti plánujeme spoluprácu s ďalšími univerzitami a inštitúciami, tak, aby existovalo viacero distribuovaných inštalácií proxy servera, so synchronizovanou databázou. Tým by sme dokázali rozšíriť používanosť proxy servera a rozšíriť a diverzifikovať tým sadu záznamov o aktivitách na webe, ktorá môže slúžiť ako základ pre zaujímavé analýzy správania a odhaľovanie nových súvislostí. Predbežne máme na tieto aktivity dobré ohlasy z univerzity *Trinity College of Dublin*, na ktorej plánujeme nasadiť druhú inštaláciu proxy platformy.

Podakovanie. Tento príspevok vznikol vďaka čiastočnej podpore grantov KEGA 345-032STU-4/2010, KEGA 028-025STU-4/2010 a v rámci OP Výskum a vývoj pre projekt: Podpora dobudovania Centra excelentnosti pre Smart technológie, systémy a služby II, ITMS: 26240120029, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Barla, M., Bieliková, M.: Personalizácia “divokého” webu: adaptívny proxy server. In: Proc. of the 4th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2009). pp. 48–51. Equilibria (2009)
2. Barricelli, B.R., et al.: Personalized web browsing experience. In: Proc. of the 20th ACM Conf. on Hypertext and Hypermedia. pp. 345–346. ACM (2009)
3. Holub, M., Bieliková, M.: Estimation of user interest in visited web page. In: Proc. of the 19th Int. Conf. on World Wide Web. pp. 1111–1112. ACM (2010)
4. Kramár, T., et al.: Disambiguating search by leveraging the social network context based on the stream of user’s activity. In: Proc. of the 18th Int. Conf. on User Modeling, Adaptation, and Personalization. pp. 387–392. Springer (2010)
5. Noll, M.G., Meinel, C.: Web search personalization via social bookmarking and tagging. In: Proc. of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference. pp. 367–380. Springer-Verlag (2007)

Hry s účelom objavovania sémantiky na webe

Jakub Šimko, Michal Tvarožek a Mária Bieliková

Fakulta Informatiky a Informačných technológií,
Slovenská Technická Univerzita v Bratislave, Ilkovičova 3, 842 16, Bratislava
{jsimko, tvarozek, bielik}@fiit.stuba.sk

Abstrakt. Kľúčovou súčasťou webu (so sémantikou) sú metadáta tvoriace opis webových zdrojov a modely domén. Ich získavanie je však náročnou a aj nákladnou činnosťou pokiaľ to vykonáva človek. Stroje dnes vedia získať sémantiku len veľmi obmedzene. Ako alternatíva (alebo doplnok) sa ponúka koncept hier s účelom – GWAP („Game with a Purpose“), ktorého základnou filozofiou je využitie činnosti ľudského mozgu pri hraní hier. Hráč pre dosiahnutie víťazstva generuje znalosti v podobe hernej stratégie a taktiky. Pri vhodnej formulácii pravidiel možno tieto znalosti prebrať a použiť pri riešení reálnych problémov. V príspevku uvádzame príklady hier s uplatnením pre web so sémantikou: „Google Image Labeller“ a „Little Google Game“ a diskutujeme potenciál hier s účelom v tejto oblasti.

1 Úvod

Získavanie anotácií webových dokumentov či modelovanie domén je kľúčovou úlohou a zároveň úzkym hrdlom praktického rozšírenia pokročilých aplikácií na webe, resp. webe so sémantikou. V súčasnosti dominujú tri hlavné skupiny prístupov a ich kombinácie: automatické (extrakcia charakteristických pojmov dokumentov, tvorba doménových ontológií s využitím spracovania prirodzeného jazyka), kolaboratívne (anotácia dokumentov, tvorba folksonómii) a manuálne (využitie ľudských expertov pre modelovanie domén, anotácia dokumentov ich autormi).

Hoci automatické prístupy umožňujú postihnúť veľa dokumentov, nedostatočne postihujú heterogenitu divokého webu, sú nepresné pre nepostačujúce spracovanie prirodzeného jazyka a nedostatočne spracovávajú multimediálny obsah. Kolaboratívne prístupy síce stále vykazujú široký záber a zvládajú dobre aj multimediálny obsah, nedosahujú však dostatočnú špecifickosť v zmysle anotácií (značky v portáli *Flickr*¹) a sémantické štruktúry majú slabú úroveň organizácie (napr. folksonómia *Delicious*²). Problémom manuálnych prístupov sú rapídne rastúce náklady pri spracúvaní väčších korpusov a podrobnom modelovaní domén. Nedostatky uvedených prístupov bránia efektívnemu rozvoju webu so sémantikou.

¹ <http://flickr.com>

² <http://del.icio.us>

Keďže na základe (implicitných) vstupov od používateľov možno vytvoriť aj rozsiahle bázy znalostí, veľa výskumného úsilia sa v posledných rokoch orientuje na spôsoby motivovania používateľov odovzdávať svoje znalosti dobrovoľne. Príklad portálu *Delicious* ukazuje potenciál kolaboratívneho generovania sémantiky ako „vedľajšieho produktu“ činnosti, ktorú používateľ vykonáva dobrovoľne.

2 Hry s účelom

Jedným z typov aplikácií, ktoré ľudia používajú dobrovoľne a radi sú hry. Pred niekoľkými rokmi sa v okruhu počítačovej vedy prvýkrát objavil koncept „hry s účelom“ (GWAP, z anglického „Game With A Purpose“). Ide o počítačovú hru, v ktorej používatelia generujú znalosti ako súčasť ich herných stratégií, čím prispievajú k riešeniu reálnych problémov. Zároveň sú však uspokojovaní zábavou, ktorú im hra poskytuje. Potenciál hier závisí najmä od miery ich atraktívnosti [1].

Aby bolo možné objavovanie znalostí, musia byť pravidlá hry vopred navrhnuté tak, aby implikovali použitie víťazných stratégií totožných s charakterom reálneho problému ktorý hra rieši [1]. Zábavná stránka hry môže mať pritom celkom iný charakter. Príkladom je úspešná hra *Google Image Labeller*, v ktorej sa hráč snaží zhodnúť sa na pojme opisujúcom daný obrázok, so svojim anonymným partnerom, ktorý hrá hru v rovnakom čase. Zábavným prvkom je v tomto prípade pocit neznáma v druhom hráčovi. Reálnym prínosom hry je však kolaboratívna anotácia obrázkov pojmami ktoré hráči generujú počas hry [3].

Kolaborácia je v hrách s účelom častým prvkom, pričom rovnaké úlohy sú zvyčajne kladené viacerým hráčom pre rozpoznanie znalostí o ktorých panuje zhoda a elimináciu rozporov. Kolaborácia je pritom v synergii so súťažou – atraktivitu hry (a teda aj jej potenciálny záber) zvyšuje súperenie medzi jej účastníkmi.

Princíp hry s účelom sme realizovali v projekte *Little Google Game*³, ktorého účelom je vytváranie folksonómie súvislostí pojmov, využiteľnej pri vyhľadávaní informácií na webe so sémantikou [4, 5].

Little Google Game je hra na formulovanie textových vyhľadávacích dopytov. Úlohou hráčov je minimalizovať počet výsledkov vrátených vyhľadávačom. Dopyty musia byť formulované v špeciálnom tvare, napríklad: „jaguar –animal –car –company“. Pojem „jaguar“ je daný hrou, ostatné výrazy generuje hráč. Prefix mínusového znamienka je direktívou webovému vyhľadávaču, vynechať z pôvodnej množiny výsledkov (získanej pomocou dopytu „jaguar“) tie dokumenty, v ktorých sa dané pojmy nachádzajú. Preto, aby bol hráč úspešný a redukoval počet výsledkov čo najviac, musí hľadať negatívne pojmy často sa vyskytujúce v dokumentoch spoločne so zadaným slovom ergo, pojmy s ním súvisiace [4].

Zábavnou stránkou hry *Little Google Game* je jednak moment sebaaprekonávania pri vymýšľaní lepších pojmov a moment súťaživosti, v ktorom hráčov porovnávame a vytvárame ich rebríček na základe skóre. „Úžitok“ z hry potom charakterizuje dolovanie súvislostí pojmov, ktoré hráči sformulujú v dopytoch [4].

³ <http://mirai.fiit.stuba.sk/LittleGoogleGame/LittleGoogleGameTestPage.html>

Princíp tvorby siete pojmov pomocou hry sme overili v experimentoch s približne 200 používateľmi, ktorí odohrali viac ako 3000 hier. Správnosť prepojení vo výslednej folksonómii sme overili prostredníctvom dotazníka s dosiahnutím úspešnosti 91% (zúčastnilo sa 20 respondentov, všetci posudzovali rovnaké prepojenia, 91% vzťahov v sieti bolo označených za správne) [4].

3 Otvorené problémy hier s účelom

Doposiaľ chýba ucelená metodológia tvorby hier s účelom. Na základe predchádzajúcich prác zaoberajúcich sa hrami s účelom [2] a vlastnými skúsenosťami s *Little Google Game* sme identifikovali tieto problémové okruhy, na ktoré by takáto metodológia mala odpovedať v rámci procesu transformácie reálneho problému na hru s účelom.

1. **Definícia problému, ktorý má hra riešiť (účel).** Pred vytvorením hry musí existovať formálny opis problému, resp. podmienky jeho riešenia. Máme síce k dispozícii formalizmy ako predikátová logika či OWL, voľný zápis problému ich prostriedkami však môže ľahko znemožniť transformáciu opisu na pravidlá hry. Potrebujeme preto špecializovaný formalizmus spolu so súborom odporúčaní (v ideálnom prípade obmedzení) tvorby formálnych opisov problémov.
2. **Klasifikácia problémov a ich vhodnosť riešenia pomocou hry** chýba. Spomínaný *Google Image Labeller* ako aj ďalšie hry (*Tag a Tune*, *PopVideo*⁴) sa pomerne úzko orientujú na rozpoznávanie sémantiky (multimediálneho) obsahu webu. Možných typov problémov je však viac a klasifikácia by mala predovšetkým pomôcť pri určovaní vhodnosti problému na riešenie hrou. Hlavnými východiskami klasifikácie by mohol byť súbor ťažko riešiteľných problémov a objavovanie znalostí predovšetkým na webe a pre web (napr. klasifikácia, budovanie ontológií, anotácia dokumentov, správanie agentov, expertné systémy), a úspešné herné motívy modelujúce podobné situácie (napr. obchodné a výstavbové stratégie, labyrinty, slovné úlohy využívajúce prirodzený jazyk).
3. **Vymedzenie herného priestoru, podmienok víťazstva a pravidiel hry.** V súčasnosti sa deje skôr ad hoc a tvorca hry potrebuje značnú dávku kreativity (napriek pokusom vytvoriť „recept“ tvorby hier s účelom pre oblasť anotácie webu [1]). Pravidlá hry nesmú len motivovať k riešeniu problému, ale aj vytvárať zábavné a zaujímavé situácie pre hráčov a obmedzovať možnosti ich zneužívania, či už podvádzaním z hľadiska *fair-play* alebo generovaním nefunkčných riešení (v prípade *Little Google Game* boli tieto dva druhy deformácie spojené snahou hráčov uvádzať v dopytoch tzv. stop slová, na základe ktorých neférovovo získavali dobré skóre a tiež generovali nesprávne prepojenia vo folksonómii) [1, 4]. V tomto smere sa ukazuje ako perspektívne odhaľovanie osvedčených vzorov v pravidlách, ktoré vyvážia aspekty účelu hry, atraktívnosti a zamedzenia deformáciám.
4. **Zvyšovanie atraktivity hry.** Opäť uskutočňované skôr ad hoc. O zhrnutie vhodných praktík v tomto smere sa pokúsil Ahn vo svojej práci [1], kde

⁴ Dostupné na <http://www.gwap.com>

identifikoval typy motivácií hrať hru: sociálna interakcia, výzva prekonať seba a iných hráčov. Taktiež spomína potrebu brať do úvahy krivku učenia sa – teda (zlepšujúcu sa) úroveň hráčových schopností. Tieto princípy sa uplatňujú v hernom priemysle a tu vidíme potenciál v zameraní na ich vyčerpávajúce prevedenie do špecifickej oblasti tvorby hier s účelom v podobe súboru vzorov a odporúčaní.

5. **Zabraňovanie deformáciám hry.** K deformáciám dochádza pokiaľ hráči objavia v pravidlách „dieru“. Preto vidíme ako vhodné vypracovať algoritmickejší prístup detekcie deformačného správania sa hráčov hľadaním výherných stratégií nezodpovedajúcim účelu hry alebo reálne dosiahnuteľnému skóre.

4 Záver a budúca práca

Nedostatky doterajších prístupov k tvorbe metadát na webe možno aspoň čiastočne riešiť pomocou hier s účelom. Využívajú sa pre riešenie strojovo ťažko riešiteľných úloh, napr. anotácie multimediálneho obsahu ale ich potenciál sa neobmedzuje len na tento problém, čo demonštrujeme príkladom vlastného projektu *Little Google Game*.

Vývoj hier s účelom nie je jednoduchý a vyžaduje sklbenie často protichodných potrieb. Dostiaľ neexistuje spoľahlivý rámec vývoja týchto hier, ktoré sú vytvárané ad hoc. Našou víziou je vytvorenie všeobecného aj špecializovaného rámca riešenie problémov v kontexte webu (so sémantikou) a návrh metodológie tvorby hier s účelom, ktorá bude zahŕňať klasifikáciu reálnych problémov s mapovaním na vzory osvedčených pravidiel hier. Chceme posunúť tvorbu hier z účelom od „náhodných“ projektov k cieľavedomej činnosti začínajúcej vždy definíciou problému.

PodĎakovanie. Tento príspevok vznikol vďaka čiastočnej podpore grantov VEGA VG1/0508/09, KEGA 028-025STU-4/2010 a v rámci OP Výskum a vývoj pre projekt: Podpora dobudovania Centra excelentnosti pre Smart technológie, systémy a služby II, ITMS: 26240120029, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Ahn, Lv, Dabbish L.: Designing games with a purpose. *Communications of the ACM* 51. 58–67 (2008)
2. Chen, L., Wang, B., Chen K. The design of puzzle selection strategies for GWAP systems. In: *Concurrency and Computation: Practice & Experience*. 890–908 (2010)
3. Seneviratne, L.: An Interactive Framework for Image Annotation through Gaming. In: *Electronic Engineering*. 517-526. (2010)
4. Šimko, J., Tvarožek, M., Bieliková, M.: Little Google Game: Tvorba siete pojmov prostredníctvom vyhľadávacej hry. In: *Datakon 2010*. (2010)
5. Šimko, J., Tvarožek, M., Bieliková, M.: Semantic History Map: Graphs Aiding Web Revisitation Support. In: *21st International Workshop on Database and Expert Systems Applications*, pp. 206–210. IEEE Computer Society, Los Alamitos (2010)

Príklad využitia webových technológií pre internetový marketing

Adela Tušanová, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Letná 9, 042 00 Košice,
adelatusanova@gmail.com, jan.paralic@tuke.sk

Abstract. V posledných rokoch sa internet stáva ďalším kanálom na realizovanie marketingových praktík, pričom množstvo spoločností si zatiaľ neuvedomuje prednosti tohto média. V tomto príspevku je na konkrétnom príklade ukázaný prínos vhodných webových technológií a internetovej reklamy pre malú firmu, vďaka ktorej môže získať konkurenčnú výhodu. V tomto prípade sa jednalo o vytvorenie webovej stránky, implementáciu internetového obchodu a realizáciu internetových reklám pre firmu Q-System. Jednotlivé riešenia sú v závere zhodnotené z pohľadu prínosov a ďalšieho možného využitia pre danú firmu.

Keywords: internetový marketing, e-shop, Google Adwords, Google Analytics

1 Úvod

Internetový marketing (e-marketing, on-line marketing) predstavuje marketing produktov a služieb na internete ako súhrn aktivít zameraných na oslovenie a získanie zákazníka. Hlavné výhody internetového marketingu sú nižšie náklady, vyššia presnosť, merateľnosť efektivity a dosiahnutých výsledkov, možnosť jednoducho viesť kampane zamerané na vybrané lokálne trhy alebo aj na globálny trh. Medzi hlavné nevýhody internetového marketingu patria nižšia penetrácia internetu najmä v prípade staršej generácie, nižšia dôveryhodnosť správ na internete, nízka dôvera pri on-line nakupovaní tovarov a služieb.

Internet marketing využíva najmä nasledovné nástroje:

- Optimalizácia pre vyhľadávače (SEO – Search Engine Optimization) – súhrn techník ktoré zabezpečia web stránke lepšiu pozíciu vo výsledkoch vyhľadávania, čo znamená viac návštevníkov.
- E-mail marketing – slúži na posilňovanie lojality už existujúcich návštevníkov. Do tejto kategórie patrí posielanie newsletterov (elektronický spravodaj) a reklamných emailov.
- Web copywriting – písanie textu pre web stránky spôsobom, ktorý je zaujímavý pre čitateľa a takisto obsahuje vybrané kľúčové slová, ktoré zabezpečia vysoké pozície vo výsledkoch vyhľadávania. Je súčasťou

optimalizácie pre vyhľadávače.

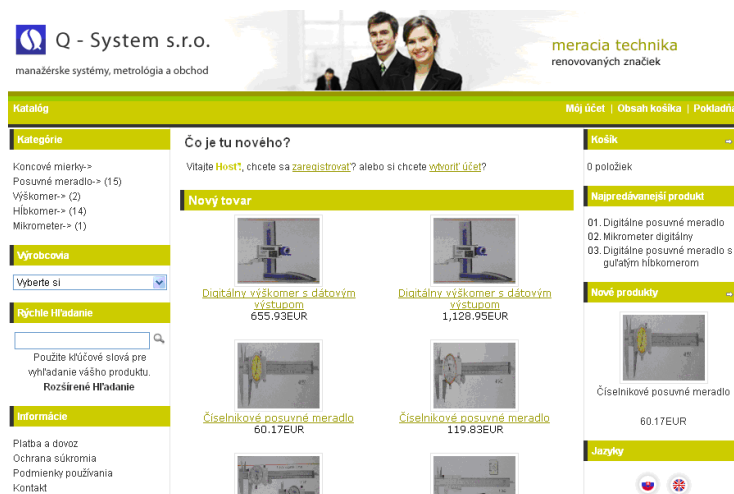
- Platená internetová reklama – napr. banerová reklama, PPC (Pay Per Click) kampane. PPC reklama je špecifická tým, že inzerent neplatí za zobrazenie, ale za jednotlivé kliknutia. Najbežnejšie formy PPC kampaní sú reklama vo vyhľadávaní, kontextová reklama a reklama cielená na umiestnenie.

2 Návrh a implementácia elektronického obchodu

Spoločnosť Q-System je konzultačno–poradensko-obchodná spoločnosť so zameraním na kvalitu a metrológiu. Bola založená vo februári 2007 so sídlom v Košiciach, ale pôsobí v rámci celého Slovenska. Z analýzy aktuálnej činnosti tejto spoločnosti vyplynula potreba realizácia internetovej propagácie v rôznych formách, napr. vytvorenie internetového obchodu, prostredníctvom ktorého by spoločnosť mohla predávať ponúkané tovary a služby.

Implementačnej fáze predchádzal výber vhodného riešenia spomedzi existujúcich open source systémov na základe stanovených kritérií reprezentujúcich potreby spoločnosti Q-System. Kritériá hodnotenia boli rozdelené do štyroch okruhov:

- Technické (podporovaný operačný systém, databázy a pod.)
- Podpora používateľov (dokumentácia, on-line demo, komunita vývojárov a používateľov)
- Podporované platobné systémy (Paypal, platobné systémy používané v SR)
- Ostatné (podpora slovenčiny, podpora viacerých mien a ďalšie)



Obr. 1 Internetový obchod spoločnosti Q-System

V čase analýzy bolo na trhu dostupných sedem open source eCommerce systémov: Eclime, Freeway, Magento Community, OpenCart, osCommerce, osCss a Zen Cart.

Z porovnania na základe stanovených kritérií vyplynulo, že najvhodnejším systémom v prípade spoločnosti Q-System bude systém osCommerce. Jeho implementácia je jednoduchá a spĺňa takmer všetky stanovené kritériá. Po implementácii a konfigurácii systému boli doplnené do databázy produkty, ktoré spoločnosť ponúka. Analýza prínosov a vhodnosti vytvoreného riešenia bola realizovaná prostredníctvom nástroja Google Analytics [1], ktorého kód bol použitý pri finalizácii aplikácie (vzhľad úvodnej stránky, možno vidieť na Obr. 1).

3 Internetový marketing

Internetový marketing predstavuje ďalší krok v propagácii firmy prostredníctvom internetu, ktorý vychádza z tradičnej webovej stránky alebo elektronického obchodu. V prípade firmy Q-System bol dôraz kladený najmä na platenú reklamu vo vyhľadávateľoch pomocou Google AdWords [2], pričom nebola zanedbaná ani optimalizácia web stránok pre vyhľadávače.

Google AdWords predstavuje celosvetovo najpoužívanejší reklamný systém, ktorý vo väčšine krajín dominuje domácemu trhu. V čase, keď vznikli úvahy o prvej kampani v Google AdWords, vyhlásila spoločnosť Google druhý ročník medzinárodnej on-line marketingovej súťaže. Táto súťaž je určená pre 5-6 členné tímy študentov z celého sveta. Každý tím dostane 200\$, ktoré môže použiť na on-line kampane lokálnej spoločnosti, ktorá ešte predtým takýmto spôsobom neinzerovala. Keďže spoločnosť Q-System spĺňala túto podmienku, úspešne získala 200\$ na svoje reklamné kampane zásluhou študentského tímu z KKUI, FEI TU v Košiciach zloženého prevažne zo študentov bakalárskeho a inžinierskeho študijného programu Hospodárska informatika.

Tabuľka 1 Prehľad AdWords kampaní

Kampan'	Rozpočet	Kliknutia	Zobrazenia	Miera Prekliku (MP)	Priem. Cena za kliknutie	Cena	Priem. pozícia
Car industry	33,00 US\$	9	1934	0,47%	0,05 US\$	0,43 US\$	1,9
ISO	33,00 US\$	89	8482	1,05%	0,05 US\$	4,23 US\$	3,6
Metrology	33,00 US\$	25	1007	2,48%	0,10 US\$	2,51 US\$	3,5
školenia	1,00 €	16	799	2,00%	0,16 €	2,52 €	1,3
e-Shop	1,00 €	125	4116	3,04%	0,14 €	17,93 €	4,3

V rámci súťaže "Google on-line challenge" [3] študenti pripravili a pre spoločnosť spustili tri kampane: Car industry, ISO a Metrology. Prehľad všetkých kampaní uvádzam v Tabuľka 1. Všetky tri boli zamerané na propagáciu webovej stránky s cieľom podporiť tzv. „branding“, t.j. zvýšenie povedomia o firme. Tento ukazovateľ je náročné odmerať, nakoľko nevieme, koľko ľudí zaregistrovalo zmienku o značke (firme). Určitým ukazovateľom môže byť počet zobrazených reklám vo vyhľadávaní a ich umiestnenie. Používateľ totiž nemusel kliknúť na reklamu no registruje, že existuje firma, ktorá ponúka ním vyhľadávané služby. Výhodou pre firmu je fakt, že

neplatí za zobrazenia, ale za kliknutia.

Reklamy všetkých troch kampaní sa v sledovanom období zobrazili spolu 17 137 krát, pričom najúspešnejšia bola kampaň ISO s takmer 73% podielom, za ňou kampaň Car industry s takmer 17% podielom a nakoniec kampaň Metrology s vyše 10% podielom. Vzhľadom na priemernú pozíciu sa najviac darilo inzerátom kampane Car industry, kde sa reklamné texty zobrazovali priemerne na 1,8-om mieste. Úspech kampane je však lepšie nemerať počtom zobrazení reklamy, ale počtom návštevníkov, ktorých reklama priniesie firme (a v prípade definovaných cieľov aj tým, koľko z návštevníkov konvertuje, t.j. realizuje ponúkanú transakciu na web stránke danej firmy).

Výkonnosť kampane sledujeme pomocou miery prekliku, čo je vlastne podiel kliknutí na reklamu k celkovému počtu jej zobrazení. Z pohľadu tohto ukazovateľa sa ako najúspešnejšia javí kampaň Metrology s 4,66% mierou prekliknutia. Zaujímavým ukazovateľom je počet návštevníkov, ktorí využili kontaktný formulár. V čase bežania kampane formulár využili štyria používatelia, pričom dva dopyty boli okamžite zrealizované následným predajom žiadaného tovaru.

Okrem webových stránok spoločnosť Q-System propagovala pomocou Google Adwords aj internetový obchod. Boli spustené dve reklamné zostavy v rámci jednej kampane: prvá s názvom „Všeobecné promo e-shopu“ a druhá z názvom „Posuvné meradlá“. Počas bežania prvej reklamnej zostavy sa nepredpokladal veľký počet zákazníkov, nakoľko v databáze ešte neboli doplnené všetky produkty, ktoré má firma na predaj. Taktiež kľúčové slová a reklamné texty boli zvolené všeobecnejšie, bez spomenutia konkrétnych produktov. Cieľom bolo hlavne dať do pozornosti nový internetový obchod, zvýšiť návštevnosť, získať registrovaných používateľov a nájsť prípadné chyby v systéme.

Obr. 2 Popredné umiestnenie vytvorenej reklamy vo vyhľadávači Google

Návštevnosť sa vďaka realizovanej reklame rapídne zvýšila – oproti predchádzajúcemu obdobiu o 150%, pričom najvyšší 56% podiel na tomto náraste mala práve popísaná forma internetovej reklamy. Z celkových 3734 zobrazení kliklo na reklamu 65 používateľov, čo je len 1,74% miera prekliknutia. Priemerná cena za kliknutie bola 0,15 €, čo je o 0,5 € menej, ako bola firma Q-System ochotná za preklik zaplatiť. Najvýkonnejšími kľúčovými slovami, ktoré priniesli nových používateľov na stránku boli „meter“, „váhy“, „meracie prístroje“, „meradlo“ a „vahy“.

Druhá reklamná zostava bola zameraná na konkrétnu kategóriu produktov a to na posuvné meradlá. Toto krátke časové obdobie bolo rozdelené na dve etapy z dôvodu technických problémov, keď bolo najlepším rozhodnutím kampane pozastaviť a po vyriešení problémov opäť spustiť. Na Obr. 2 je uvedený príklad reklamy v rámci tejto kampane.

4 Záver

Internetová propagácia je v dnešnej dobe nezanedbateľnou súčasťou každodennej činnosti firmy, kde existuje viacero nástrojov, ktoré je možné využiť pre zvýšenie informovanosti o firme a jej produktoch. Dôležitým faktorom je ich prínos a návratnosť z pohľadu firmy. V tomto prípade je nutné si stanoviť ciele, ktoré by malo navrhnuté riešenie naplniť.

V prípade firmy Q-System bol jedným z hlavných cieľov zvýšenie návštevnosti nových zákazníkov webovej stránky o 10%. Na základe uskutočnených experimentov bol preukázaný až k takmer 70%-ný nárast návštev a takmer 42%-ný nárast v počte zobrazených stránok. Naopak klesol pomer zobrazených stránok na návštevu o viac ako 16%. Aj napriek tomu môžeme zhodnotiť, že cieľ bol splnený. V prípade internetového obchodu bolo stanovené kritérium získania minimálne 10 nových zákazníkov. Tento cieľ sa však nepodarilo splniť, zrejme aj vzhľadom na úzky profil ponúkaných produktov spoločnosti a aktuálnu situáciu v danej oblasti na trhu (tá je na Slovensku zatiaľ nerozvinutá, úplne na začiatku). Posledným stanoveným cieľom bola návratnosť investícií. Tento ukazovateľ nadobudol hodnotu 4,48%, ktorá dokumentuje, že finančné prostriedky, ktoré investovala firma Q-System do vytvorenia webových aplikácií a realizácie marketingových kampaní na internete sa jej vrátili.

Bližšie informácie o realizovaných experimentoch a implementačných úlohách je možné nájsť v [4].

PodĎakovanie. Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

Referencie

1. Clifton, B.: Google Analytics. Brno: Computer Press, a.s., 2009. 334 s.: il. ISBN 978-80-251-2231-0
2. Beck, A.: Google Adwords. Praha: Grada Publishing, a.s., 2009. 232 s. il. ISBN 978-80-247-2898-8
3. Google. STUDENT GUIDE 2009 [online]. Dostupné na internete: <http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/sk/onlinechallenge/files/student_guide.pdf>
4. Tušanová, A.: Návrh, realizácia a analýza dopadov elektronického obchodu a internetového marketingu pre zvolenú firmu. Diplomová práca, KKUI FEI TU v Košiciach, 88s., 2010.

Použitie Solr na indexovanie a vyhľadávanie dát

Zoltán Balogh, Emil Gatial, Ladislav Hluchý

Ústav informatiky, Slovenská akadémia vied,
Dúbravská cesta 9, 845 07 Bratislava, Slovensko
{Zoltan.Balogh, Emil.Gatial}@savba.sk

Abstrakt. Solr je nástroj, ktorý slúži na indexáciu a vyhľadávanie v rozsiahlych dátových zdrojoch. Tento príspevok popisuje postup nasadenia tohto nástroja na aplikáciu s konkrétnymi dátami v slovenskom jazyku. V príspevku popisujeme základnú inštaláciu rámca Solr, prípravu dát, indexáciu ako aj príklady vyhľadávania. Tiež ukážeme jednoduchú implementáciu ich aplikácie, ktorá využíva Solr REST rozhranie. Záverečne zhrnieme výhody ako aj obmedzenia použitia nástroja Solr na dátami v slovenskom jazyku.

Kľúčové slová: indexovanie, vyhľadávanie, Solr, Lucene.

1 Úvod

Cieľom tohto príspevku je zhodnotiť možnosti a obmedzenia nástroja Solr [1] na indexáciu rozsiahlych dátových zdrojov. Ako spôsob testovania a vyhodnotenia týchto možností sme zvolili nasadenie nástroja nad rozsiahlymi reálnymi dátami. Ako dátový zdroj pre indexáciu sme zvolili dáta o verejných obstarávaní (v.o.) v SR. Tieto dáta sú verejne dostupné na portáli verejného obstarávania [2]. Tento príspevok v skratke popisuje inštaláciu nástroja Solr a prípravu dát na indexáciu. Jednoduchý príklad indexácie vzorových dát sa nachádza na stránke projektu Solr a popisujú inštaláciu balíka nástroja na stránkach Solr ale chýba je jednoduchý ucelený postup, ako nástroj jednoducho prispôbiť pre vlastné dáta. V tomto príspevku predkladáme krátky postup, ktorý v niektorých krokoch predstaví inštaláciu, indexovanie, upravenie schémy, ako aj samotné použitie nástroja Solr nad konkrétnymi dátami. Pri popise uvádzame aj postrehy, ktoré môžu byť užitočné pri rozhodnutí použiť nástroj Solr pre nasadenie a prácu z hľadiska bezpečnosti, alebo spracovania textu v slovenskom jazyku. Tento príspevok má za cieľ byť aj návodom na nasadenie nástroja Solr nad vlastnými dátovými zdrojmi.

2 Základy Solr

Solr je nástroj, ktorý umožňuje indexovať rozsiahle dátové zdroje na základe striktno typových definícií atribútov. Solr rozširuje vlastnosti full-textového vyhľadávania nástroja Lucene [3]. Práve vlastnosťou indexovať typové atribúty je

jednou z vlastností, ktorou sa Solr od Lucene odlišuje. Všeobecne je možné prehlásiť, že kým prístup k Lucene je výnimočne treba riešiť na implementačnej úrovni (napr. v Jave), tak Solr poskytuje vyššiu úroveň prístupu, pri ktorej sa komunikuje so cez rozhranie webovej služby. Aj pri používaní nástroja Solr je však vždy v jadre používaný Lucene vyhľadávací nástroj. Solr primárne indexuje dáta vo formáte XML v nasledovnej jednoduchej štruktúre:

```
<add>
  <doc>
    <field name="id">1557</field>
    <field name="atr1">text1</field>
    ...
    <field name="atr1">84.2</field>
  </doc>
</add>
```

Všetky dáta, ktoré chceme indexovať a následne prebádať, je potrebné konvertovať do takejto podoby. Element <doc> obsahuje jednotlivé atribúty, ktoré popisujú objekt alebo dokument. Solr indexuje len dáta, ktoré sú vložené do elementov <field>, t. j. ako atribút neindexuje žiadne iné dátové zdroje ako napr. obsah nalinkovaného externého dátového súboru. Okrem primárneho formátu XML umožňuje Solr aj indexáciu dát uložených vo formáte CSV, alebo z databázy. Element <add> v príklade vyššie reprezentuje príkaz, ktorý žiada Solr o pridanie (alebo nahradenie) dát do indexu. V prípade veľkého množstva inštancií dokumentov je možné do elementu <add> vložiť viac <doc> elementov. Atribút s názvom „id“ má špeciálne postavenie, pretože jednoducho identifikuje záznam v indexe. Ak pridáme nový záznam s atribútom „id“ zhodným so záznamom v indexe, bude existujúci záznam prepísaný. Hlavne v prípade rozsiahlych distribuovaných implementácií Solr sa odpočítava ako atribút „id“ vygenerovaný algoritmom (napr. MD5) hash reťazec, ktorý zabezpečuje jednoznačnosť hodnoty. Záznamy nie sú hne po pridaní v dotazoch viditeľné. Pre „zviditeľnenie“ záznamov je potrebné Solr poslať operáciu <commit>. Inou užitočnou operáciou je operácia <delete>, použitím ktorej je možné cielene vymazať požadovaný dokument z indexu, napr. nasledovná operácia vymaže z indexu dokument s atribútom id = 1557:

```
<delete><query>id:1557</query></delete>
```

Na komunikáciu so Solr sa používajú dve rozhrania: jedno na aktualizáciu (update URL), s ktorým sa udržiava index a druhý na výberové dotazy (select URL). V predvolenej inštalácii sú tieto rozhrania nasledovné:

```
http://localhost:8983/solr/update
http://localhost:8983/solr/select
```

Tieto rozhrania sú dostupné pomocou Curl [4], ktoré je implementované tak pre príkazový riadok, ako aj vo forme knižníc pre programovacie jazyky Ruby, PHP, Java alebo Python. Aktualizácia indexu je možné uskutočniť zaslaním HTTP POST požiadavky na update URL. Naopak dotaz na vyhľadanie v indexe sa používa http GET požiadavka na select URL.

3 Nasadenie nástroja Solr

Po rozbalení inštalného archívu je možné hne začať Solr používať, ale len s dátami dodávanými v inštalnom archíve. Základná inštalácia pozostáva z rozbalenia archívu:

```
$ tar zxvf apache-solr-1.4.1.tgz
$ ln -s apache-solr-1.4.1 solr
```

Jedinou požiadavkou inštalácie je Java 1.5 alebo vyššia verzia. Pre svoju prácu potrebuje Solr servlet kontajner (napr. Tomcat, Jetty). Solr je štandardne dodávaný s predinštalovaným Jetty [5] kontajnerom. Pre vytvorenie a používanie Solr s vlastnými dátami môžeme použiť štruktúru adresára examples/. Najprv je potrebné vytvoriť "vlastný adresár pre svoj projekt, v našom prípade napr. adresár vestnik/:

```
$ cd solr; mkdir vestnik; cd vestnik
```

Do vytvoreného adresára je potrebné nakopírovať niekoľko adresárov, ktoré sú pre fungovanie Solr spolu s dodávaným Jetty kontajnerom dôležité:

- x etc – obsahuje konfigurčné súbory predinštalovaného Jetty kontajnera jetty.xml a webdefault.xml
- x logs – obsahuje záznamy o prístupoch Jetty servera
- x solr – koreový adresár vytvorenej Solr inštalácie
- x webapps – obsahuje war súbor nástroja Solr

Domovským adresárom pre Solr je adresár solr/, ktorý má nasledovnú štruktúru:

- x bin – obsahuje spustiteľné komponenty pre Solr;
- x conf – obsahuje konfiguráciu Solr, dôležité sú najmä súbor solrconfig.xml, kde je možné nastaviť systémovú konfiguráciu a schema.xml, ktorá popisuje atribúty používaného dátového modelu;
- x data – je adresár určený pre fyzické uloženie dátových súborov a indexov.

Solr pridáva do modelu Lucene dátovú schému, ktorá vopred presne typovo definuje a popisuje jednotlivé atribúty. Atribúty sa definujú v súbore schema.xml. Najvyššou úrovňou schémy je element, ktorého atribút „name“ definuje názov schémy:

```
<schema name="vestnik" version="1.2">
```

alej je potrebné zoznam deklarácií typu <fieldtype> v rámci elementu <types>. Niekoľko preddefinovaných typov je v súbore už definovaných. Napr. Takto vyzerá deklarácia typu string:

```
<fieldType name="string" class="solr.StrField"
  sortMissingLast="true" omitNorms="true"/>
```

Atribút name="string" deklaruje názov typu, class="solr.StrField" je trieda typu zo základného Solr balíka, sortMissingLast="true" znamená, že položky bez hodnoty budú triedené až po položkách s hodnotami a omitNorms="true" zakáže normalizáciu hodnoty a zrýchlenie indexácie. Pre textové typy budeme zväčša potrebovať nakonfigurovať analýzu obsahu – tá sa konfiguruje jednoducho pridaním elementu <analyzer>:

```
<fieldType name="text_ws" class="solr.TextField"
  positionIncrementGap="100">
<analyzer>
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
```

```

</analyzer>
</fieldType>

```

Vo vyššie uvedenom kóde sa používa tokenizer, ktorý rozdelí slová a podzrier pre hľadanie identických výskytov slov. Použitejšiu analýzu obsahu je možné použiť niekoľko tokenizerov a filtrov. Je tiež možné analýzu obsahu zvlášť nakonfigurovať pre indexáciu ako aj pre vyhľadávanie.

```

<fieldType name="textgen" class="solr.TextField"
positionIncrementGap="100">
<analyzer type="index">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.StopFilterFactory"
ignoreCase="true" words="stopwords.txt"
enablePositionIncrements="true"/>
<filter class="solr.WordDelimiterFilterFactory"
generateWordParts="1" generateNumberParts="1"
catenateWords="1" catenateNumbers="1" catenateAll="0"
splitOnCaseChange="0"/>
<filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
<analyzer type="query">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.SynonymFilterFactory"
synonyms="synonyms.txt" ignoreCase="true"
expand="true"/>
<filter class="solr.StopFilterFactory"
ignoreCase="true" words="stopwords.txt"
enablePositionIncrements="true"/>
<filter class="solr.WordDelimiterFilterFactory"
generateWordParts="1" generateNumberParts="1"
catenateWords="0" catenateNumbers="0" catenateAll="0"
splitOnCaseChange="0"/>
<filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
</fieldType>

```

Vo vyššie uvedenom kóde je definované zariadenie rôznych prístupov na spracovanie a analýzu textu a to tak pre indexáciu (analyzer type="index") ako aj vyhľadávanie (analyzer type="query"). Potom ako máme nadefinované všetky potrebné typy, je možné definovať jednotlivé položky a ich mapovanie na definované typy. Typická definícia položky je nasledovná:

```

<field name="id" type="string" indexed="true"stored="true"/>

```

a definuje položku s názvom „id“, ktorá je typu „string“. Atribút indexed="true" určuje, že danú položku bude Solr indexovať. Podobne atribút stored="true" znamená, že Solr si uloží danú položku (teda nielen dáta vygenerované indexáciou, ale aj samotné pôvodné dáta).

4 Aplikácia

Ako vzorový dátový zdroj pre indexáciu sme zvolili dáta o verejných obstarávaníach (v. o.) v SR. Dáta sme získali z internetu z príslušného portálu

pomocou parametrizovaného dotazovania skriptov, ktoré tieto dáta generujú. Keďže dáta boli vo formáte HTML, pretransformovali sme ich do textovej podoby pomocou príkazu lynx a následne prekonvertovali do jednotného kódovania UTF-8 pomocou príkazu iconv. Keďže dáta o v.o. majú zákonom špecifikovanú štruktúru so záväzným pomenovaním, bolo pomerne jednoduché tieto dáta rozčleniť a následne vytvoriť ich XML reprezentáciu vo formáte požadovanom nástrojom Solr (viď prílohu 2). Ešte pred generovaním XML reprezentácie dát sme si vytvorili definíciu položiek:

```
<field name="id" type="string" indexed="true" stored="true"/>
<field name="typ" type="textgen" indexed="true" stored="true"/>
<field name="obstaravatel" type="textgen" indexed="true"
  stored="true"/>
<field name="zmluva" type="textgen" indexed="true" stored="true"/>
<field name="popis" type="textgen" indexed="true" stored="true"/>
<field name="ine" type="textgen" indexed="true" stored="true"/>
<field name="ponuk" type="textgen" indexed="true" stored="true"/>
<field name="datum" type="textgen" indexed="true" stored="true"/>
<field name="suma" type="textgen" indexed="true" stored="true"/>
<field name="dodavatel" type="textgen" indexed="true"
  stored="true"/>
```

Po vytvorení schémy a jej uložení do súboru solr/schema.xml sme spustili Solr server príkazom:

```
$ java -jar start.jar
```

Následne bola spustená indexácia dátových XML súborov. Pre zrýchlenie spracovávania súborov je vhodnejšie namiesto veľkého množstva malých súborov s jednou <doc> inštanciou, vygenerovať jeden veľký súbor, do ktorého sa zapíše viac <doc> inšancií. Indexácia bola spustená dodávaným skriptom post.sh:

```
@!/bin/sh
FILES=$*
URL=http://localhost:8983/solr/update

for f in $FILES; do
  echo Posting file $f to $URL
  curl $URL --data-binary @$f \
    -H 'Content-type:text/xml; charset=utf-8'
  echo
done

curl $URL --data-binary '<commit/>' \
  -H 'Content-type:text/xml; charset=utf-8'
echo
```

Pre pripojenie sa k Solr rozhraniam z príkazového riadku je potrebný nástroj curl. Po spustení sa najprv v prvom cykle zindexujú súbory v aktuálnom adresári a nakoniec sa pošle príkaz <commit/>, ktorý zviditeľní indexované dáta pre vyhľadávanie. Pre optimalizáciu vyhľadávania v indexoch poskytuje Solr grafické webové rozhranie (Obr. 1). V tomto rozhraní je možné sledovať a nastavenie filtrov pri analýze tak indexovania ako aj vyhľadávania. Inými užitočnými nástrojmi Solr Admin rozhrania je analyzátor dopytov, schema browser, ktorý umožňuje vizualizáciu podrobnú štatistiku a položkách, s najvyššie indexovanými termínmi alebo štatistika Solr.

4 Záver

Solr je výkonný a užitočný nástroj na indexáciu veľkého množstva dát. Poskytuje rozhrania na jednoduchú indexáciu ako aj vyhľadávanie dát. Úroveň sofistikovanosti indexov závisí od konfigurácie filtrov a iných nástrojov pre analýzu indexovaných dát. V prípade, že neexistuje požadovaný filter, je možné ho implementovať a jednoducho do Solr zaintegrovať (podobne ako pri Lucene). Výhodou Solr je možnosť jednoduchšej indexácie veľkého množstva dát. V prípade potreby Solr poskytuje elasticitu, vďaka ktorej je možné Solr prevádzkovať na niekoľkých serveroch s roztrúsenými indexmi – takáto konfigurácia však má určité obmedzenia.

Nevýhodou Solr je, že nepodporuje indexáciu dokumentov, tabuliek a dát v inom formáte ako v XML. V prípade ak je potrebné takéto dáta indexovať, najprv potrebujeme ich extrakciu a konverziu do XML formátu. Výsledný XML výstup odporúčame generovať v UTF-8 kódovaní. Je možné však indexovať dáta z SQL databáz a CSV súborov (jednoduché súbory, ktorých položky sú oddelené bodkami a čiarkou alebo dvojbodkou). Vo vzťahu k slovenskému jazyku Solr „trpí“ rovnakým nedostatkom filtrov, resp. lematizátorov ako Lucene. Aj keď členmi študentského projektu [6] bol vyvinutý slovenský stemmer, ktorý dosahuje úspešnosť okolo 90%, bolo by vhodné vytvoriť slovenský lematizátor pre Lucene s vyššou úspešnosťou, čím by sa umožnila efektívnejšia indexácia slovenských textov aj s nástrojom Solr. Inou oblasťou, ktorú je potreba pri nasadení nástroja Solr riešiť bezpečnosť a prístup k rozhraniám, ktoré samotný nástroj nedisponuje žiadnymi používateľskými právami alebo inou formou zabezpečenia.

Táto práca bola podporená z nasledovných projektov RECLER ITMS: 26240220029, SMART ITMS: 26240120005, SMART II ITMS: 26240120029.

Použitá literatúra

1. Domovská stránka projektu Apache Solr [cit. 2010-10-17].
Dostupné na internete: <<http://www.apache.org/solr/FrontPage>>
2. Úrad pre verejné obstarávanie [cit. 2010-10-17].
Dostupné na internete: <<http://www.uvo.gov.sk/>>
3. Domovská stránka Lucene [cit. 2010-10-17].
Dostupné na internete: <<http://lucene.apache.org/>>
4. Domovská stránka Curl [cit. 2010-10-17]. Dostupné na internete: <<http://curl.haxx.se/>>
5. Domovská stránka Jetty [cit. 2010-10-17].
Dostupné na internete: <<http://jetty.codehaus.org/jetty/>>
6. Hana Pifková: Slovenský stemmer [2010-10-17]. Dostupné na internete:
<http://vi.ikt.ui.sav.sk/Projekty/Projekty_2008%2F%2F2009/Hana_Pifkov%C3%A1_-_Stemer>

Distribúované spracovanie dát nad MapReduce architektúrou (Hadoop a Hive)

Martin Šeleng

Automatizované vytváranie používateľských formulárov

Emil Gatiaľ a Zoltán Balogh

Sociálne siete a grafy

Modelovanie a analýza malej komunitnej sociálnej siete

Gabriel Tutoky, Ján Paralič

Graph Transformations for Semantic Email Search

Marcel Kvassay, Michal Laclavík, Štefan Dlugolinský, Ladislav Hluchý

Využitie sociálnych sietí pri vyhľadávaní v emailoch

Michal Laclavík a Ladislav Hluchy

Sémantika a Ontológia

Sémantická sieť ako spojitý systém

Stanislav Dvorščák, Kristína Machová

Application Ontology Manager for Hydra

Ján Hreňo, Peter Kostelník, Martin Sarnovský

An Ontology Driven Approach to Software Process Engineering

Miroslav Líška, Pavol Návrat

Dolovanie informácií a znalostí

Využitie JBowl knižnice pri riešení úloh dolovania znalostí z textov

František Babič, Štefan Bašista, Roman Dudek, Roman Mihaľ, Peter Savčák

Data mining for fog prediction

Peter Bednár, František Albert

**Discovering occurrences of user-defined patterns in
historical data representing collaborative activities in
virtual user environment**

Jozef Wagner, Ján Paralič, František Babič

Postre

Web Information Integration in Knowledge Discovery

Kristína Machová, Dominika Fodorová

Použitie alternatívnych prístupov pre plánovanie výrobného procesu

Tomáš Kasanický, Ján Zelenka

Dolovanie údajov v hydrometeorologických aplikáciách

Martin Šeleng, Peter Krammer, Ondrej Habala a Ladislav Hluchý

**Text Document Retrieval by Document Space
Dimension Reduction with Feed-Forward Neural
Networks**

Lenka Skovajsová, Igor Mokriš

ICT-based Toolbox in OCOPOMO Project and Potential Methods for Integration

Peter Butka, Marián Mach, Tomáš Sabol, Karol Furdík

Multi-agent-based conception of modern aircraft design

Dmytro Konotop, Ivana Budinska, Valeriy Zinchenko, Emil Gatjal

**Podniková inteligencia, analytika a proces objavovania
znaností v databázach**

Jozef Kovač

Authors Index

Babič František, 131
Babík Marián, 62
Balogh Zoltán, 41, 114, 120
Barla Michal, 34
Bieliková Mária, 34, 37
Budinská Ivana, 41, 68
Budinský Miloš, 85
Butka Peter, 71

Džupka Peter, 23

Eckhardt Alan, 124

Forgáč Radoslav, 41, 81, 84
Furdík Karol, 29, 93

Garabík Radovan, 2
Gatial Emil, 41, 114

Hluchý Ladislav, 41, 45, 68
Hreňo Ján, 71

Kaliská Markéta, 78
Kasanický Tomáš, 57
Kostelník Peter, 16
Krajčí Stanislav, 93, 99

Laclavík, 41, 45, 68, 84
Líška Miroslav, 135

Mokriš Igor, 41, 102

Novotný Róbert, 99

Oravec Viktor, 41, 110

Paralič Marek, 19
Pázmán René, 116

Rusko Milan, 6

Sabol Tomáš, 16
Sarnovský Martin, 16
Skokan Marek, 23
Skovajsová Lenka, 102, 106

Šaloun Petr, 78
Šefránek Ján, 13
Šeleng Martin, 41, 45, 68

Tomášek Martin, 29
Tvarožek Michal, 37

Velart Zdeněk, 78
Vojtáš Peter, 124
Všetečka Petr, 88

Wagner Jozef, 19, 131

Michal Laclavík, Ladislav Hluchý
Editors

Proceedings

**5th Workshop on Intelligent and Knowledge Oriented
Technology**

1st Edition, 70 copies, Published by Institute of Informatics SAS

Printed by EQUILIBRIA, s.r.o.

2010

ISBN 978-80-970145-2-0
EAN 9788097014520