5th Workshop on Intelligent and Knowledge oriented Technologies

WIKT 2010 Proceedings

Michal Laclavík
Ladislav Hluchý (Eds.)

November 11 - 12, 2010
Bratislava, Slovakia

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

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

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. novembera 2006 v Bratislave.

Po 4 rokoch sa workshop opäť koná na Ústavu 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

Slovak National Corpus tools and resources

Radovan Garabík

Ľ. 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 Ľ. 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).

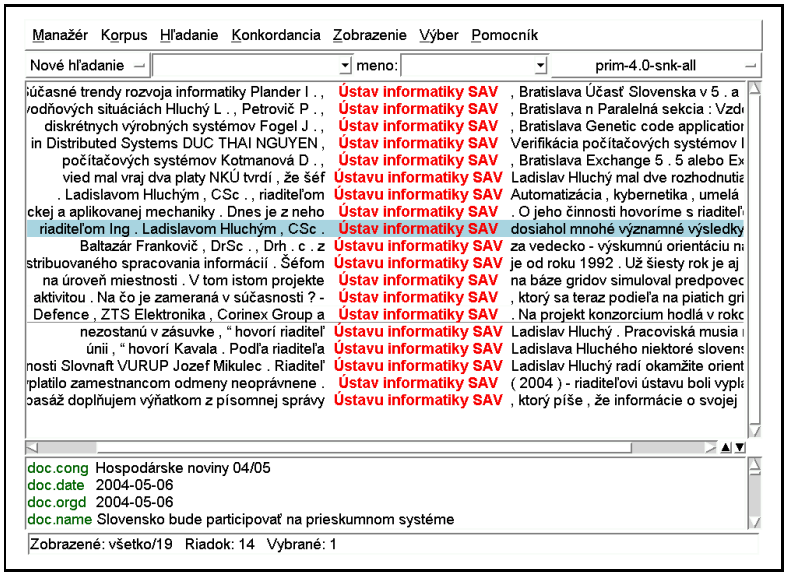


Fig. 1. Screenshot of *Bonito* client

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 explicite 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 addional visual representation of transcription and annotation, as well as links to sound recordings themselves.

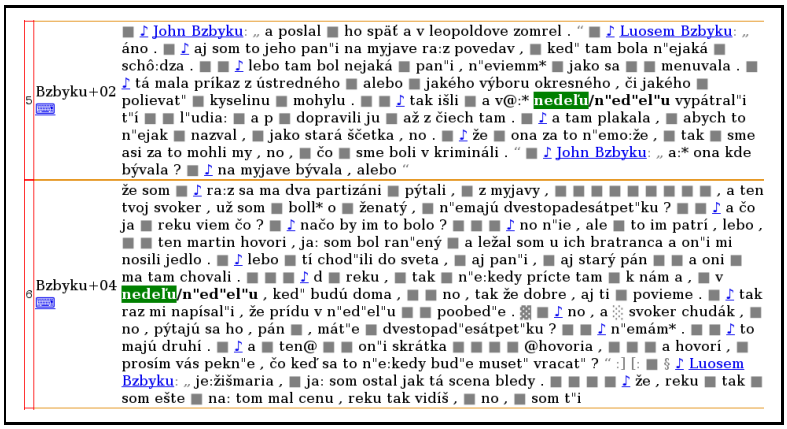


Fig. 2. Screenshot of Corpus of Spoken Slovak web interface

```
<Event desc="poz" type="noise" extent="instantaneous"/>
 niekedy/n"e:kedy príďe/prícte 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"/>
```

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

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.

```
== Lema ==
mať

== Paradigma ==
VIe+: mať
VKesa+: mám
VKesb+: máš
VKesc+: má
VKepa+: máme
VKepb+: máte
VKepc+: majú
VMesb+: maj
VMepa+: majme
VMepb+: majte
VHe+: majúc
VLesam+: mal
VLesaf+: mala
VLesan+: malo
VLepah+: mali

== Homonymia ==
[[mať]]

----
KategóriaVerbá
```

Table 2. Paradigm of the verb *mat'*

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	nejsť	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 dictionary[1]. 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 Wikiédie, 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. Náš 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 definovane 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 štatistík 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 zobraziť 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 cieli 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 prispôsobená aj navigácia. Príkladom 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 odporúčať 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 končiaca 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 odporúčať zaujímavé dokumenty a prispôsobovať 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 spojito (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 zhliadnuté 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 odlíšiť 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.

Poďakovanie. 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 odporúčaniu obsahu. Klasifikácia metód odporúčania 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 prispôsobuje študentom podľa toho, či preferujú učenie od všeobecného konceptu ku konkrétnemu, alebo naopak. Rôzne metódy odporúčania obsahu môžu vyhovovať rôznym štýlom učenia. V [5] opisujeme metódu odporúčania 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 zhlukovania 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.

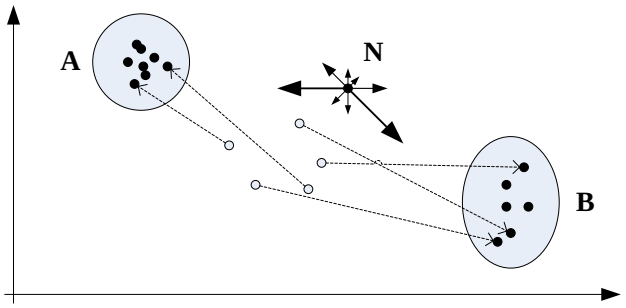
2 Hybridné metódy odporúčania

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 prispôsobovania, 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 počiatočnej 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 prispôsobovania 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 nastavia 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 porovnávanie, 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).

Poďakovanie. 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

Mária Bieliková, Michal Kompan, Dušan Zeleník

Učení příklady – personalizovaný adaptivní Web

Jan Nekula, Petr Šaloun, Zdeněk Velart, Petr Klimánek

Internet a Technológie

Adaptívny proxy server: prevádzka a skúsenosti po roku

Tomáš Kramár, Michal Barla, Mária Bieliková

Hry s účelom objavovania sémantiky na webe

Jakub Šimko, Michal Tvarožek a Mária Bieliková

Príklad využitia webových technológií pre internetový marketing

Adela Tušanová, Ján Paralič

Použitie Solr na indexovanie a vyhľadávanie dát

Zoltan Balogh a Emil Gatial

Distribuované spracovanie dát nad MapReduce architektúrou (Hadoop a Hive)

Martin Šeleng

Automatizované vytváranie používateľských formulárov

Emil Gatial 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 Ontológie

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 Zeleneka

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 Gatial

Podniková inteligencia, analytika a proces objavovania znalostí v databázach

Jozef Kovač

Authors Index

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