ORDINARY WEB PAGES AS A SOURCE FOR METADATA ACQUISITION FOR OPEN CORPUS USER MODELING

Michal Barla and Mária Bieliková
Institute of Informatics and Software Engineering, Faculty of Informatics
and Information Technologies, Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
{Name.Surname}@fiit.stuba.sk

ABSTRACT

Personalization and adaptivity of the Web as we know of today is often "closed" within a particular web-based system. As a result there are only a few "personalized islands" within the whole Web. Spreading the personalization to the whole Web either via an enhanced proxy server or using an agent residing on a client-side brings a challenge how to determine metadata within an open corpus Web domain, which would allow for an efficient creation of overlayed user model. In this paper we present our approach to metadata acquisition for open corpus user modeling applicable on the "wild" Web, where we decided to take into account metadata in the form of keywords representing the visited web pages. We present the user modeling process (which is thus keyword-based) built on the top of an enhanced proxy server, capable of personalizing user browsing sessions via pluggable modules. The paper focuses on comparison of algorithms and third-party services which allow for extraction of required keywords from ordinary web pages, which is a crucial step of our user modeling approach.

KEYWORDS

keyword extraction, user modeling, proxy.

1. INTRODUCTION

The field of adaptive web-based systems is nowadays well established and pretty mature (Brusilovsky, et al., 2007). However, most of the proposed approaches are concentrating on personalization of one particular system, either by incorporating user modeling and adaptive features into a web application or by creating an adaptive layer on the top of an existing application. As a result, there are only a few "personalized islands" within the whole Web, where majority of content and information services are provided using a failing "one-size-fits-all" paradigm.

The reason why the majority of adaptive approaches is built on the top of a closed corpus domain is that every adaptive system must track user's attitudes (such as knowledge or interest) towards domain elements often realized in the form of an overlayed user model. Closed corpus domain can provide a detailed, often manually prepared and non-changing conceptualization, which is easily used for user modeling purposes. In the case of an open corpus or vast and dynamic domain, we cannot track user's relations to all documents or other pieces of information which exist within the domain. The solution is to incorporate and use external models that exist beyond the hyperspace of interlinked documents and provide a meaningful abstraction of the domain (Brusilovsky & Henze, 2007), i.e., to provide a metadata model and a mapping between domain items and this metadata model, which serves also as a bottom layer for an overlayed user model. An example of such an approach, which maps user's interests to a set of metadata, can be seen in (Barla, et al., 2009).

When considering the whole Web as our domain of interest, it seems that we can leverage the Semantic Web as both a source of metadata and mapping inherently present within it. However, when considering findings from (Sabou, et al., 2008; Fernandez, et al., 2008) that

- *existing semantic systems are restricted to a limited set of domains* – they use a set of a priori defined ontologies covering one specific domain without proper linking to other ontologies,
- *the overall Semantic Web does not adequately cover specific terminology* and

- *many online ontologies have a weak internal structure* – few online ontologies contain synonyms or non-taxonomic relations,

we come to conclusions that we must find additional sources of metadata, which would provide good-enough results for user modeling purposes.

In this paper we present our approach to metadata acquisition for open corpus user modeling applicable on the "wild" Web, where we decided to take into account metadata in the form of keywords representing the visited web pages. We present briefly the user modeling process (which is thus keyword-based) and focus on comparison of algorithms and third-party services which allows for extraction of required keywords from ordinary web pages.

The paper is organized as follows. Section 2 discusses related works, in section 3 we present briefly our approach to user modeling using a proxy server. In section 4 we describe algorithms and services currently employed for metadata acquisition. In section 5 we present an evaluation of selected algorithms' behavior on different web pages in terms of content language and writing style. Finally we give our conclusions.

2. RELATED WORKS

The related works can be seen either from the point of view of keyword- (or tag-) based user modeling or of the actual extraction of keywords from the plain text (i.e., content of web pages).

Tags and keywords gained interest of user modeling community very quickly. Already in 2001, the (Shepherd, et al., 2001) proposed an approach, which was combining keywords extracted from web pages with user's explicit rating of the page in order to build a keyword-based user profile.

An approach to user interest recognition applicable in open information space was presented in (Badi, et al., 2006). The authors defined user interest as either a term, document or their abstractions such as term vector and metadata. Next, they focused on implicit interest indicators and their mutual comparison.

The real boom came with the rise of Web 2.0, when users started to tag and organize content on the Web to alleviate themselves its later retrieval. In (Schwarzkopf, et al., 2007) as well as in (Zhang & Feng, 2008), the authors proposed a user model resulting in analysis of user's tag space within a web-based tagging system and thus a closed information space. Approach proposed in (Carmagnola, et al., 2007) does not rely only on reasoning of specific tags semantics but considers also user dimensions which could be inferred from the action of tagging, such as user's interactivity level or interest. It seems that majority of research oriented on keyword-based user models are built in a manner that it is the *user* herself who *provides the keywords* to her model by annotating web content with tags. In our approach, we acquire keywords automatically from the visited pages, similarly to (Shepherd, et al., 2001).

The actual extraction of keywords or terms from the plain text (called also ATR – automatic term recognition) has evolved from pure research task into a mature application domain, with various available services and libraries devoted to this task, serving for various purposes in various domain from information retrieval to automated domain model construction for personalized systems (Šimko & Bieliková,2009). They are based on linguistic processing (e.g., part-of-speech tagging) and statistical models used to select relevant terms from the text (Zhang, et al., 2008).

However, the main drawback of linguistic and statistical approaches is that they often require a corpus including all documents of the domain and are thus applicable only on closed and not-ever-changing information spaces. If we want to extract relevant keywords from ordinary web pages, we need to employ and combine other techniques such as named-entity recognition leveraging linked data and other semantic knowledge.

3. KEYWORD-BASED USER MODELING FOR THE WEB SEARCH DOMAIN

A keyword-based user modeling, albeit being a rather simple approach is giving satisfactory results and seems to be good enough for capturing user interests (Kramár, et al., 2010). If we want to employ such model for the purpose of "wild" web personalization, we need an ability to acquire keywords from documents visited by the users. Because the Web is an open information space, we need to track down and process every page a user has visited in order to update her model appropriately.

To achieve this, we developed an enhanced proxy server, which allows for realization of advanced operations on the top of requests flowing from user and responses coming back from the web servers all over the Internet (Barla & Bieliková, 2009). Figure 1 depicts the user modeling flow supported by our platform. When a web server returns a web page as a response for a user's request, the proxy injects a specialized tracking javascript into it and passes the page to the client user agent. At the same time, it initializes a process of metadata extraction from the acquired page.

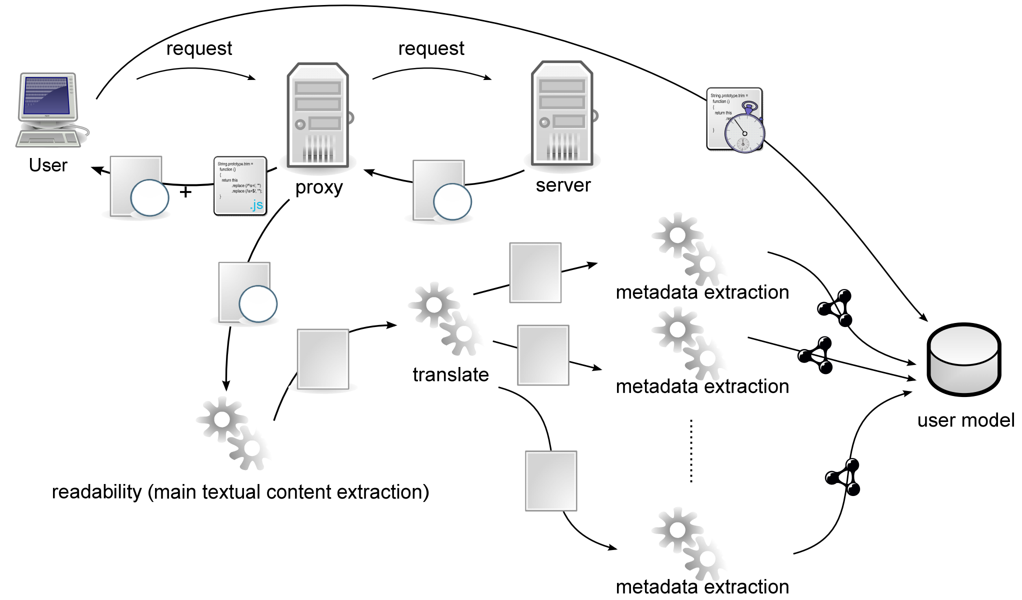

Figure 1. User Modeling process based on an enhanced proxy platform.

First, HTML page is processed by a *readability* module[1] which strips-off the HTML markup and leaves only a main textual content, omitting navigational parts, banners etc. Second, the text is translated into English (if it is not already in this language) using Google's translate service. This step is required as majority of metadata extraction algorithms and services (which are fired after the translation) work correctly only with English language. Extracted metadata are stored in a user model altogether with corresponding URL and timestamp. The tracking javascript, which was inserted to the response and passed to the user, supplies additional information about the user's activity within the page and thus adds an implicit feedback which determines a weight of contribution of just-discovered metadata to the whole user model.

The aforementioned process gathers metadata for every requested web page and creates a basic (evidence) layer of a user model. Naturally, as the time flows, the keywords which represent long-term user interests occur more often than the others. Therefore, by considering only *top K* most occurring keywords, we get a user model which can be further analyzed and serves as a basis for adaptation and personalization.

4. SERVICES FOR KEYWORD EXTRACTION

In order to allow for combination of different approaches to keyword extraction, we developed an extensible software library *JKeyExtractor*, which wraps several different metadata extractors and is able to provide results, which are composed of different extractors. These extractors could be either online web services available through different kinds of API (usually REST-based) or locally invoked algorithms and libraries. Our platform is currently able to use several following metadata extractors.

[1] reimplemented solution available at http://lab.arc90.com/experiments/readability/

JATR library

Java Automatic Term Recognition Toolkit[2] is a Java implementation of several term extraction algorithms (e.g., CValue, TermEx, GlossEx, Weirdness, TF-IDF – their comparison can be found in (Zhang, et al., 2008). The library comes with a simple corpus for determining term frequency and a pack of NLP resources from *OpenNLP*[3] which serves as a basis for sentence detection, tokenization, pos-tagging, chunking and parsing.

Delicious

We are taking advantage of Delicious API and its tag suggestion feature to determine tags for the given page (if any). Advantage of this metadata extractor (or rather provider) is that it provides metadata (tags) filled-in by humans, which might bring an added value, as not all of the acquired tags must be present in the text itself.

OpenCalais

OpenCalais[4] is a web service which automatically creates rich semantic metadata for the submitted content. It resolves named entities (such as persons, companies, cities, countries etc.), facts and events. While ATR-based approaches do not distinguish between a name and any other noun within the text and might easily declare a name as non-relevant, OpenCalais does recognize them, so we can use them to statistically track user's interest.

tagthe.net

tagthe.net is a simple REST web service that returns a set of tags based on a posted textual content. Similarly to OpenCalais, it tries to distinguish retrieved tags into categories such as topics, persons or locations.

Alchemy

Alchemy[5] is yet another web-based API which provides named entity extraction, term extraction, topic categorization and automatic language identification services. Apart from the language identification service, the three remaining services are highly relevant to our work as they provide different kinds of metadata extracted from the given text.

dmoz

dmoz[6] is an open directory project – the largest human-edited directory of the Web maintained by volunteers from all over the world. Visitors having interest in a particular category can find pages relevant to this category by navigating through the created hierarchy. However, dmoz provide also RDF dump of their database, which can be easily used for an inverse search (page → category). Similarly to delicious, if we find a page in dmoz, we get an added-value of man-made categorization (which, in this case is even hierarchically organized) ideal for modeling of user interests.

5. EVALUATION

We have conducted an experiment on a small dataset of English and other language (slovak and czech) web-pages (having a representative of technical and non-technical texts and different writing styles) to evaluate efficiency of some of aforementioned extractors. We compared the extracted metadata with those which were defined manually *prior* to the automated extraction. In order to distinguish between closely related results and those which were totally different, we compared the results (each term from manually prepared set

[2] JATR, available at http://www.dcs.shef.ac.uk/~ziqizhang/#tool

[3] OpenNLP Models, http://opennlp.sourceforge.net/models.html

[4] Open Calais, http://www.opencalais.com

[5] Alchemy, http://www.alchemyapi.com/

[6] dmoz, http://www.dmoz.org

against each term from automatically acquired set) using Wordnet-based similarity provided by Wordnet::Similarity module[7]. Only in case when we have found that one term is a substring of another we did not employ the wordnet similarity measure. In such a case, we declared them as similar. When comparing the terms, we penalized those terms, which had similarity of 0 (no similarity at all) with all terms from the other set.

Apart from comparison of automatically extracted keywords against manually prepared ones, we also compared (following the same principles) two sets of manually prepared keywords, each coming from a different evaluator, in order to acquire a natural baseline for further interpretation of our apriori evaluation.

The results (Table 1) show us that algorithms and services were able to find an overlap with human annotator in every case, even if the achieved similarity was very low in quite a few cases. However, by examining the results for different types of web pages, we can identify the strengths and weaknesses of used algorithms and services.

Table 1. Comparison of selected metadata extractors against manually chosen metadata by normalized wordnet similarity

Domain	JATR	OpenCalais	tagthe.net	alchemy	Another human evaluator
institution homepage in Slovak (fiit.stuba.sk)	0.26	0.24	0.32	0.15	0.17
institution homepage in English (l3s.de)	0.09	0.14	0.12	0.08	0.03
news article in Slovak (sme.sk)	0.03	0.09	0.07	0.03	0.15
news article in English (bbc.co.uk)	0.02	0.16	0.15	0.02	0.22
technically-oriented text in Czech (root.cz)	0.25	0.21	0.41	0.46	0.35
technically-oriented blog in English (railstips.org)	0.13	0.0	0.17	0.003	0.17

The most accurate results (according to human annotator) were achieved on technical texts, the best one on the case of a tutorial-like article on *root.cz*, which was very specialized in particular technology and contained many acronyms relevant to that technology. The second technical text which was in English gained worse results, where, for instance, OpenCalais did not manage to get any meaningful term. We can explain this by the fact that the article was devoted to very new technologies and their usage, which were probably hard to distinguish in the text.

The worst results were achieved on complex news articles, especially the one which was translated automatically from Slovak to English. In the one coming from *bbc.co.uk*, at least OpenCalais and tagthe.net were able to achieve a similarity of more than 10%. In addition, news articles were the only case, where two human annotators achieved significantly higher level of agreement on keywords than a human annotator with any other algorithmic approach. The biggest difference is in the case of the text, which was machine-translated into English prior to further processing.

There is no clear winner among the evaluated approaches and it seems that they are, in effect, mutually eliminating their weakness, which means that they could be effectively used in a combination. For instance, when pure NLP oriented JATR fails, the semantically-enhanced services such as OpenCalais or tagthe.net are able to achieve fair results and vice versa.

Apart from the mentioned experiment, we also deployed the combination of OpenCalais and tagthe.net extractors to the enhanced proxy platform described in Section 3 to determine the efficiency of the solution in real-world usage. As the proxy solution can be, apart from logging, easily used to improve user experience with ordinary web sites, we used the on-the-fly created keyword-based user models to optimize and disambiguate queries given to a web search engine and to improve navigation within our faculty website. More, we provided users with a wordle-based[8] visualization of their user profiles and collected a precious feedback, which helped us to determine “web stop-words”, i.e., words which occur often on web pages but do not make any sense from the user's interests point of view. An example of such a user profile is displayed in Figure 1.

[7] Wordnet::Similarity project, http://wn-similarity.sourceforge.net

[8] Wordle tag cloud generator, http://www.wordle.net/

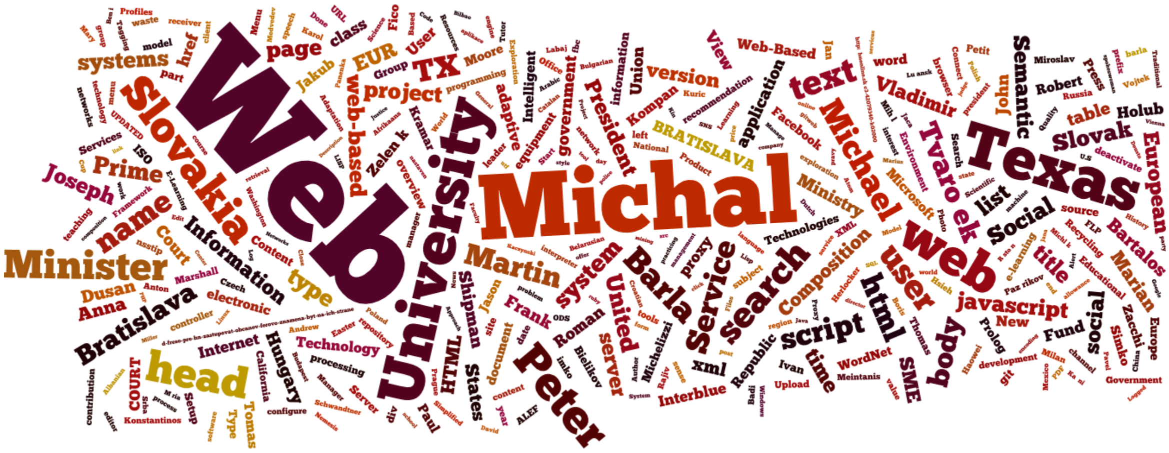

Figure 2. A wordle visualization of a keyword-based user profile collected by our enhanced proxy platform

.

The already mentioned improvement of navigation is based on combining behavioral analysis for deriving user’s interest in a web page he currently visits with collaborative filtering used for actual content recommendation (Holub & Bieliková, 2010). User's interest is derived from data acquired by tracking javascript inserted by the proxy server. An experiment with visitors of our faculty website proved that we are able to estimate correctly user's interest in a particular page by considering time spent actively on page along with additional implicit feedback indicators (i.e., scrolling, clipboard usage) and their comparison against values coming from other users.

User's of proxy server also benefited from an enhanced googling experience, as our proxy server was proposing (apart from ordinary search results) also results coming from optimized and disambiguated queries. This time, we used metadata acquired from visited web pages to automatically construct user models, which served as a basis for construction of social networks and virtual communities within them. These communities, along with user's current context (metadata acquired from web pages visited in current session), are used to infer new keywords which would optimize and disambiguate search queries (Kramár, et al., 2010). We observed that users *clicked and stayed* on the results coming out from the extended queries in 54.7% of cases, which is a significant improvement against normal googling pattern of our users without any recommendations, where they stayed only on 27.4% of all clicked results, which mean that our recommendations were meaningful. As the whole search optimization was driven by automatically constructed open corpus user model using metadata extracted from the visited web-pages, we can conclude that our approach is able to extract useful metadata from the web pages and to produce a good enough user models.

6. CONCLUSIONS

In this paper, we presented an approach to metadata extraction based on a combination of various methods. Key advantage and contribution is a move towards the “wild web” where personalization based on manually created domain models is impossible. Presented metadata extraction is a base for our new approach to open corpus user modeling. Experiments show that a user model constructed in this way can perform similarly to tag-based models constructed from user's tag space within a particular tagging system.

We described our platform of an enhanced proxy server, focusing on a way how it supports our user modeling process, which itself is highly dependent on extraction of metadata. The enhanced proxy server is an ideal way of experimenting with web applications, as we can log all relevant data along with overall browsing context (i.e., what was a user browsing in parallel to using our web application) without forcing the user to change his or her browsing behavior (such as using a proprietary browser). The basic evaluation of few of such services against a manually annotated dataset did not show a clear success (the best achieved score was 46%). However, we must emphasize that we were performing a strict *apriori* evaluation with strong penalization of any mismatched terms. Manual *posteriori* inspection of acquired results showed that most of the extracted keywords are meaningful and would definitely yield (after a proper statistical processing) a suitable user model.

However, only the usage within retrieval, recommendation or adaptation engines can really prove viability of this approach. We deployed our solution to automatic user model construction based on available metadata to our proxy platform, used for everyday web browsing and built different personalization services on the top of it (site-specific navigation recommendation and search query disambiguation). Both personalization services were evaluated as successful, which means that the underlying user modeling part is able to collect adequate information related to interests of particular users.

ACKNOWLEDGEMENT

This work was supported by the Scientific Grant Agency of SR, grant No. VG1/0508/09, the Cultural and Educational Grant Agency of SR, grant No. 345-032STU-4/2010, and it is a partial result of the Research & Development Operational Program for the project Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 25240120029, co-funded by ERDF.

REFERENCES

Brusilovsky, P. et al., 2007. *The Adaptive Web, LNCS 4321*. Springer, Berlin, Heidelberg.

Brusilovsky, P. & Henze, N., 2007, Open Corpus Adaptive Educational Hypermedia. *The Adaptive Web*, *LNCS 4321*. Springer, pp.671–696.

Barla, M et al, 2009. Rule-Based User Characteristics Acquisition from Logs With Semantics for Personalized Web-based Systems. *Computing and Informatics*, vol. 28, no. 4, pp. 399–427.

Sabou, M. et al., 2008. Evaluating the Semantic Web: A Task-Based Approach. *The Semantic Web*, *LNCS 4825*. Springer, pp. 423–437.

Fernandez, M. et al., 2008. Semantic Search Meets the Web, *Int. Conf. on Semantic Computing*, pp. 253–260.

Shepherd, M. et al., 2001. Browsing and Keyword-based Profiles: A Cautionary Tale. *Int. Conf. on System Sciences HICSS 2001, Volume 4*. IEEE Computer Society, pp. 4011.

Badi, R. et al., 2006. Recognizing User Interest and Document Value from Reading and Organizing Activities in Document Triage. *Int. Conf on Intelligent User Interfaces IUI 2006*. ACM, pp. 218–225.

Schwarzkopf, D. et al., 2007. Mining the Structure of Tag Spaces for User Modeling. *Data Mining for User Modeling On-line Proceedings of Workshop held at the Int. Conf. on User Modeling UM2007*, Corfu, Greece, pp. 63–75.

Zhang, Y. and Feng, B., 2008. Tag-based User Modeling Using Formal Concept Analysis, *Int. Conf. on Computer and Information Technology*. IEEE, pp. 485–490.

Carmagnola, F. et al., 2007. Towards a Tag-Based User Model: How Can User Model Benefit from Tags? *User Modeling 2007*, *LNCS 4511*.Springer, pp. 445–449.

Šimko, M. and Bieliková, M., 2009. Automated Educational Course Metadata Generation Based on Semantics Discovery. *Technology Enhanced Learning, EC-TEL 2009*, *LNCS 5794*. Springer, pp. 99-105.

Zhang, Z: et al. 2008, A Comparative Evaluation of Term Recognition Algorithms, *Int. Conf. on Language Resources and Evaluation LREC 2008*. pp. 2108–2113.

Barla, M. and Bieliková, M., 2009. “Wild” Web Personalization: Adaptive Proxy Server, *Workshop on Intelligent and Knowledge oriented Technologies, WIKT 2009*. Equilibria, Košice, pp. 48–51, (in Slovak).

Kramár, T. et al, 2010, Disambiguating Search by Leveraging the Social Network Context. *User Modeling, Adaptation and Personalization, UMAP 2010, LNCS 6075,* Springer, pp. 387-392

Holub, M.. and Bieliková, M., 2010. Estimation of User Interest in Visited Web Page, *WWW 2010. ACM,* pp. 1111-1112.