# Leveraging Lightweight Semantics for Search Improvement

Marián Šimko
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technology
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
simko@fiit.stuba.sk

Supervisor: Mária Bieliková
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technology
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
bielik@fiit.stuba.sk

## 1. PROBLEM AND MOTIVATION

As long as the machines do not understand the content of large information spaces (such as the Web), only a small fragment of knowledge contained within can be actually utilized. To be able to leverage such a huge amount of information, it is necessary to provide meaningful *metadata* about information resources. Metadata representing the semantics enable intelligent resource processing resulting in advanced services such as *recommendation* or personalized *search*.

Depending on the particular application, the form of metadata varies from semantically weak structures (e.g., interlinked keywords or folksonomies) to complex domain ontologies. Although ontologies are the most suitable for intelligent processing of information resources and reasoning, the number of domains they currently cover is relatively small. It is difficult to describe dynamically changing environments appropriately, especially at the desired level of granularity. Hence, for some domains we will probably never be able to provide ontology description complex enough.

We address the problem of metadata availability resulting from authoring complexity. Creation of metadata manually is in general almost impossible due to the vast number of metadata entities. Not only the concepts identification, but also reasonable interconnections creation is demanding task in particular.

The first goal of our research is to devise metadata representation sufficient for storing emergent semantics easily acquirable from the resources' content or yet existing (social) services from the Web. Secondly, our long-term goal is effective utilization of obtained semantics in order to improve current methods of information search and recommendation.

## 2. OUR APPROACH

The solution of the addressed problem we see in an automatic creation of resource *lightweight* metadata layer bridging the gap to richer semantics. We introduce a concept map-based structure as a simplified domain model. We understand concepts as fundamental semantic units exceeding the semantics of terms (basic units in traditional IR approaches), but not achieving the semantic power of ontological concepts (having taxonomies and multiple relationship types defined). Concept map is a flat structure with weighted relations determining concepts relatedness.

The advantage of described resource metadata modeling lies in its simplicity. Such simple representation enables automatic generation of the model. We have proposed several approaches how to generate metadata automatically. As sources containing relevant knowledge for metadata generation we consider resources' content, resources' link topology and annotations provided by users.

First group of approaches we utilize is related to the processing of the resources' *content*. By comprehensive text analysis we are able to recover semantics implicitly encoded in resources' textual representation. Although the semantics is generally difficult to acquire, we extract a metadata skeleton and obtain concept labels (not concepts themselves).

By analysis of the resources' *link topology* we discover relationships between concepts. Viewing the concept map as a graph, we apply graph algorithms (spreading activation, random walk variants) to rank concepts according to their relevance in the network.

Valuable source of knowledge contained within resources are annotations provided by users. We recognize two types: *explicit* and *implicit*. Explicit annotations are assigned by users using some of social services very popular in the Web 2.0 age (e.g., delicious.com). Implicit annotations are queries submitted by users and related to pages users chose among search results.

The work related to concept hierarchy creation is a subtask of ontology learning field [4]. Relationship discovery is typically associated with precise linguistic analysis. Most of the approaches rely on lexical or syntactical annotations, the presence of powerful POS taggers, already existing domain ontologies, or other external semantic resources (e.g., WordNet) [1][2]. Related work is also presented in [3] where similar concepts are discovered by latent semantic indexing (LSI) and K-means clustering of documents.

The novelty of our approach lies in the combination of the *easily* creatable metadata representation with the proposed graph-based algorithms *not yet been applied* in this context. Proposed method of metadata generation is, unlike the majority of state-of-the-art approaches, fully automatic, unsupervised and *not* dependant on external semantic resources, while yielding at least comparable results.

## 3. EVALUATION

We evaluated proposed approach of resource content processing and topology analysis by performing several experiments in e-learning domain. In Functional programming course we processed available learning objects (represented by web pages) and using proposed method we generated the map of interlinked concepts (Figure 1). Using precision and recall measures we assessed the accuracy of generated concept relationships. The experiments yielded resulting F-measure equal to 53.4% (65.2%).

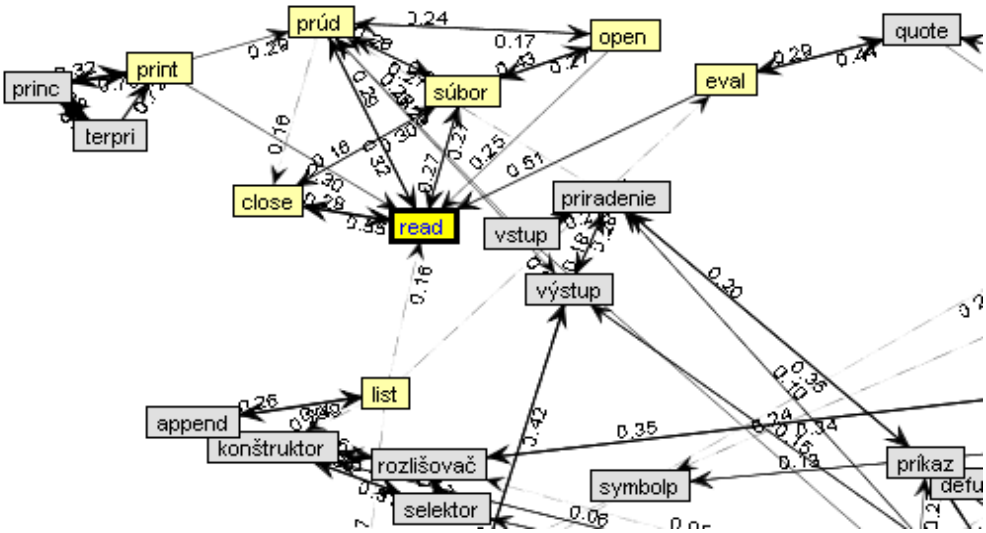


**Figure 1. Example of a metadata fragment we obtained after automatic metadata generation method application. Concepts (here in Slovak language) are connected to each other via simple *relatedness* relationship.**

Results seem promising with regards to unsupervised character of the method.

We also proposed a method of lightweight semantic search for open corpus considering two aspects of metadata: topological and statistical. The search utilizes concept map structure for query concept scoring computation that is used to sort retrieved documents according to the relatedness to the query. Topological as well as statistical features of metadata representation are considered.

Partial results of this project have been already presented at WIKT'08 (Smolenice, Slovakia), Knowledge'09 (Brno, Czech Republic), SemSearch Workshop at WWW'09 (Madrid, Spain).

## REFERENCES

[1] Buitelaar, P., Olejnik, D., and Sintek, M. A protégé plug-in for ontology extraction from text based on linguistic analysis. In Proc. of the 1st European Semantic Web Symposium (ESWS), 2004.

[2] Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S. Learning Taxonomic Relations from Heterogeneous Evidence. In Proc. of ECAI Workshop on Ontology Learning and Population, 2004.

[3] Fortuna, B., Grobelnik, M., Mladenic, D. Semi-automatic Construction of Topic Ontology. In Semantics, Web and Mining, Joint Int. Workshop, EWMF 2005 and KDO 2005, Porto, Portugal, October 3-7, 2005.

[4] Maedche, A., Staab, S. Ontology Learning for the Semantic Web, IEEE Intelligent Systems, Vol. 16, No. 2, pp. 72-79, 2001.

## STUDENT INFORMATION

Name: Marián Šimko

Email: simko@fiit.stuba.sk

Phone: HIDDEN

Address: HIDDEN

Type of study: Graduate

Affiliation: Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technology, Slovak University of Technology

Academic advisor: Professor Mária Bieliková

Poster preview:
`http://www.fiit.stuba.sk/~simko/ht-acm-poster.png`