# Discovering Keyword Relations

Peter KAJAN*

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

When working with keywords, there are some issues that one has to deal with. For instance, in the search process problems can be caused by: homonyms (wrong entities are found), synonyms (entities are not found) and the abstraction level (queries are too specific/general). If these relations are identified, search queries may be adapted to achieve more relevant results.

Analyses of folksonomies created in the tagging process are commonly used in the approaches revealing keyword relations [2]. But folksonomy like data can also be obtained from analysis of the Web usage and therefore larger datasets are produced. Users visit pages which can be described by keywords that are automatically extracted from the page content. These data (further called logs) cover more domains and are "cheaper" to acquire since document browsing is a more frequent action than tagging.

Another advantage of logs is a relevant timestamp attribute specifying the order of the visited documents and their keywords. We decided to design a method using this timestamp attribute and formulated the following hypotheses for revealing relations:

- *Similarity hypothesis:* If a significant number of users visit a page with keyword A and right after they visit a page with keyword B it means that these keywords may describe similar concept.
- *Hierarchy hypothesis:* Child keyword occurs more likely after parent keyword in browsing sessions.

In terms of [1], we define browsing session as a *set of documents visited by the user with particular information needs*. Our method consists of two steps (see Fig. 1). Logs are analysed to reveal similarity and parent-child relations in the first step. First, browsing sessions have to be identified in this step. Document similarities are calculated from the browsing sessions and then, they are used for keyword similarities calculation. Last, parent-child relations are identified according to the second hypothesis. Keywords are mapped to Linked data in the second step. There is a high probability that the relation between the keywords can be named if the similarity of two keywords is high. We decided to use the Linked data for naming these relations.

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

*Spring 2012 PeWe Workshop, April 1, 2012, pp. 63–64.*

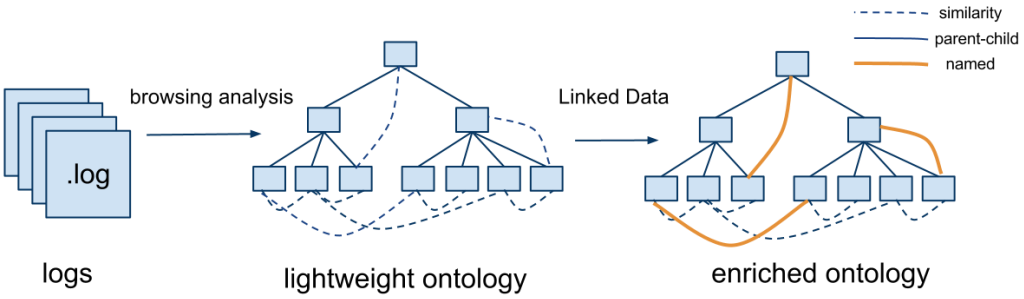


*Figure 1 Method of discovering keyword relations*

We evaluate the proposed method by implementing a prototype which analyses the logs of Web browsing activity from PeWe proxy[1]. The computations have been distributed using Hadoop[2] to achieve acceptable performance. The actual prototype is able to identify the keyword similarities. For illustration, chosen clusters of similar keywords are displayed in table 1.

*Table 1. Similar keywords grouped into clusters.*

| Cluster | First 10 keywords |
|---|---|
| 1 | Electrical, faculty, Electrical Engineering, Informatics, Industrial, Control, industrial informatics, Robotics, cybernetics, Technical |
| 2 | LG LCD Monitor, Sapphire Radeon HD, intel core 2 quad, Kingston HyperX XMP, Intel P45 Memory, box 2.66 ghz, ii 500w hdd, brand new price, CPU Cooler |
| 3 | Adriana Barraza, Idris Elba, Jaimie Alexander, Chris Hemsworth, Kat Dennings, Natalie Portman, Rene Russo, Anthony Hopkins, Ray Stevenson, Clark Gregg |

The goal of this work is to explore the importance of the ordering information for revealing relations. The idea of finding relations from document browsing seems to have a potential. In future work we plan to evaluate the results using qualitative experiment in which participants will rate discovered relations.

*Acknowledgement.* This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

## References

[1] Gayo-Avello, D.: A survey on session detection methods in query logs and a proposal for future evaluation. (2009), 1822-1843.

[2] Mika, P.: Ontologies are us: A unified model of social networks and semantics. *International semantic web conference,* (2005), pp. 522–536.

[1] http://peweproxy.fiit.stuba.sk/proxy/
[2] http://hadoop.apache.org/