EXPLOITING PUBLICATION REFERENCES FOR KEYWORD RELATEDNESS ASSESSMENT WITH EXPLICIT SEMANTIC ANALYSIS

Lukáš Žilka
Master Degree Programme (2), FIT BUT
E-mail: xzilka07@stud.fit.vutbr.cz

Supervised by: Lubomír Otrusina
E-mail: iotrusina@fit.vutbr.cz

Abstract: We aim to improve academic search by aiding the query expansion. For this purpose the keyword relatedness relation is defined as a model. It is based on the citation graph in academic publications. Explicit Semantic Analysis (ESA) is used as the implementation of this model for estimating the relatedness. ESA is trained on a corpus of automatically generated texts, that are created from a corpus of randomly selected research papers. Keyword relatedness is determined as the cosine similarity score of the ESA vectors of the keywords in the generated document collection.

Keywords: EEICT, Explicit Semantic Analysis, Keyword Relatedness, Information Retrieval

1 INTRODUCTION

A significant proportion of the work of scholars requires exploratory search of content stored in scientific databases. While keyword search is usually effective in finding content which is in the researcher's domain of expertise, it is not well suited for content exploration, because it is often difficult for the user to compose a good query. In this paper, we describe a novel approach for keyword relatedness that can be applied to improve exploratory search experience. It can serve both for the interaction with the user when entering the search query, and for the full automatic expansion of the query. The method is based on Explicit Semantic Analysis, which is trained on a special, automatically generated corpus of keyword documents. Those documents are generated from a set of merged research papers by a keyword extraction system.

The contributions of this papers are the introduction of the keyword relatedness measure in the research publications, and application of ESA as a model for approximating the keyword relatedness.

Explicit Semantic Analysis (ESA) is a method for computing semantic relatedness of two texts devised by [1]. It aims to represent any given text (input text) by a vector of weighed concepts (ESA vector). Relatedness of those texts is computed as the similarity of those vectors. So far, it gives superior results in computing semantic similarity, when compared to methods like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), or bag-of-words (BOW) [1]. We use it in a slightly different way to achieve our goal.

2 KEYWORD RELATEDNESS

We state the hypothesis for the keyword relatedness as follows: "The keywords k_1 and k_2 are related, if there exists a paper that cites some papers, where k_1 and k_2 are mentioned; or k_1 is in a paper and k_2 is in a paper that the paper (that contains k_1) cites."

Formally, *keyword relatedness* is a binary relation. This formal definition is simplified for the sake of clarity, and also because it is not yet sure what is the best measure that describes the contribution of one relation to the overall relatedness of the keywords. Without the simplification, the "$\sim$" relation should assign a numerical value that lies in the interval $[0; 1]$ to each pair of the keywords. Now, just

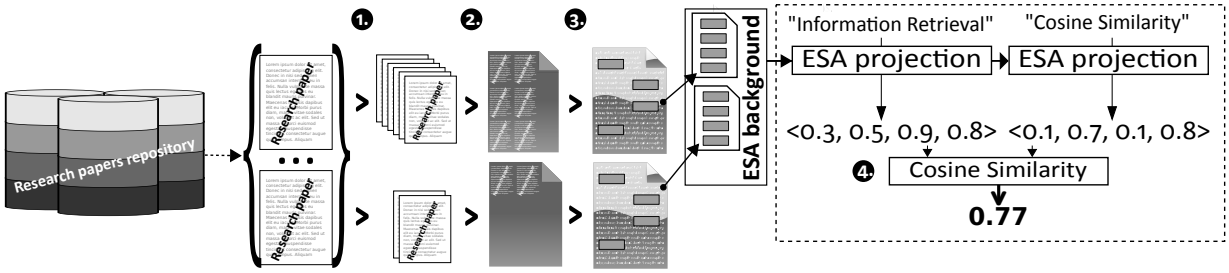

Figure 1: Illustration of our method for keyword relatedness assessment. The numbers correspond to the numbered steps in the text.

the fact that the keywords are related is defined: Let D be a set of research papers, let $d \in D$ be a research paper, and let $\rho_d \subseteq D$ be a subset of the research papers that are mentioned in the related work section of the paper d, then *keyword relatedness* "$\sim$" is defined as:

$$k_1 \sim k_2 \Leftrightarrow \exists d \in D : \{k_1, k_2\} \cap \bigcup_{d_r \in \rho_d} keywords(d_r) \neq \emptyset \quad (1)$$

It is based on the fact that every paper apart from introducing new ideas also puts into relation the research that has been made so far. Therefore, it puts into a relation also the concepts, and transitively, because the concepts are represented by the keywords, it puts into a relation also the keywords.

As every research paper should contain citations and a section about related work, a lot of training materials for learning the relatedness relation exist. The research publications are stored in the repositories and thus can be used for training the relatedness estimation methods.

3 METHOD

This section describes our ESA-based keyword relatedness assessment method. Using ESA we implement the keyword relatedness hypothesis from the previous section. In summary, for each input publication we build a keyword-document. It is done by resolving the text of the citations of the input document (i.e. find the text of the documents that it cites). Then, on each of the cited documents, the keyword extraction algorithm is run, and the keywords are accumulated. ESA background is then built from the keyword-document set. The following steps are taken and are illustrated in Figure 1:

1. **Meta-document Synthesis:** From a sample set of the research papers construct the meta-documents.
2. **Keyword Extraction:** Extract keywords from the meta-documents and create the keyword-documents.
3. **ESA Training:** Prepare a special ESA background based on the keyword-documents.
4. **Keyword Relatedness Assessment:** Utilize the built ESA background for computing the keyword relatedness.

4 INITIAL EXPERIMENTS

The research papers are taken from our database of research papers (that are used in the ReResearch project). The database consists of 11,000 research articles from the domain of Computer Science, extracted from the CiteSeerX site (http://citeseerx.ist.psu.edu/).

The database contains only about 3,200 research papers whose citation references can be resolved within the database and their text used. We were able to obtain 5,000 research papers by following up on the references. This is not ideal, but it is enough for an initial evaluation.

In this evaluation, we are interested to see what keywords does the system suggest for the given input keyword. The keywords that contain the words that are contained in the query string itself are filtered

out, and the keywords that contain a word that tends to recur in the results are grouped together by picking the one with the strongest ESA similarity score.

The program operates with the list of keywords for Computer Science obtained from MS Academic Search (http://academic.research.microsoft.com/). When a query is entered, it computes the ESA similarity between the query and each of the keywords. Then, it sorts the resulting list according to the similarity score and removes the keywords mentioned in the previous paragraph.

Here are some of the manually conducted experiments (suggestions that our system gives for a particular query):
Query: *Information Retrieval and Cosine Similarity* **Suggestions:** Index Structure, Range Minimum Query, Image Indexing, Automatic Indexing, Digital Documents, Combination Index, Human Development Index, Key Word Index, Index Words, Selection Index, Algorithm Theoretical Basis Document, Text Indexing, Index Selection Problem, Additional Index Words, Document Frequency, Query View Transformation, Document Image Processing, Query By Image Content, Document Image Analysis, Inverse Document Frequency

Query: *Information Retrieval* **Suggestions:** Content-based Indexing, Automatic Relevance Determination,Document Classification, Text Extraction, Text Database, Document Representation, Semantic Query Optimization, Document Structure,Text Indexing, Document Similarity, Data Documentation Initiative, Document Database, Structured Query Language

5 CONCLUSIONS AND FUTURE DIRECTIONS

The keyword relatedness has been defined as a citation-based measure in a document collection. We described our approach for its estimation, based on Explicit Semantic Analysis, along with our keyword extraction algorithm. Also, an initial evaluation of the system's implementation has been made, yielding promising results.

A systematic evaluation of the relatedness assessment system and the user-based testing are currently lacking and both will be a subject of the future research. Also, an efficient and user-friendly way of presenting the suggested keywords should be investigated. When the keyword suggestions are incorporated into a working system, the user feedback should be collected and used for improving the metrics and finding other features that are not captured by our model. The collected data can be used for the collaborative filtering to individually improve the performance of the system by the user's suggestions.

5.1 RELATED WORK

The closest work similar to ours is [2]. The authors utilize ESA, WordNet and collocation index (EWC) to expand the queries for information retrieval. We fundamentally differ from their approach in several ways: they aim to define a general-purpose measure for determining term relatedness, whereas we aim specifically for the relatedness for information retrieval in research papers; they utilize ESA the original way, based on Wikipedia background, for determining the term's similarity, whereas we propose an unusual application; and they assume different relatedness hypothesis.

REFERENCES

[1] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007.

[2] V. Klyuev and Y. Haralambous. Query expansion: Term selection using the ewc semantic relatedness measure. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 195–199. IEEE, 2011.