The *TSRM* Approach in the Document Retrieval Application*

Efthymios Drymonas, Kalliopi Zervanou, and Euripides G.M. Petrakis

Intelligent Systems Laboratory, Electronic and Computer Engineering Dpt.,
Technical University of Crete (TUC), Chania, Crete, Greece
max@softnet.tuc.gr, {kelly,petrakis}@intelligence.tuc.gr
http://www.intelligence.tuc.gr

Abstract. We propose *TSRM*, an alternative document retrieval model relying on multi-word, rather than mere single key-word domain terms, typically applied in traditional IR.

Key words: term extraction, term similarity, information retrieval

The *TSRM* Approach

TSRM is built upon the idea of computing the similarity among multi-word terms using internal (lexical) and external (contextual) criteria, while taking into consideration term variation (morphological and syntactic). *TSRM* can be viewed as a three phase process:

A. Corpus processing:	1. Corpus Pre-Processing 2. Term Extraction (C/NC value) 3. Term Variants Detection (FASTR) 4. Term Similarity Estimation
B. Query processing:	1. Query Pre-Processing 2. Query Term Extraction (C/NC value) 3. Query Term Expansion by term variants (FASTR)
C. Document Retrieval:	1. Similarity Computation (TSRM/A or TSRM/B) 2. Document Ranking

The C/NC-value method [1] is a domain-independent and hybrid (linguistic/statistical) method for the extraction of multi-word and nested terms. The candidate noun phrase termhood is represented by NC-value. FASTR [2] identifies term variants based on a set of morpho-syntactic rules. Based on the Nenadic et al. [3] study, term similarity (TS) in *TSRM*, is defined as a linear combination measure of two similarity criteria referred to as lexical similarity and contextual similarity. In ***TSRM/A*** document similarity is calculated as

$$Similarity(d,q) = \frac{\sum_i \sum_j q_i d_j TS(i,j)}{\sum_i \sum_j q_i d_j}, \quad (1)$$

* This work was supported by project TOWL (FP6-Strep, No. 26896) of the EU.

where i and j are terms in the query q and the document d respectively. The weight of a term (q_i, d_j respectively) is computed as the NC-value of the term. Eq. 1 takes into account dependencies between non-identical terms. In ***TSRM*/B** document similarity can be computed as:

$$Similarity(d,q) = \frac{1}{2}\{\frac{\sum_{i\in q} idf_i \max_j TS(i,j)}{\sum_{i\in q} idf_i} + \frac{\sum_{j\in d} idf_j \max_i TS(j,i)}{\sum_{j\in d} idf_j}\} \quad (2)$$

TSRM has been tested on OHSUMED, a standard TREC collection. The results in Fig. 1 demonstrate that *TSRM*, with both formulae for document matching, outperform Vector Space Model (VSM) in most cases (i.e., VSM performs well only for the first few answers).

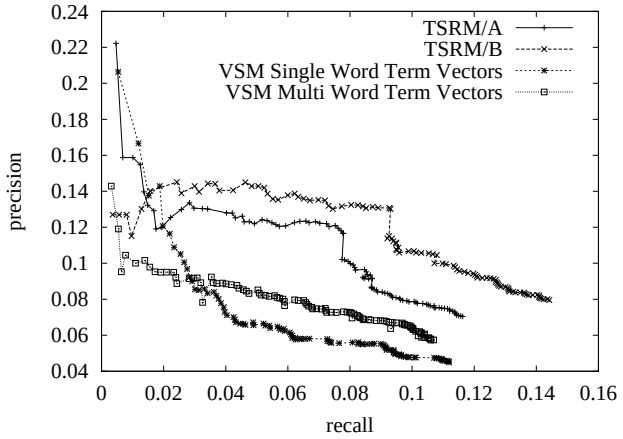

Fig. 1. Precision-recall diagram for retrieval on OHSUMED using *TSRM* and VSM.

Conclusions

We have discussed on the potential improvements to traditional IR models related to document representation and conceptual topic retrieval and proposed *TSRM* to demonstrate our ideas.

References

1. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: The C-Value/NC-value Method. Intl. J. of Digital Libraries **3**(2) (2000) 117–132
2. Jacquemin, C., Klavans, J., Tzoukermann, E.: Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In: 35^{th} Annual Meeting of the Assoc. for Comp. Linguistics, Spain (1997) 24–31
3. Nenadic, G., Spasic, I., Ananiadou, S.: Automatic Discovery of Term Similarities Using Pattern Mining. Intl. J. of Terminology **10**(1) (2004) 55–80