From terms to concepts: a revisited approach to Local Context Analysis

Annalina Caputo, Pierpaolo Basile and Giovanni Semeraro

Department of Computer Science
University of Bari
70126 Bari, Italy
{basilepp,acaputo,semeraro}@di.uniba.it

Abstract. Pseudo-Relevance Feedback (PRF) is a widely used technique which aims to improve the query representation assuming as relevant the top ranked documents. This should results in better performance as, after the expansion and re-weigh of the original query, the resultant vector should contain all those worth features able to express utterly the user's information need. This paper presents the application of a pseudo-relevance feedback technique, called Local Context Analysis (LCA), to SENSE (SEmantic N-levels Search Engine). SENSE is an IR system that tries to overcome the limitations of the ranked keyword approach by introducing semantic levels which integrate (and not simply replace) the lexical level represented by keywords. The evaluation shows that this PRF technique is able to work worthily on both the lexical level represented by keywords and the semantic level represented by WordNet synsets.

1 Introduction and Background

LCA [6] is a PRF technique which exploits the context of query words in a collection of documents, by analyzing which words in the top ranked documents simultaneously co-occur with the most of query terms. This paper presents an extension of LCA in SENSE [2], an IR system which aims to be a step forward traditional keyword-based systems. The main idea underlying SENSE is the definition of an open framework to model different semantic aspects (or levels) pertaining document content. Two basic levels are available in the framework: The *keyword level*, the entry level in which the document is represented by the words occurring in the text, and the *word meaning level*, represented through *synsets* obtained by WordNet, a semantic lexicon for the English language. A synset is a set of synonym words. Word Sense Disambiguation algorithms are adopted to assign synsets to words. Analogously, several different levels of representation are needed for representing queries. In this model also the notion of relevance of a document d in the collection for the user query q is extended to several levels of representation. A *local similarity function* computes the document relevance for each level, according to feature weights defined by the corresponding local scoring function. Then, a *global ranking function* is needed to merge all the result

lists that come from each level in a single list of documents ranked in decreasing order of relevance. In the same way, the PRF technique should be able to work over all the levels involved in our model.

2 nLCA

LCA proved its effectiveness on several test collections. This technique combines the strength of a global relevance feedback method like PhraseFinder [4] while preventing its drawbacks. LCA selects the expansion terms directly from the collection on the basis of their co-occurrences with query terms. Differently from PhraseFinder, this method computes this statistics on the basis of the top-ranked documents that are assumed to be the relevant ones, with a considerable gain in efficiency. Then, LCA joins the advantage of a global technique with the efficiency of a *local* one. This technique is grounded on the hypothesis that terms frequently occurring in the top-ranked documents frequently co-occur with all query terms in those documents too. Our work exploits the idea of LCA in the N-levels model. In that model, LCA is integrated into two representation levels: keyword and word meaning. The challenge lies in the idea that the LCA hypothesis could also be applied to the word meaning level, in which meanings are involved instead of terms. The original measure of co-occurrence degree is extended to encompass the weight of a generic feature (keyword or word meaning) rather than just a term.

We modify the orginal formula introducing two new factors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ (in bold in following formulae):

$$codegree(f, q_i) = \frac{log_{10}(co(f, q_i) + 1) \cdot idf(f)}{log_{10}(n)} \tag{1}$$

codegree is computed starting from the degree of co-occurrence of the feature f and the query feature q_i ($co(f, q_i)$), but it takes also into account the frequency of f in the whole collection ($idf(f)$) and normalizes this value with respect to n, the number of documents in the top-ranked set.

$$co(f, q_i) = \sum_{d \in S} tf(f, d) \cdot tf(q_i, d) \cdot \boldsymbol{\theta} \tag{2}$$

$$idf(f) = min(1.0, \frac{log_{10} \frac{N}{N_f}}{5.0}) \tag{3}$$

where $tf(f, d)$ and $tf(q_i, d)$ are the frequencies in d of f and q_i respectively, S is the set of top-ranked documents, N is the number of documents in the collection and N_f is the number of documents containing the feature f. For each level, we retrieve the n top-ranked documents for a query q and then we rank the feature belonging to those documents by computing the function *lca*, as follows:

$$lca(f, q) = \prod_{q_i \in q} (\delta + \boldsymbol{\gamma} \cdot codegree(f, q_i))^{idf(q_i)} \tag{4}$$

θ and γ transfer the importance of a query term into the weight of words it co-occurs with. In fact, θ takes into account the frequency of a query term (qf) in the original query ($\theta = 1 + log(qf(q_i))$), while γ takes into account a boost factor associated with a specific query term ($\gamma = 1 + log(boost(q_i))$). lca is used to rank the list of features that occur in the top-ranked documents, δ is a smoothing factor, while the power is used to raise the impact of rare features. The new query q^* is given by the sum of the original query q and the expanded query q', where $q' = (w_{f_1}, ..., w_{f_k})$ and $w_{f_i} = 1.0 - \frac{0.9i}{k}$ is the weight of the i-th feature f_i. Hence, the new query is re-executed to obtain the final list of ranked documents for each level. Differently from the original work, we applied LCA to the top ranked documents rather than passages[1].

3 Setting the scene

We evaluate our technique on the CLEF Ad-Hoc Robust Task collection [1]. The CLEF collection is composed by 166,717 documents and 160 topics. In this collection both documents and topics are disambiguated by the task organizers. Topics are structured in three fields: $Title$, $Description$ and $Narrative$. All query fields are exploited in the search phase with a different boost factor: $Title = 8$, $Description = 2$ and $Narrative = 1$. We use the Okapi BM25 [5] as local similarity functions for both meaning and keyword levels. In particular, we adopt the BM25-based strategy which takes into account multi-field documents. Documents in CLEF collection are represented by two fields: HEADLINE and TEXT. The multi-field representation reflects this structure. We set the BM25 parameters as follows: $b = 0.7$ in both levels, $k_1 = 3.25$ and 3.50 in keyword and meaning levels respectively. We tested several n, k, and δ values, and we set $n, k = 10$ and $\delta = 0.1$. To compute the global ranking function we adopt the CombSUM [3] strategy, giving a weight of 0.8 to the keyword level and 0.2 to the meaning level. All parameters (boosting factors, BM25 and global ranking function) are set after a tuning phase over a set of training topics provided by organizers. In order to compare our approach we consider the Mean Average Precision (MAP) and the Geometric Mean Average Precision (GMAP).

4 Results and Remarks

We performed two experiments in which one level at a time is considered and then the two lists are merged producing a single list of ranked documents. We explored two strategies involving LCA: The first strategy (lca) is based on the formula proposed in [6]. In the second strategy (lca-n), we took into account also the meaning level and we decided to expand only synsets referring to nouns. The second strategy tries to overcome a limit of Word Sense Disambiguation algorithms which, in general, have better performance with nouns. The latter strategy (lca-n-$\theta\gamma$) is based on lca-n, but with the introduction of θ and γ factors. The results of our evaluation are depicted in Table 1.

[1] In the original work, passages are parts of document text of about 300 words

Table 1. Results on CLEF Ad-Hoc Robust collection

	Run	***MAP***	***GMAP***
one-level (no-expansion)	keyword	.4207	.1900
	synset	.3119	.1197
n-levels	no-expansion	.4253	.1973
	lca-n	.4304	.1945
	lca-n-$\theta\gamma$	.4532	.2114

While the synset level alone is not able to reach the performance of the keyword level, the combination of these two levels without expansion strategies (*no-expansion*) improves performance in both MAP and GMAP. All *lca* strategies exploited in this paper outperform our baseline (*no-expansion*). However, it is worth to highlight here that the expansion on synset level produces slightly better results with respect to the standard metod *lca* when it involves only nouns (*lca-n*). The introduction of θ and γ parameters results in the best performance. This result supports the claim that the weight of query terms is important also to weigh the expansion terms. Future work will include the comparison in the N-levels model of the proposed approach with other PRF, such as Rocchio, Divergence from Randomness and Kullback-Leibler language modeling.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Otegi, A.: CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task. In: Peters, C., Di Nunzio, G., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) Multilingual Information Access Evaluation, Vol. I: Text Retrieval Experiments. Lecture Notes in Computer Science, Springer (2009)
2. Basile, P., Caputo, A., Gentile, A.L., Degemmis, M., Lops, P., Semeraro, G.: Enhancing semantic search using N-levels document representation. In: Bloehdorn, S., Grobelnik, M., Mika, P., Tran, D.T. (eds.) Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, June 2nd, 2008. CEUR Workshop Proceedings, vol. 334, pp. 29–43. CEUR-WS.org (2008)
3. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: TREC. pp. 243–252 (1993)
4. Jing, Y., Croft, W.B.: An association thesaurus for information retrieval. In: RIAO 94 Conference Proceedings. pp. 146–160 (1994)
5. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. pp. 42–49. CIKM '04, ACM, New York, NY, USA (2004)
6. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. 18(1), 79–112 (2000)