Semantically Enhanced Term Frequency

Christof Müller and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Computer Science Department,
Technische Universität Darmstadt, Germany, http://www.ukp.tu-darmstadt.de

Abstract. In this paper, we complement the term frequency, which is used in many bag-of-words based information retrieval models, with information about the semantic relatedness of query and document terms. Our experiments show that when employed in the standard probabilistic retrieval model BM25, the additional semantic information significantly outperforms the standard term frequency, and also improves the effectiveness when additional query expansion is applied. We further analyze the impact of different lexical semantic resources on the IR effectiveness.

Key words: Information Retrieval, Semantic Relatedness

1 Introduction and Approach

The majority of information retrieval (IR) models is based on the bag-of-words paradigm. The performance of these models, however, is limited among other things due to the polysemy and synonymy of terms. The importance of the semantic relations or associations between terms has therefore long been recognized. Several approaches have been proposed to improve IR effectiveness employing methods like query expansion, document expansion [1], or topic models [2]. In this paper, we complement the term frequency (tf), which is widely used in IR models, with information about the semantic relatedness (SR) of query and document terms. Our hypothesis is that this additional knowledge will enable IR models to estimate the document relevance more accurately, as thereby also information about the meaning of non-query terms in the documents is taken into account. In our experiments, we evaluate this approach using a standard probabilistic model, i.e. BM25 [3]. For computing SR, we use *Explicit Semantic Analysis* (ESA), which was introduced by Gabrilovich and Markovitch [4].

The BM25 model estimates the relevance of a document d and a query q as

$$r(d,q) = \sum_{t_q \epsilon q} \frac{(k_1+1)\, tf_{t_q,d}}{k_1 \left((1-b) + b\frac{l}{l_{avg}}\right) + tf_{t_q,d}} \cdot \frac{(k_3+1)\, tf_{t_q,q}}{k_3 + tf_{t_q,q}} \cdot \log \frac{(N - df_{t_q} + 0.5)}{df_{t_q} + 0.5}$$

where t_q is a term in q, $tf_{t_q,q}$ or $tf_{t_q,d}$ is its tf in q or d, l is the document length, l_{avg} is the average document length in the collection, N is the collection size, and df_{t_q} is the number of documents containing t_q. k_1, k_3, and b are parameters. In order to enhance tf, we substitute $tf_{t_q,d}$ in the above equation with $tf_{t_q,d} + s \cdot \sum_{t_d \epsilon d, t_d \neq t_q} tf_{t_d,d} \cdot sr\,(t_q, t_d)$ where $sr\,(t_q, t_d)$ is an estimation of the SR of query term t_q and document term t_d, and s is a parameter that controls

the impact of SR on the final tf value. For $sr(t_q, t_d)$, we use the score computed by ESA after applying a predefined threshold to take into account only strong SR values. We additionally experiment with binary values, setting $sr(t_q, t_d)$ to either 0 or 1 depending on whether the ESA score is below or above the threshold.

ESA requires a lexical semantic resource (LSR) for which Gabrilovich and Markovitch originally employed Wikipedia (WP). Terms are represented as vectors of their tf.idf values in WP articles which are taken as textual representations of concepts. Texts are represented as the centroid vectors of the terms' concept vectors. The SR of a pair of terms or texts is then computed by using the cosine similarity measure. ESA has shown very good effectiveness for assessing the SR of terms and texts. However, when applied to IR, retrieval effectiveness was improved only when ESA scores were linearly combined with the relevance scores of bag-of-words based IR models [5, 6]. Egozi et al. also found that additionally employed centroid vectors of small passages of the documents were crucial for the IR effectiveness. They argue that otherwise "ESA tries to "average" several topics and the result is less coherent". In our approach, the ESA scores are directly integrated into the bag-of-words based model and no passage-based index is built, as the ESA scores are computed for pairs of query and document terms.

2 Experiments and Discussion

Besides using the German WP as LSR, we follow Zesch et al. [7] and employ Wiktionary (WKT) and GermaNet[1] (GN), as they have shown good performance on estimating the SR of terms. We use the German GIRT corpus (consisting of titles, abstracts, and meta data of social science texts) and a collection of German newspaper articles (NEWS) as test collections and LSRs.[2] With the goal of reducing noise and computational costs, we set small values for concepts to zero if they are below an empirically set pruning threshold after the concept vector was normalized by its length. We tested several thresholds and found that 0.015 performs best for all LSRs. Table 1 shows the document collections used as LSR, the number of contained terms and documents (concepts), and the average number of concepts per term with non-zero values before and after pruning.

Besides standard preprocessing steps like tokenization and stop word removal, we perform stemming and compound splitting. We employ the IR framework Terrier[3] and select the parameters b and k_1 of BM25, as well as the parameters of SR by using simulated annealing on a training set.[4] We perform two sets of experiments where we enhance tf with SR (i) only for documents that contain at least one query term, and (ii) for all documents. For the optimized parameter configurations, we additionally apply query expansion (QE) with the term weighting model Bo1 which uses the Bose-Einstein statistics and is one of

[1] `http://www.sfs.uni-tuebingen.de/GermaNet`

[2] Both collections were used at CLEF. See `http://clef-campaign.org` for details.

[3] `http://ir.dcs.gla.ac.uk/terrier`

[4] training set: 75 topics of CLEF'03–05 for GIRT; 100 topics of CLEF'01–02 for NEWS
test set: 75 topics of CLEF'06–08 for GIRT; 60 topics of CLEF'03 for NEWS

Table 1. Document collections used as LSR for ESA.

LSR	#terms	#concepts	avg. #concepts per term	
			unpruned	*pruned*
GIRT	348,308	151,319	34.03	19.32
NEWS	874,637	294,339	41.17	21.94
WP	4,185,730	530,886	31.38	12.63
WKT	195,705	113,341	6.11	6.05
GN	44,879	42,014	5.86	5.85

the most effective weighting models based on the *Divergence From Randomness* framework [8]. The two parameters for QE, i.e. the number of terms to expand a query with and the number of top-ranked documents from which these terms are extracted, are also optimized using simulated annealing. We use SR however only for the initial step of retrieving documents from which to extract expansion terms and not for the final retrieval with the expanded query. Otherwise, computing SR for the large number of (possibly erroneous) expansion terms (up to 100) causes a strong topic drift and retrieval effectiveness decreases.

Table 2 shows the mean average precision (MAP) for each LSR on the two test collections with and without QE using the topics of the test set. Except for WKT on the GIRT collection, the enhancement of tf with SR increases MAP for all LSRs on both test collections. Especially for the NEWS test collection, the improvements are statistically significant. We found that using the original ESA score consistently performs better than substituting it with a binary value. Without employing QE, the test collection itself is the best performing LSR. SR shows similar improvements for BM25 as when employing QE, but only when the test collection is used as LSR. Other resources either have a lower coverage of query and document terms or contain too general term relations which can cause a topic drift for some of the queries. Anderka and Stein [9] found that ESA also performs well with other document collections than WP, which do not necessarily fulfill the requirement that each document describes exactly one concept as was originally the idea behind ESA. Our experimental results suggest a similar conclusion, as the NEWS and GIRT collections perform similar or even better than WP. We also linearly combined the relevance scores of the BM25 model and the original ESA model of Gabrilovich and Markovitch [4], but found that our approach consistently results in a higher MAP for all configurations.

Employing SR improves QE for all configurations. The quality of expansion terms is increased as SR ranks relevant documents higher during the initial retrieval step, and the top-ranked documents contain a larger number of terms that are strongly related to the query terms. Of all LSRs, WKT and GN contain the lowest number of terms and term relations, as they do not consist of long texts, but rather short lexicographic entries. In our experiments, they show the lowest MAP, except for GN on the GIRT collection, where QE is surprisingly improved by GN the most.

The consideration of documents that do not contain any of the query terms when computing SR, improves the retrieval effectiveness in most cases, although not dramatically. This comes, however, with an increase in computational costs.

Table 2. MAP and difference to BM25 without SR in percent. Statistically significant improvements (paired t-test, $\alpha = 0.05$) are marked with *. Highest MAP is in bold.

	GIRT		with QE		**NEWS**		with QE	
LSR	MAP	% diff	MAP	% diff	MAP	% diff	MAP	% diff
—	0.3609	—	0.4076	—	0.3487	—	0.4156	—
computing SR for documents that contain at least one query term								
GIRT	0.3986	+10.45*	0.4110	+0.83	0.3933	+12.79*	0.4421	+6.38*
NEWS	0.3693	+2.33	0.4128	+1.28	0.4116	+18.04*	0.4435	+6.71*
WP	0.3742	+3.69	0.4126	+1.23	0.3881	+11.30*	0.4458	+7.27
WKT	0.3575	–0.94	0.4076	0.00	0.3814	+9.38*	0.4308	+3.66
GN	0.3612	+0.08	0.4092	+0.39	0.3712	+6.45*	0.4326	+4.09
computing SR for all documents								
GIRT	**0.4148**	**+14.93***	0.4135	+1.45	0.3871	+11.01*	0.4210	+1.30
NEWS	0.3754	+4.02*	0.4145	+1.69	**0.4166**	**+19.47***	**0.4494**	**+8.13***
WP	0.3850	+6.68*	0.4118	+1.03	0.3901	+11.87*	0.4472	+7.60*
WKT	0.3548	–1.69	0.4073	–0.07	0.3817	+9.46*	0.4351	+4.69
GN	0.3621	+0.33	**0.4158**	**+2.01**	0.3726	+6.85*	0.4391	+5.65

The efficiency of our approach can in general be increased by precomputing or caching the SR values.

The results of our experiments need to be further analyzed and substantiated on other test collections. However, they are very promising, and as tf is widely used, this approach allows a simple while effective integration of SR into existing IR models. Preliminary results on employing the enhanced tf in the PL2 model suggest similar performance improvements as were yielded in the BM25 model.

Acknowledgements This work was supported by the Volkswagen Foundation (grant I/82806) and the German Research Foundation (grant GU 798/1-3). We thank Aljoscha Burchardt, György Szarvas, and the anonymous reviewers for their helpful comments.

References

1. Tao, T., Wang, X., Mei, Q., Zhai, C.: Language Model Information Retrieval with Document Expansion. In: Proc. of HLT-NAACL'06
2. Yi, X., Allan, J.: A Comparative Study of Utilizing Topic Models for Information Retrieval. In: Proc. of ECIR'09
3. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Information Processing and Management **36**(6) (2000)
4. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proc. of IJCAI'07
5. Egozi, O., Gabrilovich, E., Markovitch, S.: Concept-Based Feature Generation and Selection for Information Retrieval. In: Proc. of AAAI'08
6. Müller, C., Gurevych, I.: Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access. Volume 5706 of LNCS. Springer (2009)
7. Zesch, T., Müller, C., Gurevych, I.: Using Wiktionary for Computing Semantic Relatedness. In: Proc. of AAAI'08
8. Amati, G.: Probability Models for Information Retrieval based on Divergence from Randomness. PhD thesis, Dept. of Computing Science, Univ. of Glasgow (2003)
9. Anderka, M., Stein, B.: The ESA Retrieval Model Revisited. In: Proc. of SIGIR'09