A Logical Inference Approach to Query Expansion with Social Tags

Christina Lioma[1], Roi Blanco[2], and Marie-Francine Moens[1]

[1] Computer Science, Katholieke Universiteit Leuven, 3000, Belgium
[2] IRLab, Computer Science Department, A Coruña University, Spain
christina.lioma@cs.kuleuven.be, rblanco@udc.es, sien.moens@cs.kuleuven.be

Abstract. Query Expansion (QE) refers to the Information Retrieval (IR) technique of adding assumed relevant terms to a query in order to render it more informative, and hence more likely to retrieve relevant documents. A key problem is how to identify the terms to be added, and how to integrate them into the original query. We address this problem by using as expansion terms social tags that are freely available on the Web. We integrate these tags into the query by treating the QE process as a logical inference (initially proposed in [3]) and by considering the addition of tags as an extra deduction to this process. This work extends Nie's logical inference formalisation of QE to process social tags, and proposes an estimation of tag salience, which is experimentally shown to yield competitive retrieval performance.

1 Introduction

Query Expansion (QE) is an Information Retrieval (IR) technique that aims to expand queries with assumed relevant terms in order to render them more informative and hence facilitate the retrieval of relevant documents. Typically, the terms used for expansion are fetched from some collection or thesaurus, and weighted in different ways. There exists a variety of different ways to do so, overviewed in [5]. We present an approach to QE that uses social tags to expand queries. We collect these tags from the Web, and estimate their salience before using them to expand user queries. We formalise this as a logical inference approach, following Nie's original such formalisation of QE [3]. We extend Nie's logical inference representation by adding an additional estimation of the content salience of social tags. This work contributes an illustration of the ease with which Nie's treatment of QE as logical inference can be extended to accommodate further sources of QE, social tags in this case. In addition, this work contributes a novel estimation of tag salience. Illustrative experiments show our proposed tag-QE approach to yield competitive retrieval performance.

Section 2 presents the logical inference approach to QE with social tags, Section 3 presents illustrative experimental results, and Section 4 summarises this work and outlines future research directions.

2 Query Expansion with Social Tags as Logical Inference

Let K represent a knowledge system upon which all inference is made. Let d denote a document, and q denote a query. Then, the relevance of d to q with respect to this system can be expressed as $K \vdash d \rightarrow q$. If one can prove that $K \vdash d \rightarrow q$, then the document is said to be relevant to the query, otherwise the document is said to be irrelevant to the query. Nie [3] applies this representation to model QE, by defining a new query q' that constitutes an expanded expression of the original query q. Then, by applying classical logic transitivity, the evaluation of $K \vdash d \rightarrow q$ can be done as follows (K is removed henceforth): $d \rightarrow q' \wedge q' \rightarrow q \vdash d \rightarrow q$. This relation means that the new query q' is satisfied (implied) by the document, in which case the original query q is also satisfied by the document. Because q' can be any query expression, the above deduction can be written as: $\vee_{q'}(d \rightarrow q' \wedge q' \rightarrow q) \vdash d \rightarrow q$. Interpreting this formula in a context that involves uncertainty, the following function P can be defined:

$$P(d \rightarrow q) = P(\vee_{q'}(d \rightarrow q' \wedge q' \rightarrow q)) \tag{1}$$

where $P(d \rightarrow q')$ measures the degree of direct satisfaction of query q' to document d, and $P(q' \rightarrow q)$ measures the degree of relatedness of query q' to the original query q. Eq. 1 can be interpreted as the probability $P(R|q,d)$ that a document d is relevant to a query q as follows: $P(R|d,q) = \sum_{q'} P(R,q'|d,q) = \sum_{q'} P(R|d,q,q')P(q'|d,q)$. Assuming that q' is a good approximation of q leads to: $P(R|d,q,q') = P(R|d,q')$. The derivation of q' depends only on q, not on d, hence $P(q'|d,q) = P(q'|q)$. Based on this, we get the following expression:

$$P(R|d,q) = \sum_{q'} P(R|d,q')P(q'|q) \tag{2}$$

where $P(R|d,q')$ denotes the relevance estimation of the document to the derived query, and $P(q'|q)$ denotes the relationship between the original query q and the derived query q'. Eq. 2 can be rewritten in order to express QE on the basis of individual terms, rather than whole queries, as follows (see [3] for the full derivation):

$$P(R|d,q) = \sum_{t'} P(R|d,t')P(t'|q) \tag{3}$$

where t' denotes a term in the expanded query. This formula allows us to consider the uncertainty of the correspondence between the expansion terms and the original query terms as a factor in the estimation of relevance.

Eq. 3 has two components. The first component, $P(R|d,t')$, may be interpreted as the term weight within a document, and can be estimated by various different ranking models, for instance with Okapi's BM25 [4], which we use in this work. The second component, $P(t'|q)$, may be interpreted as the term importance of a query, and has to be estimated in a way that reflects the probability of finding an expansion term in the query. Applied to our case of QE with tags, $P(t'|q)$ denotes the probability of finding a tag (denoted τ) in the query. This

probability must be estimated in a way that reflects the salience of the tag. We propose the following IDF-like approximation:

$$P(\tau|q) = \frac{N}{n_\tau} \tag{4}$$

where N is the number of documents in the collection, and n_τ is the number of documents in the collection that contain the tag τ. The aim of Eq. 4 is to discriminate between tags on the basis of how many documents within a large collection are associated to them (hence 'tagged' by them). Eq. 4 is one suggestion for estimating tag salience, which we evaluate experimentally in Section 3. Further alternative estimations are possible, for instance by relatively straight-forward extensions to IDF, such as RIDF [2], or by more elaborate approximations of tag topicality, such as Zhou et al.'s approach [6] that uses Bayesian Inference.

3 Experimental Evaluation

We present an illustrative evaluation of our proposed QE with tags, which is organised as follows: The baseline is standard retrieval without QE. This baseline is compared against our proposed QE with tags. In order to contextualise this comparison, we further compare these results to a state-of-the-art retrieval with conventional QE (i.e. QE that uses weighted terms for expansion). At all times, retrieval is realised with BM25, whose parameter b is tuned separately for Mean Average Precision (MAP) and Precision at 10 (P10). For conventional QE with terms, we use DFR's Bo1 model [1]. Both conventional QE and our proposed QE include as parameters (i) the number of terms (resp. tags) used for expansion, and (ii) the number of documents from which these terms (resp. tags) are drawn. We tune these parameters by varying them between 1-30 (for terms or tags) and 1-10 (for documents) separately for MAP and P10. Finally, the tags used for our proposed QE are collected by querying Del.icio.us, similarly to [6], whereas the terms used for conventional QE are collected from the same collection used for retrieval. The retrieval collection is the TREC BLOG06 collection (25GB) with queries 901-950 (title only). For our QE with tags, when applying Eq. 4, we compute N and n_τ from the BLOG06 collection, because we do not have access to the statistics of the collection used by Del.icio.us. We assume that Del.icio.us uses a very large collection, and that BLOG06 is a large enough approximation of it (in terms of size). Table 1 displays the performance of our retrieval experiments without QE, with our proposed tag-QE, and with conventional term-QE. We see that our proposed QE outperforms both the baseline and the conventional term-QE at all times, and with respect to both mean and early precision. This observation indicates that our use of tags enhances retrieval performance, not only by fetching a bigger number of relevant documents, but also by fetching more precise documents (i.e. documents of higher relevance to the query). This observation may indicate that the IDF-like formula we proposed to estimate tag salience was a successful approximation, an indication worth analysing further. Overall, these experiments indicate that social tags, when filtered appropriately, may benefit IR, a conclusion also echoed by Zhou et al. [6].

measure	BM25 - No QE	BM25+QE tags	Δ%	BM25+QE terms	Δ%
MAP	0.3517	0.3636	+3.38	0.3519	+0.06
P10	0.6220	0.6540	+5.14	0.6420	+3.21

Table 1. Mean Average Precision (MAP) and Precision at 10 (P10) using as baseline BM25 without QE. Against this baseline we compare our proposed method of QE with Del.icio.us tags. To contextualise this comparison, we also display conventional QE with terms. Δ marks the % difference from the baseline.

4 Conclusion

We presented an approach to Query Expansion (QE) that adds to a query tags collected from a free online social tagging system. By formalising QE as a logical inference process, as proposed by [3], we were able to integrate into ranking an approximation of the uncertainty that a tag is relevant to the original query terms. Specifically, we realised this approximation by proposing an IDF-like weight of tag salience, which considers how many documents in a collection are tagged by a given tag. Both the treatment of QE as logical inference, and the proposed weight of tag salience used clean and tractable estimations. An illustrative experimental evaluation with a 25GB TREC collection showed our proposed tag-QE technique to outperform a baseline of no QE, as well as a state-of-the-art QE model that uses weighted terms for expansion. This is a first positive indication that social tags, when filtered appropriately, may benefit IR. Future research will be geared toward refining the estimation of tag salience, and analysing in depth the effect of tag-QE on a per query basis.

Acknowledgements: Author 2 was partially supported by *Ministerio de Ciencia e Innovación*, FEDER and *Xunta de Galicia* under projects TIN2008-06566-C04-04 and 07SIN005206PR

References

1. G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, 2003.
2. K. W. Church and W. A. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.
3. J.-Y. Nie. Query expansion and query translation as logical inference. *JASIST*, 54(4):335–346, 2003.
4. S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In D. K. Harman, editor, *NIST Special Publication 500-236: TREC-4*, pages 73–96. Springer-Verlag, 1995.
5. J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR*, pages 4–11, 1996.
6. D. Zhou, J. Bian, S. Zheng, H. Zha, and C. Lee Giles. Exploring social annotations for information retrieval. In *WWW*, pages 715–724, 2008.