Information theoretic retrieval with structured queries and documents (DRAFT)

Claudio Carpineto[1], Giovanni Romano[1], and Caterina Caracciolo[2]

[1] Fondazione Ugo Bordoni, Rome, Italy
{carpinet, romano}@fub.it
[2] FAO, Rome, Italy
caterina.caracciolo@gmail.com

1 Introduction

Information retrieval through statistical language modeling has become popular thanks to its firm theoretical background and good retrieval performance. One goal of current research on structured information retrieval is thus to extend such models to take advantage of structure information.

As a structure may be present on documents or queries or both, we are interested in supporting not only unstructured queries on structured documents, but also structured queries on unstructured documents as well as structured queries on structured documents. Most of research work has considered the first task, i.e., unstructured queries over structured docs, while some papers have addressed using structured or semistructured queries on unstructured docs. Here we take a unified approach.

Our basic retrieval model is the well known Kullback-Leibler divergence, with backoff smoothing. In this paper we show how it can be extended to model and support structured/unstructured queries on structured/ unstructured documents. We make a very general assumption on the type of structure imposed on queries and/or documents, suitable for describing various structured data. We also study how the extended model can be efficiently computed.

We finally report on our experiments at INEX 2006, in which we used a rough approximation of the presented model. A full implementation of the model and a more significant evaluation of its retrieval effectiveness are left for future work.

2 Information-theoretic retrieval model

Given a query Q and a document D, the score of D for Q is given by the negative of the Kullback-Leibler (KL) divergence of the query language model θ_Q from the document language model θ_D:

$$score(Q,D) = -KL(\theta_Q|\theta_D) = -\sum_{w \in V} p(w|\theta_Q) log \frac{p(w|\theta_Q)}{p(w|\theta_D)} \tag{1}$$

This is a well known technique for ranking documents [4]. In order to compute expression 1, we need to estimate $p(w|\theta_Q)$ and $p(w|\theta_D)$, i.e., the probability of the word w given the query language model and the document language model, respectively.

Usually this is done by using two probability distributions, one for "seen" words that occur in the text (query or document), and one for "unseen" words that do not occur in the text. This "smoothing" is due to the fact that a given text is usually too small a sample to accurately estimate the language model. One classical smoothing technique is backoff ([3]). It is based on discounting the probabilities of the seen terms, while the probability mass recuperated in this way is redistributed over the unseen terms. Usually, the probability of seen words is given by the maximum likelihood estimate applied to the text, and the probability of unseen words is estimated from the whole document collection in the same manner.

Let $c(w,T)$ be the number of times the word w occurs in text T, $c(w,C)$ be the number of times the word w occurs in the collection C, $|T|$ the number of words in T, $|C|$ the number of words in C. The probability of the word w given the text language model is given by:

$$p(w|\theta_T) = \begin{cases} \psi \, \frac{c(w,T)}{|T|} \text{ if } w \in T \\ \xi \, p(w|C) \text{ if } w \notin T \end{cases} \tag{2}$$

This smoothing technique is very popular in the speech recogniton field and it has also been used for text retrieval ([1], [6]).

3 Structured information-theoretic retrieval model

If the collection of documents is structured, the basic information retrieval model is not satisfactory because it ignores the relationships between the documents. For instance, in order to retrieve elements (components) from XML documents it is natural to exploit the tree-based structuring of documents to enrich each element's description with the description of related elements ([2], [5]).

We assume that there is a partial ordering relation ($\leq$) over the set of documents. For each document D, let D^* be the set formed by the words that are contained in any of the documents that are implied by D according to such a relation, except for D itself; i.e., $D^* = \{w \mid w \in D_i, D \leq D_i, D \neq D_i\}$.

We smooth the original document model by two probability distributions. The first, estimated from D^*, gives importance to the terms that are logically related to D. The second, estimated from the document collection, gives non-zero probabilities to the terms that are neither in the document nor in its implied documents.

$$p(w|\theta_D) = \begin{cases} \alpha \, \frac{c(w,D)}{|D|} \text{ if } w \in D \\ \beta \, \frac{c(w,D^*)}{|D^*|} \text{ if } w \in D^* \\ \mu \, p(w|C) \text{ if } w \notin D \cup D^* \end{cases} \tag{3}$$

In order to ensure that probabilities of all terms sum to 1, the following relation must hold:

$$\alpha \;+\; \beta \;+ \sum_{w \;\notin D \cup D^*} \mu\, p(w|C) = 1 \tag{4}$$

The same approach can be also used to estimate the query language model. A query with an explicit structure, e.g. with a title, a description, and a narrative field, is usually considered as a bag of words. However, it may be not convenient to consider all the fields as equally important because some fields may just contain verbose descriptions of other, shorter fields, and thus the longer fields should be given a smaller weight.

By analogy with structured documents, we can smooth the original query model $p(w|Q)$, as determined by the query title, by two probability distribution, one estimated from the complementary query representation given by the union of description and narrative (denoted by Q^*), one estimated from the whole collection.

$$p(w|\theta_Q) = \begin{cases} \gamma\, \frac{c(w,Q)}{|Q|} & \text{if } w \in Q \\ \delta\, \frac{c(w,Q^*)}{|Q^*|} & \text{if } w \in Q^* \\ \pi\, p(w|C) & \text{if } w \notin Q \cup Q^* \end{cases} \tag{5}$$

The constraint on the sum of probability is in this case given by:

$$\gamma \;+\; \delta \;+ \sum_{w \;\notin Q \cup Q^*} \pi\, p(w|C) = 1 \tag{6}$$

Thus, in all we have six parameters (i.e., $\alpha, \beta, \mu, \gamma, \delta, \pi$) and two equations (i.e., expressions 4 and 6). In the full paper we will show how to estimate the parameters in a more compact and elegant way. We will also show that the resulting model can be computed efficiently because it does not require to compute the probabilities of all terms in the collection for each document.

4 Experiments at INEX 2006

Due to tight scheduling and limited resources, we did not have time to experiment with the full model. In our experiments we used a rough approximation of it.

We used a plain search engine to select for each topic a set of relevant documents from the Wikipedia collection. We then performed an element level analysis for each retrieved document to choose the best element(s) according to the KL divergence. The first stage amounts to performing a fast discriminative selection of candidate results using a restricted set of features (i.e., full documents instead of single elements, no information about document structure, query titles only). In the second stage, the full set of features is brought to bear to perform fine selection/reordering of the results retrieved in the first stage. More details on our experiments will be given in the full paper.

The retrieval performance of this approach was of course in the low part of INEX 2006 ranking. However, given its simplicity and its very limited computational requirements, the results are quite interesting. They can be used as a baseline for the full model.

References

1. Carpineto, C., De Mori, R., Romano, G., Bigi, B.: An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.
2. Fuhr, N., GrossJohann, K.: XIRQL: A Query Language for Information Retrieval in XML Documents: In *Proceedings of SIGIR 2001*, pages 172–180, New Orleans, LA, USA,, 2001.
3. Katz, S.: Estimation of probabilities from sparses data for language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process.* , 35:400–401, 1987.
4. Lafferty. J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research, Development in Information Retrieval*, pages 111–119, New Orleans, LA, USA, 2001.
5. Ogilvie, P., Callan, J.: Language Models and Structured Document Retrieval: In *Proceedings of the INEX 2002 Worksop*, pages 33–40, Schloss Dagsthul, Germany, 2002.
6. Zhai, C, Lafferty. J.: A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.