Query Expansion for the Language Modelling Framework Using the Naïve Bayes Assumption

Laurence A.F. Park and Kotagiri Ramamohanarao

Department of Computer Science and Software Engineering,
The University of Melbourne, 3010, Australia
{lapark,rao}@csse.unimelb.edu.au
http://www.csse.unimelb.edu.au

Abstract. Language modelling is new form of information retrieval that is rapidly becoming the preferred choice over probabilistic and vector space models, due to the intuitiveness of the model formulation and its effectiveness. The language model assumes that all terms are independent, therefore the majority of the documents returned to the ser will be those that contain the query terms. By making this assumption, related documents that do not contain the query terms will never be found, unless the related terms are introduced into the query using a query expansion technique. Unfortunately, recent attempts at performing a query expansion using a language model have not been in-line with the language model, being complex and not intuitive to the user. In this article, we introduce a simple method of query expansion using the naïve Bayes assumption, that is in-line with the language model since it is derived from the language model. We show how to derive the query expansion term relationships using probabilistic latent semantic analysis (PLSA). Through experimentation, we show that using PLSA query expansion within the language model framework, we can provide a significant increase in precision.

Keywords: query expansion, language model, naïve Bayes.

1 Introduction

Many information retrieval systems make use of the assumption that each term is independent of each other in order to achieve fast query times and small storage. Unfortunately, this assumption also reduces the quality of retrieval. By using the assumption that every term is independent of each other term, we cause the retrieval system to disregard the term relationships. This implies that only those documents that contain the query terms will be retrieved, even if other documents are relevant to the query. Therefore, by making the term independence assumption, we are placing a large importance on the process of user query term selection. If the wrong terms are chosen for the query, the wrong documents will be retrieved.

An effective method of introducing term relationships to these fast retrieval systems is to modify the query before a search takes place. Query expansion is

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 681–688, 2008.
© Springer-Verlag Berlin Heidelberg 2008

a method of introducing related terms into the users query, in order to retrieve documents containing the related terms that would have otherwise not been found. Query expansion was first introduced to the vector space model [1], and later applied to the probabilistic model of information retrieval [2] due to its simplicity and effectiveness.

Language models for information retrieval [3] are a new method of information retrieval that have found recent attention due to the intuitiveness of their formulation. To date, there have been several attempts at applying query expansion to language models, but they have only caused an increase in the language model complexity.

In this article, we introduce a new form of query expansion that is derived from within the language modelling framework. We show how to use the query expansion and also how we can generate the term relationships to use for the expansion. This article makes the following contributions:

- a method of query expansion for language models, using the naïve Bayes assumption, that fits the language modelling framework (section 3.1)
- the application of PLSA term-term probabilities for query expansion in language models (section 3.2)
- an introduction to query term compensation during query expansion (section 4)

The article will proceed as follows: Section 2 will provide a brief description of language models and their use for information retrieval, section 3 describes a simple new method of query expansion for language models and provides methods of computing term-term probabilities to use within it. Section 4 shows a problem that is inherit in deriving probabilistic term relationships and provides simple methods to overcome it. Finally, section 5 contains the experimental performance of the language model query expansion and a discussion of the results.

2 Language Models for Information Retrieval

Rather than computing the relevance of a document when given a query, the language modelling approach to information retrieval is to compute the probability of a query being generated from a given document model. By assuming term independence, we are able to decompose the language modelling method into the product of query term probabilities:

$$P(Q|M_d) = \prod_{t_i \in Q} P(t_i|M_d)$$

The value $P(t_i|M_d)$ is the probability of generating the term t_i using document model M_d, therefore it is a measure of the similarity of term t_i to document d. The language modelling approach stats that every document is generated using a document model. The text within document d is sampled from the document model, based on the probability distributions within the model. Therefore, for

us to provide $P(t_i|M_d)$, we must estimate the distribution of the terms in the document model M_d. To do so, must must use the term frequency values within the document collection; the only evidence that we have of the term distributions within the document model.

By simply using the term frequencies ($f_{d,t}$) to compute $P(t_i|M_d)$, we limit our probability estimations to the sampled terms within the document and we also assign a zero probability to those terms that were not sampled from the document model. This constraint is not valid, since there is a chance that there are many terms that are found in to document model M_d, but not found in this particular sample. To obtain a more global term probability, we could observe the frequency of the term in the document collection; this value provides us with a measure of the rarity of the term, but is not specific to the document. Therefore, to obtain a better approximation of the term distributions within the document model, a mixture of the document term frequency and the collection term frequency is used to compute $P(t_i|M_d)$:

$$P(t_i|M_d) = \lambda P(t_i|d) + (1 - \lambda)P(t_i|C) \quad (1)$$

where $\lambda \in [0, 1]$ is the smoothing parameter, $P(t_i|d)$ is the probability of choosing term t_i from document d.

3 Query Expansion within Language Models

The language modelling framework provides us with a method of computing the probability of a document generating a query, even if the query terms do not exist within the document. We showed in the previous section that this is possible by observing the global document collection term probability as well as the local document specific term probability.

Unfortunately, this method of term probability computation does not take into account the relationship of the term to any other term in the document set. The probabilities are computed based only on the frequency of the term itself. By ignoring term relationships, the language modelling approach will provide high probability to those queries who's terms appear in the given document and low probability to queries who's terms do not appear in the given document, regardless of the content of the document. Therefore the document retrieval process requires the user to use the *right* query terms, even though the user is likely to be unfamiliar to the requested information. As a simple example, a search for *corn* will retrieve documents containing the term corn, but not the equally relevant documents containing the word *maize*.

In order to retrieve documents containing related term, we must be able to:

1. use the term relationships as a query expansion within the retrieval process, and
2. identify the term relationships to use as a query expansion

There have been attempts to include query expansion in the language modelling retrieval process [4,5], but they greatly increase the complexity of the model and hence negated the simplicity that makes the language modelling method desirable.

In this section we will deduce a simple method of including a query expansion within the language modelling framework by applying the naïve Bayes assumption, and we will explore two methods of computing the term relationships from the document collection.

3.1 Query Expansion Using Naïve Bayes

In order to use term relationships within the language modelling framework, we must be able to derive a model that reflects the simplicity of a language model. A query expansion process computes the set of terms that are related to the query and then uses those terms to perform the retrieval. Put into the language model framework, we compute the probability of generating the query, given the expansion terms and the document model.

To compute the set of term probabilistic relationships, we will use the document set statistics. If we choose to use the joint probability values, we would over fit our term relationships to the document set. Therefore, to generalise the relationship modelling and hence remove the over fitting, we use naïve Bayes modelling to remove the dependence of the terms on the set of documents.

To obtain the probability of generating term t_i, given term t_j and document model M_d, we use the following equation:

$$\begin{aligned} P(t_i|M_d) &= \sum_{t_j \in T} P(t_i, t_j|M_d) \\ &= \sum_{t_j \in T} P(t_i|t_j, M_d)P(t_j|M_d) \\ &= \sum_{t_j \in T} P(t_i|t_j)P(t_j|M_d) \end{aligned} \tag{2}$$

where $P(t_i|t_j, M_d) = P(t_i|t_j)$, using the naïve Bayes assumption that t_i and M_d are conditionally independent given t_j, and T is the set of unique terms. Using this equation, we can compute the probability of document model M_d generating term t_i from the probability of document model M_d generating term t_j and the probability of term t_i given term t_j.

Equation 2 provides us with a query expansion method for language models, where $P(t_i|t_j)$ is used to compute the relationship of each term to the query term and hence the query expansion, and $P(t_j|M_d)$ is the language model term probability shown in equation 1, which is used to compute the probability of generating the expansion terms give the document model.

Note that although we use Dirichlet smoothing throughout this article, the above derived query expansion within the language modelling framework can be used with any smoothing method.

3.2 Computing the Query Expansion

Now that we have set up a general framework for query expansion within the language modelling method of information retrieval, we will examine methods of

computing the term relationships that are needed in order to perform the query expansion. In this section, we present two forms of query expansion; the first is based on the probabilities produced using language models, and the second is based on the probabilities produced using probabilistic latent semantic analysis.

Probabilistic latent semantic based query expansion. Probabilistic latent semantic analysis (PLSA) [6] is a probabilistic method of discovering hidden topics withing a document collection using maximum likelihood. Given the estimated probability of document d_i and term t_j as:

$$\hat{P}(d_i, t_j) = \frac{f_{d_i,t_j}}{\sum_{d \in D} \sum_{t \in T} f_{d,t}}$$

we want to compute the actual probability of a term and a document, given the model:

$$P(d, t) = \sum_z P(d|z)P(z)P(t|z)$$

where $P(d|z)$ and $P(t|z)$ are the probability of document d given topic z and the probability of term t given topic z respectively, and P(z) is the probability of topic z.

It was recently shown that PLSA information can be used effectively as a query expansion by observing only the $P(t|z)$ and $P(z)$ values [7]. We can show:

$$\begin{aligned} P(t_i|t_j) &= \sum_{z \in Z} P(t_i, z|t_j) \\ &= \sum_{z \in Z} P(t_i|z, t_j)P(z|t_j) \\ &= \sum_{z \in Z} P(t_i|z)P(z|t_j) \\ &= \sum_{z \in Z} P(t_i|z)P(t_j|z)P(z)/P(t_j) \\ &= \frac{\sum_{z \in Z} P(t_i|z)P(t_j|z)P(z)}{\sum_{z \in Z} P(t_j|z)P(z)} \end{aligned} \tag{3}$$

where $P(t|z)$ and $P(z)$ are computed using PLSA, and $P(t_i|z, t_j) = P(t_i|z)$ using the naïve Bayes assumption that term t_i and term t_j being conditionally independent given topic z.

4 Query Term Compensation

The set of probabilities of terms T are disjoint when given term t_j This can be seen by the property that:

$$\sum_i P(t_i|t_j) = 1$$

Table 1. PLSA query expansion within the language model framework, using PLSA add compensation with various values for the compensation factor α on the Associated Press document collection. The baseline measure (language model without query expansion) provides a MAP of 0.2749. The * and ** shows a statistically significant change in MAP at the 0.1 and 0.05 levels, compared to the language model without query expansion.

Compensation (α)	0	0.1	0.3	0.5	0.7	0.9	1
MAP	0.0669**	0.2715	0.2797	0.2803**	0.2793**	0.2788**	0.2788**

Given that $P(t_i|t_j) > 0$, we find that $P(t_i|t_j) < 1$ for all i and j, including the case where $i = j$. From this we can see that the probability of a term given itself is less than one. We may also find that the $P(t_i|t_j)$ where $i \neq j$ is greater than $P(t_i|t_i)$, implying that other terms are more related to the term than the term itself.

The effect of a term having a low probability given itself, may cause problems during a query expansion. We may find that other terms introduced from the expansion have a higher probability than the original query terms. Therefore the query terms may become lost in the expansion.

To compensate for this reduction is query term probability, we have explored the method of adding 1 to the computed probability of a term given itself. This compensation is as though we have included the original query in the query expansion, where the add method adds the expansion probability of the query terms in the expansion to the query terms.

Therefore, using the PLSA-based query expansion, we provide the following methods of compensation for the conditional probabilities:

$$\text{PLSA add:} \quad P(t_i|t_j) = \begin{cases} \frac{\sum_d P(t_i|M_d)P(t_j|M_d)}{\sum_d P(t_j|M_d)} & \text{if } i \neq j \\ \frac{\sum_d P(t_i|M_d)P(t_j|M_d)}{\sum_d P(t_j|M_d)} + \alpha & \text{if } i = j \end{cases}$$

where α is the compensation factor, and the probability for $i \neq j$ is taken from our derivation earlier in equation 3.

5 Experiments

Our set of experiments examines PLSA query expansion using add query compensation within the language model framework on a collection of 84,678 documents from the associated press found in TREC disk 1[1]. Experiments were performed using the values 0, 0.1, 0.3, 0.5, 0.7, 0.9 and 1 for the compensation factor (α). The results are shown in table 1.

We can see from the results that the MAP peaks at $\alpha = 0.5$ and that the results are statistically significant at the 0.05 level for larger values of α. We can also see that the result for $\alpha = 0$ is very poor. Using the add query compensation,

[1] http://trec.nist.gov

where $\alpha = 0$ is equivalent to using no query compensation, so we can see that it is essential to use query compensation on large and small document sets.

The significant increase in MAP shows that using PLSA query expansion with query compensation is a useful addition when used within the language model framework.

6 Conclusion

Within the field of information retrieval, language models have shown to be competitive with other models of retrieval, while offering an intuitive and simple formulation. To simplify the model, language models include the assumption that all terms are independent. This assumption places great importance on the user's choice of query terms. To introduce term relationships into the language modelling framework others have applied query expansion, but the complexity of the expansion removed the simplicity from the language model formulation.

In this article, we introduced a method of query expansion for language models which uses the naïve Bayes assumption to produce generalised probabilistic term relationships. To compute the term relationships, we examined a probabilistic latent semantic analysis (PLSA) method. Experiments on a document collection showed us that the the PLSA query expansion within the language modelling framework provided a significant increase in precision over the language model with no expansion. Therefore the PLSA query expansion was also effective for larger document sets.

References

1. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using SMART: TREC 3. In: Harman, D. (ed.) The Third Text REtrieval Conference (TREC-3), Gaithersburg, Md. 20899, National Institute of Standards and Technology Special Publication 500-226, pp. 69–80 (1994)
2. Robertson, S.E., Walker, S.: Okapi/keenbow at TREC-8. In: Voorhees, E.M., Harman, D.K. (eds.) The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Md. 20899, National Institute of Standards and Technology Special Publication 500-246, Department of Commerce, National Institute of Standards and Technology, pp. 151–162 (1999)
3. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR 1998: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275–281. ACM Press, New York (1998)
4. Cao, G., Nie, J.Y., Bai, J.: Integrating word relationships into language models. In: SIGIR 2005: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 298–305. ACM Press, New York (2005)

5. Bai, J., Song, D., Bruza, P., Nie, J.Y., Cao, G.: Query expansion using term relationships in language models for information retrieval. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 688–695. ACM Press, New York (2005)
6. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57. ACM Press, New York (1999)
7. Park, L.A.F., Ramamohanarao, K.: Query expansion using a collection dependent probabilistic latent semantic thesaurus. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 224–235. Springer, Heidelberg (2007)