Constructing Better Document and Query Models with Markov Chains

Guihong Cao Jian-Yun Nie Jing Bai
Dept. IRO, University of Montreal
C.P. 6128, succursale Centre-ville, Montreal, Quebec, H3C 3J7 Canada
caogui@iro.umontreal.ca nie@iro.umontreal.ca baijing@iro.umontreal.ca

ABSTRACT
Document and query expansions have been used separately in previous studies to enhance the representation of documents and queries. In this paper, we propose a general method that integrates both of them. Expansion is carried out using multi-stage Markov chains. Our experiments show that this method significantly outperforms the existing approaches.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms
Algorithms, Theory, Experimentation, Performance

Keywords
Information Retrieval, Language Model, Markov Chain, Expansion

1. INTRODUCTION
Statistical language modeling (LM) has been widely used in information retrieval (IR) in recent years [2, 7]. Several attempts have been made to improve either document model or query model [6, 8] by smoothing. However, these approaches still suffer from the underlying assumption of term independence, which implies that a term of a query is independent from a different term in a document. It is obvious that this assumption is not true. For instance, if "computer" appears in a query and "programming" in a document, the terms are not independent; so are the query and the document. Several studies have been conducted to relax the independence assumption by integrating the relationships between query terms and document terms [1, 2, 4]. In doing this, we are indeed making inferences according to term relationships.

In the previous studies, inference has been implemented either as document expansion or query expansion, and inference has been limited to using direct term relationships. We argue that these limitations are unnecessary. By performing both expansions and by allowing indirect inferences, we can gain larger inference capabilities. In this paper, we thus propose a general model to achieve this goal.

2. GENERAL MODEL
Traditional LM approaches to IR try to determine a relevance score according to the following formula:

$$Score(Q,D) = \sum_{w_i \in Q} P(w_i \mid Q) \log P(w_i \mid D) \qquad (1)$$

where $P(w_i \mid Q)$ is the probability of a term in the query model, usually estimated by Maximum likelihood; and $P(w_i \mid D)$ is the probability in the document model, which is usually smoothed with the collection model. These models cannot represent well the document and query contents due to the assumption of term independence. In order to create better models, we have to incorporate term relationships into the models, i.e. our desired models should contain not only the terms appearing in the document or the query, but also those that are closely related. We will create such models thought document and query expansion.

Let $\hat{P}(w_i \mid D)$ and $\hat{P}(w_i \mid Q)$ be the expanded document model and query model. Then the documents are ranked by the following cross entropy:

$$Score(Q,D) = \sum_{w_i \in V} \hat{P}(w_i \mid Q) \log \hat{P}(w_i \mid D)$$

where w_i is a word of vocabulary V. In practice, query expansion should be limited to a relatively small number (e.g. 80 in our case) of terms because of retrieval efficiency. Let E be the set of expansion terms selected, then the above equation can be simplified to:

$$Score(Q,D) = \sum_{w_i \in E \cup Q} \hat{P}(w_i \mid Q) \log \hat{P}(w_i \mid D) \qquad (2)$$

The two models in Equation (2) are defined using Markov chains (MC).

3. DEFINING MODELS USING MARKOV CHAIN
Basically, a MC is defined by two probabilities [3]: the initial probability to select a state, and the transition probability from one state to another. Accordingly, for query model, term generation in the final model is done according to the following steps:

Step 0: An initial query term w is generated according to $P_0(w \mid Q)$;

Step 1: Given the word w_j is selected at step 0, a new word w_i is generated in two different ways: it can be generated from w_j (because they are related) at probability $(1-\gamma)$, or added according to $P_0(w_i \mid Q)$ (i.e., reset to step 0) at probability γ. The generation of w_i from the related term w_j is made according to the transition probability $P(w_i \mid w_j)$. This latter is determined as in [4] according to different types of relation between *w'* and *w, i.e.* co-occurrence relations and relations in WordNet. Details are described in [4]. So the probability of generating w_i at step 1 according to both cases is:

$$P_1(w_i \mid Q) = P_1(w_i \mid w_j, Q) = \gamma P_0(w_i \mid Q) + (1-\gamma) P(w_i \mid w_j) \qquad (3)$$

Step *t*: the query model at step *t* is defined as:

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011..

$$P_t(w \mid Q) = \sum_{w' \in V} P_R(w \mid w') P_{t-1}(w \mid Q) \quad (4)$$

A stationary distribution $\pi(w \mid Q)$ [3] can be reached after a number of transitions, and this corresponds to:

$$\pi(w \mid Q) = \lim_{T \to +\infty} \hat{P}_T(w \mid Q) = \gamma \sum_{t=0}^{\infty} (1-\gamma)^t P_t(w \mid Q) \quad (5)$$

The stationary distribution $\pi(w \mid Q)$ is the best statistical model that we can construct from the information available, i.e. Q and terms relations. So, we define the final query model $\hat{P}(w \mid Q)$ as $\pi(w \mid Q)$. In the same way, the document model $\hat{P}(w \mid D)$ is also defined as the stationary probability $\pi(w \mid D)$, which is determined similarly to equation (5).

The initial distributions are now determined as follows:

$$P_0(w \mid Q) = \lambda P_{ml}(w \mid Q) + (1-\lambda) P(w \mid F);$$

$$P_0(w \mid D) = \frac{\max(c(w;D) - \delta, 0)}{|D|} + \frac{\delta |D|_u}{|D|} P_{ml}(w \mid C) \quad (6)$$

where $P(w|F)$ and $P_{ml}(w|C)$ are respectively the feedback model [8] and the collection model. λ is set to be 0.5 according to [8] and δ is set experimentally to 0.7 according to a training collection. The coefficient γ is set at 0.3, as our experiments show that the retrieval effectiveness is relatively insensitive to γ when $\gamma \in [0.1,1]$.

4. EXPERIMENTS

We use three TREC collections in our experiments: AP88-90, WSJ and SJM on TREC disks 2-3. The parameters are determined using queries 101-150 and 201-250 (title+description), and topics 51-100 are used for testing.

Table 1 shows a comparison between Unigram model (**UM**), document and query expansion models alone (**MC-DE** and **MC-QE**) and our general model (**GM**). From this table, we see that both **MC-DE** and **MC-QE** outperform **UM** to some degree. The improvement scales of **MC-QE** are much higher than **MC-DE**, because **MC-QE** has incorporated a feedback model, which has proven to be very helpful to capture the information need [8]. The most important observation for this study is that the general model **GM** is better than using **MC-DE** or **MC-QE** separately. Although the improvements of **GM** over **MC-QE** are not very large, they are consistent on all the collections. This result shows that the combination enables the system to benefit from both approaches.

In order to see the impact of multi-step expansion, we use different numbers of transitions in query expansion in Figure 1. We observe that when we increase the transition steps, the effectiveness is also improved, in particular, between 1 and 5 steps. These increases are directly attributed to the increased steps of transition. They show that multi-step transition is superior to one-step transition.

5. RELATED WORK

Query expansion and document expansion have been used in a number of previous studies, either with LM or vector space model [1, 4, 6, 8]. The experiments have produced encouraging results. In comparison with the previous attempts, in this paper, we propose to extend the expansion process further on the two following aspects: Using MC to enable multi-step inference; Using a general model that combines both document expansion and query expansion. Our experiments show that each of the above extensions has resulted in some improvements in retrieval effectiveness.

Table 1: Performance of General Model

Coll.		**UM**	**MC-DE**	%UM	**MC-QE**	**GM**	%UM	%QE
AP	MAP	0.1925	0.2138	+11.06**	0.2580	**0.2629**	+22.96**	+2.02
	Ret.	3289	3530		3994	**4064**		
WSJ	MAP.	0.2466	0.2590	+5.02*	0.2860	**0.2891**	+11.62**	+1.08
	Ret.	1659	1704		1794	**1845**		
SJM	MAP.	0.2045	0.2155	+5.37	0.2522	**0.2584**	+19.91**	+2.46
	Ret.	1417	1572		1621	**1742**		

* and ** mean statistical significance at level of p<0.05 and p<0.01.

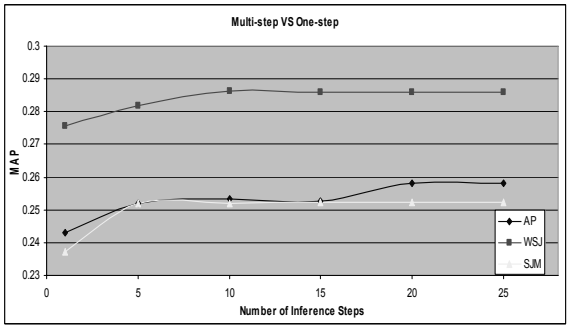

Figure 1: One-step QE Vs Multi-step MC-QE

MC has also been used in some previous studies in IR [5]. In comparison, our method is different on several aspects: first, we do not use many heuristics as in [5]. Second, [5] does not define an explicit query model, while we did. So our model provides a clearer and more principled formulation than [5].

6. CONCLUSION

Many previous studies on LM assumed independence between terms. In this paper, we proposed a general approach that integrates term relationships in both document and query models. The resulting models can better represent the document and the query. Our experiments also confirm that they lead to better retrieval effectiveness.

REFERENCES

[1] Bai, J., Song, D., Bruza, P., Nie, J.-Y. and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. *CIKM,* pp. 688-695.

[2] Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. *SIGIR*, pp. 222-229.

[3] Brémaud, P. (1999) *Markov chains: Gibbs fields, monte carlo simulations, and queues.* Springer-Verlag.

[4] Cao, G., Nie, J.-Y., and Bai, J. (2005). Integrating word relationships into language modeling. *SIGIR*, pp. 298-305

[5] Collins-Thompson, K, and Callan, J. (2005). Query expansion using random walk models. *CIKM,* pp.704-711

[6] Lavrenko, V. and Croft, W.B. Relevance-based language models. *SIGIR,* pp. 120-127.

[7] Ponte, J. and Croft, W.B. (1998). A language modeling approach to information retrieval. *SIGIR,* pp. 275-281.

[8] Zhai, C. and Lafferty, J. (2001b). Model-based feedback in the language modeling approach to information retrieval. *CIKM,* pp. 403-410.