

Boosting Relevance Model Performance with Query Term Dependence

Koji Eguchi
National Institute of Informatics
Tokyo 101-8430, Japan
eguchi@nii.ac.jp

W. Bruce Croft
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
croft@cs.umass.edu

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval — *Query formulation*

General Terms: Performance, Experimentation

Keywords: information retrieval, term dependence, language models

1. INTRODUCTION

Phrase-based queries are known to perform effectively, especially against large-scale and noisy text data such as typically appear on the Web [9, 7, 2]. These methods can capture term dependencies that appear in a query. Meanwhile, a relevance model is a probability distribution over terms that is estimated from the set of relevant documents to address important notions of synonymy and polysemy [5]. The relevance model is usually estimated with unigrams from the results of pseudo-relevance feedback, i.e., from the top-ranked documents retrieved in response to a query. The combination of phrase-based query structuring and query expansion via the relevance model is promising, because each model has its own advantages: phrase-based query structuring attempts to address problems with *bag-of-words* representations and the relevance model attempts to address problems with simple word matching. In this paper, we investigate how query structuring with term dependence can improve the performance of query expansion via a relevance model. Some preliminary results can be found in the investigations of our research group [8, 1]. In this paper, we attempt to examine the above point thoroughly.

2. TERM DEPENDENCE MODELS

Metzler and Croft developed a general, formal framework for modeling term dependencies via Markov random fields [7], and showed that the model is very effective in a variety of retrieval situations using the Indri platform [6]. Markov random fields (MRFs) are commonly used to model joint distributions succinctly. In [7], the joint distribution $P_\Lambda(Q, D)$ over queries Q and documents D , parameterized by Λ , was modeled using MRFs, and for ranking purposes the posterior $P_\Lambda(D|Q)$ was derived by the following ranking function, assuming a graph G that consists of a document node and query term nodes: $P_\Lambda(D|Q) \stackrel{\text{rank}}{=} \sum_{c \in C(G)} \lambda_c f(c)$, where $Q = t_1 \dots t_n$, $C(G)$ is the set of cliques in an MRF graph G , $f(c)$ is some real-valued feature function over clique values, and λ_c is the weight given to that particular feature function. Three variants of the MRF model were assumed: (1)*full-independence* variant assumes that query terms are independent of each other; (2)*sequential dependence* variant assumes dependence between neighboring query terms; and

(3)*full-dependence* variant assumes that all query terms are in some way dependent on each other. To express these assumptions, the following ranking function was derived:

$$P_\Lambda(D|Q) \stackrel{\text{rank}}{=} \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in O \cup U} \lambda_U f_U(c) \quad (1)$$

where T is defined as the set of 2-cliques involving a query term and a document D , O is the set of cliques containing the document node and two or more query terms that appear contiguously within the query, and U is the set of cliques containing the document node and two or more query terms appearing noncontiguously within the query.

We further developed a *two-stage term dependence model* that captures long- and short-range dependencies differently [2], which can be considered to be generalized from Metzler and Croft's model. We assumed that: (i)global dependencies between query components are explicitly separated in a query; and (ii)local dependencies occur between constituents within a compound word when the compound word appears in a query component. The phrase-based query structuring used in this paper is based on this model. To achieve the model mentioned above, we extended Eq. (1), as follows [2]:

$$\begin{aligned} P_\Lambda(D|Q) \stackrel{\text{rank}}{=} & \sum_{c_q \in T(Q)} \lambda_T f_T(c_q) + \sum_{c_q \in O(Q)} \lambda_O f_O(c_q) \\ & + \sum_{c_q \in O(Q) \cup U(Q)} \lambda_U f_U(c_q) \quad (2) \\ f_T(c_q) = & f'_T \left(\sum_{q_k \in c_q} \sum_{c_t \in T(q_k)} \mu_T g_T(c_t) \right) \\ f_O(c_q) = & f'_O \left(\sum_{q_k \in c_q} \sum_{c_t \in O(q_k)} \mu_O g_O(c_t) \right) \\ f_U(c_q) = & f'_U \left(\sum_{q_k \in c_q} \sum_{c_t \in O(q_k) \cup U(q_k)} \mu_U g_U(c_t) \right) . \quad (3) \end{aligned}$$

Here, Q consists of query components $q_1 \dots q_k \dots q_m$, and each query component consists of individual terms $t_1 \dots t_n$. $T(Q)$, $O(Q)$ and $U(Q)$ can be defined in the same manner as in Eq. (1) with the query components consisting of a whole query, while $T(q_k)$, $O(q_k)$ and $U(q_k)$ are defined with the individual terms consisting of a query component. The feature functions f'_T , f'_O and f'_U and another feature functions g_T , g_O and g_U can be given in the same manner as f_T , f_O and f_U that were defined in Eq. (1), respectively. Hereafter, we assumed that the constraint $\lambda_T + \lambda_O + \lambda_U = 1$ was imposed independently of the query, and we assumed μ_T , μ_O and μ_U were determined by $1/\#(c_t, c_q)$ for the functions f'_T , f'_O and f'_U , respectively, where $\#(c_t, c_q)$ represents the number of times the clique of terms c_t was observed in the clique of query components c_q . According to Eq. (2), we defined *glfd+* model as an instance of the two-stage term dependence

model, considering special features of Japanese [1, 2]. The *glfd+* model expresses the dependencies between query components on the basis of the *full dependence*. It also expresses the dependencies between constituent words within a query component on the basis of the *sequential dependence*. The two-stage term dependence model should be reasonable for other languages, if query components can be specified in a query, for example “ozone hole, human body”.

3. RELEVANCE MODELS

Lavrenko and Croft formulated relevance models that explicitly incorporated relevance into the language modeling [5]. Following [8], we use the relevance models as a pseudo-relevance feedback function in the framework of inference network-based retrieval models, as briefly described in the following. Given an initial query Q_{orig} , we retrieve a set of $\#docs_{fb}$ documents and form a relevance model from them. We then form Q_{rm} by wrapping the $\#combine$ operator of Indri [6], around the most likely $\#terms_{fb}$ terms from the relevance model that are not stopwords. Finally, an expanded query is formed that has the following form: $\#weight(\nu Q_{orig} (1.0 - \nu) Q_{rm})$, where $\#weight$ indicates an Indri operator [6]. In this paper, we formulate Q_{orig} using the two-stage term dependence model described in Section 2, instead of the usual *full-independence* (*fi*) model.

4. EXPERIMENTS

We used a 100-gigabyte web document collection, *NW100G-01*, which consisted of web documents gathered from the .jp domain and thus were mostly written in Japanese. The *NW100G-01* was used in a series of *NTCIR WEB task* [4, 3, 10]. In the experiments described in the following, we only used the terms specified in the title field of each topic statement. All the topics were written in Japanese, and so we performed morphological analysis using the MeCab tool¹ to segment the title portion of the topic and to add POS tags. The definition of the title is different from the one used in a series of TREC,² for instance, in the following ways: (i) the title field gives 1–3 components that were delimited by commas; and (ii) each of these components is supposed to indicate a certain concept, and so it sometimes consists of a single word, but it may also consist of a compound word [4].

For training, we experimented using the *glfd+* model with relevance model, as described in Section 3, over the relevance judgment data used in the NTCIR-3 WEB task [4], changing the weight ν , $\#docs_{fb}$ and $\#terms_{fb}$.³ For testing, we used the relevance judgment data that were used in the NTCIR-5 WEB task [10]. For the relevance model, we used the top-ranked 5 and 10 documents ($\#docs_{fb}$), and 5, 10 and 20 terms ($\#terms_{fb}$) for feedback. We used the optimized value of the weight ν , which we obtained from training, corresponding to each pair of ($\#docs_{fb}$, $\#terms_{fb}$) above. The results are shown in Table 1. In this table, ‘AvgPrec_a’ indicates the mean average precision over all the 35 topics. ‘%chg_f’ and ‘%chg_r’ were calculated on the bases of (i) each model without the relevance model (i.e., *fi* alone or *glfd+* alone); and (ii) the *fi* model alone, respectively. As shown in this table, the *glfd+* model alone worked 13% better than the *fi* model alone, as also reported in [2]. The combination of the *glfd+* model and the relevance model achieves 8–12% improvements over the *glfd+* model alone. Combining with the relevance model, the *glfd+* model worked 9–15% better than the *fi* model under the same conditions of ($\#docs_{fb}$, $\#terms_{fb}$).

¹<http://sourceforge.jp/projects/mecab/>.

²<http://trec.nist.gov/>.

³We also optimized the parameters for the *glfd+* model using the NTCIR-3 WEB test collection, as described in [2].

Table 1: Evaluation results

	ν	AvgPrec _a	%chg _f	(%chg _r)
<i>fi</i> alone	1.0	0.1405	+0.00	(+0.00)
(05, 05)	0.9	0.1577	+12.23*	(+12.23**)
(05, 10)	0.9	0.1565	+11.40*	(+11.40**)
(05, 20)	0.9	0.1553	+10.52*	(+10.52**)
(10, 05)	0.9	0.1583	+12.66*	(+12.66**)
(10, 10)	0.9	0.1575	+12.14*	(+12.14**)
(10, 20)	0.9	0.1565	+11.39*	(+11.39**)
<i>glfd+</i> alone	1.0	0.1588	+0.00	(+13.02**)
(05, 05)	0.7	0.1730	+8.96*	(+23.14**)
(05, 10)	0.7	0.1750	+10.20*	(+24.55**)
(05, 20)	0.5	0.1784	+12.33	(+26.96**)
(10, 05)	0.7	0.1728	+8.80*	(+22.97**)
(10, 10)	0.7	0.1714	+7.95	(+22.01**)
(10, 20)	0.5	0.1762	+11.01	(+25.46**)

In the left column, ‘(.,.)’ indicates ($\#docs_{fb}$, $\#terms_{fb}$). ‘*’ and ‘**’ indicate statistically significant improvements over each model without the relevance model (i.e., *fi* alone or *glfd+* alone) and over the *fi* model alone, respectively, where $p < 0.05$ with the two-sided Wilcoxon signed-rank test.

5. CONCLUSIONS

In this paper, we investigated how query structuring with term dependence could improve the performance of query expansion via a relevance model. We used the two-stage term dependence model to structure queries. We demonstrated through a series of experiments that the combination of the term dependence model and relevance model was more effective than either the term dependence model or the relevance model alone. Using the two-stage term dependence model with the relevance model, we achieved a significant 22–27% gain from the term *independence* model without using the relevance model.

Acknowledgments: We thank Donald Metzler for valuable discussions and comments. This work was supported in part by the Overseas Research Scholars Program and the Grants-in-Aid for Scientific Research (#17680011 and #18650057) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, in part by the Telecommunications Advancement Foundation, Japan, and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

6. REFERENCES

- [1] K. Eguchi. NTCIR-5 query expansion experiments using term dependence models. In *Proc. 5th NTCIR Workshop*, Tokyo, Japan, 2005.
- [2] K. Eguchi and W. B. Croft. Query structuring with two-stage term dependence in the Japanese language. In *Information Retrieval Technology: Third Asia Information Retrieval Symposium, Singapore*, LNCS 4182, pages 522–529. Springer, 2006.
- [3] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa. Overview of the Informational Retrieval Task at NTCIR-4 WEB. In *Proc. 4th NTCIR Workshop*, Tokyo, Japan, 2004.
- [4] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Overview of the Web Retrieval Task at the Third NTCIR Workshop. In *Proc. 3rd NTCIR Workshop*, Tokyo, Japan, 2003.
- [5] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. ACM SIGIR 2001*, pages 120–127, New Orleans, Louisiana, USA, 2001.
- [6] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750, 2004.
- [7] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proc. ACM SIGIR 2005*, pages 472–479, Salvador, Brazil, 2005.
- [8] D. Metzler, T. Strohman, Y. Zhou, and W. B. Croft. Indri at TREC 2005: Terabyte Track. In *Proc. TREC 2005*, 2005.
- [9] G. Mishne and M. de Rijke. Boosting web retrieval through query operations. In *Advances in Information Retrieval: 27th European Conference on IR Research, Santiago de Compostela, Spain*, LNCS 3408, pages 502–516. Springer, 2005.
- [10] M. Yoshioka. Overview of the NTCIR-5 WEB Query Expansion Task. In *Proc. 5th NTCIR Workshop*, Tokyo, Japan, 2005.