CUHK at ImageCLEF 2005: Cross-Language and Cross-Media Image Retrieval

Steven C. H. Hoi, Jianke Zhu, and Michael R. Lyu

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{chhoi, jkzhu, lyu}@cse.cuhk.edu.hk

Abstract. In this paper, we describe our studies of cross-language and cross-media image retrieval at the ImageCLEF 2005. This is the first participation of our CUHK (The Chinese University of Hong Kong) group at ImageCLEF. The task in which we participated is the "bilingual ad hoc retrieval" task. There are three major focuses and contributions in our participation. The first is the empirical evaluation of language models and smoothing strategies for cross-language image retrieval. The second is the evaluation of cross-media image retrieval, i.e., combining text and visual contents for image retrieval. The last is the evaluation of bilingual image retrieval between English and Chinese. We provide an empirical analysis of our experimental results, in which our approach achieves the best mean average precision result in the monolingual query task in the campaign. Finally we summarize our empirical experience and address the future improvement of our work.

1 Introduction

Although content-based image retrieval (CBIR) has received considerable attention in the community [1], there are so far only a few benchmark image datasets available. The CLEF (Cross Language Evaluation Forum) organization began the ImageCLEF campaign from 2003 for benchmark evaluation of cross-language image retrieval [2]. ImageCLEF 2005 offers four different tasks: bilingual ad hoc retrieval, interactive search, medical image retrieval and an automatic image annotation task [2]. This is the first participation of our CUHK (The Chinese University of Hong Kong) group at ImageCLEF. The task in which we participated this year is the "bilingual ad hoc retrieval".

In the past decade, traditional information retrieval has mainly focused on document retrieval problems [3]. Along with the growth of multimedia information retrieval, which has received ever-increasing attention in recent years, cross-language and cross-media retrieval have been put forward as an important research topic in the community [2]. The cross-language image retrieval problem is to tackle the multimodal information retrieval task by unifying the techniques from traditional information retrieval, natural language processing (NLP), and traditional CBIR solutions.

In this participation, we offer our main contributions in three aspects. The first is an empirical evaluation of language models and smoothing strategies for cross-language image retrieval. The second is an evaluation of cross-media image retrieval, i.e., combining text and visual contents for image retrieval. The last is the design and empirical evaluation of a methodology for bilingual image retrieval spanning English and Chinese sources.

The rest of this paper is organized as follows. Section 2 introduces the TF-IDF retrieval model and the language model based retrieval methods. Section 3 describes the details of our implementation for this participation, and outlines our empirical study on the cross-language and cross-media image retrieval. Finally, Section 4 concludes our work.

2 Language Models for Text Based Image Retrieval

In this participation, we have conducted extensive experiments to evaluate the performance of Language Models and the influences of different smoothing strategies. More specifically, two kinds of retrieval models are studied in our experiments: (1) The TF-IDF retrieval model, and (2) The KL-divergence language model based methods. The smoothing strategies for Language Models evaluated in our experiments [4] are: (1) the Jelinek-Mercer (JM) method, (2) Bayesian smoothing with Dirichlet priors (DIR), and (3) Absolute discounting (ABS).

2.1 TF-IDF Similarity Measure for Information Retrieval

We incorporate the TF-IDF similarity measure method into the Language Models (LM) [3]. TF-IDF is widely used in information retrieval, which is a way of weighting the relevance of a query to a document. The main idea of TF-IDF is to represent each document by a vector in the size of the overall vocabulary. Each document D_i is then represented as a vector $(w_{i1}, w_{i2}, \cdots, w_{in})$ if n is the size of the vocabulary. The entry $w_{i,j}$ is calculated as: $w_{ij} = TF_{ij} \times \log(IDF_j)$, where TF_{ij} is the term frequency of the j-th word in the vocabulary in the document D_i, i.e. the total number of occurrences. IDF_j is the inverse document frequency of the j-th term, which is defined as the number of documents over the number of documents that contain the j-th term. The similarity between two documents is then defined as the cosine of the angle between the two vectors.

2.2 Language Modeling for Information Retrieval

Language model, or the statistical language model, employs a probabilistic mechanism to generate text. The earliest serious approach for a statistical language model may be tracked to Claude Shannon [5]. To apply his newly founded information theory to human language applications, Shannon evaluated how well simple n-gram models did at predicting or compressing natural text. In the past, there has been considerable attention paid to using the language modeling techniques for text document retrieval and natural language processing tasks [6].

The KL-divergence Measure. Given two probability mass functions $p(x)$ and $q(x)$, $D(p||q)$, the Kullback-Leibler (KL) divergence (or relative entropy) between p and q is defined as

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} \tag{1}$$

One can show that $D(p||q)$ is always non-negative and is zero if and only if $p = q$. Even though it is not a true distance between distributions (because it is not symmetric and does not satisfy the triangle inequality), it is often still useful to think of the KL-divergence as a "distance" between distributions [7].

The KL-divergence based Retrieval Model. In the language modeling approach, we assume a query q is generated by a generative model $p(q|\theta_Q)$, where θ_Q denotes the parameters of the query unigram language model. Similarly, we assume a document d is generated by a generative model $p(q|\theta_D)$, where θ_Q denotes the parameters of the document unigram language model. Let $\hat{\theta}_Q$ and $\hat{\theta}_D$ be the estimated query and document models, respectively. The relevance of d with respect to q can be measured by the negative KL-divergence function [6]:

$$-D(\hat{\theta}_Q||\hat{\theta}_D) = \sum_w p(w|\hat{\theta}_Q) log p(w|\hat{\theta}_D) + (-\sum_w p(w|\hat{\theta}_Q) log p(w|\hat{\theta}_Q)) \tag{2}$$

In the above formula, the second term on the right-hand side of the formula is a query-dependent constant, i.e., the entropy of the query model $\hat{\theta}_Q$. It can be ignored for the ranking purpose. In general, we consider the smoothing scheme for the estimated document model as follows:

$$p(w|\hat{\theta}_D) = \begin{cases} p_s(w|d) & \text{if word } w \text{ is present} \\ \alpha_d p(w|\mathcal{C}) & \text{otherwise} \end{cases} \tag{3}$$

where $p_s(w|d)$ is the smoothed probability of a word present in the document, $p(w|\mathcal{C})$ is the collection language model, and α_d is a coefficient controlling the probability mass assigned to unseen words, so that all probabilities sum to one [6]. We discuss several smoothing techniques in detail below.

2.3 Several Smoothing Techniques

In the context of language modeling study, the term "smoothing" can be defined as the adjustment of the maximum likelihood estimator of a language model so that it will be more accurate [4]. As we know that a language modeling approach usually estimates $p(w|d)$, a unigram language model based on a given document d, one of the simplest methods for smoothing is based on the maximum likelihood estimate as follows:

$$p_{ml}(w|d) = \frac{c(w;d)}{\sum_w c(w;d)} \tag{4}$$

Unfortunately, the maximum likelihood estimator will often underestimate the probabilities of unseen words in the given document. Hence, it is important to employ smoothing methods that usually discount the probabilities of the words seen in the text and assign the extra probability mass to the unseen words according to some model [4].

Some comprehensive evaluation of smoothing techniques for traditional text retrieval can be found in literature [8, 4]. They have been an important tool to improve the performance of language models in traditional text retrieval. To achieve efficient implementations for large-scale tasks, three representative methods are selected in our scheme, which are popular and relatively efficient. They are discussed in turn below.

The Jelinek-Mercer (JM) Method. This method simply employs a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ to control the influence:

$$p_\lambda(\omega|d) = (1-\lambda)p_{ml}(\omega|d) + \lambda p(\omega|\mathcal{C}) \tag{5}$$

It is a simple mixture model. A more general Jelinek-Mercer method can be found in [9].

Bayesian Smoothing with Dirichlet Priors (DIR). In general, a language model can be considered as a multinomial distribution, in which the conjugate prior for Bayesian analysis is the Dirichlet distribution with parameters [4] $(\mu p(\omega_1|\mathcal{C}), \mu p(\omega_2|\mathcal{C}), \ldots, \mu p(\omega_n|\mathcal{C}))$. Thus, the smoothing model can be given as:

$$p_\mu(\omega|d) = \frac{c(\omega;d) + \mu p(\omega|\mathcal{C})}{\sum_\omega c(\omega;d) + \mu} \tag{6}$$

Note that μ in the above formula is a DIR parameter that is usually estimated empirically from training sets.

Absolute Discounting Smoothing (ABS). The absolute discounting method subtracts a constant from the counts of seen words for reducing the probabilities of the seen words, meanwhile it increases the probabilities of unseen words by including the collection language model. More specifically, the model can be represented as follows:

$$p_\delta(\omega|d) = \frac{\max(c(\omega;d) - \delta, 0)}{\sum_\omega c(\omega;d)} + \sigma p(\omega|\mathcal{C}) \tag{7}$$

where $\delta \in [0,1]$ is a discount constant and $\sigma = \delta|d|_\mu/|d|$, so that all probabilities sum to one. Here $|d|_\mu$ is the number of unique terms in document d, and $|d|$ is the total count of words in the document, i.e., $|d| = \sum_\omega c(\omega;d)$.

Table 1 summarizes the three methods in terms of $p_s(\omega|d)$ and α_d in the general form. In the table, for all three cases, a larger parameter value of λ, μ

Table 1. Summary of three smoothing methods evaluated in our submission.

Method	$p_s(\omega\|d)$	α_d	parameter
JM	$(1-\lambda)p_{ml}(\omega\|d)+\lambda p(\omega\|\mathcal{C})$	λ	λ
DIR	$\frac{c(\omega;d)+\mu p(\omega\|\mathcal{C})}{\sum_\omega c(\omega;d)+\mu}$	$\frac{\mu}{\sum_\omega c(\omega;d)+\mu}$	μ
ABS	$p_\delta(\omega\|d)=\frac{\max(c(\omega;d)-\delta,0)}{\sum_\omega c(\omega;d)}+\frac{\delta\|d\|_\mu}{\|d\|}p(\omega\|\mathcal{C})$	$\frac{\delta\|d\|_\mu}{\|d\|}$	δ

or δ means it involves more smoothing in the language model. Typically, these parameters can be estimated empirically by training sets. Once the smoothing parameters are given in advance, retrieval tasks using of the three methods above can be deployed very efficiently.

3 Cross-Language and Cross-Media Image Retrieval

In this section, we describe our experimental setup and development at the ImageCLEF 2005, in which we have participated in the bilingual ad hoc image retrieval task. In addition, we empirically analyze the results of our submission.

3.1 Experimental Setup and Development

The goal of the bilingual ad hoc retrieval task is to find as many relevant images as possible for each given topic. The St. Andrew collection is used as the benchmark dataset for the ad hoc retrieval task. There are 28 queries in total for each language. More details about the task can be found in [2].

For the bilingual ad hoc retrieval task, we have studied the query tasks in English and Chinese (simplified). Both text and visual information are used in our experiments. To evaluate the language models correctly, we employ the *Lemur* toolkit [1]. A list of standard stopwords is used in the parsing step.

To evaluate the influence on the performance of using the different schemes, we produce the results using a variety of configurations. Tables 2 shows the configurations and the experimental results in detail. In total, 36 runs with different configurations are provided in our submission.

3.2 Empirical Analysis on the Experimental Results

In this subsection, we empirically analyze the experimental results of our submission. The goal of our evaluation is to check how well the language model performs for cross-language image retrieval and what kinds of smoothing achieve better performance. Moreover, we are interested in comparing performance between the bilingual retrieval with Chinese queries and the monolingual retrieval with the normal English queries.

[1] http://www.lemurproject.org/.

Table 2. The configurations and official testing results of our submission

Run ID	Language	QE	Modality	Method	MAP
CUHK-ad-eng-t-kl-ab1	english	without	text	KL-LM-ABS	0.3887
CUHK-ad-eng-t-kl-ab2	english	with	text	KL-LM-ABS	0.4055
CUHK-ad-eng-t-kl-ab3	english	with	text	KL-LM-ABS	0.4082
CUHK-ad-eng-t-kl-jm1	english	without	text	KL-LM-JM	0.3844
CUHK-ad-eng-t-kl-jm2	english	with	text	KL-LM-JM	0.4115
CUHK-ad-eng-t-kl-di1	english	without	text	KL-LM-DIR	0.3820
CUHK-ad-eng-t-kl-di2	english	with	text	KL-LM-DIR	0.3999
CUHK-ad-eng-t-tf-idf1	english	without	text	TF-IDF	0.3510
CUHK-ad-eng-t-tf-idf2	english	with	text	TF-IDF	0.3574
CUHK-ad-eng-tn-kl-ab1	english	without	text	KL-LM-ABS	0.3877
CUHK-ad-eng-tn-kl-ab2	english	with	text	KL-LM-ABS	0.3838
CUHK-ad-eng-tn-kl-ab3	english	with	text	KL-LM-ABS	0.4083
CUHK-ad-eng-tn-kl-jm1	english	without	text	KL-LM-JM	0.3762
CUHK-ad-eng-tn-kl-jm2	english	with	text	KL-LM-JM	0.4018
CUHK-ad-eng-tn-kl-di1	english	without	text	KL-LM-DIR	0.3921
CUHK-ad-eng-tn-kl-di2	english	with	text	KL-LM-DIR	0.3990
CUHK-ad-eng-tn-tf-idf1	english	without	text	TF-IDF	0.3475
CUHK-ad-eng-tn-tf-idf2	english	with	text	TF-IDF	0.3660
CUHK-ad-eng-v	english	without	vis	Moment-DCT	0.0599
CUHK-ad-eng-tv-kl-ab1	english	without	text+vis	KL-LM-ABS	0.3941
CUHK-ad-eng-tv-kl-ab3	english	with	text+vis	KL-LM-ABS	0.4108
CUHK-ad-eng-tv-kl-jm1	english	without	text+vis	KL-LM-JM	0.3878
CUHK-ad-eng-tv-kl-jm2	english	with	text+vis	KL-LM-JM	0.4135
CUHK-ad-eng-tnv-kl-ab2	english	with	text+vis	KL-LM-ABS	0.3864
CUHK-ad-eng-tnv-kl-ab3	english	with	text+vis	KL-LM-ABS	0.4118
CUHK-ad-eng-tnv-kl-jm1	english	without	text+vis	KL-LM-JM	0.3787
CUHK-ad-eng-tnv-kl-jm2	english	with	text+vis	KL-LM-JM	0.4041
CUHK-ad-chn-t-kl-ab1	chinese	without	text	KL-LM-ABS	0.1815
CUHK-ad-chn-t-kl-ab2	chinese	with	text	KL-LM-ABS	0.1842
CUHK-ad-chn-t-kl-jm1	chinese	without	text	KL-LM-JM	0.1821
CUHK-ad-chn-t-kl-jm2	chinese	with	text	KL-LM-JM	0.2027
CUHK-ad-chn-tn-kl-ab1	chinese	without	text	KL-LM-ABS	0.1758
CUHK-ad-chn-tn-kl-ab2	chinese	with	text	KL-LM-ABS	0.1527
CUHK-ad-chn-tn-kl-ab3	chinese	with	text	KL-LM-ABS	0.1834
CUHK-ad-chn-tn-kl-jm1	chinese	without	text	KL-LM-JM	0.1843
CUHK-ad-chn-tn-kl-jm2	chinese	with	text	KL-LM-JM	0.2024

LM denotes Language Model, KL denotes Kullback-Leibler divergence based, DIR denotes the smoothing using the Dirichlet priors, ABS denotes the smoothing using Absolute discounting, and JM denotes the Jelinek-Mercer smoothing.

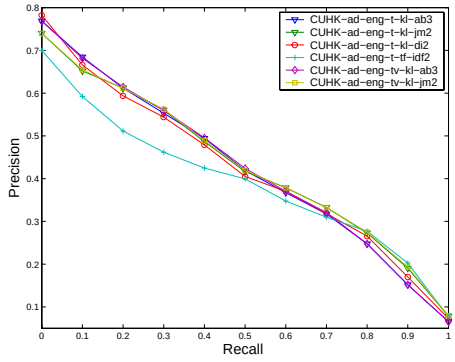


Fig. 1. Experimental Result of Precision vs. Recall with Selected Configuration

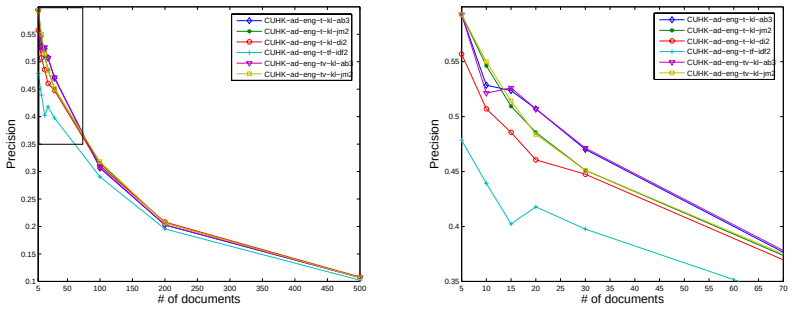
Empirical Analysis of Language Models. Figure 1 and Figure 2 plot the curves of *Precision* vs. *Recall* and the curves of *Precision* vs. *Number of Returned Documents*, respectively. From the experimental results shown in Figure 1 and Figure 2 as well as in Table 2, we can observe that the KL-divergence language model outperforms the simple TF-IDF retrieval model significantly (around 5%). In the evaluation of the smoothing techniques, we observe that the Jelinek-Mercer smoothing and the Absolute discounting smoothing yield better results than the Bayesian smoothing with the Dirichlet priors (DIR). More specifically, from Figure 2(b), we see that the Jelinek-Mercer smoothing achieves the best result when the number of returned documents is less than or equal to 13, while the Absolute discounting smoothing method achieves the best when the number of returned documents is greater than 13. Finally, from the official testing results [2], our approach achieves the best MAP (Mean Average Precision) result among all submissions on the monolingual query. This shows that the language model method is the state-of-the-art approach for text based image retrieval.

Cross-Language Retrieval: Chinese-To-English Query Translation. To deal with the Chinese queries for retrieving English documents, we first adopt a Chinese segmentation tool from the Linguistic Data Consortium (LDC) [10], i.e., the "LDC Chinese segmenter" [2], to extract the Chinese words from the given query sentences. The segmentation step is an important step toward effective query translation. Figure 3 shows the Chinese segmentation results of part queries. We can see that the results can still be improved.

For the bilingual query translation, the second step is to translate the extracted Chinese words into English words using a Chinese-English dictionary. In our experiment, we employ the LDC Chinese-to-English Wordlist [10]. The final translated queries are obtained by combining the translation results.

From the experimental results shown in Table 2, we can observe that the mean average precision of Chinese-to-English queries is about half of the mono-

[2] http://www.ldc.upenn.edu/Projects/Chinese/seg.zip.



(a) (b)

Fig. 2. Experimental Result of Precision vs. Number of Returned Documents with Selected Configuration. (a) shows the original comparison on 500 returned documents; (b) shows the detailed comparison on 70 returned documents.

lingual queries. There are many ways that we could improve this performance. One is to improve the Chinese segmentation algorithm. Some post-processing techniques may be effective for improving the performance. Also, the translation results can be further refined. Finally, one can better tune the results by adopting various Natural Language Processing techniques [11].

Cross-Media Retrieval: Re-Ranking Scheme with Text and Visual Content. In this task we study the combination of text and visual contents for cross-media image retrieval. We suggest the re-ranking scheme to combine text and visual contents. For a given query, we first rank the images using the language modeling techniques. We then re-rank the top ranked images by measuring the similarity of visual content to the query images.

In our experiment, two kinds of visual features are used: texture and color features. For texture, the discrete cosine transform (DCT) is engaged to calculate coefficients that multiply the basis functions of the DCT. Applying the DCT to an image yields a set of coefficients to represent the texture of the image. In our implementation, a block-DCT (block size 8x8) is applied on a normalized input image, which generates 256 DCT features. For color, color moment is employed to represent the images. For each image, 9 color moment features are extracted. Thus, in total, each image is represented by a 265-dimensional feature vector.

As shown in Table 2, the MAP performance of the retrieval results using only visual information is only about 6%; this is much lower than the approaches using text information, which yielded over 40%. From the experimental results, we observe that the re-ranking scheme produces only a marginal improvement compared with the text-only approaches. However, there are some reasons that explain the results. One is that the engaged visual features may not be able to discriminate between the images effectively. Another is that relevant images of

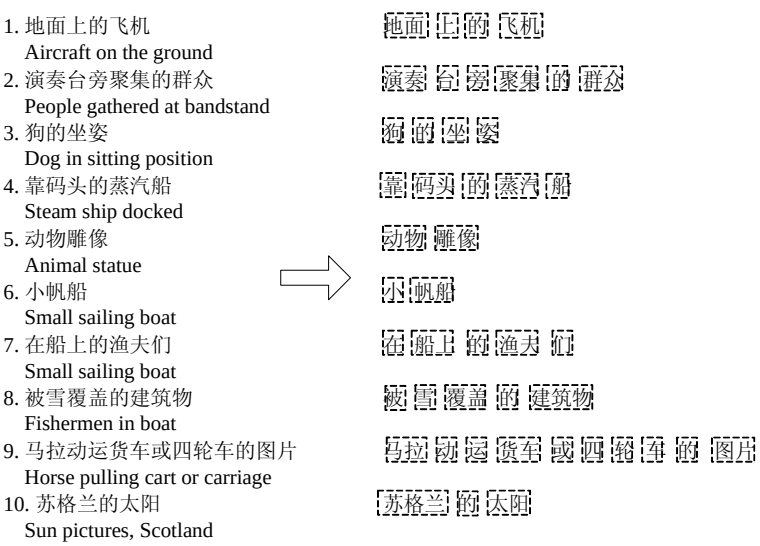


Fig. 3. Chinese segmentation results of part Chinese (Simplified) queries. Each dashed box represents a segmented Chinese word from the given English query.

the same queries in the ground truth may vary significantly in visual content, which makes it difficult for low-level features to discriminate between relevant and irrelevant images. In the future, two important research directions that could improve the performance are studying more effective techniques of low-level features, and finding more elegant methods of combining text and visual contents. Moreover, if users' logs of relevance feedback are available, that may also help the retrieval task.

Query Expansion for Information Retrieval. In general, Query Expansion (QE) refers to adding further terms to a text query (e.g. through pseudo-relevance feedback or a thesaurus) or adding further image samples to a visual query. From the experimental results in Table 2, we observe that most of the queries are greatly enhanced by adopting query expansion. The average improvement for all the queries is around 1.71%, which accounts for 4.14% of the maximum MAP of 41.35%. It is interesting to find that QE especially benefits considerably from the Jelinek-Mercer smoothing method; in this case, the mean gain with QE is about 2.49%, which accounts for 6.02% of the maximum MAP of 41.35%. Note that the number of feedback documents or samples usually strongly influences the improvement achieved with QE schemes. In our experiments, this number is estimated empirically from the official training set.

4 Conclusions

In this paper, we report our empirical studies of cross-language and cross-media image retrieval in the ImaegCLEF 2005 campaign. We address three major fo-

cuses and contributions. The first is the evaluation of Language Models and the smoothing strategies for cross-language image retrieval. We empirically show that the Language modeling approach is the state-of-the-art approach for text-based cross-language image retrieval. Among the smoothing techniques, the Jelinek- Mercer smoothing and the Absolute discounting smoothing perform better than the Bayesian smoothing with the Dirichlet priors. The second is the evaluation of cross-media image retrieval. We observe that the combination of text and visual contents gives only a marginal improvement. We can study more effective low-level features to improve this performance. The last is the evaluation of the bilingual image retrieval between English and Chinese. In our experiments, the mean average precision of Chinese-to-English Queries is about half of the monolingual queries. In future work, we can study more effective natural language processing techniques to improve this performance.

Acknowledgements

The work described in this paper was fully supported by Innovation and Technology Fund (Project No. ITS/105/03), and the Shun Hing Institute of Advanced Engineering (SHIAE) at CUHK.

References

1. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.
2. P. Clough, H. Müeller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jensen, and W. Hersh, "The CLEF 2005 cross language image retrieval track," in *Proceedings of the Cross Language Evaluation Forum 2005*. Springer Lecture Notes in Computer science, 2005.
3. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.
4. C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *ACM International SIGIR Conference (SIGIR'01)*, 2001, pp. 334–342.
5. C. E. Shannon, "Prediction and entropy of printed English," *Bell Sys. Tech. Jour.*, vol. 30, pp. 51–64, 1951.
6. C. Zhai and J. Lafferty, "Model-based feedback in the kl-divergence retrieval model," in *Proc. Tenth International Conference on Information and Knowledge Management (CIKM2001)*, 2001, pp. 403–410.
7. T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.
8. D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term," in *Proceedings 25th ACM SIGIR conference*, 2002, pp. 35–41.
9. F. Jelinek and R. Mercer, "Interpolated estimation of markov sourceparameters from sparse data," *Pattern Recognition in Practice*, pp. 381–402, 1980.
10. "http://www.ldc.upenn.edu/projects/chinese/."
11. C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.