# Language Model for IR Using Collection Information


Rong Jin
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
rong+@cs.cmu.edu

Luo Si
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lsi@cs.cmu.edu

Alex G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alex+@cs.cmu.edu

Jamie Callan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
callan+@cs.cmu.edu



## ABSTRACT

Information retrieval using meta data can be traced back to the early age of IR where documents are represented by the controlled vocabulary. In this paper, we explore the usage of meta-data information under the framework of language model. We present a new language model that is able to take advantage of the category information for documents to improve the retrieval accuracy. We compare the new language model with the traditional language model over the TREC4 dataset where the collection information for documents is obtained using the k-means clustering method. The new language model outperforms the traditional language model, which verifies our statement.


## 1. INTRODUCTION

In traditional information retrieval, a document is usually assumed to be a sequence of words, which is essentially the only evidence to be used by the retrieval system to judge. However, in the web environment, documents contain many useful meta-data such as topic information in XML format. Therefore, it becomes an interesting question as how to take advantage of the meta-data of documents to improve the retrieval accuracy.

Information retrieval using meta data can be traced back to the early age of IR where documents are represented by the controlled vocabulary. Previous works on the combination of full texts with the controlled vocabulary are mainly done within the framework of vector space model [1]. In this paper, we extend the similar study to the framework of language model. The meta-data examined in this study is the category labels for documents and the problem that we address here is how to adapt language models to make use of the category labels for documents. The basic idea underlying the new language model that we propose is to combine the evidence of the words together with category labels to help predict whether a document is relevant to the query. The documents within a category that is relevant to a query will have a better chance to be relevant than those in a category that is surely irrelevant to the query. The crucial problem with the new language model is how to compute the relevance of a category to a query and then use the computed relevance information for categories to help determine the relevance information for documents.

## 2. MODEL DESCRIPTION

The traditional language model defines the similarity between a document and a query as the probability $P(Q|D)$, i.e. the probability of generating query $Q$ given the observation of the document $D$ [2]. When category labels of documents are available, it would be appropriate to define the document-query similarity as the probability $P(Q|D,C_D)$, where $C_D$ stands for the unique category label for the document $D$.

As the confidence on the relevance of a document to a query depends on both the words in that document and the document label, we expand the probability $P(Q|D,C_D)$ as

$$P(Q \mid D, C_D) = \frac{P(Q \mid D) P(C_D \mid Q, D)}{P(C_D \mid D)} \qquad (1)$$

where probability $P(Q|D,C_D)$ is decomposed into a product of two parts, with the first part $P(Q|D)$ relying only on the document $D$ and the second part $P(C_D|Q,D)/P(C_D|D)$ depending on both collection $C_D$ and document $D$. To remove the influence of document $D$ from the second part, we can further expand probability $P(Q|D,C_D)$ as a linear combination of predictions from document $D$ and query $Q$, i.e.

$$P(C_D \mid Q, D) = \alpha P(C_D \mid Q) + (1-\alpha) P(C_D \mid D) \qquad (2)$$

where $\alpha$ is the parameter weighting the relative importance between the evidence $Q$ and evidence $D$. By substituting Equation (2) for $P(C_D|Q,D)$ in Equation (1), we have the expression for $P(Q|D,C_D)$ i.e.

$$P(Q \mid D, C_D) = (1-\alpha) P(Q \mid D)\left(\frac{\alpha}{(1-\alpha) P(C_D \mid D)} P(C_D \mid Q) + 1\right) \qquad (3)$$

Furthermore, by assuming that the assignment of document $D$ to collection $C_D$ is usually correct, we can have $P(C_D|D)$ approximated as a constant $p$ which is close to 1. Then, Equation (3) can be simplified as

$$P(Q \mid D, C_D) \propto P(Q \mid D)(\beta P(C_D \mid Q) + 1) \qquad (4)$$

where constant $\beta$ is defined as $\frac{\alpha}{(1-\alpha)p}$ and constant term $(1-\alpha)$ is ignored in Equation (4). As seen from Equation (4), probability $P(Q|D,C_D)$ is influenced by the document information through the term $P(Q|D)$ and the collection label information through the term $\beta P(C_D|Q)+1$.

To accomplish the computation in Equation (4), we need to compute both probability $P(C_D|Q)$ and $P(Q|D)$. For term $P(C_D|Q)$, we can simply apply Bayesian rule, which results in

$$P(C_i \mid Q) = \frac{P(Q \mid C_i) P(C_i)}{\sum_i P(Q \mid C_i) P(C_i)} \qquad (5)$$

The probability $P(Q|C)$ in above equation, i.e. how likely the collection $C$ is able to answer the query $Q$, can be computed by looking at the averaged likelihood for the documents in collection $C$ to be relevant to the query $Q$, or

$$P(Q \mid C) \approx \sum_{D' \in C} P(Q, D' \mid C) = \sum_{D' \in C} P(Q \mid D', C) P(D' \mid C)$$
$$\approx \frac{1}{|C|} \sum_{D' \in C} P(Q \mid D', C) \quad (6)$$

As indicted by Equation (6), the similarity between a collection and a query is measured by the document-query similarity averaged over all the documents of that collection. Finally, term $P(Q|D)$ can be computed using the traditional language model [2], which can be expressed as

$$P(Q \mid D) = \prod_{qw \in Q} (\lambda P(qw \mid D) + (1 - \lambda) P(qw \mid GE)) \quad (7)$$

where $\lambda$ is a smoothing constant.

## 3. EXPERIMENTS

To evaluate the effectiveness of the new language model, we conduct a contrastive experiment by comparing the new language model with the traditional language model. Since we don't have a collection with both manually assigned category labels and human relevance judgments for documents, we use the TREC4 collection as the test bed where the category labels are automatically computed using k-means clustering method [3]. There are altogether 100 different categories. Of course, the automatically generated category labels for documents are not of high qualities and therefore may have negative impact on the retrieval performance of the new language model.

The exact formula for the traditional language model is shown in Equation (2), which is also used to compute the $P(Q|D)$ in the new language model. The smoothing constant $\lambda$ is set to 0.5 for both language models.

Since the weight $\beta$ reflects the importance of query words in determining the document-query similarity, and in most case, query words provide much more information than the document labels, $\beta$ here is set to 19 empirically.

**Table 1**: Average precision and precisions at different recall points for both traditional language model and the new language model over TREC4 dataset

| Precision@Recall | Tradition LM | New LM |
|---|---|---|
| 0% | 0.6794 | 0.7013 |
| 10% | 0.4097 | 0.4270 |
| 20% | 0.3366 | 0.3568 |
| 30% | 0.2669 | 0.2803 |
| 40% | 0.1999 | 0.2161 |
| 50% | 0.1659 | 0.1688 |
| 60% | 0.1130 | 0.1173 |
| 70% | 0.0662 | 0.0627 |
| 80% | 0.0227 | 0.0304 |
| 90% | 0.0028 | 0.0027 |
| 100% | 0.0004 | 0.0008 |
| Avg. Precision | 0.1825 | 0.1911 |

Table 1 lists the average precision and precisions at different recall points for both the traditional language model and the new language model. As seen from Table 1, the new language model outperforms the traditional language model in terms of both average precision and precision at most recall points except the precision at 90% recall. The quality of the categories labels used in the experiments is not very high as we only use pseudo labels generated by k-means algorithms. It can be imagined that the performance will be further improved if we can incorporate high quality labels such as those generated by domain experts.

Notice that the difference between this work and the work on combining full-text and controlled vocabulary indexing is that, for the later problem, the retrieval accuracy is improved by the extra matches of query words with the manually assigned keywords, while in our work the improvement is obtained by using the estimated matching probabilities of collections with respect to the query.

## 4. CONCLUSIONS & FUTURE WORK

In this paper, we proposed a new language model for information retrieval that is able to make use of the document labels. In the experiment with TREC4 data, where the category labels for documents are simulated using k-means clustering method, the new language model outperforms the traditional language model. Therefore, we conclude that the new language model is a very promising method for exploring category information to improve the retrieve performance.

One problem left in the new framework is the parameter $\beta$, which was set to be 19 empirically. How to adjust this parameter automatically is a question for future work. In principle, the proposed method can be extended to incorporate other meta information. However, the problem of how to set the parameters such as lambdas needs to be explored in the future study. Furthermore, since we only explore the linear combination as the way of combining evidences in this work, more studies need to be carried out on different combination methods.


## 5. ACKNOWLEDGES

We thank ChengXiang Zhai for valuable discussions related to this work. This material is based on work supported by NSF grants EIA-9983253, IIS-0118767, and DUE-0085834, NSF Cooperative Agreement IRI-9817496, and Advanced Research and Development Activity (ARDA) contract MDA908-00-C-0037. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.


## 6. REFERENCES


[1] T.B. Rajashekar and W.B. Croft. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. Journal of American Society for Information Science, 1993

[2] J. Lafferty, C. Zhai. Document Language Models, Query Models and Risk Minimization for Information Retrieval. In Proc. of the 24th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.

[3] J. Xu and W.B. Croft. Cluster-based language models for distributed retrieval. In Proc. of the 22nd Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.