# Information Retrieval Using Label Propagation Based Ranking

Yang Lingpeng, Ji Donghong, Nie Yu
Institute for Infocomm Research
21, Heng Mui Keng Terrace
Singapore 119613
{lpyang, dhji, ynie}@i2r.a-star.edu.sg

## Abstract

*The I$^2$R group participated in the cross-language retrieval task (CLIR) at the sixth NTCIR workshop (NTCIR 6). In this paper, we describe our approach on Chinese Single Language Information Retrieval (SLIR) task and English-Chinese Bilingual CLIR task (BLIR). We use both bi-grams and single Chinese characters as index units and use OKAPI BM25 as retrieval model. The initial retrieved documents are re-ranked before they are used to do standard query expansion.*

*Our document re-ranking method is done by a label propagation-based semi-supervised learning algorithm to utilize the intrinsic structure underlying in the large document data. Since no labeled relevant or irrelevant documents are generally available in IR, our approach tries to extract some pseudo labeled documents from the ranking list of the initial retrieval. For pseudo relevant documents, we determine a cluster of documents from the top ones via cluster validation-based k-means clustering; for pseudo irrelevant ones, we pick a set of documents from the bottom ones. Then the ranking of the documents can be conducted via label propagation.*

*For Chinese SLIR task, experiences show our method achieves 0.3097, 0.4013 mean average precision on T-only run (Title based) at rigid, relax relevant judgment and 0.3136, 0.4071 mean average precision on D-only run (short description based) at rigid, relax relevant judgment.*

*For English-Chinese BLIR task, experiences show our method achieves 0.2013, 0.2931 mean average precision on T-only run at rigid, relax relevant judgment and 0.1911, 0.2804 mean average precision on D-only run at rigid, relax relevant judgment.*

**Keywords:** Document Re-ranking, Cluster Validation, Label Propagation, Chinese Information Retrieval, Query Expansion

## 1. Introduction

At NTCIR6, we participated in two sub-tasks in the Cross Lingual Information Retrieval (CLIR): Chinese Single Language Information Retrieval (SLIR) and English-Chinese Bilingual CLIR (BLIR). Readers are referred to [3] to get the information about NTCIR6 and the task description in detail.

For Chinese SLIR, we submitted two compulsory runs at STAGE1: a T-only run which uses field TITLE as query and a D-only run which uses field DESC as query.

For English-Chinese BLIR, we submitted two compulsory runs: T-only run and D-only run.

In NTCIR6, we use OKAPI BM25 as retrieval model and use both bi-grams and single Chinese characters as index units. The initial retrieved results are re-ranked before standard query expansion.

The document re-ranking method explores the intrinsic information among top retrieved documents [6]. This is done by using a label propagation-based learning algorithm to integrate pseudo labeled data with unlabeled data [9, 11]. This algorithm first represents labeled and unlabeled examples as vertices in a connected graph, then propagates the label information from any vertex to nearby vertex through weighted edges and finally infers the labels of unlabeled examples until the propagation process converges.

The English query in BLIR sub-task is translated to Chinese language by an online-dictionary and the translation results are weighted according to statistical search results by Google.

The rest of this paper is organized as following. In section 2, we describe the pre-processing on documents and queries. In section 3, we describe the index units and the retrieval model used in our system. In section 4, we describe our document re-ranking method based on cluster validation and label propagation. In section 5, we describe how to do query expansion in our system. In section 6, we describe the approach used in our BLIR sub-task. In section 7, we evaluate the performance of our proposed method on NTCIR6 and give out some result analysis. In section 8, we present the conclusion and some future work.

## 2. Pre-Processing

Before the normal Chinese IR process, all Chinese documents and Chinese queries are pre-processed as:

- All documents and queries are converted from BIG-5 code based to GB2312 code based so that we can save indexes space without losing too much precision. The BIG5 to GB2312 mapping is a many to one mapping because there are 13060 Chinese Characters in BIG5 representation but only 6763 Chinese Characters can be represented in GB2312 code. For those BIG5 Chinese Characters which have no mapping in GB2312 code, we assign 0xFEFE (first byte and second byte are 0xFE) as their mapping code in GB2312.

## 3. Indexing Units and Retrieval Model

While we use short terms as index units at NTCIR4 [4] and use bi-grams as index units at NTCIR5 [5], we use both bi-grams and single Chinese character as index units at NTCIR6.

For retrieval model, we use OKAPI BM25 model [8].

For the BM25 model, the relevance between the document and the query is defined in (1)-(3).

$$\sum_{t \in q} w_t \frac{(k_1+1)tf_d(t)}{K+tf_d(t)} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \quad (1)$$

$$w_t = \log \frac{(N-df(t)+0.5)}{df(t)+0.5} \quad (2)$$

$$K = k_1 \times ((1-b)+b \times \frac{dl}{avdl}) \quad (3)$$

where $w_t$, defined in (2), is the Robertson/Spark Jones weight of *t*. $k_1$, *b* and $k_3$ are parameters. $k_1$ and *b* are set as 1.2 and 0.75 respectively by default, and $k_3$ is set as 7. *dl* and *avdl* are respectively the document length and average document length measured by the number of the bi-grams.

## 4. Document Re-ranking

The document re-ranking is recast as a two-class label propagation problem. For this purpose, we need three sets of data: labelled relevant data as positive instances, labelled irrelevant data as negative instances, and unlabeled data. Since we do not have the labelled data except the query, which can be seen as a simple labelled relevant data, we try to generate some pseudo labelled data from the initial retrieval.

### 4.1 Pseudo Data for Label Propagation

Given a query *q*, suppose that we get *M* ranked documents in the initial retrieval. For irrelevant data, we simply pick *N* bottom ones as the pseudo irrelevant data. Regarding relevant data, a similar method would be to select top *K* documents as the pseudo relevant data. However, if noisy documents dominate the top ones, this method would fail. So, we turn to determine some clusters of documents among the top ones, and then select one closest to the query. After that, we take the documents in the cluster and the query itself as pseudo relevant documents.

To do that, we select top *K* (*K*<<*M*) documents from the *M* retrieved documents, and use a cluster validation-based [10] K-means clustering algorithm to determine the document clusters. First, a stability-based cluster validation approach is used to automatically determine the number of clusters. Then, the k-means clustering algorithm is used to cluster these documents. Finally, the *R* documents in the cluster closest to the query are picked as the pseudo relevant data. This is based on the assumption that all the *R* documents in the cluster most similar with the query tend to be relevant documents with higher probabilities.

The stability-based cluster validation approach [10] is capable of identifying both important feature words and true model order (cluster number). Important feature subset is selected by optimizing a cluster validity criterion subject to some constraint. For achieving model order identification capability, this feature selection procedure is conducted for each possible value of cluster number. The feature subset and cluster number which maximize the cluster validity criterion are chosen as answer.

## 4.2 Label Propagation-based Document Re-ranking

Given a query $q$ and $M$ ranked retrieved documents, we now have three datasets for label propagation: $M$ unlabeled examples, $R$ pseudo relevant examples derived from top $K$ documents and $N$ pseudo irrelevant examples as labelled data. As a result, document re-ranking can be achieved by ranking the $M$ unlabelled documents according to their similarities with the $R$ pseudo relevant documents via a label propagation algorithm as shown in Figure 1.

Following are some notations for the label propagation algorithm in document reranking:

- $q$ : the query
- $\{r_j\}$ ($1 \le j \le R$): the $R$ pseudo relevant labelled documents
- $\{n_j\}$ ($\le j \le N$): the $N$ pseudo irrelevant labelled documents
- $\{m_j\}$ ($1 \le j \le M$): the $M$ pseudo unlabeled documents, i.e. the initial $M$ retrieved documents, to be re-ranked
- $X = \{x_i\}$ ($1 \le i \le R+N+M$) refers to the union set of the above three categories of documents in the above order, i.e. $x_i$ ($1 \le i \le R$) represents the $R$ relevant labelled documents $\{r_j\}$ ($1 \le j \le R$), $x_i$ ($R+1 \le i \le R+N$) represents the $N$ irrelevant labelled documents $\{n_j\}$ ($1 \le j \le N$) and $x_i$ ($R+N+1 \le i \le R+N+M$) represents the initial $M$ retrieved documents $\{m_j\}$ ($1 \le j \le M$) to be re-ranked. That is, the first $R+N$ documents are pseudo labelled documents while the remaining $M$ documents are pseudo unlabeled documents to be re-ranked.
- $C = \{c_j\}$ ($1 \le j \le 2$) denotes the class set of documents where $c_1$ represents that a document is relevant with the query and $c_2$ represents that a document is irrelevant with the query.
- $Y^0 \in H^{s \times 2}$ ($s=R+N+M$) represents initial soft labels attached to each vertex, where $Y_{ij}^0= 1$ if $x_i$ is $c_j$ and 0 otherwise. Let $Y_L^0$ be the top $l=R+N$ rows of $Y^0$, which corresponds to the pseudo labelled data, and $Y_U^0$ be the remaining $u=M$ rows, which corresponds to the pseudo unlabeled data. Here, each row in $Y_U^0$ is initialized according the similarity of a document with the query.

In the label propagation algorithm, the manifold structure in $X$ is represented as a connected graph and the label information of any vertex in the graph is propagated to nearby vertices through weighted edges until the propagation process converges. Here, each vertex corresponds to a document, and the edge between any two documents $x_i$ and $x_j$ is weighted by $w_{ij}$ to measure their similarity. Here $w_{ij}$ is defined as follows: $w_{ij} = \exp(-d_{\mathrm{ij}}^2 / \sigma^2)$ if $\mathrm{i} \neq \mathrm{j}$ and $w_{ii} = 0$ ($1 \leqslant \mathrm{i,j} \leqslant \mathrm{l+u}$), where $d_{ij}$ is the distance between $x_i$ and $x_j$ (for example: cosine distance, Jenson-Shannon divergence distance), and σ is a scale to control the transformation. In this paper, we set σ as the average distance between labeled documents in different classes. Moreover, the weight $w_{ij}$ between two document $x_i$ and $x_j$ is transformed to a probability $t_{ij} = \mathrm{P}(\mathrm{j} \rightarrow \mathrm{i}) = w_{ij}/(\sum^{\mathrm{s}}_{\mathrm{k=1}} w_{\mathrm{kj}})$, where $t_{ij}$ is the probability to propagate a label from document $x_j$ to document $x_i$. In principle, larger weights between two documents mean easy travel and similar labels between them according to the global consistency assumption applied in this algorithm. Finally, $t_{ij}$ is normalized row by row: $\overline{t_{ij}} = t_{ij} \Big/ \sum_{k=1}^{s} t_{ik}$ . This is to maintain the class probability interpretation of Y. The s × s matrix $[\,\overline{t_{ij}}\,]$ is denoted as $\overline{T}$.

During the label propagation process, the label distribution of the labelled data is clamped in each loop and acts like forces to push out labels through unlabeled data. With this push originates from labelled data, the label boundaries will be pushed much faster along edges with larger weights and settle in gaps along those with lower weights. Ideally, we can expect that $w_{ij}$ across different classes should be as small as possible and $w_{ij}$ within a same class as big as possible. In this way, label propagation happens within a same class most likely.

This algorithm has been shown to converge to a unique solution [9] with $u=M$ and $l=R+N$:

$$\hat{Y}_U = \lim_{t \to \infty} Y_U^t = (I - \overline{T}_{uu})^{-1} \overline{T}_{ul}\, Y_L^0 \cdot$$

where $I$ is $u \times u$ identity matrix. $\overline{T}_{uu}$ and $\overline{T}_{ul}$ are acquired by splitting matrix $\overline{T}$ after the $l$-th row and the $l$-th column into 4 sub-matrices $\overline{T} = \begin{bmatrix} \overline{T}_{ll} & \overline{T}_{lu} \\ \overline{T}_{ul} & \overline{T}_{uu} \end{bmatrix}$.

In theory, this solution can be obtained without iteration and the initialization of $Y_U^0$ is not important, since $Y_U^0$ does not affect the estimation of $\hat{Y}_U$. However, the initialization of $Y_U^0$ helps the algorithm converge in practice. In this paper, each row in $Y_U^0$ is initialized according the similarity of a document with the query.

**Fig. 1 the label propagation algorithm in document re-ranking**

```
Input:
  $q$: query;
  $M$: the set/the number of ranked retrieved documents to be re-ranked;
  $R$: the set/the number of relevant documents extracted from top $K$ documents using cluster validation;
  $N$: the set/the number of irrelevant documents picked from the bottom of the ranked retrieved documents
  Algorithm: LabelPropagation($q, M, R, N$)
  BEGIN
      Set the iteration index $t$=0
      BEGIN DO Loop
          Propagate the label by $Y^{t+1} = \overline{T} Y^t$;
          Clamp the labelled data by replacing the top $l$ row of $Y^{t+1}$ with $Y_L^0$.
      END DO Loop when $Y^t$ converges;
      Re-order documents $x_h$ ($l$+1 ≤ h ≤ $l$+$M$) according to $Y_{h1}$ (probability of being a relevant document)
  END
```

## 5. Query Expansion

We use re-ranked retrieved documents to do query expansion. We use Robertson's RSV scheme [7] to select 200 bi-grams or single Chinese characters from top 20 re-ranked documents. We also make use of Rocchio's [2] formula, as improved by Salton and Buckley [1] to perform query expansion. The new query is retrieved again to get the final result.

## 6. English-Chinese BLIR

We first translate query from English to Chinese, then we use the same approach of Chinese SLIR to retrieve documents.

Generally, there are two ways in CLIR to translate query: by machine translation or by bilingual dictionary. We use an online dictionary to translate meaningful words or phrases in query from English to Chinese. Each pair of English word or phrase and its one possible Chinese translation are inputted to Google search engine to retrieve documents. Top 20 pages of returned results are used for analysis to calculate the probability of translating the English word or phrase to current Chinese translation.

## 7. Evaluation

For Chinese SLIR STAGE1, we submitted two compulsory runs: a T-only run which only uses field TITLE as query and a D-only run which only uses field DESC as query. For English-Chinese BLIR, we submitted two compulsory runs: a T-only run and a D-only run.

Table 1 lists statistical result of mean average precision (MAP) for 50 query topics on relax relevance judgment and rigid relevance judgment. Relax relevance judgment considers high relevant documents, relevant documents and partially relevant documents. Rigid relevance judgment only considers high relevant documents and relevant documents. In table 1, row [C-C-T] represents Chinese to Chinese T-only run in Chinese SLIR, [C-C-D] represents Chinese to Chinese D-only run in Chinese SLIR, row [E-C-T] represents English to Chinese T-only run in English-Chinese BLIR, and row [E-C-D] represents English to Chinese D-only run in English-Chinese BLIR; Column [rigid] represents MAP using rigid measurement, and Column [relaxed] represents MAP using relaxed measurement.

From the statistical results in Table 1, for Chinese SLIR T-only run, our group achieves 0.3097 and 0.4013 MAP on rigid and relaxed relevance judgment; for Chinese SLIR D-only run, our group achieves 0.3136 and 0.4071 MAP on rigid and relaxed relevance judgment; for English-Chinese BLIR T-only run, our group achieves 0.2013 and 0.2931 MAP on rigid and relaxed relevance judgment; for English-Chinese BLIR D-only run, our group achieves 0.1911 and 0.2804 MAP on rigid and relaxed relevance judgment.

Although both SLIR and BLIR use the same document collection, the performance of BLIR is quite poor compared with that of SLIR. Analysis shows that the quality of translation of queries in BLIR plays an important role and we need to find more effective ways to translate query from English to Chinese.

Our experiments also show that using both bi-grams and single Chinese characters as index units can produce better results than only using bi-grams as index units.

Our experiments also show document re-ranking can improve the effectiveness of retrieved documents and can help query expansion to produce better results. Our experiments show that the performance of document re-ranking on top 100 to 1000 retrieved documents (R=10, N=5) improves MAP from 9.8% to 17.2%, respectively.

Table 1 Official MAP results at NTCIR-6

| | rigid | relaxed |
|---|---|---|
| i2r-C-C-T-01 | 0. 3097 | 0. 4013 |
| i2r-C-C-D-01 | 0. 3136 | 0. 4071 |
| i2r-E-C-T-01 | 0. 2013 | 0. 2931 |
| i2r-E-C-D-01 | 0. 1911 | 0. 2804 |

## 8. Conclusion

In this paper, we introduce our CLIR approach and our experience in participating in Chinese SLIR and English-Chinese BLIR in NTCIR6. For Chinese SLIR, our system achieves 0.3097 and 0.4013 MAP on rigid and relaxed relevance judgment for T-only run and 0.3136 and 0.4071 MAP on rigid and relaxed relevance judgment for D-only run. For English-Chinese BLIR, our system achieves 0.2013 and 0.2931 MAP on rigid and relaxed relevance judgment for T-only run and 0.1911 and 0.2804 MAP on rigid and relaxed relevance judgment for D-only run.

Our experimental results show that proper document re-ranking can improve the precision of top retrieved documents and further improve the effectiveness of query expansion.

Our experimental results also show that using both bi-grams and single Chinese character as index units produces better results than only using bi-grams as index units.

Our experimental results also show that how to translate query affects seriously the performance in BLIR. In the future, we'll try other approaches to improve our system's performance in BLIR.

## References

[1]G.Salton, C. Buckley. *Improving Retrieval Performance by relevance feedback*. J. Am. Soc. Inf. Sci. 41, 288-297. 1990.

[2] J. Rocchio. *Relevance Feedback in Information Retrieval*. In the SMART Retrieval System – Experiments in Automatic Document Processing. G.Salton, Ed., Prentice Hall, Englewood Cliffs, N.J. 1971

[3] K. Kishida, K. -H. Chen, S. Lee, K. Kuriyama, N. Kando, H. -H. Chen, and S. -H. Myaeng. *Overview of CLIR task at the sixth NTCIR workshop*.In Proceedings of the Sixth NTCIR Workshop.

[4] L.P. Yang, D.H. Ji, L. Tang. *Chinese Information Retrieval Based on Terms and Ontology*. In the Fourth NTCIR Workshop.

[5] L.P. Yang, D.H. *$I^2R$ at NTCIR5*. In the Fifth NTCIR Workshop.

[6] L.P. Yang, D.H. Ji, G.D. Zhou, Y. Nie, G.Z. Xoao. *Document re-ranking using cluster validation and label propagation*. ACM Conference on Information and Knowledge Management (CIKM) 2006. Pp. 690-697. p

[7] S. E., Robertson. *On Term Selection for Query Expansion*. Journal of Documentation 46. Dec 1990, pp 359-364.

[8] S.E. Robertson, S. Walker, and M. Sparck Jones, *Okapi at TREC-3*. Proc. of Third Text Retrieval Conference (TREC-3), 1995.

[9] X. Zhu, Z. Ghahramani. *Learning from Labeled and Unlabeled Data with Label Propagation*. CMU CALD technical report CMU-CALD-02-107. 2002.

[10] Z.Y. Niu, D.H. Ji, C.L. Tan. *Document Clustering based on Cluster Validation*. In Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM-2004), Washington, DC, USA, pp.501-506

[11] Z.Y. Niu, D.H. Ji, and C.L. Tan. Word Sense Disambiguation Using Label Propagation Based Semi-supervised Learning. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05), Ann Arbor, Michigan, US, pp.395-402. 2005.