Document Re-ranking Using Cluster Validation and Label Propagation

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

Guodong Zhou, Yu Nie
Institute for Infocomm Research
21 Heng Mui Keng Terrace
119613 Singapore
{zhougd,ynie}@i2r.a-star.edu.sg

Guozheng Xiao
Center for Study of Language & Information
Wuhan University, 430072, China
xiaogz@whu.edu.cn

ABSTRACT
This paper proposes a novel document re-ranking approach in information retrieval, which 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. Evaluation on benchmark corpora shows that the approach can achieve significant improvement over standard baselines and performs better than other related approaches.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Clustering

General Terms:
Languages, Algorithms.

Keywords
Document Re-ranking, Information Retrieval, Data Manifold Structure, Label Propagation

1. INTRODUCTION
Document re-ranking (DR) in information retrieval has been a hot research topic during the last decade. Normally, it can be achieved in two ways: direct re-ranking on initial retrieved documents and indirect re-ranking via automatic query expansion (QE). While automatic query expansion assumes that top ranked documents are more likely to be relevant, the terms in these documents can be used to augment the original query and a better ranking can be expected via a second retrieval, direct re-ranking improves the rankings of the initial retrieved documents by directly adjusting their positions without a second retrieval. Normally, these two approaches can be integrated. For example, direct re-ranking can be used to improve automatic query expansion since better ranking in top retrieved document can be expected to improve the quality of the augmented query.

This paper will focus on direct document re-ranking. According to the different information sources used, traditional document re-ranking can be classified into three categories.

The first category uses inter-document relationship. For example, [1] re-ranked documents by using document distances to modify their relevance weights while [8] proposed their approach based on document clustering. The second category uses various external resources, such as manually built thesaurus [15], manually crafted grammars [2] and controlled vocabularies [6]. The third category uses specific information extracted from documents or queries. For example, [11] used the information in the document title; [4] used stemmed words in the initial retrieval and augmented un-stemmed words in queries in document re-ranking; [17, 18] made use of global and local information to re-rank the documents via local context analysis; [12] used maximal marginal relevance to adjust the contribution of relevant terms; [19, 20] used query terms, which occur in both queries and top retrieved documents, to re-rank documents.

Recently, there is a trend to explore the intrinsic structure of documents to re-rank documents. [21] proposed an affinity graph to re-rank documents by optimizing their diversity and information richness. [7] proposed a structural re-ranking approach using asymmetric relationships among documents induced by language models. [5] used score-regularization to adjust ad-hoc retrieval scores from an initial retrieval so that topically related documents received similar scores.

This paper follows the trend and proposes a novel document re-ranking approach by exploring the intrinsic information among top retrieved documents. This is done by using a label propagation-based learning algorithm to integrate pseudo labeled data with unlabeled data. 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.

Label propagation [22] is a kind of semi-supervised learning algorithms, which is characterized of using unlabeled data in the learning process. The rationale behind this algorithm is that the instances in high-density areas tend to carry the same labels.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

Label propagation has been successfully applied in NLP, such as word sense disambiguation [13] and text classification [3, 16, 22, 23].

The rest of this paper is organized as follows. In Section 2, we describe the re-ranking approach in details while the experimental results are given in Section 3. Finally we conclude this paper with some remarks in Section 4.

2. DOCUMENT RE-RANKING

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

2.1 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 [14] 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 [14] 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.

2.2 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 labeled 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
- $\{r_j\}$ ($1 \leqslant j \leqslant R$): the R pseudo relevant labeled documents
- $\{n_j\}$ ($\leqslant j \leqslant N$): the N pseudo irrelevant labeled documents
- $\{m_j\}$ ($1 \leqslant j \leqslant M$): the M pseudo unlabeled documents, i.e. the initial M retrieved documents, to be re-ranked
- $X = \{x_i\}$ ($1 \leqslant i \leqslant R+N+M$) refers to the union set of the above three categories of documents in the above order, i.e. x_i ($1 \leqslant i \leqslant R$) represents the R relevant labeled documents $\{r_j\}$ ($1 \leqslant j \leqslant R$), x_i ($R+1 \leqslant i \leqslant R+N$) represents the N irrelevant labeled documents $\{n_j\}$ ($1 \leqslant j \leqslant N$) and x_i ($R+N+1 \leqslant i \leqslant R+N+M$) represents the initial M retrieved documents $\{m_j\}$ ($1 \leqslant j \leqslant M$) to be re-ranked. That is, the first $R+N$ documents are pseudo labeled documents while the remaining M documents are pseudo unlabeled documents to be re-ranked.
- $C = \{c_j\}$ ($1 \leqslant j \leqslant 2$) 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 y_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 labeled 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: $w_{ij} = \exp(-d_{ij}^2/\sigma^2)$ if $i \neq j$ and $w_{ii} = 0$ ($1 \leqslant i,j \leqslant l+u$), 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 $t_{ij} = P(j \rightarrow i) = w_{ij}/(\sum_{k=1}^{s} w_{kj})$, 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 $s \times s$ matrix $[\overline{t_{ij}}]$ is denoted as $\overline{T}$.

During the label propagation process, the label distribution of the labeled data is clamped in each loop and acts like forces to push out labels through unlabeled data. With this push originates from labeled 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 [22] with $u=M$ and $l=R+N$:

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

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 labeled 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 ($l+1 \leqslant h \leqslant l+M$) according to Y_{h1} (probability of being a relevant document)

END

3. EXPERIMENTS AND EVALUATION

We evaluated our approach on both the NTCIR-3 (NII-NACSIS Test Collection for IR Systems) CLIR Chinese SLIR document collection (*http://research.nii.ac.jp/ntcir/index-en.html*) and TREC-8 Ad-Hoc data using the vector space model (VSM) and OKAPI BM25 model, respectively.

In VSM, each document or query is represented as a vector in a vector space and each dimension of the vector is a word for the English language or bi-gram for the Chinese language. Here, the weight of a word or bi-gram b in document d is calculated using the following TF/IDF weighting scheme:

$$w(b,d)=\log(TF(b,d)+1)*\log(P/DF(b)+1) \quad (1)$$

where $w(b,d)$ is the weight for b in d; $TF(b, d)$ is the frequency of b in d; P is the number of documents in the document set; and $DF(b)$ is the number of documents that contain b.

The weight of a word or bi-gram b in query q, $w(b, q)$, is given by the following weighting scheme:

$$w(b,q)=TF(b,q) \quad (2)$$

where $TF(b, q)$ is the frequency of b in q.

For OKAP BM25 model, we use the default parameter settings.

Table 1 gives the information about our test dataset. For comparison, the same document re-ranking algorithm was applied once to each query and the overall performance was averaged over all the queries. Moreover, we used standard Mean Average Precision (MAP) to measure the overall retrieval performance. Finally, we set K as 10, which means that we only seek the pseudo relevant documents among top 10 documents. During the cluster-validation, we suppose that the cluster number ranges from 2 to 6, which means that the most appropriate number of clusters existing in top 10 documents for all the queries falls within 2 to 6, although it may vary with different queries. Stability-based cluster validation was used to infer the most appropriate cluster number for each query. In addition, Jenson-Shannon (JS) divergence [9] is applied to measure document distance d_{ij}.

Table 1 Test Data Set

	NTCIR-3	TREC-8
Language	Chinese	English
Retrieval Model	VSM	OKAPI BM25
Indexing units	Bi-grams	Words
Number of queries	42	150

3.1 Experiments on NTCIR-3

Relaxed relevance and rigid relevance are adopted by NTCIR to measure the performance. Relaxed relevance measurement (relaxed) considers highly relevant, relevant, and partially relevant documents as relevant, while rigid relevance measurement (rigid) only considers highly relevant and relevant documents as relevant. In NTCIR-3, each query is a description of a topic in the Chinese language, e.g.

查询故宫博物院所举办之千禧汉代文物大展相关内容
(Find information about the exhibition "Art and Culture of the Han Dynasty" in the National Palace Museum)

Since the query q can be always regarded as one of labeled relevant documents, our first experiment only used 1 pseudo relevant document (query q, R=1) and 5 pseudo irrelevant documents (N=5). The experimental results show that the re-ranking doesn't improve the performance. This means that, although the last five documents provide some information about irrelevant documents, the query itself may provide too little information about relevant documents.

Table 2 evaluates the effect of pseudo relevant documents via cluster validation-based k-means clustering when different numbers of retrieved documents are to be re-ranked. Here, the number of pseudo irrelevant documents N is set to 5. In Table 2, the first column indicates number of top documents to be re-ranked, INI refers to initial retrieval result, MAP(rigid) and MAP(relaxed) represent MAP values on rigid relevance measurement and relaxed relevance measurement respectively, and +x% (-x%) denotes improvement (decrease) against baseline [INI].

Table 2 MAPs after document re-ranking (K=10, N=5)

	MAP(rigid)	MAP(relaxed)
INI	0.1688	0.2197
M=40	0.1865 +10.5%*	0.2421 +10.2%*
M=50	0.1895 +12.3%*	0.2445 +11.3%*
M=60	0.1936 +14.7%**	0.2489 +13.3%**
M=70	0.1985 +17.6%**	0.2546 +15.9%**
M=80	0.1997 +18.3%**	0.2564 +16.7%**
M=90	0.2005 +18.8%**	0.2573 +17.1%**
M=100	0.2036 +20.6%**	0.2601 +18.4%**
M=200	0.2107 +24.8%**	0.2720 +23.8%**
M=300	0.2130 +26.2%**	0.2775 +26.3%**
M=400	0.2157 +27.8%**	0.2808 +27.8%**
M=500	0.2164 +28.2%**	0.2825 +28.6%**
M=600	0.2179 +29.1%**	0.2843 +29.4%**
M=700	0.2181 +29.2%**	0.2852 +29.8%**
M=800	0.2188 +29.6%**	0.2876 +30.9%**
M=900	0.2198 +30.2%**	0.2882 +31.2%**
M=1000	0.2203 +30.5%**	0.2887 +31.4%**

Table 2 shows that document re-ranking improves MAP(rigid) and MAP(relaxed) from 10.5% to 30.5% and from 10.2% to 31.4% respectively when the number of retrieved documents increases from 40 to 1000. To see whether the improvement is significant, we conducted the paired t-test. Table 3 also shows the significance marks. Here, **, * and ~ denote p-values smaller than 0.01, in-between (0.01, 0.05] and bigger than 0.05, and mean significantly better, better and almost the same, respectively.

From Table 2, we can see that, in both MAP(relaxed) and MAP(rigid),

- Document re-ranking achieves better performance when less than 60 documents are re-ranked.
- Document re-ranking achieves significantly better performance when more than 60 documents are re-ranked.

This implies that the cluster of pseudo relevant documents does contribute to the label propagation-based document re-ranking approach, which leads to improvement of the performance.

Table 3 lists the comparison of effectiveness on precision where PreAt10(rigid) and PreAt10(relaxed) represents the precision at top 10 documents on rigid relevance and relaxed relevance measure, respectively. Table 3 shows that document re-ranking improves PreAt10(rigid) and PreAt10(relaxed) from 0.3321 to 0.3333 and from 0.4566 to 0.4595 respectively when the number of retrieved documents increases from 40 to 90.

Table 3 Precisions after document re-ranking (K=10, N=5)

	PreAt10(rigid)	PreAt10(relax)
INI	0.2595	0.3619
M=40	0.3321	0.4566
M=50	0.3333	0.4571
M=60	0.3333	0.4577
M=70	0.3333	0.4583
M=80	0.3333	0.4588
M=90	0.3333	0.4595

Table 4 evaluates the quality of pseudo relevant documents extracted from top K retrieved documents via cluster validation-based k-means clustering. Notice that when computing recall, we only consider the actually relevant documents occurring in top K documents, since the clustering is only conducted on them. Table 4 also lists the accuracy in top K, which refers to the percentage of actually relevant documents in top K ones. Comparing precision of C and accuracy of top K, we can see that the quality of the automatically acquired pseudo relevant documents is much better (20% higher in precision) than simply treating top K documents as relevant ones. This implies that cluster validation-based k-means clustering is effective in choosing pseudo relevant documents. Figure 2 shows the MAP comparison between the clustering-based pseudo relevant documents (K=10, N=5; denoted as Cluster in Figure) and top 10 documents directly as pseudo relevant documents (R=10, N=5; denoted as non-Cluster in Figure). This suggests that the selection of pseudo relevant documents by clustering is useful for improvement of MAPs.

Table 4 Quality of pseudo relevant documents

	K=10	*K*=20
Size of *C*	3.67	4.56
F-measure of *C*	0.4886	0.4476
Precision of *C*	0.5229	0.4655
Recall of *C*	0.4585	0.4310
Accuracy in top *K*	0.4228	0.3213

Fig. 2 Comparison: Cluster vs. non-Cluster

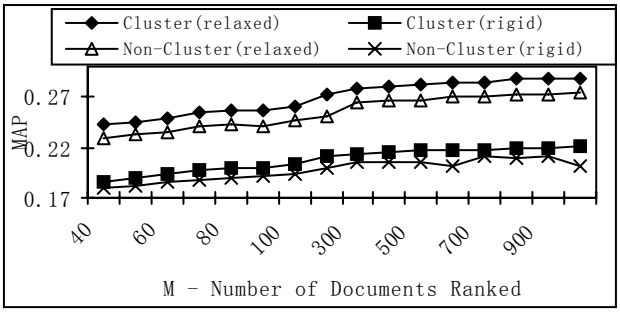


Table 5 Comparison: actually relevant documents vs. pseudo relevant documents

	MAP(rigid)	MAP(relaxed)
INI	0.1688	0.2197
M=40	0.2235 +32.4%**	0.2730 +24.3%**
M=50	0.2271 +34.5%**	0.2784 +26.7%**
M=60	0.2318 +37.3%**	0.2835 +29.0%**
M=70	0.2387 +41.4%**	0.2914 +32.6%**
M=80	0.2430 +44.0%**	0.2961 +34.8%**
M=90	0.2439 +44.5%**	0.2967 +35.0%**
M=100	0.2472 +46.4%**	0.3008 +36.9%**
M=200	0.2562 +51.8%**	0.3194 +45.4%**
M=300	0.2621 +55.3%**	0.3298 +50.1%**
M=400	0.2710 +60.5%**	0.3401 +54.8%**
M=500	0.2728 +61.6%**	0.3420 +55.7%**
M=600	0.2754 +63.2%**	0.3452 +57.1%**
M=700	0.2772 +64.2%**	0.3472 +58.0%**
M=800	0.2788 +65.2%**	0.3501 +59.4%**
M=900	0.2813 +66.6%**	0.3523 +60.4%**
M=1000	0.2816 +66.7%**	0.3525 +60.5%**

Table 5 compares the performance of using actually relevant documents in top 10 documents as relevant documents (if no actually relevant documents occur in top 10, we simple use top 3 documents to fill the gap) and 5 pseudo irrelevant documents. It shows that document re-ranking using actually relevant documents in top 10 performs much better than document re-ranking using pseudo relevant documents. Paired t-tests between them show that the performance difference is significant. This implies that, although document re-ranking using pseudo relevant documents achieves promising improvement, there are still much potential for further improvement.

To further explore the impact of *K* (the number of top retrieved documents) in determining pseudo relevant documents, we fixed the number of pseudo irrelevant documents *N* to 5 and changed *K* from 10 to 20. Figure 3 shows the performance tendency when 1000 documents are re-ranked, where INI and DR represent MAP values before and after document re-ranking, respectively. It demonstrates that, when *K* changes from 10 to 16, both DR(rigid) and DR(relaxed) increase steadily, while when *K* changes from 16 to 20, both DR(rigid) and DR(relaxed) decrease slightly. To see why, we checked the quality of the selected cluster of pseudo relevant documents, whose F-scores went up from 0.4886 to 0.4973, and then down to 0.4476. This again suggests that the quality of pseudo relevant documents is an important factor.

Fig. 3 The effect of *K* when *N*=5

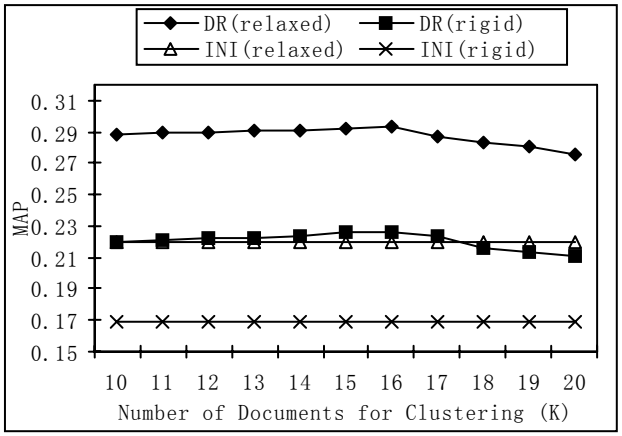
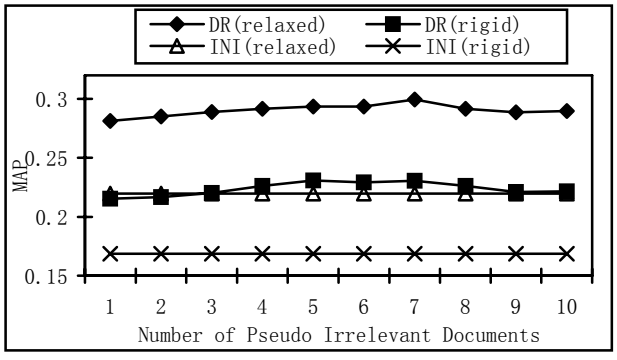


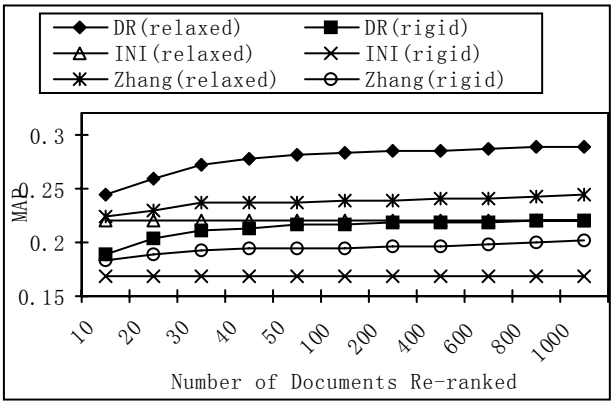
Fig. 4 The effect of *N* when *K*=10



To explore the impact of N (pseudo irrelevant documents) in document re-ranking, we fixed K to 10 and changed N from 1 to 10. Figure 4 shows the MAP tendency, which demonstrates that the performance reaches the maximum when N falls in 5 to 7. This implies it is a better choice to choose 5-7 pseudo irrelevant documents.

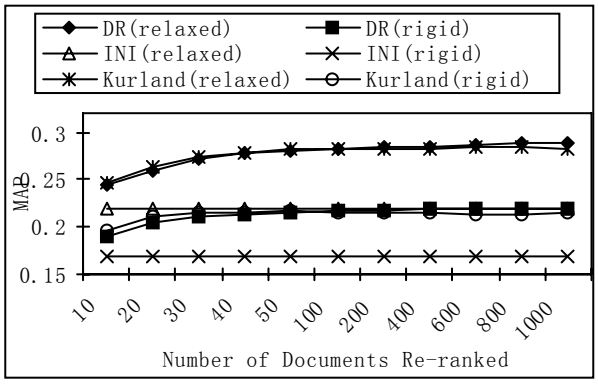
In comparison with other approaches, Figure 5 shows the performance of our approach (K=10, N=5) and Zhang et al.'s affinity graph-based approach [21] with the ranking-combination scheme as (α=0.5; β=0.5). It demonstrates that our approach achieves better performance than theirs. This may be due to the fact that their approach focuses more on diversify and information richness and cares less on precision of the retrieval results.

Fig. 5 Comparison: affinity-graph based & ours



We also compared our approach (K=10, N=5) with Kurland et al.'s structural re-ranking approach [7]. Figure 6 shows the MAP comparison where parameters of their approach are set as (R-W-In+LM: Recursive Weighted Influx + Language Model). It demonstrates that our approach gets comparable performance with theirs. However, their approach is based on language models which require large scale training data to be effective, our method is based on the intrinsic manifold structure of top retrieved documents with little training data needed, and pseudo labeled data is automatically created from the ranking list of the initial retrieval.

Fig. 6 Comparison: structural re-ranking & ours



We also compare our method with Mitra et al.'s maximal marginal relevance (MMR) method [12], which uses term correlation to re-order retrieved documents. If $\{w_1, \ldots, w_m\}$ is the set of query words presented in document d (ordered by decreasing idf), then the new ranking score between q and d is calculated by following formula:

$$Sim_{new} = idf(w_1) + \sum_{i=2}^{m} idf(w_i) \times \min_{j=1}^{i-1}(1 - P(w_i \mid w_j)) \quad (3)$$

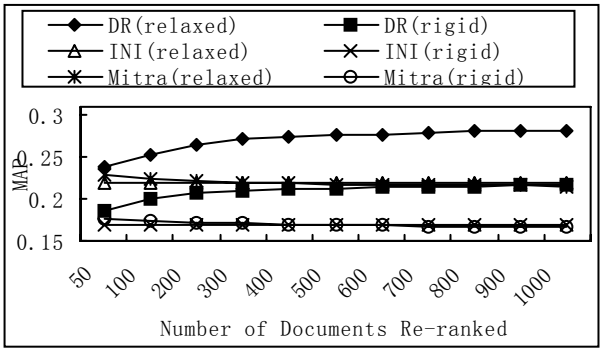
where $idf(w_i)$ is the inverse document frequency of word w_i in retrieved documents to be re-ranked, $P(w_i \mid w_j)$ is the word correlation between w_i and w_j in top K retrieved documents calculated by the formula:

$$P(w_i \mid w_j) = \frac{\text{number of documents in S containing query term } w_i \text{ and } w_j}{\text{number of documents in S containing query term } w_j}$$

where S refers to document set.

Figure 7 shows the comparison of performance between Mitra's and our document re-ranking method (K=10, N=5). In the experiment, we re-rank top 50, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 documents respectively.

Fig. 7 Comparison with Mitra's method



From Figure 7, we can see that our method (DR) achieves better performance than that of Mitra's for both MAP(rigid) and MAP(relaxed) consistently at every document number setting. On the other hand, for our method, the improvement increases in a stable way as the number of documents to be re-ranked increases, while for Mitra's method, the improvement generally decreases as the document number increases. For example, Mitra(rigid) decreases from 0.1751 to 0.1749 and 0.1676 as document number increase from 50 to 100 and 1000, while our DR(rigid) increases from 0.1895 to 0.2036 and 0.2203. Another finding is that Mitra's method is only applicable to top (50 to 100) ranking documents, as was claimed in Mitra's paper, while our method is more robust and applicable to both smaller and larger scope of documents.

Now that document re-ranking can improve the performance of initial search, it should help query expansion. To confirm it, we combined document re-ranking with standard Rocchio's relevance feedback. That is, we first re-rank top 100 retrieved documents by using (K=10, N=5), and then applied standard Rocchio's relevance feedback on re-ranked top documents. In the experiment, we selected 200 bi-grams from the top F (F=15, 20, 25 or 30) retrieved documents, and the

selected units were added to the original queries to form new ones.

Table 6 gives the MAP values for standard QE (Rocchio) and extended QE (re-ranking + Rocchio), where the first column indicates the number of the top documents used for query expansion.

From Table 6, we can see that extended QE achieved better results against standard QE with 13.9%-17.0% improvement. This indicates that the re-ranking helps query expansion to improve precision. Paired t-tests between extended QE and standard QE shows that the difference is significant, which indicates that extended QE not only improves the performance of initial retrieval, but also significantly outperforms standard QE.

Table 6 MAPs of query expansion (QE)

	Standard QE		Extended QE	
	MAP (rigid)	MAP (relaxed)	MAP (rigid)	MAP (relaxed)
F=15	0.2196	0.2836	0.2556** +16.4%	0.323** +13.9%
F=20	0.2229	0.2853	0.2579** +15.7%	0.3241** +13.6%
F=25	0.2216	0.2843	0.2586** +16.7%	0.3281** +15.4%
F=30	0.2208	0.2839	0.2583** +17.0%	0.3271** +15.2%

3.2 Experiments on TREC-8 Ad-hoc Data

To see whether the same finding on Chinese dataset comes out on English dataset, we performed experiments on the TREC-8 ad-hoc retrieval dataset, which consists of 50 *ad hoc* topics 410-450. We use all collections on Tipster disks 4&5 (no CR). Each query consists of two parts: TITLE and DESCRIPTION. We formed query vectors based on "TITLE" field with some preprocessing, such as stemming and tossing out stop words.

Table 7 shows the performance of document re-ranking on top *M* retrieved documents (*K*=10, *N*=5), which demonstrates that document re-ranking improves MAP from 9.8% to 17.2%. To see whether the improvement is significant, we also conducted paired t-tests, in which MAPs of the 150 topics are regarded as sampled observations. Table 7 also lists the significance marks. This suggests that the method also applies to English data.

In comparison with other approaches, Table 8 shows the performance of our approach (*K*=10, *N*=5) and the score regularization approach [5] with the Okapi BM25 scores.

From Table 8, our approach achieves comparable performance when more than 500 documents are re-ranked, and achieves better performance than theirs when less than 500 documents are re-ranked.

Table 7 MAPs after document re-ranking (*K*=10; *N*=5)

	MAP
INI	0.2301
M=40	0.2527 +9.8%*
M=50	0.2545 +10.6%*
M=60	0.2559 +11.2%**
M=70	0.2581 +12.2%**
M=80	0.2593 +12.7%**
M=90	0.2612 +13.5%**
M=100	0.2614 +13.6%**
M=200	0.2621 +13.9%**
M=300	0.2632 +14.4%**
M=400	0.2637 +14.6%**
M=500	0.2642 +14.8%**
M=600	0.2648 +15.1%**
M=700	0.2655 +15.4%**
M=800	0.2662 +15.7%**
M=900	0.2681 +16.5%**
M=1000	0.2697 +17.2%**

Table 8 Comparison: Score regularization based & ours

	Score regularization	DR
M=100	0.2389	**0.2614**
M=250	0.2452	**0.2627**
M=500	0.2527	**0.2642**
M=1000	0.2615	**0.2697**

4. CONCLUSION AND FUTURE WORK

This paper proposes a novel document re-ranking approach in information retrieval. It is done by using a label propagation-based semi-supervised learning algorithm to integrate labeled data with unlabeled data.

Since no labeled relevant or irrelevant documents are available in IR, we try to automatically create some pseudo labeled data from the initial retrieval. Given an initial ranked list of retrieved documents, our approach extracts a set of documents from the top ones via cluster validation-based k-means clustering as pseudo relevant data and picks a set of documents from the bottom ones as pseudo irrelevant data while recasting the whole initial ranked list of retrieved documents as unlabelled data to be re-ranked. In this way, the intrinsic

manifold structure underlying in the retrieved documents can contribute to the ranking via label propagation.

The label propagation-based approach is based on a global consistency assumption that similar examples in a high-density area should have similar labels. Our experiment demonstrates the potential of this manifold learning and shows its effectiveness in document re-ranking approach.

There are some possible improvements in the future. For example, the quality of the pseudo relevant documents is a key factor in document re-ranking, and we will explore more effective approaches to find pseudo relevant documents. As another example, there are also other semi-supervised methods being proposed to determine the manifold structure of data, we may compare these methods in document re-ordering. In addition, we may apply this method to other data sets as well.

5. REFERENCES

[1] Balinski, J., Danilowicz, C. 2005. *Re-ranking Method Based on Inter-document Distance*. Information Processing and Management 41(2005) 759-775.

[2] Bear J., Israel, D., Petit J., Martin D. *Using Information Extraction to Improve Document Retrieval*. Proceedings of the Sixth Text Retrieval Conference. 1997.

[3] Belkin, M., & Niyogi, P.. 2002. Using Manifold Structure for Partially Labeled Classification. Advances in Neural Information Processing Systems 15.

[4] Crouch, C., Crouch, D., Chen, Q. and Holtz, S. 2002. Improving the Retrieval Effectiveness of Very Short Queries. Information Processing and Management, 38(2002).

[5] Diaz, F., *Regularizing Ad Hoc Retrieval Scores.* In the Proceedings of the Fourteenth International Conference on Information and Knowledge Management (CIKM), 2005.

[6] Kamps, J. 2004. Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary. The 21th European Conference on In-formation Retrieval.

[7] Kurland O., Lee L. 2005. PageRank without Hyper-links: Structural Re-ranking using Links Induced by Language models. In the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[8] Lee K., Park Y., Choi, K.S. 2001. Document Re-ranking Model Using Clusters. Information Processing and Management. V. 37 n.1, p1-14.

[9] Lin, J. 1991. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory, 37:1, 145–150.

[10] Liu, X.Y and Croft W.B, 2004. Cluster Based Retrieval Using Language Models. In Proceedings of SIGIR, pp. 186-193.

[11] Luk, R. W. P., Wong, K. F. 2004. Pseudo-Relevance Feedback and Title Re-Ranking for Chinese IR. In Proceedings of NTCIR Workshop 4.

[12] M. Mitra., A. Singhal. and C. Buckley. 1998. Improving Automatic Query Expansion. In Proc. ACM SIGIR'98.

[13] Niu Z.Y., Ji D.H., and Tan C.L. 2005. 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.

[14] Niu Z.Y., Ji D.H., and Tan C.L. 2004. 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.

[15] Qu, Y.L., Xu, G.W.,Wang J. 2000. Rerank Method Based on Individual Thesaurus. Proceedings of NTCIR2 Workshop.

[16] Szummer, M., & Jaakkola, T.. 2001. Partially Labeled Classification with Markov Random Walks. Advances in Neural Information Processing Systems 14.

[17] Xu J., Croft, W.B. 1996. Query Expansion Using Local and Global Document Analysis. In Proc. ACM SIGIR'96.

[18] Xu J., Croft, W.B. 2000. Improving the Effectiveness of Information Retrieval with Local Context Analysis. ACM Transactions on Information Systems, 18(1):79--112, 2000.

[19] Yang L.P., Ji D.H. 2005(a). Chinese Information Retrieval Based on Terms and Relevant terms. ACM Transactions on Asian Language Information Processing. Vol. 4, Issue 3 (2005). pp. 357-374.

[20] Yang L.P. Ji D.H. and Leong M.K. 2005(b). Chinese Document Re-ranking Based on Term Distribution and Maximal Marginal Relevance. Second Asia Information Retrieval Symposium (AIRS 2005). LNCS 3689, Pp. 299-311.

[21] Zhang B.Y, .Li H., Liu Y., Ji L., Xi W., Fan W., Chen Z., Ma W. 2005. Improving Search Results using Affinity Graph. In the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

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

[23] Zhu, X., Ghahramani, Z., & Lafferty, J. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In Proceedings of the 20th International Conference on Machine Learning.