On the Robustness of Document Re-Ranking Techniques: A Comparison of Label Propagation, KNN, and Relevance Feedback

Yuen-Hsien Tseng, Chen-Yang Tsai*, and Ze-Jing, Chuang**
National Taiwan Normal University, Taipei, Taiwan, R.O.C., 106
samtseng@ntnu.edu.tw
*Fu Jen Catholic University, Taipei, Taiwan, R.O.C., 242
otto4321a@yahoo.com.tw
**WebGenie Information LTD., Taipei, Taiwan, R.O.C., 231
bala@webgenie.com.tw

Abstract

This paper describes our work at the sixth NTCIR workshop on the subtask of C-C single language information retrieval. We compared label propagation (LP), K-nearest neighboring (KNN), and relevance feedback (RF) for document re-ranking and found that RF is a more robust technique for performance improvement, while LP and KNN are sensitive to the choice and the number of relevant documents for successful document re-ranking.

Keywords:

Chinese IR, relevance feedback, unsupervised clustering, semi-supervised categorization.

1. Introduction

From the experience gained from participating in the past NTCIR workshops, we learn that the major factors that affect retrieval effectiveness are: indexing schemes, retrieval models, query expansion techniques, and document re-ranking methods. Top-performing systems often used sophisticated techniques such as pseudo relevance feedback (PRF), fine-tuned probabilistic or language retrieval model, hybrid term indexing, and document clustering for re-ranking. However, we have noticed that some techniques reported successful in one's implementation maybe fail in another's try. For example, in [1] document clustering for re-ranking can improve performance, while in [2] the improvement does not show.

This leads to the concept of robustness, by which we mean if a technique is robust, it is not sensitive in performance change by different implementation or parameter tuning.

Having this idea in mind, this year we explore document re-ranking techniques to understand their strength and weakness, especially their robustness.

2. Document Re-ranking

Document re-ranking (DR) is an idea to re-order the initial retrieved documents for better result, based on the information manifested in the retrieved set. Various methods can be used to make use of this information, such as unsupervised document clustering, semi-supervised document categorization, relevance feedback, or a combination of them. The cluster hypothesis states that relevant documents tend to be more similar to each other than to non-relevant documents [3], so that clustering the retrieved documents may further separate the relevant from the irrelevant. Furthermore, if some few "relevant" documents can be known from the initial retrieved set, either a supervised document categorization method can be used to classify the remaining documents into relevant and irrelevant classes, or the "relevant" documents can be feedback to the system for query term re-weighting or re-formulation. In many cases, we can assume those top-ranked documents to be the few "relevant" documents for classification or feedback. This makes automatic document re-ranking possible.

Since there are various re-ranking techniques available, this report focuses on a semi-supervised method, called *label propagation* (LP). As reported in [4], LP has higher performance than the other approaches, such as affinity graph-based method [5], structural re-ranking method [6], and maximal marginal relevance method [7]. We implemented the LP algorithm described in [4] and [8]. To know the performance level of our LP implementation, we also implemented the *K-Nearest Neighbor* (KNN) method and *pseudo relevance feedback* (PRF) for comparison, in the subtask of Chinese-to-Chinese Single Language Information Retrieval (C-C SLIR).

3. Label Propagation, KNN, and PRF

LP is a sort of semi-supervised learning algorithm which propagates the labels of a few known items to the unknown ones by exploiting the similarities among all the items. It can be simply considered as a kind of KNN method which labels the unknown item with the labels of its nearest known items. However, LP may break the nearest rule when a set of unknown item are close enough to each others, i.e., unknown items in the high-density area tend to converge to the same labels irrespective of each individual's nearest known neighbors.

In the LP algorithm for document re-ranking, assume there is a set of m documents in which the relevance of l documents ($l<<m$) are known (both relevant and irrelevant) and the rest m-l documents (which are to be re-ranked) are unknown. Let S be an mxm matrix having the similarities of any two documents in the set and Y be an mx2 matrix containing the label probability of each document to each of the two relevance classes: relevant and irrelevant. The initial value of the element y_{ij} of Y for $i<=l$ is as follows:

y_{ij} = 1 if (the ith-document is relevant and j=1) or (the ith-document is irrelevant and j=2);
y_{ij} = 0 otherwise.

For the cases where $i>l$, y_{ij} can be all zeroes. However, the use of the retrieved relevance score will make LP converge faster. In other words, we can initially set:

$y_{ij} = r_i$ if j=1, $i>l$
$y_{ij} = 1-r_i$ if j=2, $i>l$

where r_i is normalized relevance score of document i.

To apply the LP algorithm, the similarity matrix S needs to be transformed into a transition matrix T, as follows:

$$t_{ij} = \frac{s_{ij}}{s_j}, s_j = \sum_{k=1}^{m} s_{kj}$$

(Note: the above column normalization is used in [4] and [8]. However, we think that it should be row normalization for the LP algorithm to be physically correct. Since the similarity matrix we used is symmetric, it does not matter if the column normalization is used or the row one is applied.)

With these two matrices T and Y, the LP algorithm can be expressed as in Figure 1. The first column of the last m-l rows in the final Y(t) contains the relevance probability, which can be used to re-rank the unlabeled documents.

The choice of the l labeled documents is crucial to the success of the LP algorithm. Like the choice in [4], for irrelevant documents we choose the last n (normally n=5) ones from the initial retrieved set. For the relevant ones, the query string is considered as a candidate for relevant documents and is clustered with the top 10 documents based on the complete-link method. We then choose those documents belonging to the clusters with high intra-cluster similarities. We set a rule to choose at least 5 documents as relevant training items, because [4] suggested that too few relevant documents do not improve performance.

Under the above formulation, the KNN method can be expressed in a similar way, as shown in Figure 2. With the last m-l rows of the initial Y being all zeroes, the T*Y matrix multiplication computes the sum of the similarity-weighted label probabilities of the known items. That is, for $i>l$:

$$y_{ij} = \sum_{k=1}^{l} t_{ik} y_{kj}$$

where t_{ik} is the normalized similarity between document i and k. From Figure 1 and 2, we know that KNN is just a non-iterative variation of LP under the initial condition of y_{ij} being all zeroes for $i>l$.

Input: T, Y(t=0), l, and Error_threshold
Repeat
1. Y(t+1) = T * Y(t)
2. Normalize the rows of Y(t+1) such that the probability distribution is kept, i.e.,

$$y_{ij} = \frac{y_{ij}}{y_j}, y_j = \sum_{k=1}^{m} y_{ik}$$

3. Set the first l rows of Y(t) to those of Y(0)
4. Error = |Y(t+1) - Y(t)|
5. Set the last m-l rows of Y(t) to those of Y(t+1)

Until Error < Error_threshold
Ouput: Y(t) // the last m-l rows are the answers

Figure 1. The label propagation algorithm.

Input: T, Y // The last m-l rows of Y are all zeroes
Perform :
1. Y = T * Y
2. Normalize the rows of Y such that the probability distribution is kept, i.e.,

$$y_{ij} = \frac{y_{ij}}{sy_i}, sy_i = \sum_{k=1}^{n} y_{ik}$$

Ouput: Y // the last m-l rows are the answers

Figure 2. The KNN algorithm.

As to the PRF, fifteen best terms from six top-ranked documents retrieved by the initial query were used. These six documents were first concatenated into one text string and then the keyword extraction algorithm [9] was applied to extract maximally repeated patterns. The extracted patterns were filtered by some stop words and then sorted in decreasing order of occurrence. The first 15 terms were added to the initial query for the second

run of document retrieval. The decision on the number of best terms and the number of top-ranked documents was quite arbitrarily and was based on our impression on others' implementation.

4. Experiment Results

The above techniques were evaluated on three Chinese collections from FJU SCRC, NTCIR-3, and NTCIR-6. The FJU SCRC is a Chinese collection containing OCR converted texts. It is well documented in [10] and is freely available at [11]. Some statistics about these collections are shown in Table 1. The last three rows show the average, maximum, and minimum numbers of (relaxed) relevant documents for all the topics.

Our system used 1-grams, bi-grams, dictionary words, and key-phrases extracted by the algorithm [9] as the index terms. To search the noisy FJU SCRC collections, the query strings were segmented with all these index terms. However, 1-grams and bi-grams are not used in query segmentation when searching the other two collections.

The BM11 probability retrieval model was used as our baseline, as it showed good results in our past evaluation [12-14]. The NAP (Non-interpolated Average Precision) measures of the relaxed relevance judgment for all runs are reported in Table 2.

There are four sets of results in Table 2, each separated by a bold underline. The first set compares the effectiveness among the four techniques, i.e., the BM11 baseline, the BM11 baseline plus document re-ranking with the KNN method, the BM11 baseline plus DR with LP, and the BM11 baseline with PRF. Here the four techniques are all automatic, meaning that no manual feedback is involved. Note: for the BM11+KNN and BM11+LP runs, we only re-ranked the top 40 documents, as [4] showed that at this number the performance of LP re-ranking has improved.

However, our results show that neither KNN nor LP improves the performance on any of the three collections and both are inferior to PRF. This may attribute to our inferior approach in selecting the relevant documents for label propagation.

To see how the performance changes when true relevant documents are selected, we conducted more experiments by supplying r true relevant documents and n true irrelevant documents (both from the relevance judgment file) to the KNN and LP methods, where r and n both range from 1 to 5, with other parameters and implementation unchanged. The results are in the second and third sets in Table 2, where the number in the parenthesis of the RunID inidicates the value of r and n. For comparison, we also supplied PRF with at most r true relevant documents from the top forty in the initial ranked list. If no any relevant item occurs in the top forty, the top r items are used to fill the gap. The results of this true relevance feedback (TRF) are in the fourth set.

As can be seen, only one true relevant item does not help for both KNN and LP. In contrast, only one true relevant document is enough to consistently boost the performance for relevance feedback. This may explain why PRF is so robust, while KNN and LP are not. Another observation is that the advantage of more iterative label propagation computation in the LP algorithm is not clear, as it only outperforms KNN on the FJU SCRC collection. On the other two collections, the performance difference is not significant.

Table 1. Statistics of the three test collections.

	FJU SCRC	NTCIR-3	NTCIR-6
Sources	News from Mainland China, Hong Kong, and Taiwan	News from Taiwan news agencies	News from Taiwan news agencies
Year range	1950-1976	1998-1999	2000-2001
Documents	8438	381679	901446
Topics	30	42	50
Field used	Title	Description	Title
Avg. Rel.	30.03	64.39	39.68
Max. Rel.	125	249	400
Min. Rel	4	6	15

Table 2. Performance (relax) of different runs.

RunID	FJU SCRC	NTCIR-3	NTCIR-6
Bm11	0.4436	0.2335	0.2608
Bm11+KNN(a)	0.3921	0.2206	0.2455
Bm11+LP(a)	0.3826	0.2219	0.2493
Bm11+PRF*	0.4674	0.3017	0.3103
Bm11+KNN(1)	0.3798	0.2433	0.2524
Bm11+KNN(2)	0.4607	0.2644	0.2731
Bm11+KNN(3)	0.4836	0.2817	0.2915
Bm11+KNN(4)	0.5126	0.2937	0.3022
Bm11+KNN(5)	0.5360	0.3026	0.3088
Bm11+LP(1)	0.4004	0.2490	0.2656
Bm11+LP(2)	0.4807	0.2721	0.2837
Bm11+LP(3)	0.5073	0.2826	0.2982
Bm11+LP(4)	0.5317	0.2970	0.3056
Bm11+LP(5)	0.5493	0.3068	0.3123
Bm11+TRF(1)	0.5016	0.3314	0.3279
Bm11+TRF(2)	0.5421	0.3569	0.3507
Bm11+TRF(3)	0.5542	0.3747	0.3617
Bm11+TRF(4)	0.5603	0.3779	0.3695
Bm11+TRF(5)	0.5559	0.3860	0.3722

* Our official run used this option. However, due to a bug in our output, our results are excluded from evaluation. Thus all runs in this table are done after submission due.

5. Conclusions

Observing the phenomenon of inconsistent reports for the same kind of IR techniques, we evaluated the robustness of some document re-ranking techniques. Our results show that PRF consistently helps in effectiveness on different collections. It is thus safe to say that PRF is a robust technique to improve performance while KNN and LP remain to be validated. However, our results also show that there is much room to improve the PRF technique or the baseline technique, as there is a big performance difference in the PRF and TRF.

Acknowledge

This work is supported in part by NSC under the grant numbers: NSC 95-2221-E-003-016- and NSC 95-2524-S-003-012-.

References

[1] A. Tombros, R. Villa and C. J. Van Rijsbergen (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. Information Processing and Management, 38, pp. 559-582.

[2] M.-H. Hsu, H.-H. Chen, “National Taiwan University at Terabyte Track of TREC 2005”, Proceedings of the Fourteenth Text REtrieval Conference, 2005.

[3] Marti A. Hearst, Jan O. Pedersen, “Reexamining the cluster hypothesis: scatter/gather on retrieval results”, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '96, pp. 76 – 84.

[4] Lingpeng Yang, Donghong Ji, Guodong Zhou, Yu Nie, Guozheng Xiao, “Document re-ranking using cluster validation and label propagation” Proceedings of the 15th ACM international conference on Information and knowledge management CIKM '06, pp. 690 – 697.

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

[6] 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.

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

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

[9] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[10] Yuen-Hsien Tseng and Douglas W. Oard, "Document Image Retrieval Techniques for Chinese" Proceedings of the Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 151-158.

[11] Yuen-Hsien Tseng, “FJU Test Collection for Evaluation of Chinese OCR Text Retrieval”, http://blue.lins.fju.edu.tw/~tseng/Collections/Chinese_OCR_IR.html, Sep. 20, 2002.

[12] Da-Wei Juang and Yuen-Hsien Tseng, "Uniform Indexing and Retrieval Scheme for Chinese, Japanese, and Korean," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.137-141.

[13] Yuen-Hsien Tseng, Da-Wei Juang and, Shiu-Han Chen "Global and Local Term Expansion for Text Retrieval," Proceedings of the Fourth NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, June 2-4, 2004, Tokyo, Japan.

[14] Yuen-Hsien Tseng, Yu-Chin Tsai, and Chi-Jen Lin "Comparison of Global Term Expansion Methods for Text Retrieval," Proceedings of the Fifth NTCIR Workshop on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Dec 6-9, 2005, Tokyo, Japan, pp. 150-155.