Clustering and OCCC approaches in Document Re-ranking

Chong Teng[1] , Yanxiang He[1], Donghong Ji[1], Yixuan Geng[2], Zhewei Mai[2], Guimin Lin[2]
1 School of Computer Science, Wuhan University, Wuhan 430072, China
2 International School of Software, Wuhan University, Wuhan 430072, China
Email: tengchong@whu.edu.cn

Abstract

In this paper, we describe our approach for information retrieval for question answering (IR4QA) of NTCIR-8 tasks. For improving information retrieval performance, we focus mostly on the document re-ranking technique, which locates between the first retrieval documents and query expansion. In this paper, we employ two approaches in document re-ranking. One is based on entropy clustering, a kind of unsupervised learning technology. Relevant documents from top initial retrieval result can be automatically clustered same class according to information entropy values. That is a continuation of our previous work. The other is One Class Co-Clustering (OCCC) approach. it aims to detect topical terms, and compute document's topicality score. The method is simple and performs well. The experiment result shows using the two approaches in Document Re-ranking, Clustering and OCCC, can improve information retrieval performance.

Keywords: *NTCIR, Document re-ranking, Entropy Clustering, OCCC*

1. Introduction

We participated in the IR4QA (Information Retrieval for Question Answering) task of the ACLIA (Advanced Cross-lingual Information Access) task in NTCIR-8. Our research focuses on the document re-ranking technology in information retrieval. We use both bi-gram and single Chinese Character as index units and OKAPI BM25 as retrieval model. We apply document re-ranking technology between the initial retrieval and the query expansion phrase. We test two document re-ranking methods on the NTCIR-8 data collection and submit five runs.

The rest of this paper is organized as following. In section 2, we briefly describe the initial retrieval and query expansion phrases of our system. In section 3, the document re-ranking with entropy clustering approach is discussed. In section 4, we introduce another re-ranking method, namely, OCCC. The performance of our methods on NTCIR-8 is evaluated and analyzed in section 5. In section 6, we draw conclusions and present future work.

2. Initial Retrieval and Query Expansion

Like our previous work, we use Okapi BM25 model [7] as the retrieval model in initial retrieval, with documents indexed by bi-gram and single Chinese character. 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.

We use re-ranked retrieved documents to do query expansion, and use Robertson's RSV scheme [6] to select 200 bi-grams or single Chinese characters from top 20 re-ranked documents. We also make use of Rocchio's [5] formula, as improved by Salton and Buckley [3] to perform query expansion. The new query is retrieved again to get the final result.

3. Document Re-Ranking With Entropy Clustering Approach

Since no labeled relevant or irrelevant documents are generally available in information retrieval system. Our approaches try to explore relevant document. Firstly, we regard top documents after initial retrieval as relevant ones, or likely relevant ones. The hypothesis result in some noises, which will be carried into relevant documents, and affect next work. It is one of our previous researches.

In this paper, we make use of entropy clustering approach to cluster relevant documents, which aims to

improve the pseudo-labeled document construction process. Entropy clustering approach merges elements with high similarity into the same group and separates elements having low similarity to different groups. The value of within-cluster entropy determines whether two classes should be merged, and the value of between-cluster entropy determines how many clusters are reasonable. In essence, entropy clustering is a kind of unsupervised learning approach. That is to say, this process is conducted with providing training data. We focus on documents in initial retrieval result for the clustering strategy.

In Entropy clustering Approach, two definitions must be clarified: within-cluster entropy and between-cluster entropy. According to Robert Jenssen [4], within-cluster entropy is the entropy calculated based on points in the same cluster and is computed for each cluster respectively; between-cluster entropy is used to calculate the double sum runs over all data points. We use within-cluster entropy to evaluate the consistency of information within a single cluster and use between-cluster entropy to evaluate difference between two clusters. Within-cluster entropy[1] is represented as $H(C_k)$, while between-cluster Entropy between any two classes C_i and C_j is represented as $H(C_i,C_j)$. they are defined in (4) and (5).

$$H(C_k) = -\log \int f^2(x)dx$$
$$= -\log \frac{1}{(2\pi)^n k^2 h^{2n}} \sum_{i=1}^{k}\sum_{j=1}^{k} \exp\left(-\frac{\sum_{p=1}^{n}(w_{ip}-w_{jp})^2}{2h^2}\right) \quad (4)$$

$$H(C_i,C_j) = -\log \frac{1}{(2\pi)^n 2k^2 h^{2n}} \sum_{i=1}^{k}\sum_{j=1}^{k} M(d_i,d_j) \exp\left(-\frac{\sum_{p=1}^{n}(w_{ip}-w_{jp})^2}{2h^2}\right)$$

$$M(d_i,d_j) = \begin{cases} 1, d \in C_i, d \in C_j \quad or \quad d_j \in C_i, d \in C_j \\ 0, \quad other \end{cases} \quad (5)$$

Where h is a parameter which is set as 0.5. C_k is a document cluster including top k documents from first retrieval, which is defined as $C_k=\{d_1,d_2,\ldots,d_k\}$, $C_K \in R^k$. R^k is k dimensional vector space. d_i ($1 \le i \le k$) is a document element in the document set and is represented by key terms extracted from itself, which is which is denoted respectively as $d_i=\{w_1,w_2,\ldots,w_n\}$, $d \in R_n$, meaning n dimensional vector space composed by weights, which is measured by TF·IDF scheme. w_{ip} and w_{jp} is the weight of the pth term of document d_i and d_j respectively.

The whole clustering process can be formulated as follows:

1) Treat each document as an initial class for clustering.

2) With respect to any two classes, assuming merging them together, records the difference in within-cluster entropy, which is indicated by *f (Ci,Cj)*.

$f(C_i,C_j)$ is a evaluation function to calculate change in within-cluster entropy as follows:

$$f(C_i,C_j) = H(C_i \cup C_j) - H(C_i) - H(C_j) \quad (6)$$

3) Find out the smallest value of $f(C_i,C_j)$, and then merge corresponding C_i and C_j to one new class.

4) Go back to step 2) and repeat the process until all the documents are clustered into one.

During the clustering process, the between-class entropy between any two classes C_i and C_j is calculated and the minimum one is recorded. A record containing all minimum between-class entropy for each cluster step is generated after completing the whole process.

Input:
rankDocs, initial retrieval ranking of documents
K, number of documents in the top document set
Output:
re-rank, re-ranking of documents

```
for i=1:k
        topDocs[i]= rankDocs[i];
end
for i=1:k
        calculate f(Ci,Cj);
        merge clusters;
        clusters[i][]=update clusters;
        betweenEntropys=calculate H(Ci,Cj)
        minEntropys[i]=min(betweenEntropys);
end
        reasonableStep=max(minEntropys);
for i=1: length of clusters[reasonableStep]
        similarity(clusters[reasonableStep][i],
query);
        update maxSimilar;
end
coreCluster= clusters[reasonableStep][ maxSimilar];
foreach document in coreCluster
        extract 5 key terms;
        update keys[];
end
pseudoDoc=constructPseudoDoc(keys[]);
foreach document in rank
        score=similarity(document, pseudoDoc);
        update scores[];
end
re-rank=sortBy(rank, scores[]);
```

Figure 1 Re-ranking with entropy clustering

Finally, reasonable clustering result can be found in the position where there exists a significant value of $H(C_i,C_j)$. Because according to the understanding in information entropy, a combination of text information experiences a transform from disorder to order at that position. That is, documents which have high similarity are clustered into together while others which have low similarity are divided into different clusters.

Based on information entropy clustering technology, clustering result is just an intermediate stage during document re-ranking process. What to do next is to find a cluster from clustering result which is most similar with query to construct pseudo labeled document (Figure 1). The pseudo-labeled document represents

relevant information which matches user's query. Then compare such pseudo-labeled document with all the documents for retrieval as our previous methods [2]. At last, the output of document re-ranking is ranged according to their relevance between document and user query.

4. One Class Co-Clustering (OCCC) Re-ranking

Intuitively, more topic related terms a document has, more relevant to the topic it is. Based on the assumption, Ron Bekkerman el al. [9] proposed an OCCC (One-Class Co-Clustering) method to retrieve more topic-related documents. One step of our document re-ranking method is inspired from the algorithm.

Assuming that the top 20 documents are relevant documents and the top 1000 documents are general Chinese, we firstly extract all terms, including both bi-grams and single Chinese characters, from top 20 documents from first retrieval to construct a term set (S_w). And score the topicality of every term in S_w by adopt the function $\gamma(w) = \frac{P_l(w)}{P_g(w)}$, where $P_l(w)$ is w's occurrence probability in the top 20 documents from first retrieval (i.e. w's TF in top 20 documents divided by the sum of the sum of all terms' TF in top 20 documents), and $P_g(w)$ is w's occurrence probability in top 1000 documents from first retrieval (i.e. w's TF in top 1000 documents divided by the sum of all terms' TF in the top 1000 documents). Then, we represent top 1000 documents as bags-of-words (BOW) over S_w. At last, we apply Max-KL Algorithm, a simple information theoretic algorithm, for score the topicality of top K documents which will be re-ranked. The KL distance is defined as:

$$\mathrm{KL}(d_i) = \sum_{w \in S_w} \frac{P_d(d_i, w)}{\sum_{w \in S_w} P_d(d_i, w)} \log \frac{P_l(w)}{P_g(w)} \qquad (7)$$

Where $P_d(d, w)$ is w's occurrence probability in the document d (i.e. w's TF in the document divided by the length of the document).

5. Evaluation

We submitted five runs to NTCIR8: WHUCC-CS-CS-01-T, WHUCC-CS-CS-02-T, WHUCC-EN-CS-01-T, WHUCC-EN-CS-02-T, WHUCC-EN-CS-03-T.

Table 1 list statistical results of MAP (Mean Average Precision), MQ (Mean Q-measure), MnDCG (Mean Normalized Discounted Cumulated Gain) for all query topics [8].

Table 1. statistical results

lan	meas	min	max	ave	WHUCC
CS-CS	MAP	0.2647	0.4407	0.3641	0.3964
	MQ	0.2871	0.4779	0.3997	0.4323
	MnDCG	0.4819	0.6761	0.6121	0.6433
EN-CS	MAP	0.1036	0.4139	0.2748	0.4139
	MQ	0.1204	0.4499	0.3069	0.4499
	MnDCG	0.2810	0.6509	0.5073	0.6509

Row [CS-CS] represents simplified Chinese to simplified Chinese, row [EN-CS] represents English to simplified Chinese. Column [min] represents the minimum among all participants, column [max] represents the maximum among all participants, column [ave] represents the average of all participants, and column [WHUCC] represents our group's best result.

From statistical results, for CS-CS run, our group achieves 0.3964, 0.4323 and 0.6433 based MAP, MQ and MnDCG respectively; for EN-CS run, our group achieves the best evaluation result among all teams, which query employs the keywords of QA analysis result in APQA-EN-CS-01-T.

For CS-CS run, Topic 0048 and 0049 get better evaluation. However, we find we get poor results on individual query topics, such as topic 0062 and topic 0030. We list these three query topics as following:

<TOPIC ID="ACLIA2-CS-0048">
<QUESTION LANG="CS">
<![CDATA[谁是陈信安？]]>
</QUESTION>
<NARRATIVE LANG="CS">
<![CDATA[使用者想知道陈信安可能有的称号。]]>
</NARRATIVE>
</TOPIC>

<TOPIC ID="ACLIA2-CS-0049">
<QUESTION LANG="CS">
<![CDATA[谁是龙应台？]]>
</QUESTION>
<NARRATIVE LANG="CS">
<![CDATA[使用者想知道龙应台可能有的称号或担任过的职务。]]>
</NARRATIVE>
</TOPIC>

<TOPIC ID="ACLIA2-CS-0062">
<QUESTION LANG="CS">
<![CDATA[当切尼担任美国副总统时，谁是总统？]]>
</QUESTION>
<NARRATIVE LANG="CS">
<![CDATA[用户想了解在切尼担任美国副总统时的总统是谁，他的全名不需要。]]>
</NARRATIVE>

```
</TOPIC>

<TOPIC ID="ACLIA2-CS-0030">
<QUESTION LANG="CS">
 <![CDATA[ 请列出张艺谋的印象系列。]]>
</QUESTION>
 <NARRATIVE LANG="CS">
   <![CDATA[ 分析者想知道张艺谋的印象系列有哪些。]]>
 </NARRATIVE>
 </TOPIC>
```

Anlyzing the results, we find question about definition can get better results. Maybe little keywords decrease noise in retrieval procedure. However, some feature terms don't get enough weightiness when question is complex.

6. Conclusion and Future

In this paper, we introduce our approach for information retrieval system and our experience in participating in IR4QA task in NTCIR-8.

In our information retrieval system, firstly, we use both bi-grams and single characters as index units. The initial retrieval generates ordering 1000 documents. Secondly, we focus document re-ranking technique which it is implemented between the first retrieval and query expansion. We attempt two methods: Entropy Clustering Approach and One Class Co-Clustering to improving re-ranking. Lastly, we use re-ranked retrieved documents to do query expansion.

The evaluation results show proper re-ranking technology can explore more relevant information for unlabeled document, and improve the precision of retrieval system. In future, we will find an semi-supervised learning approach for improving information retrieval system greatly.

Acknowledge

This paper is supported by the National Nature Science Foundation of China (No. 60773011 and No. 90820005) and "985" 2nd phase development project (No. 985yk001 and 985yk004). Thanks to NTCIR and LDC for offering experiment dataset.

References

[1] A. Renyi, *On Measures of Entropy and Information*, Fourth Berkeley Symposium on Mnrhemoiicol Simistics and Probability, 1960, pp. 547-561.

[2]Chong Teng, et al. Yanxiang He, Donghong Ji, Guiming Lin, Zhewei Mai. *A Study on Pseudo Labeled Document Constructed for Document Re-ranking*, proceedings of 2009 International Conference on Artificial Intelligence and Computational Intelligence (AICI'09), shanghai during 7-8, Nov. 2009:377-380

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

[4] J. Robert, et al. *Clustering using Renyi's entropy,* Proceedings of the International Joint Conference, July 2003, Vol. 1, pp. 523-528.

[5] 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

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

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

[8] Tetsuya Sakai, Hideki Shima, Noriko Kando, Ruihua Song, Chuan-Jie Lin, Teruko Mitamura, Miho Sugimoto, *Overview of the NTCIR-8 ACLIA IR4QA*

[9] R. Bekkerman, K. Crammer. *One-Class Clustering in the Text Domain*. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 41–50, Oct, 2008.