Information Retrieval Using PU Learning Based Re-ranking

Chong Teng[1,2] , Yanxiang He[2], Donghong Ji[2], Han Ren[2],Lingpeng Yang[3], Wei Xiong[4]
1 Computer Center, Wuhan University, Wuhan 430072, China
2 School of Computer Science, Wuhan University, Wuhan 430072, China
3 Snooway (Wuhan)Tech. Co. Ltd, Wuhan 430074, China
4 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) on simple Chinese language of NTCIR-7 tasks. Firstly, we use both bi-grams and single Chinese characters as index units and use OKAPI BM25 as retrieval model. Secondly, we re-rank all documents' orders for the first retrieval documents. We focus mostly on the document re-ranking technique. We address probabilistically labeling relevant degree between the first retrieval documents and query topics. In other words, we want to know the probability of a document belongs to relevance/irrelevance class. We employ PU（positive and unlabeled） learning to solve this problem, and use Bayesian classifier and EM algorithm in process of computing the probability. Consequently, those relevant documents with high probability are updated rank. Lastly, we use re-ranked retrieved documents to do query expansion. Evaluation at NTCIR-7 shows that our group achieves 0.3862 and 0.3806 MAP based on pseudo-qrels and real qrels respectively.

Keywords*: NTCIR, Chinese information retrieval, PU learning, Document re-ranking*

1. Introduction

At NTCIR-7, we participated in the IR4QA (Information Retrieval for Question Answering) task of the ACLIA (Advanced Cross-lingual Information Access) task cluster. Readers are referred to [1, 8] to get the information about NTCIR-7 and the task description in detail.

For IR4QA, we submitted two runs: WHUCC-CS-CS-01-T and WHUCC-CS-CS-02-T. In our Chinese information retrieval system, we use both bi-grams and single Chinese characters as index units. We use OKAPI BM25 as retrieval model. The initial retrieval generates ordering 1000 documents. In our paper, we focus document re-ranking technique which it is implemented between the first retrieval and query expansion. We regard document re-ranking as classification problem, and attempt to put those ideas of PU learning into re-ranking process. Lastly, we use re-ranked retrieved documents to do query expansion.

The rest of this paper is organized as following. In section 2, we describe the initial retrieval and retrieval model. In section 3, we present concrete approach in process of document re-ranking. In section 4, we describe roughly query expansion to get the final result of information retrieval system. In section 5, we evaluate the performance of our proposed method on NTCIR-7 and analyze reason for experimental results. In section 6, we present conclusion and some future work.

2. Initial retrieval and retrieval model

Firstly, we use both bi-grams and single Chinese characters as index units.

For retrieval model, we use OKAPI BM25 model[7]. For the BM25 model, the relevance between the document and the query id defined in (1)-(3).

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

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

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

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.

3. Document re-ranking using PU learning

As a middle step of information retrieval system, the goal of re-ranking is not to find all relevant documents, but to update the rank for those relevant documents. At the same time, it supports the better input for query expansion of next step.

The document re-ranking is considered as a text classification problem in our experiment. Classification is a well-studied problem in machine learning such as PU (positive and unlabeled) learning. Li et al. [9] proposed a LPLP technique (Learning from Probabilistically Labeled Positive examples) based on PU algorithm when the number of training examples in the positive set P is small. Our document re-ranking method bases on the algorithm.

The input of document re-ranking is top 1000 documents from first retrieval and queries which are with the use of answer type analysis from CCLQA task, which there are 97 topics. For 1000 documents of the initial retrieval, we regard the top 20 documents as positive set P. The remaining 980 documents compose an unlabeled set U.

We use term extraction method introduced in NTCIR-4[5]to extract key terms from top 1000 retrieved documents.

3.1. Building the pseudo labeled document

The pseudo labeled document is a document vector which is made of some representative terms and their scores. For each key term in positive set P, we compute their scores and sort by descending. The function of computing scores adopts TF· IDF idea, we intend to find those terms that frequently occur in P but seldom occur in the whole corpus P and U. In general, 5-15 key terms would be sufficient so as to lessen unnecessary noise in identifying the likely positive document from unlabeled set. So we select top 15 key terms as representative terms, and build pseudo labeled document.

All documents in unlabeled set are denoted as document vectors which are compared with pseudo relevant document vector using cosine similarity. The bigger cosine value is, the more similarity the document in U set and the pseudo relevant document. All documents in set U are sorted by similarity. If the similarity value is more 0, the document will be put into subset U1. Because these documents are similar with the pseudo relevant document, we have enough excuse to conclude they are likely positive documents.

3.2 Document re-ranking with improved PU learning

If we classify these documents according to relevant with query topic, the retrieval document will be classified as relevant (signed with "+") and irrelevant (signed with "-") document with query. However, we can not estimate for most documents they are relevant or irrelevant. We can do is, how extent these retrieval documents are relevant with query topic. Therefore, we hope to get a probability P(cj|di), cj∈{"+", "-"} for every retrieval document. P(+|di) denotes the probability that a document di belongs to relevant class, and P(-|di)) denotes the probability that a document di belongs to irrelevant class. The class with the higher probability is assigned as the class of the document di. At the same time, we may know how extent a retrieval document is relevant with query topic depending on the order of probability for one class. The naïve Bayesian method is an effective technique for text classification. Assuming that the probabilities of all terms are independent given the class, we use naïve Bayesian framework to compute P(cj|di).

```
Input :
   Q:  the query string;
   P:  the set/the number of the relevant documents
           from the document set M;
   U:  the set/the number of unlabeled document
           from the document set M;

Algorithm: PUReranking(Q, P, U)
   BEGIN
    Initial (P ,U) after the first retrieval;
    Compute Score for each key term in document
        set P;
    Using top 15 key terms to build the pseudo
        labeled document rd, and take each
        document di in U as document vectors to
        compute similarity (rd, di);
   Split set U into subset U1 and U2;
   Initial P(cj|di) which represent the probabilities of
        a document belongs to a class(+,-);

   Build an NB-C classifier C using P ,U1 ,U2 based
        on the EM algorithm;
    BEGIN DO Loop
      Compute P(cj) according to the posterior
        probability of P(cj|di);
      Compute P(ti |cj) which represent the
        probabilities of a term t when a document
        belongs to some a class;
      Compute P(cj|di) which represent the
        probabilities of a document belongs to a
        class, using the value P(cj) and P(ti |cj);
    END DO Loop when NB classifier converge;
   OUTPUT: sorting the documents di by P( + | di )
        according to the value after iterating;
  END
```

Figure 1. PUReranking algorithm in document re-ranking

The naïve Bayesian method needs a precondition in classification problem, that is, we need to know total probability distribution of each class. However, we have not it. We use Expectation-Maximization (EM) algorithm to solve the problem. The EM algorithm [2] is a popular method of iterative algorithms when the data is incomplete. It iterates over two basic steps, the Expectation step and the Maximization step. The Expectation step basically fills in the missing data, while the Maximization step estimates the parameters.

As previous presentation, we already split data set into three subsets: P, U1 and U2. The subset of U1 and U2 compose set U. Initial posterior probabilities P(cj|di) are different if documents come from different subsets. Initial posterior probabilities are shown in Table 1. If we directly use these initial values, EM algorithm will not build an accurate classifier. Then we compute prior probability P(+) and P(-), posterior probability P(t|+) and P(t|-) which represent the probabilities of a term t when a document belongs to some a class. Of course, the term t must appear in this document. These values can help to revise initial probabilities P(c|di), then iteratively employ the revised posterior probabilities to build a better NB classifier until the probabilities of documents converge. These are a process of EM algorithm.

At last, we rank the probabilities which all documents are relevant (or irrelevant) class. The ranked probabilities show how relative retrieval document and query, the output of document re-ranking is the new list of documents which are reordered by probability values.

4. Query expansion

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 [4] formula, as improved by Salton and Buckley [3] to perform query expansion. The new query is retrieved again to get the final result.

5. Evaluation

Table 1 lists statistical results by MAP(mean average precision) for 97 topics and CS (simple Chinese) runs based on two evaluation method: pseudo-qrels and real qrels. The pseudo-qrels are constructed with "majority votes", the systems are ranked by "popularity" (how closely they resemble the other systems) rather than effectiveness. The pseudo-qrels don't use any manual relevance assessments.

Table 1. MAP results

	MAP	
	pseudo-qrels	real qrels
min	0.2597	0.1117
max	0.5199	0.6337
ave	0.3888	0.4554
whucc	0.3862	0.3806

Row [min] represents the minimum among all participants, row [max] represents the maximum among all participants, row [ave] represents the average of all participants, and row [whucc] represents our group's result.

From statistical results, our group achieves 0.3862 and 0.3806 MAP, based on pseudo-qrels and real qrels respectively.

Figure 2 gives comparison of per-topic average precision. Topic 68 gets the best MAP which is 0.8824. However, we find we get poor results on individual query topics, such as topic 48, topic 78 and topic 350. We list the three query topics as following:

```
<TOPIC ID="ACLIA1-CS-T48">
- <QUESTION LANG="EN">
- <![CDATA[ List the attitudes of the leaders of other countries to the Indonesian anti-Chinese incident. ]]>
  </QUESTION>
- <QUESTION LANG="CS">
- <![CDATA[ 列举其他领导人对印尼排华事件的态度。 ]]>
  </QUESTION>
</TOPIC>

<TOPIC ID="ACLIA1-CS-T78">
- <QUESTION LANG="EN">
- <![CDATA[ List the disputes triggered by the France World Cup. ]]>
  </QUESTION>
- <QUESTION LANG="CS">
- <![CDATA[ 列举法国世界杯引发的争议。 ]]>
  </QUESTION>
- </TOPIC>

<TOPIC ID="ACLIA1-CS-T350">
- <QUESTION LANG="EN">
- <![CDATA[ What is Regenerative medicine? ]]>
  </QUESTION>
- <QUESTION LANG="CS">
- <![CDATA[ 什么是再生医学？ ]]>
  </QUESTION>
- </TOPIC>
```

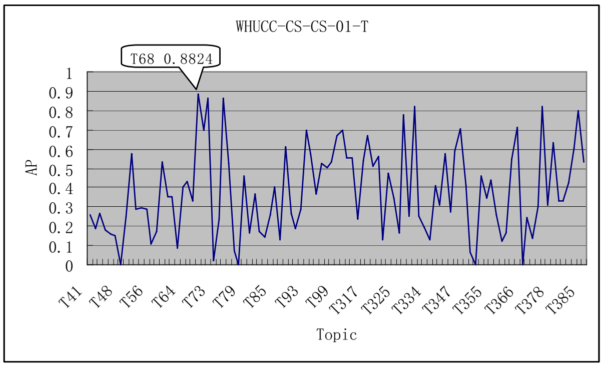

Figure 2. Per-topic average precision

Analyzing the results, we find the main reason is maybe that we don't employ different methods for different kinds of question (DEFINITION, BIOGRAPHY, RELATIONSHIP and EVENT), which causes some feature terms not getting enough recognition.

6. Conclusion and future

In this paper, we introduce our approach for Chinese information retrieval system and our experience in participating in simple Chinese IR4QA task in NTCIR-7. Our system achieves 0.3862 and 0.3806 MAP based on pseudo-qrels and real qrels respectively.

In our Chinese information retrieval system, firstly, we use both bi-grams and single Chinese 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 regard document re-ranking as classification problem, and attempt to put those ideas of PU learning into re-ranking process. Lastly, we use re-ranked retrieved documents to do query expansion.

The evaluation results show there is much room to improve the retrieval system. But we think it is valuable using both bi-grams and single Chinese characters as index units, and applying classification approach into document re-ranking. In future, we will attempt to propose a new algorithm of document re-ranking for improving simple Chinese information retrieval system.

Acknowledge

This paper is supported by the National Nature Science Foundation of China (No. 60773011 and No. 60703008) and by National Natural Science Foundation of China Youth project (No. 60803080). Thanks to NTCIR and LDC for offering experiment dataset.

References

[1] *Overview of the NTCIR-7 ACLIA: Advanced Cross-Lingual Information Access*

[2] A. P. Dempster, N. M. Laird, D. B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. 1997.

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

[4] 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

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

[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] T. Sakai, N. Kando, C. J. Lin, T. Mitamura, D. H. Ji, K. H. Chen, E. Nyberg: *Overview of the NTCIR-7 ACLIA IR4QA task*, Proceedings of NTCIR-7, to appear, 2008.

[9] X. L. Li, B. Liu, S. K. Ng, *Learning to Classify Documents with Only a Small Positive Training Set.* ECML 2007, LNAI 4701, pp. 201-213, 2007.