Redundancy Removal to Selectively Diversify Information Retrieval Results *

Xiao-Lin Wang[1,2], Hai Zhao[1,2], Bao-Liang Lu[1,2†]
[1]Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dong Chuan Rd., Shanghai 200240, China
arthur.xl.wang@gmail.com, {zhaohai; blu}@cs.sjtu.edu.cn

ABSTRACT

The document ranking subtask of NTCIR-9 Intent track is to rank retrieved documents to better satisfy users' multiple intents. We propose a redundancy removal algorithm to approach this task. The implemented system achieves average performance according to the official evaluation. Furthermore, analysis on evaluation results indicates the proposed algorithm does improve the diversity of retrieval results but slightly hurts the relevance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Performance, Experimentation

Keywords

Information Retrieval, Diversity, Redundancy Removal, BM25F

1. INTRODUCTION

In real-world applications of web page retrieval, users' queries are often ambiguous and/or underspecified [2], while traditional research has mostly focused on satisfying clearly specified information needs. NTCIR-9 Intent task held at 2010 – 2011 is to address this issue that users' ambiguous queries contain multiple intents [6]. In addition to NTCIR-9 Intent task, this issue has also attracted attentions from many other researchers [1, 7, 12].

NTCIR-9 Intent task comprises the Subtopic Mining subtask (given a query, output a ranked list of possible subtopic strings) and the Document Ranking subtask (give a query, output a ranked list of URLs that are selectively diversified) noted as Intent-DR. One idea behind Intent-DR is that when the retrieval system has no or little knowledge of the user, the best it can do is to produce output that reflects several interpretations (or intents) of such queries [11, 10]

Notetablely, NTCIR-9 Intent-DR employ large-scale datasets whose purpose might be to encourage participants to develop scalable algorithms. The dataset used in the Chinese subtask is the public dataset of SogouT, which contains 135.4 million Web pages from 5.3 million Chinese Web sites, and the total uncompressed storage size is 5.0 TBytes [1]. Processing such a large dataset itself is a meaningful challenge for most participants.

Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University (SJTU-BCMI) particiapte at NTCIR-9 Intent-DR Chinese subtask. This paper describes SJTU-BCMIąś's system and discusses its evaluation results. The rest of paper is organized as follows: the system is presented in Sec. 2; then the submitted runs are described and discussed in Sec. 3; finally this paper in concluded in Sec. 4.

2. SYSTEM DESCRIPTION

The core of SJTU-BCMI's system for NTCIR-9 Intent-DR is to employ a re-rank algorithm of redundancy removal (RedRem) to selectively introduce diversity into the normal results of web page search. Figure 1 presents its framework where RedRem is taken as a post processing after page retrieval and page rank.

The system works as follows:

1. Find the top-N relevant web pages by BM25 similarity;
2. Re-rank with page rank scores;
3. Evaluate redundancy and re-rank web pages iteratively with the RedRem Algorithm.

The following three subsections present the related details.

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 61170114 and Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), the Science and Technology Commission of Shanghai Municipality (Grant No. 09511502400), and the European Union Seventh Framework Programme (Grant No. 247619).
†Corresponding author

[1] http://www.sogou.com/labs/dl/t-e.html

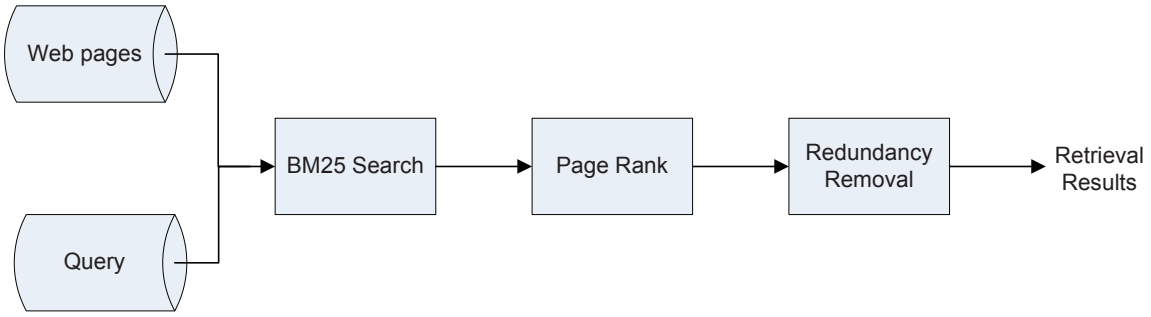

Figure 1: Framework of SJTU-BCMI's system for NTCIR-9 Intent-DR subtask

2.1 Page Retrieval with BM25/BM25F

The page retrieval component is a modified Lucene search engine. Lucene is an open source search engine powered by the Apache Software Foundation [2]. The default similarity model of Lucene is a combination of a boolean model and a vector space model [3]. Standard BM25 and BM25F similarity as 1 is implemented based on Lucene by us [8, 13], which slightly raises the accuracy according to our pilot experiments.

The implement of BM25F follows the Formula 1 presented at [5].

$$BM25F(d,q) = \sum_{t \in q} \frac{tf(t,d)}{k_1 + tf(t,d)} \cdot idf(t) \quad (1)$$

$$tf(t,d) = \sum_{c \in d} w_c * tf_c(t,d) \quad (2)$$

$$tf_c(t,d) = \frac{occur_c(t,d)}{1 - b_c + b_c \frac{l_{d,c}}{l_c}} \quad (3)$$

$$idf(t) = log \frac{N - n(t) + 0.5}{n(t) + 0.5} \quad (4)$$

where d is a document, q is a query, t represents a word, c represents a field contained in d, $occur_c(t,d)$ is the occurrences of t in c of d, $l_{d,c}$ is the length of c at d, l_c is the average length of c, N is the number of documents, n(t) is the number of documents containing t, and the rest notations of k_1,w_c and b_c are parameters. $tf_c(t,d)$ is called the field term frequency function, $tf(t,d)$ is called term frequency function, and $idf(t)$ is called inverse document frequency.

A Chinese tokenizer named Institute of Computing Technology – Chinese Lexical Analysis System(ICTCLAS) is employed to preprocess the Chinese text of both web pages and queries [14] [4]. ICTCLAS is an integrated Chinese lexical analysis system based on multi-layer HMM, including word segmentation, Part-Of-Speech tagging and unknown words recognition [5]. Following is an example of ICTCLASA:

沙拉/n 制作/v 是/vshi 将/p 各种/rz 凉/ng 透/v 了/ule 的/ude1 熟料/n 或是/c 可以/v 直接/ad 食用/v 的/ude1 生料/n 加工/v 成/v 较/d 小/a 的/ude1 形状/n。

[2] http://lucene.apache.org/

[3] http://lucene.apache.org/java/2_9_0/api/core/org/apache/lucene/search/Similarity.html

[4] http://ictclas.org/

[5] http://morphix-nlp.berlios.de/manual/node12.html

2.2 Re-rank with Page Rank Scores

SogouT-Rank is one of the resources released with the SogouT which is the data set of NTCIR-9 Intent-DR Chinese subtask [6]. SogouT-Rank contains the page rank scores for each pages in SogouT. SJTU-BCMI's system employ these scores to re-rank the result of page retrieval.

The Formula 5 is used in SJTU-BCMI's system, which is a straight combination of BM25F scores and page rank scores.

$$score(d,q) = \frac{BM25F(d,q)}{\max_{d' \in D} BM25F(d',q)} \quad (5)$$

$$+\lambda \frac{PR(d)}{\max_{d' \in D} PR(d')} \quad (6)$$

where $PR(d)$ is the page rank score of document d, D is the set of retrieved documents, and λ is a parameter.

2.3 Redundancy Removal Algorithm (RedRem)

RedRem is a iterative re-rank algorithm (see Figure 2). In each iteration, it evaluates the documents which have not been put into the final results according to their original scores and information redundancies against the final results.

A core component of RedRem is to measure the information redundancy between one document and a collection of other documents, noted as $f_{RED}(p,U)$ in Figure 2. The Formula 7 is taken as the measurement of information redundancy, which counts the overlapped and non-overlapped unigrams based on the bag-of-word model.

$$f_{RED}(d,U) = \alpha \frac{|d \cap U|}{|d|} + \beta \frac{|\{w | w \in d, w \notin U\}|}{|d|} \quad (7)$$

where α and β are two parameters, U is a set of accepted documents, $|d|$ is the number of words in d, $d \cap U$ is the number of d's words found in a member of U, and $d \cap \bar{U}$ is the number of d's words not found in any member of U.

3. EXPERIMENTS

3.1 Experimental Settings

[6] http://www.sogou.com/labs/dl/t-rank.html

Require: a list of retrieved documents $S = \{(d_i, s_i)|i = 1, \ldots, n\}$ where d_i is a document and s_i is the normalized confidence score subjected to $s_1 = 1$ and $s_i \geq s_{i+1}$.
Ensure: a re-ranked list of documents $U = \{(p_{k_j}, u_j)|j = 1, \ldots, n\}$ where d_{k_j} is the j-th page and u_j is its updated score.
$U \leftarrow \{(d_1, s_1)\}$
for all j, $j = 2, \ldots, n$ **do**
  for all t, $t = i, \ldots, n$ **do**
    u_t=$s_t - f_{RED}(d_t, U)$
  end for
  $k_j \leftarrow argmax_t(u_t)$
  add (d_{k_j}, u_j) into U
end for
return U

Figure 2: Redundancy Removal (RedRem) algorithm

Table 1: Description of SJTU-BCMI submitted runs to Intent-DR Chinese subtask

Runs	Page retrieval	Page rank?(λ)	RedRem?(α,β)
SJTUBCMI-D-C-1	BM25F[a]	Yes(0.4)	Yes(0.1,-0.9)
SJTUBCMI-D-C-2	BM25F[a]	Yes(0.4)	No
SJTUBCMI-D-C-3	BM25F[a]	No	No
SJTUBCMI-D-C-4	BM25[b]	Yes(0.4)	Yes(0.1, -0.9)
SJTUBCMI-D-C-5	BM25[b]	No	Yes(0,-0.9)

[a] The parameters are k_1=4, b_{title}=0.25, w_{title}=4, b_{main}=0.25,w_{main}=1.
[b] The parameters are k_1=4, b=0.25.

Five runs of SJTU-BCMI are submitted to NTCIR-9 Intent-DR Chinese subtask. Table 1 summarizes the approach of each run. The parameters are tuned to maximum performances on the evaluation dataset. BM25F considers that a document consists of a title and a main body, which are the output of our HTML parser. In contrast, BM25 combines the title and the main body into a single piece of text.

3.2 Experimental Results

NTCIR-9 INTENT-DR employ three evaluation matrices: intent recall(I-rec), D-nDCG which measures overall relevance across intents and D#-nDCG which is a linear combination of intent recall [11, 9, 6]. Among these three matrices, D#-nDCG is taken as the primary matrices, and it is a simple average of I-rec and D-nDCG as the default setting of the NTCIREVAL [9]

The evaluation results of SJTU-BCMI's runs are presented at Table 2. On the aspect of diversity ($I-rec$), RedRem does provide higher performances as expected, as scores of RUN1, RUN4 and RUN5 are all higher than those of RUN2 and RUN3. On the aspect of relevance ($D-nDCG$), RedRem slightly hurts the performances.

4. CONCLUSIONS

In this paper, we proposed a Redundancy Removal (RedRem) method to stress the problem of intent search. The implemented system achieves average performances at NTCIR-9 Intent-DR Chinese subtask. The proposed RedRem improves the quality of search results on the aspect of intent search, but slight hurts the on the aspect of global relevance.

The core of Redundancy Removal is to measure the information redundancy between a document and a set of other documents. In the future we will put more effort on this issue. A instant attempt is to employ bag-of-ngrams model instead of bag-of-word(unigrams) model, which is a popular technique of evaluating semantic overlapping between natural language texts in such domains as machine translation and automatic summarization [4, 3]. Other attempts might be to leverage hierarchical clustering models and topic models.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] C. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings*, volume 5766, page 188. Springer-Verlag New York Inc, 2009.

[3] C. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

[4] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[5] J. Pérez-Agüera, J. Arroyo, J. Greenberg, J. Iglesias,

Table 2: Evaluation of SJTU-BCMI submitted runs to NTCIR-9 Intent-DR Chinese subtask

Run	I-rec@10	D-nDCG@10	D#-nDCG@10
SJTUBCMI-D-C-1	0.6038	0.2654	0.4346
SJTUBCMI-D-C-2	0.6008	**0.3317**	**0.4663**
SJTUBCMI-D-C-3	0.5856	0.3288	0.4572
SJTUBCMI-D-C-4	0.6108	0.2756	0.4432
SJTUBCMI-D-C-5	**0.6228**	0.2816	0.4522

and V. Fresno. Using bm25f for semantic search. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 1–8. ACM, 2010.

[6] T. S. M. K. Y. L. M. S. Q. W. R. Song, M. Zhang and N. Orii. Overview of the ntcir-9 intent task. In *NTCIR-9 Workshop Meeting*, page to appear, 2011.

[7] F. Radlinski, P. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*, volume 43, pages 46–52. ACM, 2009.

[8] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST, 1995.

[9] T. Sakai. Ntcireval: A generic toolkit for information access evaluation. In *Information Technology 2011*, page to appear, 2011.

[10] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Lin. Simple evaluation metrics for diversified search results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)*, 2010.

[11] T. Sakai and R. Song. Evaluating diversified search resultsusing per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1052, 2011.

[12] M. Sanderson. Ambiguous queries: test collections need more sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 499–506. ACM, 2008.

[13] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In *Proceedings of TREC 2004*. Citeseer, 2004.

[14] H. Zhang, H. Yu, D. Xiong, and Q. Liu. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics, 2003.