Introduction of the HTRDP Chinese IR Evaluation Task

Le Sun Junlin Zhang
Institute of Software, Chinese Academy of Sciences
P.O. Box 8718, Beijing, 100080, P. R. China
sunle@iscas.cn

Abstract

In this paper, we give a brief introduction of the HTRDP Chinese Information Retrieval Evaluation, which is sponsored by HTRDP (High Technology Research and Development Program of China, namely 863 Program). The web data collection, query design, evaluation metrics, and the evaluation procedures will presents in detail. Like TREC, NTCIR and CLEF, its purpose is to provide the infrastructure necessary for large-scale evaluation of Chinese information retrieval methodologies and to help advance the state of the art in information retrieval technology. We conclude our paper with results analysis and future works.

Keyword: *Chinese IR evaluation, web data collection, evaluation metrics.*

1 Background

The HTRDP Chinese Information Retrieval Evaluation is sponsored by HTRDP (National High Technology Research and Development Program of China, namely 863 Program).This evaluation is one of the ongoing series of evaluations of Chinese Information Processing and Intelligent Human-Machine Interface technologies[1]. Its purpose is to provide the infrastructure necessary for large-scale evaluation of Chinese information retrieval methodologies and to help advance the state of the art in information retrieval technology. When designing the evaluation the organizers take both the difficulties of the current IR technology and the characteristics of Chinese into consideration.

Each participating group conducts research and experiments using the same data provided by the IR evaluation organizer with the various models and approaches they prefer. Following TREC[2], NTCIR[3], and CLEF[4], the HTRDP evaluation of Chinese information retrieval provides test corpora (data sets usable for experiments) and unified evaluation procedures for all experiment results which are handed in by the participating groups. However, there is a difference of this evaluation at 2003 and 2004 with TREC and NTCIR that all participating groups must conduct their experiment on-site rather than off-site.

The importance of reusable large-scale standard test corpora for Chinese IR has been widely recognized in IR research community in China[5]. An evaluation workshop is hold to facilitate these IR researchers by providing a forum for research idea exchange and technology transfer.

The first HTRDP IR evaluation was performed at October, 2003. There are three groups from different universities took party in this tasks and they all submitted their results. The second HTRDP IR evaluation was performed at October, 2004. There are four groups registered for the tasks and submitted the results for one or more tasks (document task and

passage task). The third HTRDP IR evaluation was performed at October 2005 and one month later a workshop was held in Beijing and five groups submitted the work notes.

The remainder of the paper is organized as follows: firstly, we will describe the HTRDP Chinese IR evaluation task. Then the date collection and topic design are showed respectively in section 3 and 4. The result analysis is given in section 5. We conclude our paper in section 6 with future works.

2 Task Descriptions

There are two IR evaluation tasks in HTRDP at 2003. One is small corpus task (200M) and the other is large corpus (2G), all the data is collected automatically from Chinese website. After 2003, we focus on the large web corpus (about 15G). There are two evaluation tasks in 2004:

- Document level retrieval: IR system should return results on document level
- Paragraph level retrieval: IR system should return results on paragraph level. Participating groups can determine the definition and the length of a paragraph.

Participating groups can choose their interested tasks. In the 2005 IR evaluation, there is only one task, that is, **Relevant Web Page Retrieval.** In this task, IR systems are required to retrieve and rank relevant Web pages for a set of given topics from a large collection of Web pages (about 100G, see detail in section 3).

There are six steps in the 2005 IR Evaluation:

- Step 1: Register to participate. Each group desiring to participate in the Chinese web IR evaluation must register no later than the deadline for registration.
- Step 2: Receive the test collection. Evaluation Group entrusts the Computer Network and Distributed Systems Laboratory of Peking University to release the corpus, namely CWT100g.
- Step 3: Receive the topics. Evaluation Organizers will send the topics to each participant via E-mail.
- Step 4: Submit the results. Each participant should submit the results no later than the deadline for submission.
- Step 5: Receive the evaluation results. Evaluation Organizers will send the evaluation results to each participant via E-mail
- Step 6: Attend the evaluation workshop. HTDRP sponsors a follow-up evaluation workshop where evaluation participating groups and government sponsors meet to review evaluation results; share knowledge gained, and plan for the next evaluation. Evaluation participating groups are expected to describe their technology and research in detail.

3 Date Collection

The evaluation data contains only test corpus, namely CWT100g (The Chinese Web Test collection with 100 GB web pages), which is provided by the Computer Network and Distributed Systems Laboratory of Peking University [9, 10].

CWT100g consists of 5,712,710 Web pages (about 90GB in size) crawled from 17,683 websites in China in June, 2004. Every page in the collection has a "text/html" or "text/plain" MIME type returned from the corresponding HTTP server.

4 Topic Design

A topic is a statement which describes the users' need. Following TREC and NTCIR, there are four sections in the topic: an identifier, a title, a description, and a narrative. We use a standard XML format for all the topics. The encoding for Chinese character is GB2312. Three example topics are given as follows:

<topic>
<num>编号：003
<title>周杰伦演唱会
<desc>描述：周杰伦2003年9月12日北京演唱会的

相关介绍及评论
<narr>叙述：2003年9月12日，“小天王”周杰伦在北京工人体育场举办了个人演唱会。关于这次演唱会的所有报道与评论均在检索之列。周杰伦其他时间、其他地点的演唱会不在检索之列。
</topic>

<topic>
<num> 编号：017
<title>奥运会吉祥物
<desc> 描述：2008年奥运会吉祥物的设计和征集工作有关内容介绍
<narr> 叙述：奥运会吉祥物是代表奥运形象的一个重要因素，每一届奥运会吉祥物都成为当届奥运会的亮点，因此2008北京奥运会吉祥物的设计征集引起了世界的关注。查询与2008奥运会吉祥物的设计要求，设计理念或者征集方案等相关文档，对奥运会申办等其他方面的介绍文章视为无关。
</topic>

<top>
<num>编号：042
<title>治理教育乱收费
<desc>描述：介绍我国在治理教育乱收费方面采取的措施
<narr>叙述：搜索我国就治理教育乱收费采取的各方面措施，网页内容包括各个地区政府针对不同形式的教育乱收费采取的各种对策，制定的相关法律法规，对违规的学校进行的处理等内容。其他由于教育乱收费导致学生失学等社会影响问题方面的网页不在检索范围之内。
</top>

There are totally 50 topics in 2005's evaluation. It's not easy to design topics manually by different people without any rules. According to the objections of our evaluation task, we set up some rules for our topic design group as followings:

- In order to simulate the Chinese user's real needs under large-scale web collection, the length of title of each topic should not longer than 5 Chinese words and no less than 2 Chinese words. The description section of topic includes 1 or 2 Chinese sentences which describe the user's need. The narrative section of a topic shows the user's need in details and also gives the information not related to this topic.
- The content of these topics should cover as many as fields, such as, politics, economics, entertainment, culture, physical training, sciences, and so on.
- The difficulties of the topic for IR system should be carefully designed. The simple topics should no less than 50% of the whole topics.
- The relevant documents for each topic should not too much (for example above 100) and should not too little (for example less than 5).

The participating groups are permitted not only automatically to generate query from topic but also manually to generate query from topic. All the information in the topics can be used for query.

5 Results Analysis

5.1 Evaluation Metrics

There are several evaluation metrics [6, 7, 8] used in our evaluation, such as MAP, R-Precision, and P@10. Very simple descriptions of them are given in the following.

MAP (Mean Average Precision)

Average precision for a single topic is the mean of the precision after each relevant document is retrieved. *Mean Average Precision* (MAP) for a set of topics is the mean of the average precision scores for each topic. This is a single-valued measure that reflects the performance over all relevant documents. It favors systems that retrieve document quickly (highly ranked). When a relevant document is not retrieved at all, its precision is assumed to be 0.

R-Precision

The R-Precision for a single topic is the

precision after R documents have been retrieved, where R is the number of relevant documents for the topic. The average R-Precision for a set of topics is computed by taking the means of R-Precisions of the individual topics.

P@10

The P@10 for a single topic is the precision after ten documents have been returned. The P@10 for a set of topics is computed by taking the means of P@10's of the individual topics.

5.2 Relevance Judgments

During the first HTRDP evaluation in 2003 and 2004, we manual find the relevant documents for each topic with the help of some tools. However, in 2005, the test collection is about 100G, so it is impossible to do complete relevance judgments, that is, a relevance judgment decision is made for every document in the collection for each topic. Instead, in the 2005 IR evaluation we use pooling method to create a subset of the documents (the "pool") to judge for a topic. Each document in the pool for a topic is judged for relevance by the person who designed this topic. Documents that are not in the pool are assumed to be irrelevant to this topic.

The judgment pools are created step by step as follows. When participating groups submit their retrieval results to evaluation organizer in the order they prefer. Evaluation organizer chooses a number of results and merged into the pool. For each selected result, the top n documents (usually, n=100) per topic are added into the topics' pools. Since the retrieval results are ranked by decreasing relevance to the topic, so the top documents are the documents most likely to be relevant to the topic.

In the 2005 IR Evaluation, binary relevance judgment is adopted which means that a document is either relevant to a topic or not relevant. A document is relevant to a topic only if:

- The document contains information that the "title" section of the topic indicates
- The document contains appropriate information that is in accord with the topic restricts in the "desc" and "narr" sections.

5.3 Results Analysis

In 2003, there are 3 systems in our evaluation. The results are showed in table 1.

Table 1 Chinese IR Evaluation Result in 2003

	small-scale collection(F1)	large-scale collection(P@10)
System1	0.292	/
System2	0.294	0.592
System3	0.040	0.662

In 2004, these are 2 tasks, one is document retrieval and the other is passage retrieval. The results are showed in table 2 and table 3 respectively.

Table 2 Chinese IR Evaluation Results in 2004(Document level)

	Precision	Recall	F1	AP
System1	0.0833	0.0761	0.0747	0.0463
System2	0.1204	0.1711	0.1297	0.0801
System3	0.0213	0.0690	0.0312	0.0247
System4	0.0770	0.6082	0.1316	0.2383

Table 3 Chinese IR Evaluation Results in 2004(Passage level)

	Precision	Recall	F1	R-Precision
System1	0.0039	0.0418	0.0116	0.0073

In 2005, these are two groups at our evaluation. One is for the results use manually formed query, the other is for the results use automatically formed query. The results are showed in table 4 and table 5 respectively [11, 12, 13, 14, and 15].

Table 4 Chinese IR Evaluation Results in 2005(Manually query)

	MAP	R-precision	P@10
System1	0.3257	0.3826	0.5580
System2	0.1705	0.2327	0.4640
System3	0.3538	0.4078	0.6840
System4	0.2673	0.3185	0.4800
System5	0.3671	0.4140	0.7040

Table 5 Chinese IR Evaluation Results in 2005(Automatically query)

	MAP	R-precision	P@10
System1	0.2727	0.3320	0.5300
System2	0.1862	0.2554	0.5180
System3	0.3107	0.3672	0.6240
System4	0.3175	0.3605	0.5540
System5	0.2858	0.3293	0.6280

Compared with the 2003 and 2004's evaluation result, the performance of these systems attended the 2005 Chinese IR evaluation has got much increase. We thought the following factors maybe contribute to the better performance in 2005:

- Since the corpus has been expanded to 100G and much more information such as link information are provided, the participating groups could use some advanced relevant evaluation technology such as the link analysis, anchor text analysis etc, which leads to more accurate search results;
- With last year's evaluation data as a training corpus, the participating groups could make use of these training set to effectively overcome some difficulties of Chinese IR system such as Chinese name entity recognition. They can also obtain more stable and effective retrieval model by adjusting the system parameters;
- Effective use of relevant feedback and re-ranking method also help to improve the search results;

After the concrete analysis of the evaluation result, we can draw the conclusion that there are still much room for one Chinese IR system to increase performance by solving the following problems:

- Chinese segmentation error, especially the segmentation error for name entity;
- Chinese new word or abbreviations recognition;
- Mismatch of query words and document words with the same semantic meaning;
- Weight schema of query words.

5 Conclusions and Future Work

The HTRDP evaluation of Chinese Web IR, sponsored by National High Technology Research and Development Program (HTRDP, also called "863" Program), is one of the most influential IR evaluations in China. This evaluation has been hold three times at 2003, 2004 and 2005. This evaluation has enhanced the communication among IR research community in China. Through the HTRDP evaluation of Chinese web IR system, several related techniques, such as Chinese segmentation, name entity, especially designed for IR system, have drawn much attention. In the near future, we wish to simulate the real users' needs on a more large-scale collection (about 200G). The evaluation topics will choose from Chinese real users' query logs. We wish attract more national groups to attend this evaluation, and we welcome international groups to attend this evaluation, too.

6 Acknowledgements

As we known, this work is done by a team. Other researchers in this team include Prof. Qun Liu, Dr. Yang Liu and many others. With special thanks to Dr. Hongfei Yan for provided the original web data.

References

[1]LIU Qun, WANG Xiangdong, LIU Hong, SUN Le, TANG Sheng, XIONG Deyi, HOU Hongxu, LV Yuanhua, LI Wenbo, LIN Shouxun, QIAN Yueliang，Introduction to HTRDP evaluations on Chinese information processing and intelligent human-machine interface, Frontiers of Computer Sciences in China, Vol.1, No.1, Feb.2007

[2] The Text Retrieval Conference, http://trec.nist.gov/

[3] NII Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/

[4] Cross Language Evaluation Forum, http://www.clef-campaign.org/

[5] Junlin Zhang etc. Research on the 863 Chinese Information Retrieval Evaluation. Journal of Chinese Information Processing(vol.20).2006.7.

[6] Harman D. The first text retrieval conference (TREC-1). Information Processing and Management, 29(4), pp.411-414, 1993.

[7] Ellen M. Voorhees, “Overview of TREC 2005”, In Proceedings of the Text REtrieval Conference (TREC). Gaithersburg, Maryland, November 2005.

[8] Shah,C and Croft,WB, "Evaluating High Accuracy Retrieval Techniques", Proceedings of SIGIR '04, 2004.8.

[9] http://e.pku.edu.cn.

[10] H. Yan, J. Li, J. Zhu, and B. Peng, "Tianwang Search Engine at TREC 2005: Terabyte Track," in TREC 2005.

[11] Yuxin Cheng etc. 863 Web Track Experiments at ICST-PKU. Journal of Chinese Information Processing(vol.20).2006.7.

[12]Zhao Le etc. 2005 THUIR Report for 863 Information Retrieval Evaluation. Journal of Chinese Information Processing(vol.20).2006.7.

[13]Zhichang Zhang etc. Technology Report of HIT-IRLab for Evaluation 2005 of 863 Information Retrieval. Journal of Chinese Information Processing(vol.20).2006.7.

[14]Weiran Xu etc. PRIS Information Retrieval System Report. Journal of Chinese Information Processing(vol.20).2006.7.

[15]Bibo Lv etc, 863 Information Retrieval Evaluation –Institute of Automation. Journal of Chinese Information Processing(vol.20).2006.7.