Web Search Experiments Using OASIS

Vitaliy KLUEV
The Core and Information Technology Center
The University of Aizu
Tsuruga, Ikki-machi, Aizu-Wakamatsu city, Fukushima, 965-8580, Japan
vkluev@u-aizu.ac.jp

Abstract

The major aim of participating in the Web Retrieval Task was to test OASIS to see how well the system supports Japanese Web search. The 10-gigabyte data set was used in our experiments. Because OASIS is a distributed system we simulated distributing data. The collection was divided into ten independent subsets. Results showed that some improvements in the indexing strategy have to be done.
Keywords: *search engine, phrasal indexing, vector space model, full text searching.*

1 Introduction

There are dozens of powerful search tools on the net available without cost. Nevertheless, searching for information is still inefficient. How is it possible to find appropriate information easily and quickly? How can researchers and users compare systems to each other? Which methods are more preferable to design a new search system? The Web Retrieval task should help to find answers to these and other questions.

OASIS is a distributed search system in the Internet [1]. It was designed to support multiple languages. These tests checked its ability of searching for Japanese Internet data. The following main issues were considered in our experiments:

- indexing technique applied to a large data set
- result merging methods for distributed searching

The paper is organized as follows. The system description is discussed in section 2. Retrieval results are described in section 3. A failure analysis is presented in section 4. Final remarks are put into section 5.

2 System Description

Our system participated in subtask II-A1 (the topic retrieval subtask). The 10-gigabyte (small) collection was applied in the tests. Results of two official runs were submitted as an outcome of these tests. They are OASIS11 and OASIS12. The following search strategy was utilized:

- The dataset has been located in the ten different directories. They are 00,...,09. Ten independent index files according these directories were produced;
- Queries were obtained automatically from the topics provided. Their number is 47. Each query was processed twice. The first search retrieved N documents from each index independently (N=200 for run OASIS11 and N=1000 for run OASIS12). After that all retrieved documents were merged, a small collection was created and its documents were indexed. It was utilized in run OASIS11. The second search produced the final retrieval. The 1000 best results were presented in the result file. Another technique was used in run OASIS12. The basic idea according to [2] looks like this: Authors of this approach assume that document collections have been indexed using the same model and document scores obtained as a result of the search are comparable across all collections. These scores are then used to merge search results into a single list.

The OASIS server dedicated for the experiments was equipped by the following hardware: a PC compatible computer, the Intel Pentium III 1 GHz processor, Intel STL2 Server Board Motherboard, 2 GB RAM, 36GB 10,000rpm internal disk, 500GB RAID unit, SCSI Ultra Wide 68pin PIN type connector (Single ended). OS Linux 7J was running on this computer. The organizers of the Web Retrieval task provided it.

The description of the system has been presented in Table 1.

Table 1. System description

Parameter	Description
Subtask	II-A1 (the topic retrieval subtask)
Topic Part	Description
Link Info	Only content
Index Unit	Combination of bi-words and phrases. Overlapped bi-gram words were selected from every indexing document. Phrases (virtual word collocations) consisted up to 4 Japanese characters were automatically determined. Hiragana characters were used as word boundaries. Katakana sequences were considered as words. Hiragana characters were discarded.
Index Technique	We did not use any NLP technique. The dynamic window (from 2 to 4 Japanese characters) was shifted through texts.
Index Structure	Inverted index
Query Method	Automatic
Query Unit	The same as the Index Unit: combination of bi-terms and phrases.
IR Model	Vector Space Model
Ranking	TF*IDF
Query Expansion	No query expansion
Filtering	No filtering
SearchTime	about 26 hours for all queries (for run OASIS11); about 42 hours 30 minutes for all queries (for run OASIS12)

3 Retrieval Results

Figure 1 and Figure 2 present average results of the search for both runs. As we can see, the approach used in the OASIS12 run produced a slightly better outcome.

Estimations according to the DCG metric have been presented in Figure 3 and Figure 4. Score have been computed in the following manner:

- DCG[1]: (Highly relevant, Relevant, Partially relevant) = (3,2,1)
- DCG[3]: (Highly relevant, Relevant, Partially relevant) = (3,2,0)

From the aforementioned figures we can see: results are practically the same.

Because retrieval results for both runs are very close to each other, we collected in Table 2 the best (all relevant documents were retrieved) and worst (nothing relevant were produced by the system) outcome for run OASIS12 according to the "gprel" metric. The number of queries is equal to 45. Average precision exceeded 0.42 for query 58.

There is one more lesson we taught from our experiments: Row score merging produced slightly better results compared to the intermediate indexing and retrieval approach. The first search in run OASIS12 retrieved 1000 documents from each index compared to 200 documents in run OASIS11. It explains why this approach is time consuming. We gained only a little improvement in our tests.

Table 2. The best and worst retrieval results

Best	Worst
22, 27, 30, 43, 47, 57, 58, 60	14 (from 9 docs), 62 (from 3 docs)

4 Failure Analysis

Our system produced relatively low results. What is the reason? We see several roots:

- We met difficulties during indexing some documents. Several files were very large. There is only one example: File 096959 from directory 096 consists of 269914 KB. Due to some reasons (indexing method [5]), the system did not process such files. They were discarded (about 13% of the documents). As result, some relevant documents were not indexed. In connection with this issue the following note is important: Any parser which is designed to run on the entire Web must be capable of handing a huge array of possible errors. These include typos in the HTML tags, kilobytes of zeros in the middle of tags, HTML tags nested hundreds deep, etc [3].
- We paid attention to word collocations to catch a content. Sequences consisted of only one Japanese character did not indexed. The outcome of this is as follows: A lot of key words were lost.

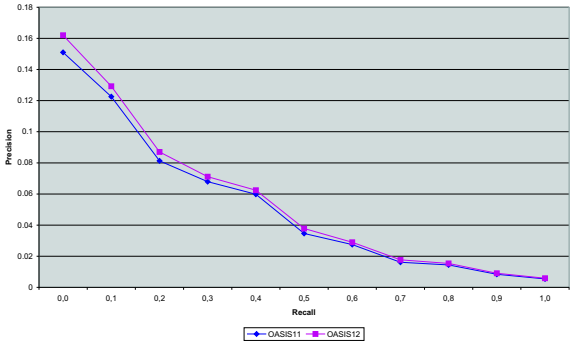

Figure 1. Retrieval results according to the "gprel" metric

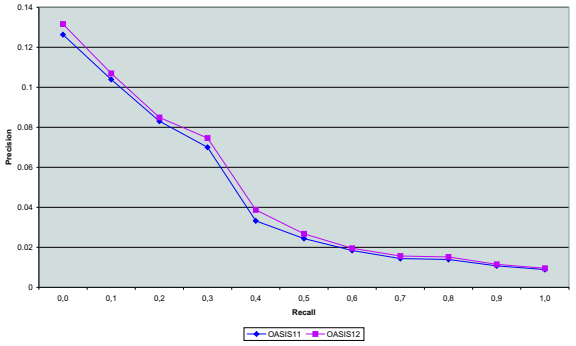

Figure 2. Retrieval results according to the "grel" metric

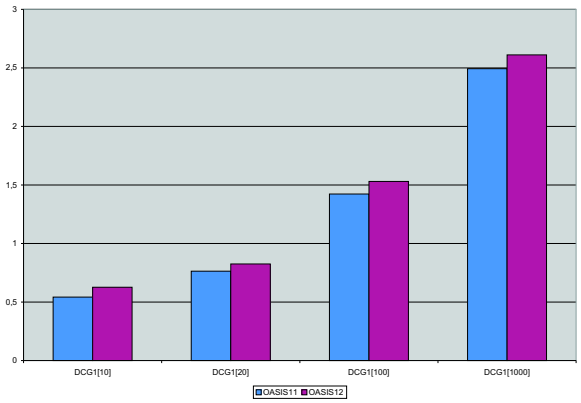

Figure 3. DCG[1] estimation

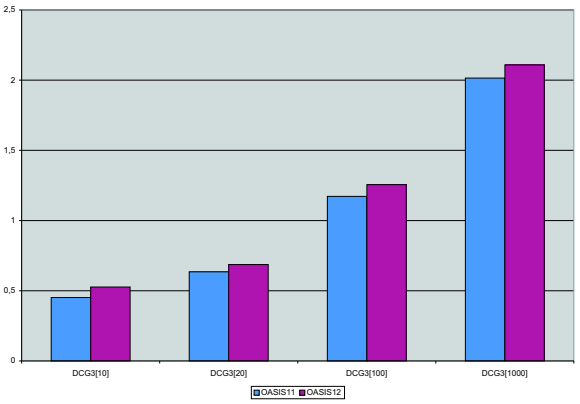

Figure 4. DCG[3] estimation

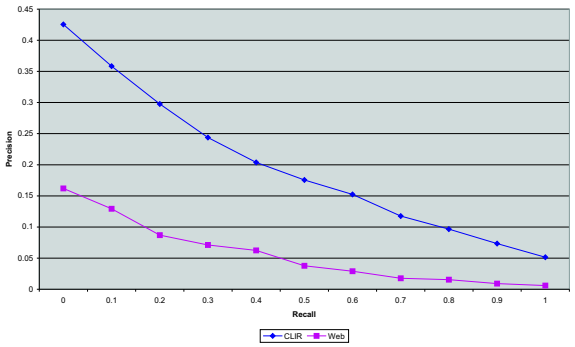

Figure 5. Web and CLIR tasks

We can compare retrieval results for CLIR [4] and Web task obtained by our system. See Figure 5. They have to be close to each other because the same indexing technique has been used. But the outcome of the Web task is worse compared to this one of the CLIR task.

5 Conclusions

Participating in the Web task showed that our system can index and retrieve the real Japanese Internet data. We used the simplest approach to index: variable-gram (two - seven). We tested two techniques to merge results retrieved from different sources. The idea behind one of them is to index intermediate data and process the second search. The second approach aims to adopt the row score method. Both techniques produced practically the same results.

Now we have a clear knowledge what should be done to improve the quality of the search.

References

[1] A. Patel, L. Petrosjan and W. Rosenstiel, editor. *OASIS: Distributed Search System in the Internet*. St. Petersburg State University Published Press, St. Petersburg, Russia, 1999. (ISBN: 5-7997-0138-0).

[2] Kwork K. L., Grunfeld L., and Lewis D. D. Trec-3 ad-hoc: Routing retrieval and thresholding experiments using pircs. In *Proceedings of TREC-3*, 1995.

[3] Sergey Brin and Lawrence Page. The anatomy of a large–scale hypertexual web search engine. In *Proceedings of 7th International World Wide Web Conference (WWW-7)*, 1998.

[4] V. Kluev. Oasis at ntcir3: Monolingual ir task. In *Proceedings of the Third NTCIR Workshop Meeting*, Tokyo, Japan, 2002.

[5] V. Kluev, M. Bessonov, and V. Dobrynin. Ntcir experiments using the oasis system. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, 2001. (ISBN: 4-92-4600-96-2).