R^2D^2 at NTCIR-4 Web Retrieval Task

Teruhito Kanazawa
KYA group Corporation
5–29–7 Koishikawa, Bunkyo-ku, Tokyo 112–0002, Japan
tkana@kyagroup.com

Tomonari Masada
University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
masada@nii.ac.jp

Atsuhiro Takasu Jun Adachi
National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan
{takasu, adachi}@nii.ac.jp

Abstract

We evaluated the Relevance-based Superimposition Model at NTCIR 4 Web task A (survey retrieval) and B (target retrieval). We developed a distributed indexing / searching engine for treating the large amount of documents in a practical processing time. Some improvements of the retrieval precisions were achieved algorithmically.

1 Introduction

We have proposed a method named the Relevance-based Superimposition (RS) model to solve the semantic ambiguity problem in information retrieval. A query usually provides only a very restricted means to represent the user's intention. Query expansion is a method for semantic disambiguation on query issuing phase. It includes index terms related to the original query expression, thus assisting novice users who have limited vocabulary in the target field. However, it is difficult to choose terms that represent the user's intention automatically and carefully. Therefore, pragmatically effective retrieval can only be achieved by adjusting many parameters depending on the database [11].

Document feature vector modification is one of the methods that use information extracted from the documents for semantic disambiguation in index generation phase. We believe it achieves higher recall without losing precision of retrieval, because documents usually have much more information than a query.

For evaluation the RS model, we have developed a retrieval system using the model, named R^2D^2(RetRieval system for Digital Documents) and have participated in the NTCIR 1 / 2 ad-hoc task [6, 9]. The major focus at the NTCIR 3 Web task is on processing the large amount of documents, and on achieving higher precisions on the Web documents.

2 System Description

2.1 Hardware Specs

We designed the new indexing / search engine as a distributed process on a PC cluster. The evaluation system consists of 10 Linux PCs, connected by 1Gbps Ethernet.

2.2 Software Architecture

2.2.1 Overview

$R^2D^2$is a full-text retrieval system designed based on the vector space model. The RS model, our proposed method, improves the retrieval effectiveness by solving the semantic ambiguity caused by variance of expression among the documents.

The inverted index file was splitted into segments, and each back-end search process handled the segmented part of index. For the NTCIR 4 Web documents, the inverted file was about 60GB and we splitted it into 34 segments.

Figure 1 depicts the process flow of the indexing engine. Parsing and indexing of NW100G-01 (NTCIR Web corpus) required about 2 weeks, where parsing

was done by 8 parallel processes and indexing was done by a single process. Searching required about 100 seconds for each query.

2.2.2 The RS Model

The proposed RS model is designed using the document feature vector modification approach This model partitions the documents so that the relevant documents dealing with the same topic fall into the same cluster. However, the idea is different from the traditional cluster-based methods [1, 2] in which the document clusters are usually mutually exclusive. These methods assume that documents can be classified into orthogonal topics; however, it is natural to assume that a document can belong to several topics. This difference in assumptions will reflect on the retrieval.

The details of the RS model has already been reported in [8] and [7]. We have evaluated the model using NTCIR 1/2 test set consisting of scientific papers and TREC San Jose Mercury consisting of news articles. The experimental results showed that the RS model improves the average precisions by 7%, which can be considered significant (5–10% is generally required for significant improvement [14]).

In the RS model, each document is represented by a feature vector. Term frequencies are often used as the features. Suppose that a document database contains a set of documents $\{\mathsf{d}_1, \mathsf{d}_2, ..., \mathsf{d}_n\}$ and their feature vectors are $\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n$.

In the RS model, documents in the database form clusters $C_1, C_2, ..., C_m$ which represent topics. Note that a document may be contained in more than one cluster in the RS model, whereas clusters in other methods are often mutually exclusive. At this point, we must decide what type of relevance we will use to make clusters. The principle of the RS model is independent of the source of relevance information, and our choice will depend on the type of database and the types of elements in it. For instance, the following elements included in the database can be candidate sources for relevance information and used for document clustering:

- keywords given by the authors or automatically extracted;
- references, hyperlinks;
- bibliographic information, such as author name, publication date, and journal title.

When clusters representing topics are given, the document feature vector is modified in two steps: (1) representative vector (RV) generation for each cluster, and then, (2) feature vector modification by RVs. We can design a statistical method so that the RV can be considered to accurately represent overall characteristics of the documents that belong to the same cluster. Next, the modification method should properly perform the superimposition of features represented by RVs so that the topics of each cluster are reflected in the modified document feature vectors, thereby reducing the ambiguity of retrieval caused by expressional mismatches between the query and the documents.

2.2.3 Automatic Keyword Extraction

In the previous section, we described how to make document clusters using the well-chosen keywords given by the authors of the documents. However, we must also consider archives where no explicit keywords are given for clustering.

There are two possible answers: one is automatic unsupervised keyword extraction and the other is to find another clue of relevance. We investigated the former approach in the evaluation. Details are described in [10].

2.2.4 Dictionary Improvement by Compound Noun Detection

The accuracy of word splitting significantly affects the effectiveness of Japanese document retrieval. Especially for Web search, developing the new vocabulary is important.

We hired the MeCab [3] and IPADIC for Japanese morphological analysis. The IPADIC contains about 8,500 loan words expressed in *katakana* letters, and the number seems too small for analyzing Web documents that consists of assorted topics.

We hired Nakagawa's compound noun detection algorithm [12, 13]. This method ranked the possible compound nouns based on their expressional expandability, therefore it was easy to eliminate noise patterns. Table 1 shows an example of compound noun detection. We extracted 14,000,000 possible compound noun patterns from the NTCWEB corpus, and chose about 200,000 patterns as the indexing terms. It proved by sample survey that about 60% from the chosen 200,000 patterns were genuine compound nouns.

2.2.5 Domain Ranking and Domain Core Page Detecting for Target Retrieval Task

At Web task B, we examined a method for the target retrieval. The method utilizes the URLs (unified resource locators) to cluster the Web pages. The method outline is as follows.

1. obtain 1,000 pages for each query by R^2D^2.
2. calculate the score of each domain:

$$\text{score}(\text{domain } D) = \max\left(\text{score}(\text{page in } D)\right) + \frac{1}{\#\text{ of pages in } D}\sum\left(\text{score}(\text{page in } D)\right)$$

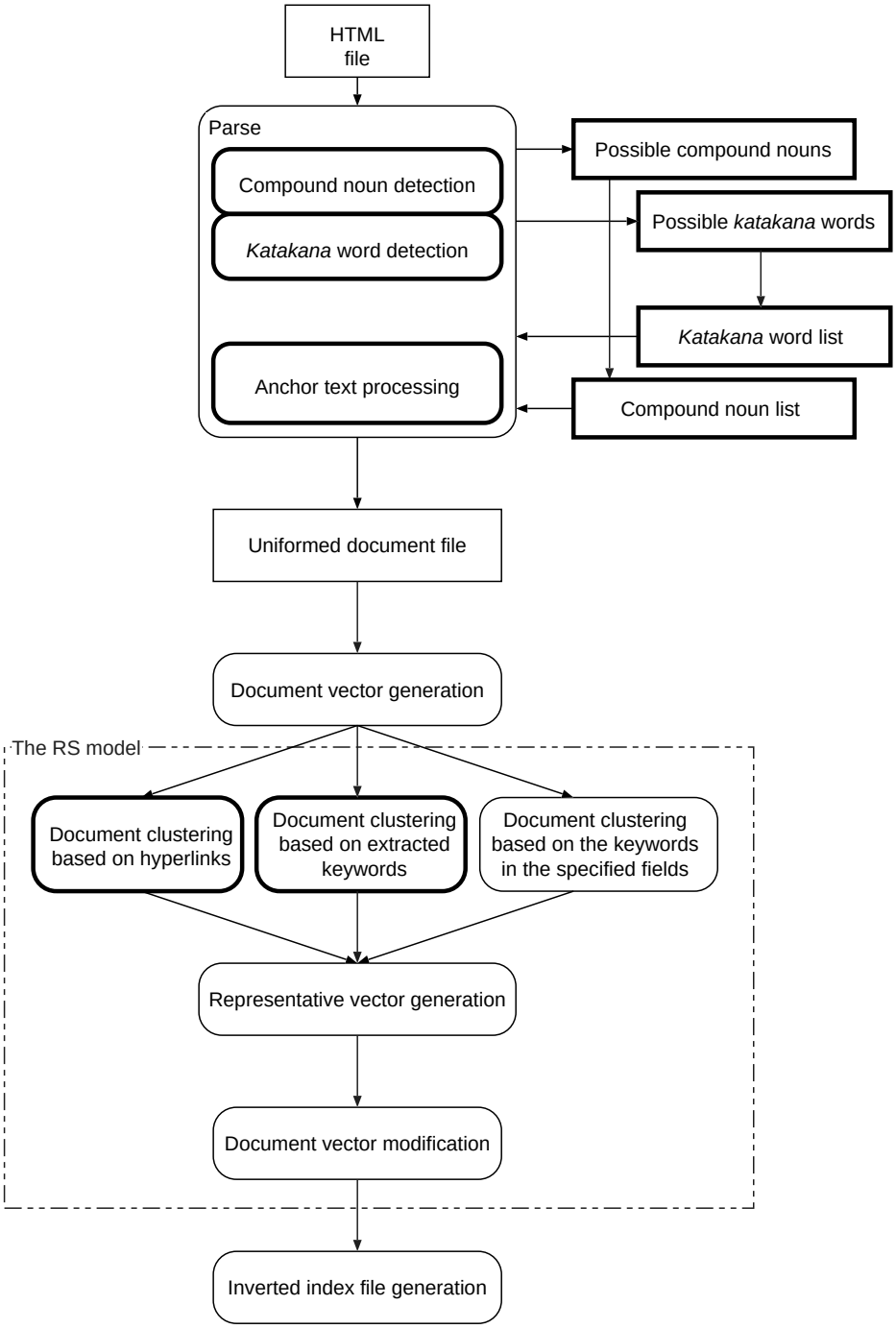

Figure 1. Process Flow of R^2D^2 indexer
The emphasized steps and files are introduced for the Web document retrieval.

Table 1. Example of compound noun detection.

The subpatterns of 情報処理振興事業協会 (Information technology Promotion Agency) and their scores. 情報処理 (information processing) was recognized as the most informative compound noun.

possible compound noun pattern	score
情報処理	2401279
事業協会	233846
振興事業	212927
情報処理振興	76927
処理振興	53701
情報処理振興事業	16212
処理振興事業	13614
振興事業協会	5318
処理振興事業協会	0

3. choose the representative page of each domain. The criteria are:
   - the representative page is at the highest hierarchy (ex. `/1/foo.html` is more desirable than `/1/2/foo.html`),
   - at the same level of hierarchy, files of special names are more desirable, such as `index.html`.

3 Evaluation

3.1 Dictionary Improvement

The table 2 shows the precisions of the RS model on the NTCIR 3 Web. The improved dictionary increased the retrieval effectiveness by 20.3-24.7%.

For further improvement, it is necessary to respond to the expressional variety such as prolonged sound symbol (*ex.* コンピュータ / コンピューター) and alternative kanji (*ex.* 国 / 國). The treatment of phrases containing particles is a difficult problem that would require the analysis of modification relations.

3.2 Document Clustering with Automatic Keyword Extraction

The table 3 shows the effectiveness of the RS model using the automatic keyword extraction for document clustering.

Due to the failure of the PC cluster, we couldn't obtain the baseline result on the NTCIR 4 web. The improvement achieved by the RS model on the NTCIR 3 web test set was less than ones on the NTCIR 2 / TREC. It seems that the parameter optimization for the automatic keyword extraction is required.

Table 2. The Effectiveness of the Dictionary Improvement

The average precisions are for the NTCIR 3 Web test set.

	baseline	RS model (ratio)
standard dict.	0.1112	0.1158 (+4.1%)
improved dict.	0.1387	0.1393 (+0.4%)
ratio	+24.7%	+20.3%

3.3 Long Query

In NTCIR 1/2, R^2D^2made use of the term cooccurrence statistics for calculating the importance of query terms. This technique improved the retrieval precision especially for the long queries that contain more than three words.

In NTCIR 4 evaluation, we disabled this function for time cost reason. This disadvantage notably degraded the retrieval precisions when using the DESC field of query.

The sign test proved that the DESC run result was inferior to the TITLE by probability of 99.9%.

On the other hand, the Wilcoxon test proved that there was no significant difference between the results of shorter TITLE that contains less than or equal to 3 words and the results of longer TITLE that contains more than 3 words by probability of 98.5%.

Most of the longer TITLE are composed of compound nouns, while the DESC are sentences including trivial words. The compound nouns are dealt with desirably by using the improved dictionary, however, other long expressions such as one in the DESC field require another approach to solve.

3.4 Target Retrieval

There is room for improvement in the method for detecting the domain core page, by using hyperlinks and anchor text. The method for calculating the representative score of each domain may be also improved.

There are many targets expressed by the undefined vocabulary such as partly-English expressions (*ex.* NTT インターコミュニケーションセンター) and *hiragana* words (*ex.* みのもんた). It is necessary to develop a suitable vocabulary-improving method for each kind of undefined words.

4 Conclusion

The compound noun detection and the automatic keyword extraction increase the effectiveness of the

Table 3. The Effectiveness of the Automatic Keyword Extraction
The scores are average precisions.

	baseline	RS (i)	RS (ii)
NTCIR 2 EE	.2984	.3160 (+5.9%)	.3211 (+7.6%)
TREC SJM	.1773	—	.2051 (+15.7%)
NTCIR 3 Web	.1112	—	.1158 (+4.1%)
NTCIR 3 Web /w the improved dictionary	.1387	—	.1393 (+0.4%)
NTCIR 4 Web /w the improved dictionary	—	—	.1420

RS(i): using keywords given by the authors.
RS(ii): using extracted keywords.
SJM: San Jose Mercury.

Table 4. The query types and the precisions

	TITLE	DESC	ALT0	ALT1	ALT2	ALT3
avg. prec.	.1420	.0928	.0963	.1089	.0994	.0346
ratio	1.00	0.65	0.68	0.77	0.70	0.24

RS model at the Web search. The experimental results of the target retrieval proved that the vocabulary may admit of development.

Acknowledgment

This work is supported by Information-technology Promotion Agency Japan, Exploratory Software Project FY2002 & 2003 [5, 4].

References

[1] R. Burgin. The Retrieval Effectiveness of Five Clustering Algorithms as a Function of Indexing Exhaustivity. In *J. American Society for Information Science*, volume 46, pages 562–572, 1995.

[2] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *SIGIR '96*, 1996.

[3] `http://chasen.aist-nara.ac.jp/~taku/software/mecab/`.

[4] `http://www.ipa.go.jp/jinzai/esp/index.html`.

[5] `http://www.ipa.go.jp/SPC/report/02fy-pro/report/865/paper.pdf`.

[6] T. Kanazawa. R^2D^2 at NTCIR: Using the Relevance-based Superimposition Model. In *NTCIR Workshop 1 Proc.*, pages 83–88, Tokyo, Aug. 1999.

[7] T. Kanazawa, A. Aizawa, A. Takasu, and J. Adachi. The Effects of the Relevance-based Superimposition Model in Cross-Language Information Retrieval. In *Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 312–324, Darmstadt, Sept. 2001.

[8] T. Kanazawa, A. Takasu, and J. Adachi. A Relevance-based Superimposition Model for Effective Information Retrieval. *IEICE Transactions*, E83-D(12):2152–2160, Dec. 2000.

[9] T. Kanazawa, A. Takasu, and J. Adachi. R^2D^2 at NTCIR 2 Ad-hoc Task: Relevance-based Superimposition Model for IR. In *NTCIR Workshop 2 Proc.*, pages 204–210, Tokyo, March 2001.

[10] T. Kanazawa, A. Takasu, and J. Adachi. Improving the Relevance-based Superimposition model for IR with automatic keyword extraction. In *RIAO 2004*, May 2004.

[11] M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In *SIGIR '98*, pages 206–214, 1998.

[12] H. Nakagawa. Automatic Term Recognition based on Statistics of Compound Nouns. In *Terminology*, volume 6, pages 195–210, 2000.

[13] H. Nakagawa and T. Mori. Automaic Term Recognition based on Statistics of Compound Nouns and their Components. In *Terminology*, volume 9, pages 201–219, 2003.

[14] E. Voorhees. Variations in Relevance Judgements and the Measure of Retrieval Effectivenss. In *SIGIR '98*, pages 315–323, 1998.