Patent Search: A Case Study of Cross-DB Associative Search

Yoshiki Niwa, Toru Hisamitsu, Shingo Nishioka, Osamu Imaichi, and Masakazu Fujio
Central Research Laboratory. Hitachi, Ltd.
2520 Akanuma, Hatoyama, Saitama 350-0395, Japan
{yniwa, hisamitu, nis, imaichi, m-fujio@harl.hitachi.co.jp}

Abstract

An associative searching method was applied to the cross-database search between newspaper articles and patent texts. The results were compared with the results by an ordinary type of search using the narrative field and description field of each query. As for the associative searching using newspaper articles, we conducted two runs and compared their results. One is fully automatic type of run where whole texts of given articles were used as queries, and the other one is a manual type of search where we selected from each given query article only the related parts of texts, considering the purpose of each query described in the narrative field.

The overall performances of these runs were in the following order: (narrative + description) > (article with manual selection) > (article - using whole texts automatically). (these three types will be noted hereafter as *narrative*, *article-manual*, and *article-auto*, respectively.) We examined in detail the difference between *narrative* and *article-manual*, and the difference between *article-manual* and *article-auto*.

Keywords: associative search

1. Introduction

We applied our associative searching method to the NTCIR-3 patent retrieval task, and experimentally evaluated the usefulness of the method for this characteristic task --- searching patent texts using related newspaper articles as queries.

The associative searching method we applied may be fairly standard type, using vector space model with word based index file. We adopted SMART-*like* measure for calculating the relatedness between word-frequency vectors. As for the Japanese morphological analyzer, we used ChaSen developed by the Computational Linguistic Laboratory of NAIST (http://chasen.aist-nara.ac.jp/).

Probably the most characteristic part of the experiments we conducted this time is in the task itself, searching patents using newspaper articles, that is to say, crossing quite different types of text databases. In this sense, this was a precious opportunity for us to evaluate the usefulness of associative searching methods when they are applied to cross-database searching. Here, we would like to thank to the organizers for setting this type of task.

Considering the task's characteristics, we organized four runs listed below, in which the central one is the automatic search using the given newspaper articles themselves.

a) auto Article (whole) (+supplement)
b) auto Description field only
c) auto Narrative (+Description)
d) manual Article (selected part) (+supplement)

The last one is a manual type, where we selected from each given newspaper article only the related parts of texts, considering the purpose of each query described in the *narrative* field.

1.1 Summary of results

First, we show the performances of four submitted runs (Figure 1). Here the performance is measured by the average of maximum precision-recall, where precision-recall value is calculated by (prec * rec / prec + rec). The average is taken over all 31 queries.

In our experiments, the top score was attained by the case of using both description and narrative fields as queries. The next position was shared by two cases, one is the manual run of article & supplement, the other is the case of using description field only. The automatic run of article & supplement was the lowest score.

In the following sections of this note, we will see more details about the comparison between these runs. Here is the summary.

Although *narrative* over-performed *article-manual* as a whole, their performances were almost even when the manually extracted texts are not very short, or more precisely, when they are longer than or as long as the texts of narratives. This shows the usefulness of cross-database associative search under the condition that the texts used as queries are describing the subject of searching to a certain extent.

In the comparison between the manual and automatic cases using newspaper articles, the manual case over-performed the automatic case as is normally expected, but the difference was not so large as we expected.

What was surprising was that their performances were almost even, when the ratio of manually deleted part is lower than 60%. This means that even if more than half of an article is spared for irrelevant things such as the history or personality of the inventors, the associative search is not affected so much.

The influence of deletion becomes remarkable when the ratio of deletion exceed 80 %. In that case, the influences are mostly positive, but in some unfortunate cases, big negative influences were detected. They seem to be caused by the weakened stability which is un-escapable when the query text is very short.

In section 2, we compare narrative with article-manual. Next in section 3, we compare manual with auto for *article* cases.

2. *Narrative* vs. *Article*

In this section, we compare *narrative* with *article*. As we mentioned before, there are two article-based runs, auto and manual. We chose *manual* here as the competitor of *narrative*, because manual case is considered to be a more natural situation when newspaper articles are used as queries.

As we will show later, an newspaper article, even though its main topic is concerning a newly invented patent, that article also contains large portion of non-relevant descriptions such as the history of the company, or the inventor, or the personality of the inventor. If we use the article in the real situation as a query for searching related patents, probably we will omit these non-relevant parts.

The overall performance, as is mentioned before, is better for *narrative*. But it is not always the case for each search topic. There are not small portion of topics for which the *manual-article* is advantageous. Table 1 classifies the topics by the run-type (*narrative* or *article*) which won.

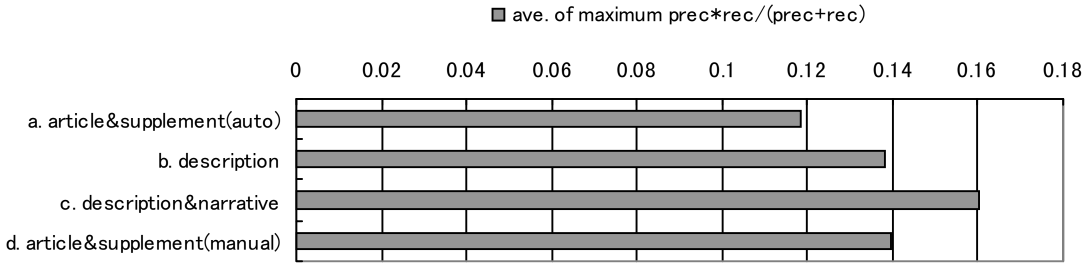


Figure 1. Average recall-precision of four runs. (Relevance = A)

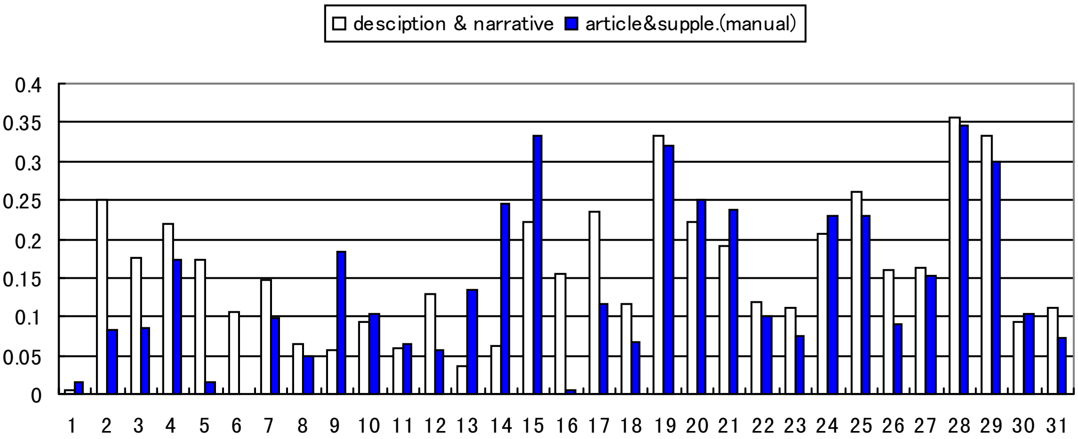


Figure 2. *Narrative* vs. *Article* per topic

Table 1. *Narravie* vs. *Article*

run	Topics for which the left run won	# win	#big win
description & narrative	**2**,**3**,4,**5**,**6**,7,8,**12**,**16**,**17**,18,19, 22,23,25,26,27,28,29,31	20	7
article & supple. (man)	1, **9**,10,11,**13**,**14**,**15**,20,21,24,30,	11	4

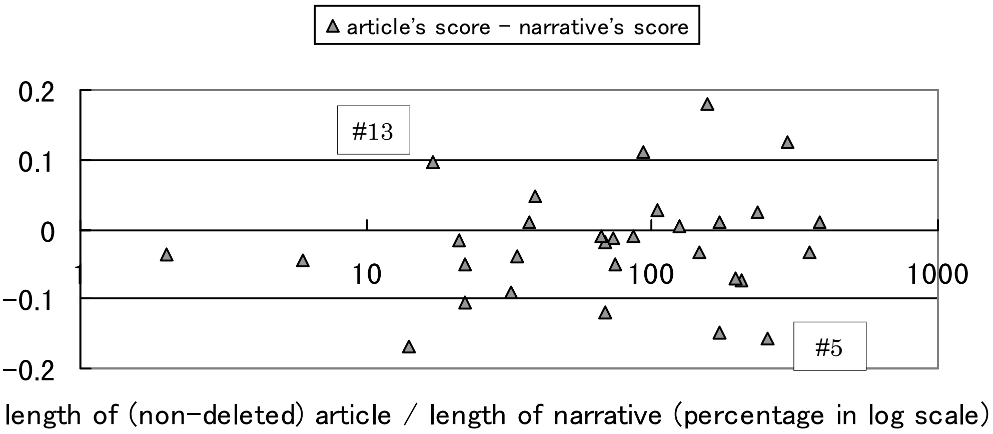


Figure 3. Score-Difference vs. Query-Length-Ratio

This table again shows the advantage of *narrative*. But we should be careful about the quality of newspaper articles as queries for searching aimed patents. Some articles describe the target inventions only as a side topic in their small part and we could only use small sizes of texts as queries.

From this point of view, the graph in Figure 3 was made to show the effect of the quality of newspaper articles. The horizontal scale shows the ratio of length of article text divided by the length of *description* and *narrative* text. Since we chose manual case of *article*, the length of article is the size of remaining texts after deletion of non-relevant part. The vertical line shows the difference of scores: *article* minus *narrative*.

Although the overall correlation in this graph seems unclear, we can see that *narrative* is clearly advantageous over *article* in the region where the lengths of articles are shorter than narratives (the part x axis is smaller than 100). But we should also pay attention to the almost evenness between *narrative* and *article* if the lengths of relevant parts of articles are longer than or almost as same as those of narratives.

As exceptional two cases, we marked topic no. 5 and 13 on the graph. Topic 5 was a case for which the performance by *article* was much lower, though the length of remaining part of the article was long enough. Topic no. 5, on the other hand, is a lucky case for *article*, which over-performed *narrative,* though the remining article was less than 20% of narrative.

Figure 4 shows the detailed results of topic no. 13 for both *narrative* and *article*. The word "Supply chain (サプライチェーン)" in the narrative may be the main cause of bad results.

3. Article: Auto vs. Manual

In this section, we compare the automatic run and manual run using newspaper articles as queries. As we mentioned before, the manual run used parts of articles after deleting non-relevant parts of articles.

The graph in Figure 5 shows the scores of these two cases for each search topic. In order to make more clear the advantage or disadvantage of these two cases, we classified the queries by the case which gave higher score. The results are shown in the next table 2.

First ten search results with their relevance

Description and narrative results

1. × 計画作成方法およびシステム
2. × *サプライチェーン*を横断する拡張企業体プランニングシステム及び方法
3. × 食材・加工調理品宅配業界における、食材宅配*サプライチェーン*システム
4. × リソース割当のための動的最適化装置 ...
5. × 拠点立地評価方法
6. × 企業システムの定義情報を設定する.....
7. × 物品の移動を最適化する方法および .....
8. × ロジスティクスチェーンシミュ ......
9. × 生産計画の上流側の工程に対する .....
10. × ビジネスイベントサーバ

query

Description:
トヨタ自動車の「かんばん方式」(生産方式)に関連するビジネス方法特許。

Narrative:
トヨタ自動車の「かんばん方式」(生産方式)とは、原料調達から製品配達までを一括管理し、システム全体の最適化、効率化を図る生産方式である。最近はビジネス方法特許がブームになったが、このトヨタ自動車の「かんばん方式」とＩＴ化を合体させた方式がＳＣＭ（*サプライチェーン*マネジメント）と呼ばれるビジネス方法である。具体的には原料調達から製品配達までの全生産管理システムをＩＴ化し、常に最適なルートを設定する、最適化条件を見つけるためのビジネス方法を言う。トヨタ自動車は他社のビジネス方法特許の権利行使の排除、防御的な目的で「かんばん方式」特許を出願したと言われている。

Article and supplement results

1. A マトリクス式生産設備
2. × プレニル2リン酸合成酵素をコードするDNAを連結させた ......
3. A 物品納入管理システム
4. × IDカード, 生産機器管理装置, ...
5. A 部品の納入指示方法
6. × 薄板用アーク溶接装置
7. × 薄膜デバイスの製造方法
8. A 部品納入指示方法および部品納入指示装
9. × 生産ライン管理装置
10. × 浮体式メガフロート生産設備

query

Article: (after deleting non-relevant part)
現地採用社員に自らトヨタ生産方式を教え込み、生産効率を飛躍させた。

Supplement:
「かんばん方式」で知られるトヨタ生産方式

Topic #13 “Toyota’s Kanban method”
Narrative is much longer than *article and supplement*. But the results is better (4A’s – none) for *article*.
The word “Supply chain (サプライチェーン)” may be the main cause of bad results.

Figure 4. Topic #13 “Toyota Kanban method” (left = Narrative, right = Article)

As we expected, the manual run won for much larger number of topics, 16 vs. 6. But here remains a question. For some queries, very large parts of the query texts were deleted, whereas for other queries, only small parts of the texts were deleted. Therefore we need to make clear the relation between the ratio of deletion and the gain-or-loss of performance.

The loss or gain between auto and manual may be dependent on the percentage of deletion from the original text. We plotted their relation on the graph of Figure 6.

What is surprising is that up to almost 60 percent of deletion has only little effect on the performance. The effects are also small when the deletion is less than 80 percent. The effect of deletion is remarkable only when the percentage of deletion is larger than 80 percent.

In that case, the effects are mostly positive, but there are several cases where the effects are negative. Big gains were seen for query no. 21 and 29, whereas big losses were seen for query no. 6. Here we would like to analyze the details of the results of no. 6 (Figure 7).

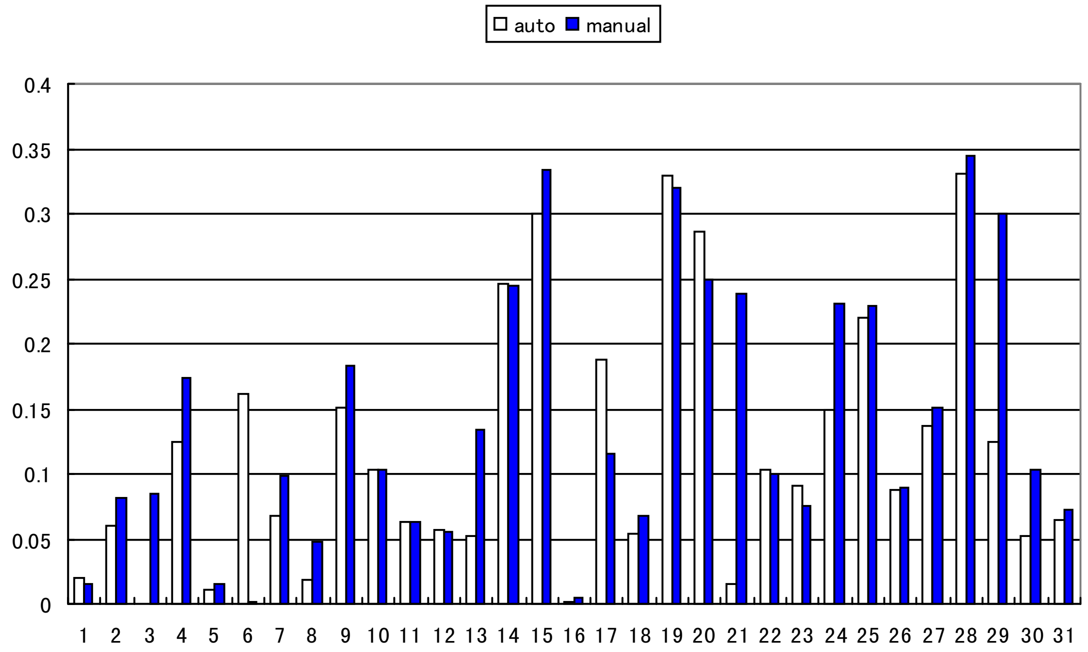


Figure 5. Article-Auto vs. Article-Manual

Table 2. Article-Auto vs. Article-Manual

Man > auto (man >> auto)	Man < auto (man << auto)	Man ~ auto
2,5,7,9,15,18,27,31 (**3**,4,8,**13**,**21**,24,**29**,30)	1,19,20,23 (**6**,17)	10,11,12,14,16,22,25,26,28
16 (8)	6 (2)	9

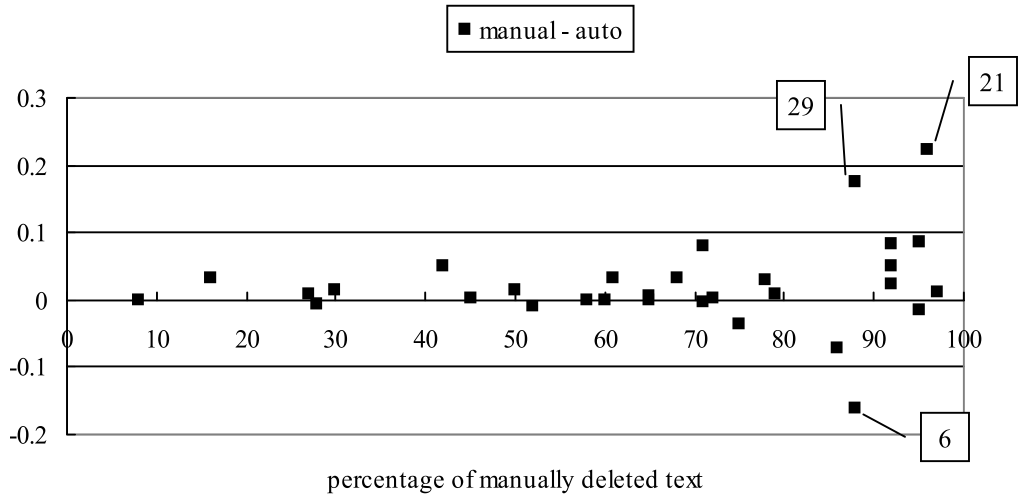
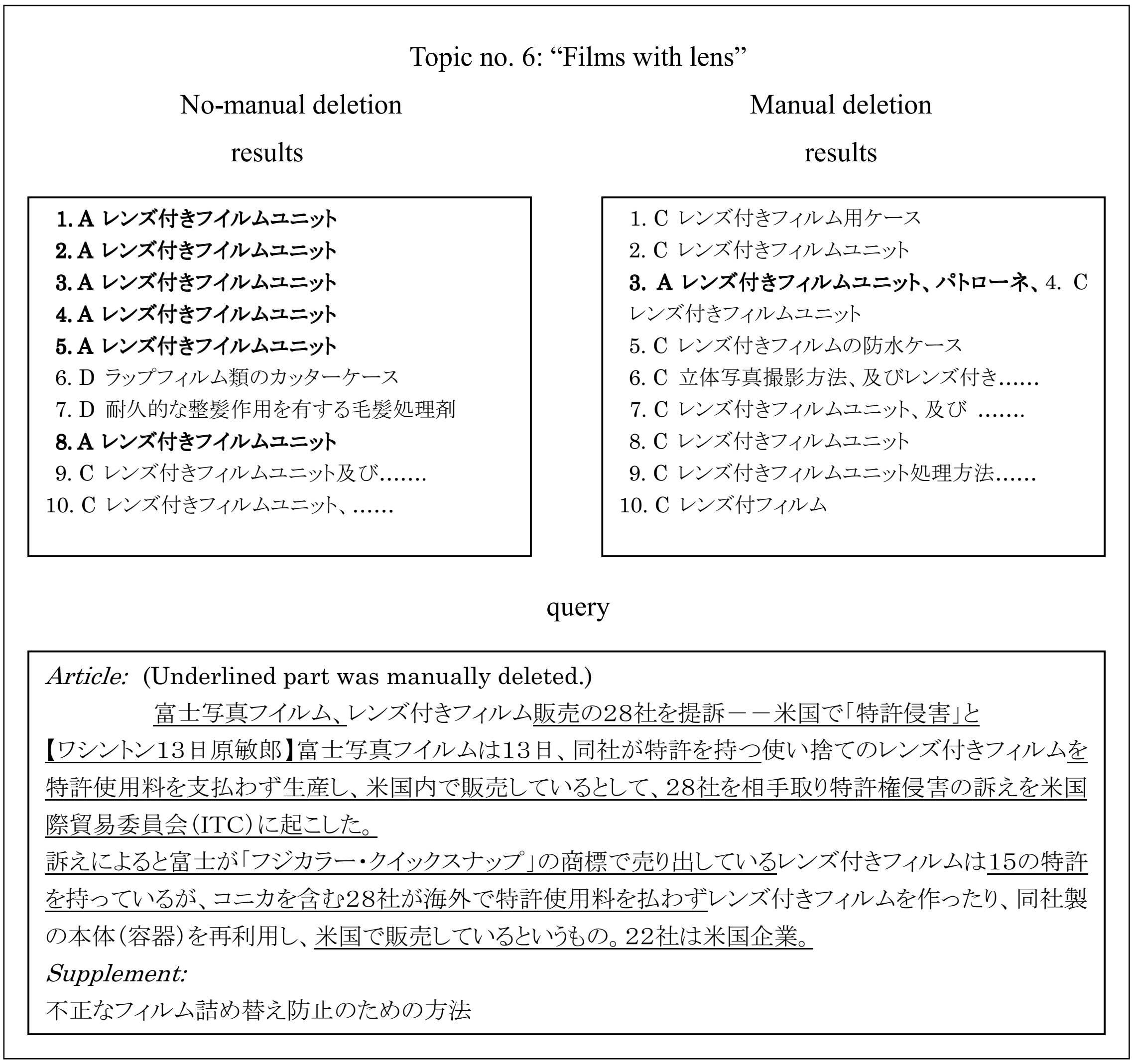


Figure 6. Score-difference vs. Text-deletion-ratio



Topic no. 6: “Films with lens”

No-manual deletion results

1. A レンズ付きフイルムユニット
2. A レンズ付きフイルムユニット
3. A レンズ付きフイルムユニット
4. A レンズ付きフイルムユニット
5. A レンズ付きフイルムユニット
6. D ラップフィルム類のカッターケース
7. D 耐久的な整髪作用を有する毛髪処理剤
8. A レンズ付きフイルムユニット
9. C レンズ付きフィルムユニット及び.......
10. C レンズ付きフィルムユニット、......

Manual deletion results

1. C レンズ付きフィルム用ケース
2. C レンズ付きフィルムユニット
3. A レンズ付きフィルムユニット、パトローネ、4. C レンズ付きフィルムユニット
5. C レンズ付きフィルムの防水ケース
6. C 立体写真撮影方法、及びレンズ付き......
7. C レンズ付きフィルムユニット、及び ....
8. C レンズ付きフィルムユニット
9. C レンズ付きフィルムユニット処理方法......
10. C レンズ付フィルム

query

Article: (Underlined part was manually deleted.)

富士写真フイルム、レンズ付きフィルム販売の28社を提訴－－米国で「特許侵害」と

【ワシントン13日原敏郎】富士写真フイルムは13日、同社が特許を持つ使い捨てのレンズ付きフィルムを特許使用料を支払わず生産し、米国内で販売しているとして、28社を相手取り特許権侵害の訴えを米国際貿易委員会（ITC）に起こした。

訴えによると富士が「フジカラー・クイックスナップ」の商標で売り出しているレンズ付きフィルムは15の特許を持っているが、コニカを含む28社が海外で特許使用料を払わずレンズ付きフィルムを作ったり、同社製の本体（容器）を再利用し、米国で販売しているというもの。22社は米国企業。

Supplement:

不正なフィルム詰め替え防止のための方法

Figure 7 Topic no. 6 “Film-with-lense” (left = Auto, right = Manual)

Figure 7 shows the results for both automatic and manual cases of topic no. 6 and their used queries. In the query, the underlined part was deleted as ‘non-relevant’ part. Their difference may not be clear from the titles appearing in the top 10 rankings. But what was unfortunate situation for the manual case is that the first five A-ranked patents in the results of automatic case are *Fuji-Film*’s patents, and these A-ranked patents were missed in the top ranking of manual case. This is because the all appearances of the company’s name (富士写真フイルム) were (deliberately) deleted as they were considered in the non-relevant parts. So this case may be considered as a special case.

4. Conclusion

An associative searching method was applied to the cross-database search between newspaper articles and patent texts. The results were compared with the results by an ordinary type of search using the narrative field and description field of each query. As for the associative searching using newspaper articles, we conducted two runs and compared their results. One is fully automatic type of run where whole texts of given articles were used as queries, and the other one is a manual type of search where we selected from each given query article only the related parts of texts, considering the purpose of each query described in the narrative field.

The overall performances of these runs were in the following order: (narrative + description) > (article with manual selection) > (article - using whole texts automatically). (these three types will be noted hereafter as *narrative*, *article-manual*, and *article-auto*, respectively.) We examined in detail the difference between *narrative* and *article-manual*, and the difference between *article-manual* and *article-auto*.

Although *narrative* over-performed *article-manual* as a whole, their performances were almost even when the manually extracted texts are not very short, or more precisely, when they are longer than or as long as the texts of narratives. This shows the usefulness of cross-database associative search under the condition that the texts used as queries are describing the subject of searching to a certain extent.

In the comparison between the manual and automatic cases using newspaper articles, the manual case over-performed the automatic case as is normally expected, but the difference was not so large as we expected.

What was surprising was that their performances were almost even, when the ratio of manually deleted part is lower than 60%. This means that even if more than half of an article is spared for irrelevant things such as the history or personality of the inventors, the associative search is not affected so much.

The influence of deletion becomes remarkable when the ratio of deletion exceed 80 %. In that case, the influences are mostly positive, but in some unfortunate cases, big negative influences were detected. They seem to be caused by the weakened stability which is un-escapable when the query text is very short.

References

[1] M. Iwayama, Y. Niwa, S. Nishioka, A. Takano, T. Hisamitsu, O. Imaichi, H. Sakurai, and M. Fujio. The Effect of Document Clustering in Interactive Relevance Feedback. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization,* pages 5--90-97, 2001.

[2] Y. Niwa, M. Iwayama, T. Hisamitsu, S. Nishioka, A, Takano, H. Sakurai, and O. Imaichi. Interactive document search with *DualNAVI*. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 123 – 130, 1999.

[3] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21 – 29, 1996.