The Effect of Document Clustering in Interactive Relevance Feedback

Makoto Iwayama, Yoshiki Niwa, Shingo Nishioka, Akihiko Takano†,
Toru Hisamitsu, Osamu Imaichi, Hirofumi Sakurai and Masakazu Fujio

Central Research Lab., Hitachi, Ltd.
2520 Hatoyama, Saitama 350-0395, Japan
{iwayama,yniwa,nis,hisamitu,imaichi,hirofumi,m-fujio}@harl.hitachi.co.jp

† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
takano@acm.org

Abstract

We examined the two cluster-based representations of search results in a relevance feedback environment. Through interactive runs by human subjects, we found that both representations, "ranking with categories" and "dendrogram," could not outperform the conventional relevance ranking from the standpoint of average precision. However, by analyzing the interactions between the human subjects and the system, we found the dendorgram to be more effective than the conventional relevance ranking for the human subjects to easily find a group of similar documents.

1 Introduction

Interactive relevance feedback is an effective method for improving the result of document searches. The system retrieves documents based on the user's query and displays them in order of perceived relevance. The user looks through the ranking and judges the actual relevance of one or more of the retrieved documents. These judgements are fed back to the system for updating the current query.

In this feedback cycle, the amount of improvement of the system depends on the judgements of relevance provided by the user [1]. We have been trying to attain an efficient interface for users to be able to find many relevant documents at little cost [6, 4]. In NTCIR-2, we examined the use of two advanced representations of search results, which are "ranking with categories" and "dendrogram." Both are based on the automatic clustering of retrieved documents.

In the "ranking with categories" representation, the clustering algorithm finds three major categories in the initial search result and then each retrieved document is classified into these categories. This representation is similar to that of Scatter/Gather [2] or that used by CLARITECH [3]. By focusing on a relevant category, users can find desired documents efficiently.

In the "dendrogram" representation, the clustering algorithm calculates the local similarity between every pair of documents, and places similar documents side by side. If users judge a document to be relevant, very similar documents around the seed document would also be relevant, so users could find a bunch of relevant documents easily.

We assume that these advanced representations would be helpful for users to find many relevant documents, which are in turn effective in relevance feedback. To confirm this assumption, we executed several interactive runs in this NTCIR, and this paper reports the results and discussions.

2 Automatic Runs

To investigate the baseline performance of our system, we executed the following automatic runs.

ID	query type	query expansion
DOVE1	D	none
DOVE2	D+N	none
DOVE3	D	pseudo feedback
DOVE4	D+N	pseudo feedback

"D" denotes "<DESCRIPTION>" and "N" denotes "<NARRATIVE>". We did not use "<CONCEPT>". Our experimental system is

based on the standard vector space model. Here is the summary of the system description:

term unit	word (by ANIMA [7])
term weighting	Lt.Lnc [8]
term re-weighting in pseudo feedback	Modified Rocchio $\alpha = 8, \beta = 16, \gamma = 0$

In the pseudo relevance feedback, the top 100 terms from the top 10 retrieved documents were automatically fed back to the system for updating the initial query. Negative feedback was discarded, so γ was equal to zero.

We used GETA [1] to implement the experimental system.

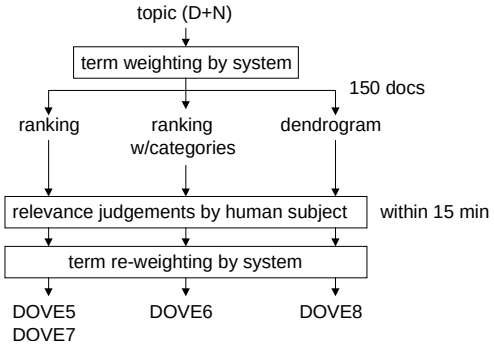


Figure 1: Overview of interactive runs

3 Interactive Runs

Each of the interactive runs consists of the following steps.

1. The system constructs the initial query from each topic. Concretely, ANIMA extracts all words except stop words from a topic description (D+N), and the Lt.Lnc term weighting formula is used to weight the extracted terms.

2. The system retrieves the top 150 documents from the initial query and displays them to a human subject in one of the following forms:

   - ranking (see Sec. 3.1)
   - ranking with categories (see Sec. 3.2)
   - dendrogram (see Sec. 3.3)

   In every form, a title is used as a surrogate of a document.

3. The subject looks through the search result and tries to collect as many relevant documents as possible within 15 minutes. The subject can interact with the system through several commands which will be described later. All interactions between the subject and the system are recorded with the timing information.

4. The system updates the initial query by using the top 300 terms from the relevant documents judged by the subject, and retrieves a new set of documents (i.e., 1,000 documents) from the updated query.

Figure 1 shows the overview of these steps.

[1] GETA is a collection of C libraries to facilitate flexible and large-scale documents/words analysis including document searches. GETA also contains clustering libraries used in our experiments. GETA will be available from IPA's web site "`http://www.ipa.go.jp/STC/dokusou.html`" this April.

3.1 Ranking

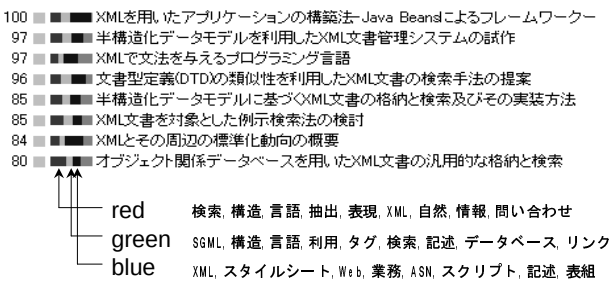
This is the conventional relevance ranking, where documents are ranked in order of decreasing relevance calculated by the system. In this representation, subjects can interact with the system through the following commands.

AVISIT	View the full text of the selected document
ASEL	Mark the selected documents as relevant
AUNSEL	Un-mark the selected documents

This "ranking" representation forms the baseline of the following two advanced representations.

3.2 Ranking with Categories

In the "ranking with categories" representation, each retrieved document has additional information of categories to which the document belongs. Although the initial order of retrieved documents remains the same as the conventional relevance ranking, subjects can re-sort the documents according to these categories. Here is an example of this representation for topic "0108".



The RGB spectrum bar next to the title represents the likelihood of the document belonging to

each of the three major categories automatically constructed from the 150 retrieved documents. The system uses a hierarchical clustering algorithm [5] to divide the 150 retrieved documents into three clusters and regards them as categories. For each retrieved document, the system calculates the similarity to each category, and displays the similarities in the RGB spectrum. Here R(ed) corresponds to the primary category, G(reen) the secondary one, and B(lue) the tertiary one. Human subjects can investigate the contents of each category by using the "VCAT" command.

VCAT	View the representative words of the selected category

If a subject is interested in a category and tries to focus on it, he/she can collect only the documents which are closely related to the category by executing the "GCAT" command.

GCAT	Re-sort the search result by the similarity to the selected category

After executing the GCAT command, the top-ranked documents become the representative documents for the selected category. Here is the result of executing the GCAT command in the above example. In this case, we focused on the primary category (i.e., "red").

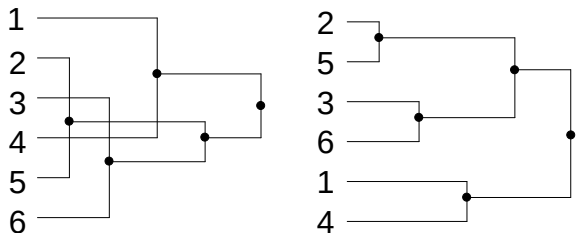
97 半構造化データモデルを利用したXML文書管理システムの試作
85 半構造化データモデルに基づくXML文書の格納と検索及びその実装方法
85 XML文書を対象とした例示検索法の検討
80 オブジェクト関係データベースを用いたXML文書の汎用的な格納と検索
76 Web文書に対する言語処理の問題点と言語処理を援助するタグセットについ
76 Web文書に対する言語処理を援助するタグセット
72 文書構造化言語XMLえお利用した文書管理手法の提案
70 XML応用の最近の動向：文書・データから、オブジェクト・知識表現まで

Subjects can also use the "AVISIT", "ASEL", and "AUNSEL" commands for viewing, selecting, and de-selecting a document.

By providing these categories for interactions, we expect that subjects can effectively focus on the relevant topic, and, in consequence, can find relevant documents easier than by using the conventional relevance ranking.

3.3 Dendrogram

The "dendrogram" representation is closely related to the hierarchical clustering algorithm we used. At the beginning of the algorithm, each of the 150 retrieved documents is a singleton cluster, and the algorithm merges the most similar pairs of clusters step by step. As a result, the following tree, called dendrogram, is constructed.



The left-hand side is the naive form of the dendrogram. Here the order of the documents remains in the original order (i.e., "1", "2", "3", ...), many branches are entangled and it is hard to grasp the relationship between documents. The right-hand side is the disentangled version of this dendrogram, where similar pairs of documents are re-located side by side.

Our experimental system only shows the order of documents in the disentangled dendrogram and omits the tree structure. However, this ordering itself reflects the local similarity between documents well. If a subject finds document "3" to be relevant, then "6" next to "3" might be relevant because "6" is very similar to "3." The subject can therefore find a set of similar documents easier than if documents were ranked by their relevance scores, because relevance scores do not reflect local similarity between documents.

The following example for topic "0119" shows the effect of the "dendrogram" representation.

ranking

84 [k]家庭科における学習が食生活に対する意識や価値観の形
a 84 生活価値観の変化に伴う新しい住要求に関する研究その2高
83 [k]浦東地区開発計画に伴う価値意識の変化に関する研究-E
:
82 東広島市における留学生の環境認識・評価に関する研究その
b 81 生活価値観の変化に伴う新しい住要求に関する研究その1研
81 パタン・ランゲージの方法による農村地域活性化のための生
:
77 [k]東京とロンドンとの空間構造と都市交通に関する比較研究
c 76 生活価値感の変化に伴う新しい住要求に関する研究:その4.
76 [k]職業特性の比較研究と価値志向の動向把握
:
71 在日外国人の住まい方に関する予備的研究
d 71 生活価値観の変化に伴う新しい住要求に関する研究その3パ
70 [k]「新・日本人の国民性調査」のための基礎的研究

dendrogram

62 過疎地域への転入定住者の実態と価値意識について山形県
a 84 生活価値観の変化に伴う新しい住要求に関する研究その2高
c 76 生活価値感の変化に伴う新しい住要求に関する研究:その4.
b 81 生活価値観の変化に伴う新しい住要求に関する研究その1研
d 71 生活価値観の変化に伴う新しい住要求に関する研究その3パ
80 東京の都市空間のイメージ特性に関する研究外国人との比較
71 在日外国人の住まい方に関する予備的研究
67 アメリカに居住する日本人の住様式(第1報)-履床様式につい

In the "ranking" representation, a bunch of similar documents, "a," "b," "c," and "d" (a series of

papers by the same authors), are scattered according to their relevance scores. On the other hand, the "dendrogram" collects the bunch of documents and display them side by side. Subjects can successfully interpret these documents as a series of papers.

In this form, subjects can also use the "AVISIT", "ASEL", and "AUNSEL" commands.

3.4 Experimental Setting

Seven subjects, the authors, participated in the experiments. The purpose of the experiments was to compare the advanced representations ("ranking with categories" and "dendrogram") with the baseline representation ("ranking").

To keep the consistency of each topic, the same subject conducted two interactive runs for the same topic: one for the baseline representation and the other for one of the advanced representations. Subjects were instructed to leave at least one week interval between the two runs to avoid the memory effect. The order of the two runs was randomized. We conducted the following four interactive runs based on the above constraints.

DOVE5	ranking (baseline of DOVE6)
DOVE6	ranking with categories
DOVE7	ranking (baseline of DOVE8)
DOVE8	dendrogram

4 Results and Discussions

4.1 Overall Results

Table 1 shows the average precision for all runs. Although all the interactive runs outperformed the automatic runs, we could not see the explicit advantage of the advanced representations (DOVE6 and DOVE8) to the conventional rankings (DOVE5 and DOVE7 respectively). The "dendrogram" (DOVE8) was slightly more effective against the "ranking" (DOVE7), but the difference was small.

Figure 2 shows the average precision in each interactive run at timed intervals. Here the performance of the advanced representations was lower than their baseline performance. The "dendrogram" could catch up with and slightly outperform the baseline at the end of the session.

4.2 Relevance Judgements

Table 2 shows the agreement of the DOVE relevance judgements with the NII official judgements.

In every run, the agreement was over 70%, which is sufficient. From the table, we can also see that the advanced runs (DOVE6 and DOVE8) could not find a larger number of relevant documents than the baseline runs (DOVE5 and DOVE7) could. However, for the query-averaged ratio of the correct judgements to all the judgements, the "dendrogram" was a 4% improvement over the baseline for the S+A case. The "ranking with categories" was also effective, but the difference was subtle (only a 1% improvement). For S+A+B judgements, neither of the advanced runs was effective.

Table 3 shows the coverage of the DOVE judgements over the relevant documents. Let's assume a case where there are 20 S-judged documents in the 150 retrieved documents. If a DOVE subject determines that only 15 of the 20 documents are relevant, then the coverage is 15/20.

In Table 3, the coverages were all relatively small, i.e., lower than 40%. This indicates that the DOVE subjects missed many relevant documents due to several reasons: misunderstandings, lack of time, etc. If the coverage were 100%, that is, the subjects could find all the relevant documents contained in the 150 documents, then the average precision would become 0.5166 for S+A judgements, and 0.4775 for S+A+B judgements. These form the upper-bounds of our interactive runs. All the interactive runs (DOVE5 $\sim$ DOVE8) achieved 77% $\sim$ 79% of the upper-bound for S+A judgements, and 82% $\sim$ 85% for S+A+B judgements.

Compared with the baseline runs, both of the advanced runs cover fewer officially relevant documents out of the 150 documents, because the advanced runs found fewer relevant documents than the baseline runs did (see Table 2).

In all the runs, the coverage of B-judged documents was small. This is because of the instruction offered to the subjects, which stated "find as many totally relevant (i.e., S+A) documents as possible." In consequence, the subjects might have judged B-level documents to be irrelevant.

4.3 Interaction Log Analysis

In this section, we investigate the log files of interactions between the subjects and the system to see the effect of the advanced representations.

Our assumption is that subjects can find a group of relevant documents efficiently by using the advanced representations because these representations juxtapose similar documents. In "ranking with categories", this is done by using a focused category to re-sort the retrieved documents. In "dendrogram", a pair of similar documents are located side by side at the outset. In both representations,

ID	query type	search type	other info	average precision S+A	S+A+B
DOVE1	D	auto	—	0.2261	0.2088
DOVE2	D+N	auto	—	0.2821	0.2827
DOVE3	D	auto	pseudo feedback	0.2800	0.2678
DOVE4	D+N	auto	pseudo feedback	0.3119	0.3219
DOVE5	D+N	inter	baseline of DOVE6	0.4095	0.4020
DOVE6	D+N	inter	ranking with categories	0.3996	0.3943
DOVE7	D+N	inter	baseline of DOVE8	0.4052	0.3976
DOVE8	D+N	inter	dendrogram	0.4069	0.3891

Table 1: Average precision for all runs

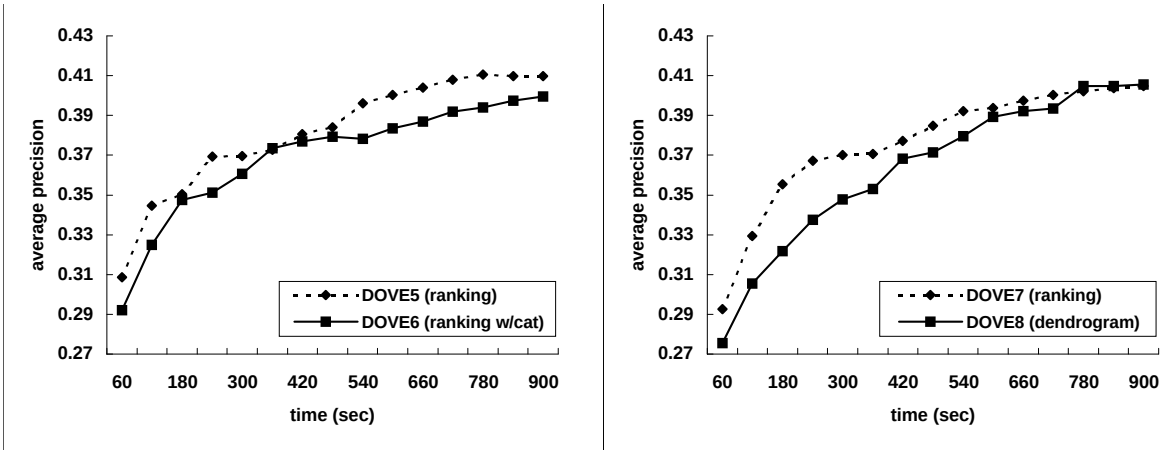


Figure 2: Average precision at timed intervals

ID	ASEL	S	A	B	C	(S+A)/ASEL query-averaged	(S+A+B)/ASEL query-averaged
DOVE5 (ranking)	688	120	383	62	122	0.7264	0.8371
DOVE6 (categories)	598	94	325	50	129	0.7336	0.8202
DOVE7 (ranking)	671	99	387	66	116	0.7201	0.8372
DOVE8 (dendrogram)	543	87	319	40	94	0.7496	0.8368

Table 2: Agreements of the DOVE judgements to the official judgements

ID	S	A	B	S+A	S+A+B
DOVE5 (ranking)	0.2915	0.4150	0.1287	0.4483	0.4048
DOVE6 (categories)	0.2768	0.3591	0.1687	0.4092	0.3807
DOVE7 (ranking)	0.2617	0.4142	0.1420	0.4429	0.3912
DOVE8 (dendrogram)	0.2798	0.3761	0.1254	0.4090	0.3627

Table 3: Query-averaged coverage of the DOVE judgements to the relevant documents in the 150 retrieved documents

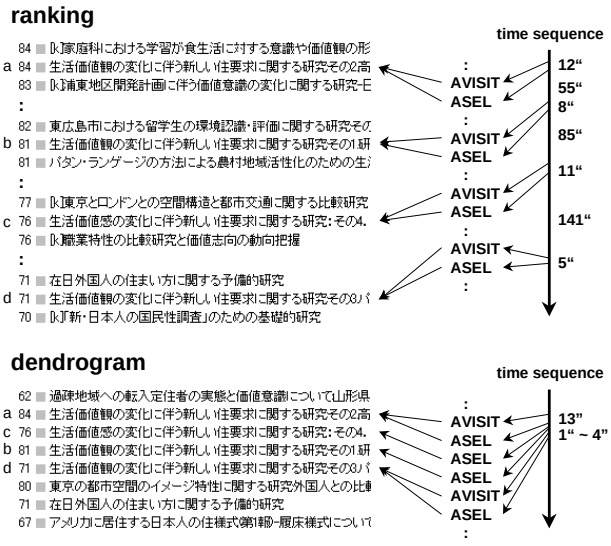


Figure 3: Example of transactions: “ranking” and “dendrogram” for topic “0119”

if a subject determines one of those similar documents to be relevant, then other similar documents around the seed document would also be relevant and could be successively found.

Figure 3 shows examples of subject-system transactions. In the “ranking” representation, subjects generally look through documents from the top rank to the downward direction one by one and thus encounter many irrelevant documents. In the figure, let’s focus on a series of papers, “a,” “b,” “c,” and “d” written by the same authors. In the “ranking” representation, these papers were at various rankings according to their relevance scores. As a result, the intervals of the subject encountering these papers were 55, 85, 141 seconds for each. These intervals are long enough for the subject to forget the contents of the previously visited paper. Actually, the subject read the each of the four papers. In “dendrogram,” on the other hand, all four papers were grouped in one place and the subject could successfully interpret the four papers as a series of papers. The subject first encountered “a” and read its full text. Then the subject selected “c” and “b” without reading their full texts. As for the last one, “d,” the subject verified the full text, but the reading time was less than 4 seconds.

Table 4 shows the number of “ASEL ASEL” succession (ASEL followed by ASEL) in each run, and Table 5 shows the number of documents which are judged to be relevant without browsing their full texts (i.e., ASEL without AVISIT). Large values for these numbers represent that the subjects could find a series of relevant documents at little cost, that is, by simply glancing at their titles. Both tables show the advantage of using “dendrogram,” which had about twice the value of the baseline for both cases. “Ranking with categories” was not effective.

ID	
DOVE5 (ranking)	65
DOVE6 (categories)	57
DOVE7 (ranking)	50
DOVE8 (dendrogram)	135

Table 4: Number of ”ASEL ASEL” succession

In Figure 4, we plot the ASEL/AVISIT ratio (query averaged) at timed intervals. This ratio also measures the efficiency of the judgements. Although the plots for “ranking with categories” were almost same as the baseline, the plots for “dendrogram” were totally different from the baseline. Note that after about 100 seconds, the baseline falls gradually, whereas the “dendrogram” curve stays at about a 0.5 ratio of ASEL/AVISIT. This means that in the “ranking” representation, subjects read larger numbers of irrelevant documents as the session progressed. In the “dendrogram” representation, subjects read irrelevant documents of course, but they also encountered many relevant documents, meaning that the “dendrogram” representation did not lose its effectiveness at the end of the session. This is a reason why the “dendrogram” representation overtook the baseline in Figure 2 (average precision at timed intervals).

5 Conclusions

In NTCIR-2, we investigated two clustering-based representations of search results to achieve efficient relevance feedback. Although we were not convinced of the advantage of these representations from the standpoint of average precision, the “dendrogram” representation helped human subjects to easily select a group of similar relevant documents. For the “ranking with category” representation, we could not see its explicit advantage in any of the evaluations. We believe that the current interface for “ranking with categories” is complex and need to be refined. For example, we provided several re-sorting algorithms for categories, subjects might have been confused by those algorithms and could not select the best one.

ID	ASEL	ASEL without AVISIT		
		total	S+A	S+A+B
DOVE5 (ranking)	688	76 (0.1105)	62	66
DOVE6 (categories)	598	65 (0.1087)	56	58
DOVE7 (ranking)	671	51 (0.0760)	38	40
DOVE8 (dendrogram)	543	107 (0.1971)	70	79

Table 5: Documents which are judged to be relevant without browsing their full texts

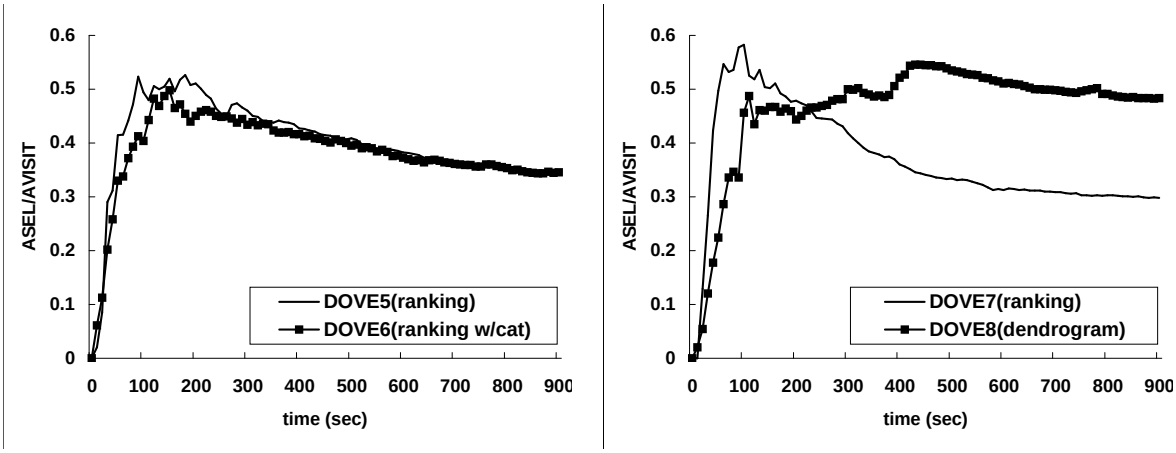


Figure 4: ASEL/AVISIT ratio at timed intervals

References

[1] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, 1994.

[2] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[3] D. A. Evans, A. Huettner, Tong X., P. Jansen, and J. Bennett. Effectiveness of clustering in ad-hoc retrieval. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999.

[4] M. Iwayama. Relevance feedback with a small number of relevance judgements: Incremental relevance feedvack vs. document clustering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–16, 2000.

[5] M. Iwayama and T. Tokunaga. Cluster-based text categorization: A comparison of category search strategies. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, 1995.

[6] 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.

[7] H. Sakurai and T. Hisamitsu. A data structure for fast lookup of grammatically connectable word pairs in japanese morphological analysis. In *International Conference on Computer Processing of Oriental Languages (ICCPOL'99)*, pages 467–471, 1999.

[8] 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.