データベースシステム 124−17
情 報 学 基 礎 62−17
（２００１．５．２２）

対話的文書検索における文書クラスタリングの役割

岩山真, 丹羽芳樹, 西岡真吾, 高野明彦†, 久光徹, 今一修, 櫻井博文, 藤尾正和
(株) 日立製作所中央研究所
† 国立情報学研究所
iwayama@harl.hitachi.co.jp

適合性フィードバックにおいて，ユーザが効率良く適合文書を判定できることを目的に，二つの検索結果表示法「カテゴリーバー表示」「デンドログラム表示」を提案して評価した．これらはいずれもクラスタリングアルゴリズムを用いて検索結果を自動分類表示する．残念ながら再現率/精度の評価尺度では，両者の有効性を確認することはできなかった．しかし，ユーザとシステムとの対話ログを調べた結果,「デンドログラム表示」ではユーザが効率良く適合文書のまとまりを見つけていることがわかった．

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 Laboratory, Hitachi Ltd.
† National Institute of Informatics
iwayama@harl.hitachi.co.jp

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 はじめに

対話的な文書検索の多くは，関連度フィードバック (relevance feedback) という手法を用いて，ユーザとシステムが対話的に情報を交換しながら検索結果の向上を試みる．具体的には，検索結果の幾つかの文書に対してユーザが適合/不適合性の判定を行い，これらの判定を使ってシステムは検索要求を更新し新たな検索を行う．

関連度フィードバックによる検索精度向上の割合は，ユーザがシステムに与えた判定の数に大きく依存する [2]．我々は，ユーザが効率良くできるだけ多くの適合文書を選べるようなインターフェイスについて研究してきた [7, 5]．本論文では，文書クラスタリングを利用して検索結果を自動分類表示することの有効性を調べる．実際には「カテゴリーバー表示」と「デンドログラム表示」の二つの表示法について評価した．

カテゴリーバー表示では，クラスタリングアルゴリズムを用いて検索結果から 3 個の主カテゴリを見つけ，各文書とそれら主カテゴリとの距離をカラーバー表示する．ユーザは各々のカテゴリに注目して検索結果を並べかえることもできる．この表示法は，Scatter/Gather [3] で提案されている表示法や，Evans 等によって用いられた表示法 [4] と似たものでり，適切な分類に注目することで多くの適合文書を効率良く集めることができる．

デンドログラム表示では，階層的クラスタリングの結果をそのまま表示する．類似している文書対はなるべく近い場所に表示されるため，ある適合文書を種にして，その文書に類似する文書群も芋づる式に見つけることができる．

本研究では，NTCIR-2 [1] での実験を介して，上記二つの表示法を評価した．実験では実際にユーザとシステム間の対話を記録しているため，再現率/精度だけではなくユーザの行動も考慮した評価を行った．

2 検索結果の表示法

本研究で評価した検索結果の表示法を説明する．

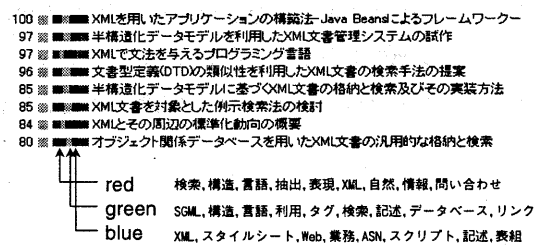
2.1 ランキング表示

検索要求との適合度が高い順に検索結果をならべて表示する．多くの検索システムで用いられているため，本研究でも提案手法のベースラインとしてランキング表示を用いる．ユーザは以下の操作を行うことができる．

AVISIT	指定した文書のフルテキストを表示する
ASEL	指定した文書 (複数可) に適合マークを付ける
AUNSEL	適合マークをはずす

2.2 カテゴリーバー表示

検索結果の各文書には，カテゴリーへの所属の度合を示すカテゴーバーが付いている．文書の初期並びはランキング表示と同じ適合度の順である．以下にカテゴリーバーの例を示す．これは，検索トピック「XML を用いた自然言語処理に関する論文」に対する検索結果の一部分である．



タイトルの横にある RGB スペクトルがカテゴリーバーである．それぞれの色 (R:赤，G:緑，B:青) は，検索結果 (実験では上位 150 文書) を要約する 3 個の主カテゴリに相当する．システムは階層的クラスタリングアルゴリズム [6] を使って，検索結果の文書群を 3 個のクラスタに分割する．これら 3 個のクラスタを主カテゴリとみなす．次に，各文書と 3 個のクラスタ間の距離を計算し，正規化の後に RGB スペクトルに変換する．よって，各色が占める割合は，対応するカテゴリーへの所属度の強さの割合に対応している．ここで，ユーザは各カテゴリの代表語を見ることができる．

VCAT	選択したカテゴリの代表語を見る

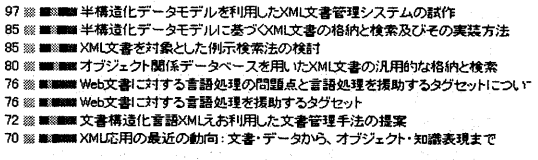
あるカテゴリに興味を持った場合，ユーザはそのカテゴリに注目して現在の検索結果を並べかえることができる．

GCAT	選択したカテゴリに注目して検索結果を並べかえる

この並べかえにより，注目しているカテゴリを代表する文書が上位に集まる．現在並べかえの方法として，

1. 注目カテゴリへの所属度が他カテゴリへの所属度よりも大きい文書のみを集める．ただし順位は検索要求との適合度の順である．
2. 単純に注目カテゴリの色の長さでソートする．

の2種類が選択可能である．以下は，上記の例に対して赤カテゴリに注目して並べかえを行った結果である．並べかえの方法は1.の方法を用いた．

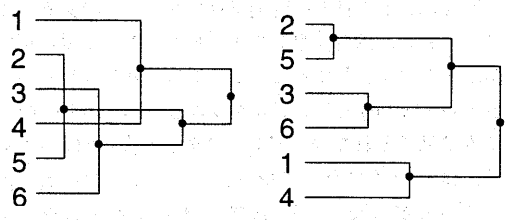


このように，カテゴリを介したインタラクションにより，ユーザは効果的に検索結果を絞りこむことができる．

なお，カテゴリーバー表示では，この他にも前述のAVISIT, ASEL, AUNSELコマンドが利用できる．

2.3 デンドログラム表示

デンドログラム表示では，階層的クラスタリングアルゴリズムの適用結果をそのまま表示する．階層的クラスタリングアルゴリズムでは，まずクラスタリングの対象文書それぞれを別々のクラスタとして設定する．次に一番近いクラスタ対をマージする．このマージを繰りかえすと最終的には以下のような木ができあがる．この木のことをデンドログラムと呼ぶ．

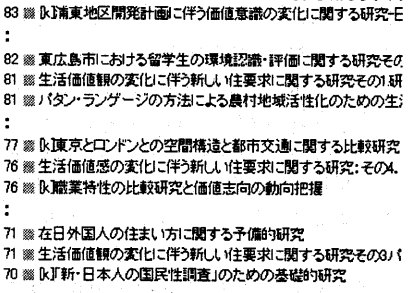


上図で左側のデンドログラムは文書の順序を整列しないで描いた木で，ここでは多数の枝が交差していて文書間の類似性が見にくい．右側の図は交差をほぐして文書を並べかえたデンドログラムである．類似する文書はできるだけ近くに配置されている．

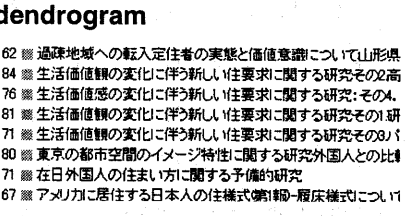
本研究では，デンドログラムの木構造は表示せず，並べかえ後のデンドログラム(上図右)における文書の順序のみをユーザに提示する．とはいっても，この順序そのものは有用であり多くの情報を含んでいる．例えば，あるユーザが文書3を適合文書として選んだ場合，すぐ隣りの文書6もおそらく適合文書である．なぜなら文書3と文書6は類似しているからである．このようにして，種文書の近くにある文書を見ることで，種文書に類似する多くの文書を見つけることができる．

似ている文書が近くに配置されることでタイトル間の類似/差異といった一覧性も良くなり，ユーザはタイトルを見るだけでこれらの文書の関係をとらえやすくなる．例えば以下の例を見てみる．これは検索トピック「日本人の生活価値観の変化」に対する検索結果の一部分である．

ranking



dendrogram



今，文書a, b, c, dに注目する．タイトルを見ればわかるように，これらは同じ著者による一連の論文である．ランキング表示では，適合度の値がばらついているため，これらの文書群が離れた位置に表示されているのに対し，デンドログラム表示では，まとまって近くに表示されている点に注目してほしい．よって，デンドログラム表示では文書a, b, c, dがシリーズを成していることは一目瞭然である．

デンドログラム表示では，AVISIT, ASEL, AUNSELコマンドが使用可能である．

3 実験環境

NTCIR-2 [1]の日本語検索に参加して，クラスタリングに基づく分類表示の評価を行った．以下は，

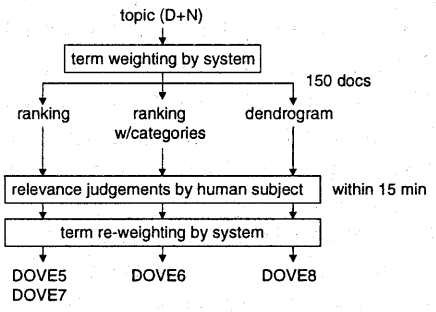


図 1: 対話的 run の概要

NTCIR-2 に提出した結果の作成手順である.

1. システムは各トピックから初期検索要求を作る. 具体的には, 各トピックの <DESCRIPTION> と <NARRATIVE> フィールドからストップワードを除く全ての単語を抽出して検索タームとした. 今回は <CONCEPT> フィールドは使わなかった. 形態素解析プログラムには ANIMA [8] を, 単語の重み付け法には Lt.Lnc 法 [9] を用いた. よって, 検索システムはベクトル空間モデルに基づいていることになる.
2. システムは初期検索要求から 150 文書を検索し, 以下のいずれかの表示法で各被験者に表示する.
   - ランキング表示 (2.1 節参照)
   - カテゴリーバー表示 (2.2 節参照)
   - デンドログラム表示 (2.3 節参照)
3. 各被験者は提示された検索結果から 15 分以内にできるだけ多くの適合文書をマークする. この間になされた操作は全て時刻付きで記録する.
4. 被験者がマークした適合文書をシステムにフィードバックして, システムは初期検索要求を更新する. 具体的には, 適合文書から上位 300 タームをフィードバックして改良 Rocchio 法により検索要求を更新した. 改良 Rocchio 法のパラメータは $\alpha = 8, \beta = 16, \gamma = 0$ とした. つまり, 負のフィードバックは行わなかった.
5. 更新した検索要求に基づいてシステムは再検索を行う. 検索結果の 1,000 文書を評価対象として提出した.

図 1 に概略を示す.

7 人の被験者 (著者ら) が実験に参加した.

実験目的は, ベースラインのランキング表示とクラスタリングに基づく二つの表示法とを比較することである. よって, 各検索トピックでの一貫性を保つために, 同じ被験者には二つの表示法について実験してもらった. 一つはベースラインのランキング表示で, もう一つはカテゴリーバー表示, デンドログラム表示のいずれかである. ただし, 二つの試行を行う順番が問題であるため, どちらを先に行なうかはランダムに決めた. 更に, 念のため, 二つの試行の間には最短でも一週間の間隔をおいてもらった. まとめると, NTCIR-2 には以下の 4 つの対話的 run を提出した.

DOVE5	ランキング表示 (DOVE6 のベースライン)
DOVE6	カテゴリーバー表示
DOVE7	ランキング表示 (DOVE8 のベースライン)
DOVE8	デンドログラム表示

4 実験結果と考察

4.1 総合結果

表 1 に平均精度 (average precision) を示す. ここで, S, A, B は適合性のレベルで, S は「特に適合」, A は「適合」, B は「部分的に適合」を意味する.

図からもわかるように, 残念ながら, カテゴリーバー表示 (DOVE6), デンンドログラム表示 (DOVE8) 共に, ベースラインのランキング表示を有為に上回ることができなかった. デンドログラム表示 (DOVE8) は, かろうじてベースラインを上回ったがその差は小さい.

図 2 は, 平均精度の時間推移である. 各時刻までにマークされた適合文書をフィードバックして得た平均精度をプロットしてある. ここでも, カテゴリーバー表示 (DOVE6) とデンドログラム表示 (DOVE8) の優位性は見てとれない. デンドログラム表示 (DOVE8) は, 15 分近くになってやっとベースラインに追い付いているが, ほとんどの時刻においてベースラインを下回っている.

4.2 適合判定について

表 2 に, 被験者が行った適合性判定のどれだけが正規の判定と一致したかを示す. つまり被験者が行った判定の精度である.

全 run において, 精度は 70%を上回っている. よっ

ID	other info	平均精度	
		S+A	S+A+B
DOVE5	ランキング表示 (DOVE6 のベースライン)	0.4095	0.4020
DOVE6	カテゴリーバー表示	0.3996	0.3943
DOVE7	ランキング表示 (DIVE8 のベースライン)	0.4052	0.3976
DOVE8	デンドログラム表示	0.4069	0.3891

表 1: 平均精度 (average precision)

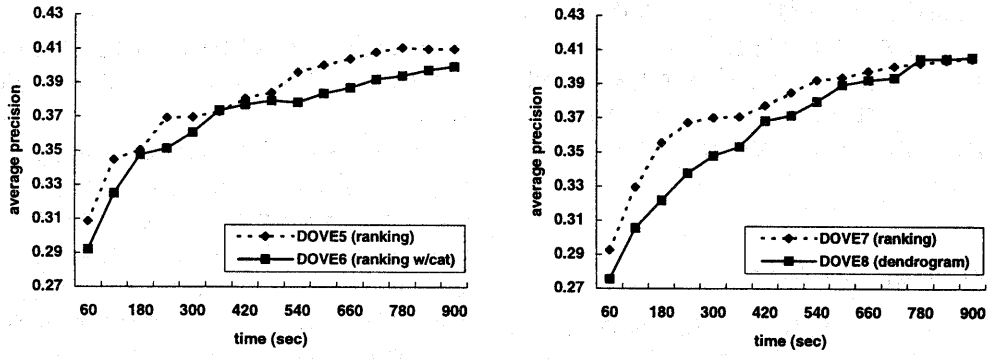


図 2: 平均精度の時間推移

ID	ASEL	S	A	B	C	(S+A)/ASEL query-averaged	(S+A+B)/ASEL query-averaged
DOVE5 (ランキング)	688	120	383	62	122	0.7264	0.8371
DOVE6 (カテゴリーバー)	598	94	325	50	129	0.7336	0.8202
DOVE7 (ランキング)	671	99	387	66	116	0.7201	0.8372
DOVE8 (デンドログラム)	543	87	319	40	94	0.7496	0.8368

表 2: 被験者が行った適合性判定の精度

て，被験者の判定はそれほど間違っていなかったことがわかる．また，S ランクと A ランクの正解文書に関しては，ランキング表示を用いるよりも，カテゴリーバー表示やデンドログラム表示を用いるほうが精度が高いこともわかる．ただし，差は大きくない．B ランクの正解文書も含めるとカテゴリーバ表示，デンドログラム表示共にベースラインに劣る．

一方，表 3 は，被験者による判定の再現率である．つまり，提示された 150 文書に含まれている正解文書のうち，どれだけを実際に見つけることができたかである．

図 3 からもわかるように，再現率は押しなべて低い．全てにおいて 45%を割り込んでいる．よって被験者は多くの正解文書を判定し逃していることになる．現在原因を調査中であるが，多くの場合は検索トピックの解釈の相違のようである．もし判定の精度/再現率が 100%ならば，検索の平均精度は S+A 判定で 0.5166， S+A+B 判定で 0.4775 に達っし，これらの値は本実験での上限に相当する．

また，いずれの run においても B 判定の再現率が低いのは，実験のインストラクションで「S および A 判定に相当する適合文書を探す」ことを指示したためであろう．

4.3 対話ログの解析

クラスタリングを用いた自動分類表示の利点は，お互いに類似している文書を簡単に見つけることができる点である．例えばカテゴリーバー表示では，あるカテゴリーに着目して検索結果を並べかえることで，そのカテゴリーという観点で検索結果が整理できる．デンドログラム表示では，近い文書対は近く配置されるため類似する文書が一覧できる．本節では，自動分類表示の利点を調べるために，実際に被験者が行った操作のログを解析する．

図 3 は，検索トピック「日本人の生活価値観の変化」に関する対話ログの一部である．ここでも，同じ著者による一連の論文 a, b, c, d に注目する．これらの論文は様々な適合性のスコアを持つため，ランキング表示では離れて表示されてしまう．かつ，ユーザは通常，上位から下位に向って文書を調べていくため，これら一連の論文を見る間に多くの無関係な論文も目にはいってしまい，a, b, c, d 間のまとまりを識別することが困難である．対話ログを見てみると，被験者は a, b, c, d の順に見ているのだが，その間隔は，55 秒，85 秒，141 秒と比較的長い．よって，前に見た文書の内容を忘れ，a, b, c, d 全てにおいてフルテキストを参照している．参照時間も比較的長い．

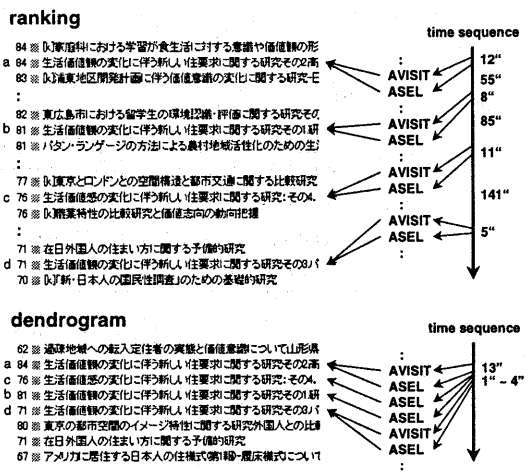


図 3: 被験者とシステムとの対話例

一方，デンドログラム表示では，a, b, c, d がまとまって表示されるため，被験者もまとまりとしてこれらの文書群が認識できる．対話ログを見ると，最上位の a についてはフルテキストの参照を行っているが，次の c, b については，タイトルを見ただけで適合性の判定を行っている．かつその間隔は非常に短い．最後の文書 d については，フルテキストを参照しているが，参照時間は 4 秒と短いため確認のための参照と言える．

表 4 に ASEL コマンドの連続数を示す．ASEL コマンドは適合性マークを付けるコマンドであるため，この連続が意味するところは，後ろの文書に対してはフルテキストの参照なしに適合と判断した可能性が高いということである．表 5 には，実際にフルテキストの参照なしに適合と判定できた文書数を示す．いずれの表でも，デンドログラム表示では，これらの値が大きい．ベースラインのランキング表示と比べると 2 倍程度である．一方，カテゴリーバー表示はここでもベースラインを上回ることができなかった．

最後に図 4 に，ASEL/AVISIT の割合を時間経緯で示す．ASEL/AVISIT の割合は，被験者が行った操作の効率性 (どれだけ無駄なフルテキスト参照がなかったか) を計っている．カテゴリーバー表示は，ランキング表示とほとんど同じ曲線を描いているのに対し，デンドログラム表示は，ランキング表示とは

ID	S	A	B	S+A	S+A+B
DOVE5 (ランキング)	0.2915	0.4150	0.1287	0.4483	0.4048
DOVE6 (カテゴリーバー)	0.2768	0.3591	0.1687	0.4092	0.3807
DOVE7 (ランキング)	0.2617	0.4142	0.1420	0.4429	0.3912
DOVE8 (デンドログラム)	0.2798	0.3761	0.1254	0.4090	0.3627

表 3: 被験者が行った適合性判定の再現率

ID	ASEL	ASEL without AVISIT		
		total	S+A	S+A+B
DOVE5 (ランキング)	688	76 (0.1105)	62	66
DOVE6 (カテゴリーバー)	598	65 (0.1087)	56	58
DOVE7 (ランキング)	671	51 (0.0760)	38	40
DOVE8 (デンドログラム)	543	107 (0.1971)	70	79

表 5: フルテキストの参照なしに適合と判断できた文書数

ID	
DOVE5 (ランキング)	65
DOVE6 (カテゴリーバー)	57
DOVE7 (ランキング)	50
DOVE8 (デンドログラム)	135

表 4: ASEL コマンドの連続数

傾向が異っている．約 100 秒を過ぎると，ベースラインが次第に下っていくのに対し，デンドログラム表示では依然として約 50%の割合を保っている．つまり，ランキング表示においては時間が増すにつれ無関係な文書を読む割合が増えるのに対し，デンドログラム表示では，セッションの後半になってもあまり多くの不適合文書を読むことがない．

5 おわりに

適合性フィードバックを効果的に行うための検索結果表示法として，二つの自動分類表示「カテゴリーバー表示」「デンドログラム表示」を提案し評価した．平均精度の観点からは両者の効果は認められなかったが，ユーザとの対話ログを調べた結果，特にデンドログラム表示の有効性が確認できた．検索結果をデンドログラム表示することにより，ユーザは類似する適合文書を効果的に集めることができた．カテゴリーバー表示に関しては，いずれの評価においても通常のランキング表示を上回ることがなかった．一つの原因として，インターフェイスが未熟でユーザが操作にとまどっていることが挙げられる．例えば，あるカテゴリーに注目した並べかえについて幾つかの手法をユーザに選ばせているが，明らかに複雑でわかりにくいインターフェイスである．今後は，カテゴリーバー表示のインターフェースを洗練化する予定である．

謝辞

本研究の一部は IPA 独創的情報技術育成事業の支援を受けて行われました．

参考文献

[1] *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[2] 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*

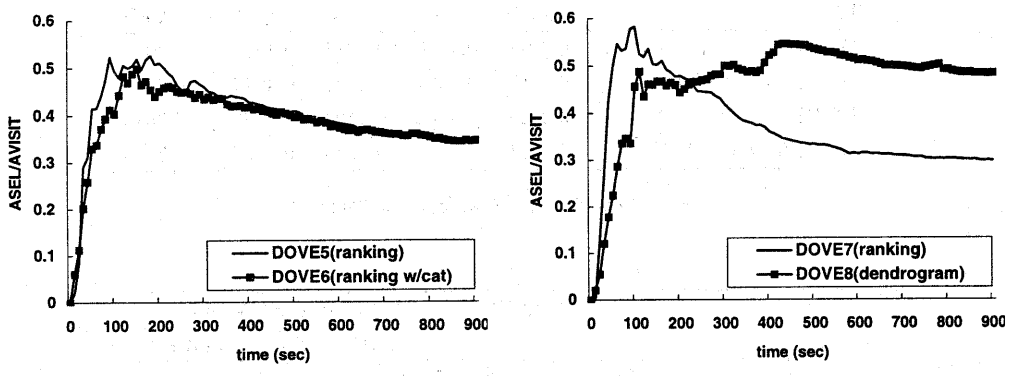


図 4: ASEL/AVISIT の時間推移

Research and Development in Information Retrieval, pp. 292–300, 1994.

[3] 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*, pp. 318–329, 1992.

[4] 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.

[5] 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*, pp. 10–16, 2000.

[6] M. Iwayama and T. Tokunaga. Hierarchical bayesian clustering for automatic text classification. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1322–1327, 1995.

[7] 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*, pp. 123–130, 1999.

[8] 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)*, pp. 467–471, 1999.

[9] 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*, pp. 21–29, 1996.