情報メディア 13－6
グループウェア 4－6
（1993. 10. 28）

情報のブロードキャッチシステム

湯浅寛子　　小島啓二

日立製作所　システム開発研究所

ネットワークを介した個人の情報収集活動の支援を目的として試作したブロードキャッチシステムとその試用結果について述べる。計算機ネットワークの整備により、種々の情報サービスからの広範囲な情報収集が可能になりつつあるが、大多数のユーザにとって、情報サービスへのアクセス、情報の検索、検索結果の整理といった操作は煩雑で、多忙さから利用されていないことが多い。本システムでは、ユーザが興味ある事柄に関するキーワードを羅列して登録するだけで、複数の情報サービスから情報を自動収集することができる。この収集結果は収集初期時には10%程度の情報収集精度であるが、収集結果を類似した情報群に随時クラスタリングする機能により、最終的には80%程度にまで向上させることができた。

An Information Broad-Catch System

Hiroko Yuasa and Keiji Kojima

Hitachi Ltd. System Development Laboratory

This paper describes an implementation and evaluation of an information filtering system which supports personal information collecting activities. Today, rapid progress of computer networks provides us a rich environment for accessing to various information sources. Many users, however, hesitate to enjoy the environment because required operations are complicated and troublesome (connecting to databases, inputting search commands, reviewing the retrieved results, and so forth). The proposed system watches information sources continually, and collects informations which match to user's interests. The system automatically classifies the collected informations when the amount of them crosses over a predetermined limit. According to a primary experimentation, this classification facility improves the precision of information filtering from 10% to 80%.

1. はじめに

計算機ネットワークの整備が急速に進み、電子メールや電子掲示板を利用した質疑応答、インタネットニュースからの情報収集、オンライン情報サービスの利用といった、いわゆる情報のブロードキャッチが行える環境が整いつつある。しかし、大多数のユーザ、特に計算機の非専門家にとって、種々の情報サービスへの接続、情報の検索、検索結果の整理といった操作は煩雑で、多忙さから利用していないことが多い。トレンドに敏感で、ネットワークを流れる最新情報の価値を認識してはいても、それらの情報をじっくり読む時間的余裕のない多忙な現代人の情報収集活動を支援するシステムが望まれている。

ネットワークを介した広域情報検索を支援するシステムとしては、WAIS[1,2] が知られている。WAIS(Wide Area Information Servers)は知りたい質問事項とアクセスする情報サービス名を登録すると、検索を行い、ヒットした文書情報にどの程度適切かという評価付けをして、そのタイトル（ポインタ）を返す。問い合わせるのに最適な情報サービスを見つける機能もあり、ネットワーク内の情報サービスのナビゲータの役割を果たす。また、関連性フィードバック機能により、検索された文書情報からユーザが所望の文書情報(または、その一部)を選んで登録すると、次回の検索をより効率良く行う。

WAISは広域情報サーバとして優れた機能を備えているが、WAISを利用しても収集される情報は膨大な量に上る。収集結果の分類、整理にも手間と時間がかかり、結局、未整理のまま有効に利用されないことが多い。

一方、情報の自動分類を狙ったシステムとしては、CONSTRUE/TIS[3]が知られている。

CONSTRUE/TIS (Topics Identification System)はデータベースに蓄積される情報を自動分類して、インデックスを自動付与し、データベースの検索に利用するものである。ロイターのニュースの自動分類に適用されて、90%以上の精度での分類を実現している。この高精度は、数人年のオーダの工数をかけて精密な分類ルールを作成することにより達成されており、個人向けのシステムへの適用は難しい。

そこで、個人が手軽に情報収集を行えるように支援することを目的として、ブロードキャッチシステムを試作した。

本システムは、ユーザが興味ある事柄に関するキーワードを羅列して登録しておけば、適合する文書情報を自動的に外部の情報サービスから収集し、収集結果の整理を行う。

本稿では、2章で試作システムの概要について説明し、3章と4章で情報収集と収集結果の整理方法について具体例を交えて説明する。5章で本システムの試用評価結果について述べ、6章で今後の課題などについて述べる。

2. システム構成

今回試作したブロードキャッチシステムはクライアント・サーバ型のシステムで、UNIX[注1] 上で稼働する(図1)。またクライアントのグラフィカル・ユーザ・インタフェース(GUI)はOSF/Motif[注2] を用いた。

現在、インタネットニュースと社内開発したCD-ROMベースの特許データベース上の文書情報を情報収集の対象としている。インタネットニュースには NNTP(Network News Transfer Protocol)によって接続しており、特許データベースには社内開発した専用のプロトコル接続している。

2.1 ブロードキャッチサーバ

ブロードキャッチサーバには大きく分けて二つの役割がある。

一つは情報収集と収集結果の整理である。

本システムのユーザには一つの「フォルダ」が割り当てられる。ユーザはこのフォ

[注1] UNIX はATTの登録商標である

[注2] OSF/Motif はOSFの商標である。

ルダに自由に興味ある事柄に関するキーワード群（インタレストと呼ぶ）を登録しておく(図2)。このフォルダの下に階層的にフォルダを作成してそれぞれにインタレストを登録してもよい(図3)。

サーバは各情報サービスに定期的にアクセスして、このインタレストに適合する新情報をフォルダに収集する。これを情報フィルタリング機能と呼ぶ。

さらに、フォルダに収集された情報数が多くなるとフォルダを自動分割することにより分類を細分化し収集結果を整理する。これをフォルダ自動分割機能と呼ぶ。

もう一つはクライアントの要求処理である。

クライアントからの情報収集結果の表示などの要求を処理する。また、クライアントと外部情報サービスの仲介役を果たす。つまり、クライアントが情報の内容表示などを要求すると、情報サービスに応じて適切なコマンドを発行して必要な情報を取得し、クライアントに提供する。したがって、ユーザは各情報サービスへの接続方法や検索方法の違いを意識せずに単一のインタフェースで利用できる。

2.2 ブロードキャッチクライアント

クライアントは本システムを利用するためのGUIを提供する。ユーザはクライアントを介して、インタレストの登録、フォルダの作成、収集結果のブラウジングを行う。図4にクライアントの画面イメージを示す。

3. 情報フィルタリング機能

本章では、情報収集を行う情報フィルタリング機能について説明する。これは、(1)各情報サービスからの新しい文書情報の取得、(2)いずれかのインタレストに適合する文書情報のフィルタリング、(3)適合した文書情報のフォルダへの格納、(4)適合した文書情報に出現するワードの頻度情報の蓄積、の四つのステップからなる。各ステップについて説明する。

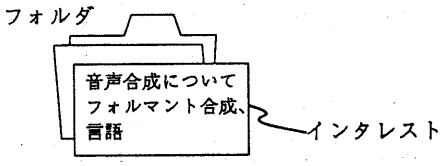


図2 フォルダとインタレストの例

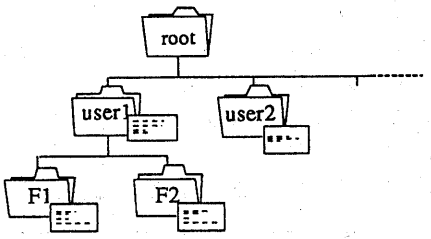


図3 フォルダの階層構造

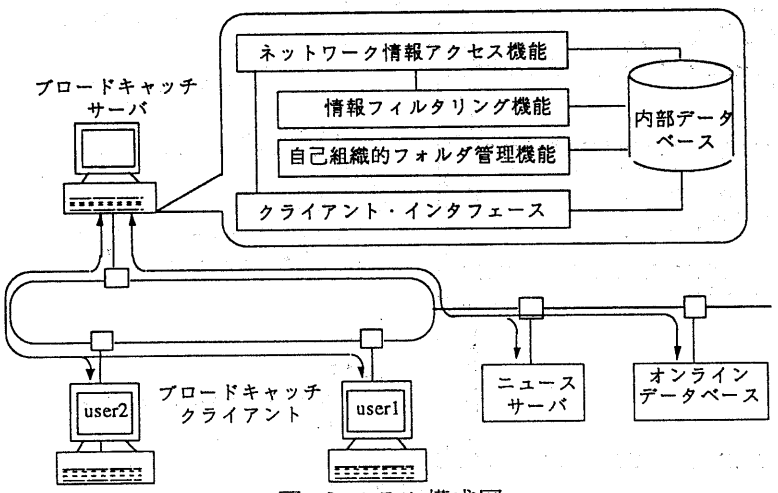


図1 システム構成図

3.1 新規文書情報の取得

ブロードキャッチサーバは、各情報サービスをウォッチして、新しく蓄積される文書情報を取得する。ドメイン毎に取得済みの文書情報番号を記録しておき、定期的に各情報サービスにアクセスして、それより新しい文書情報を取得するという方法を採っている。

たとえば、インタネットニュースでは、ドメインはニュースグループに、文書情報番号は記事番号に当たる。NNTPに従って、ニュースサーバにLISTコマンドを送ることによって、各ニュースグループに新しい記事が蓄積されているか調べて、もし新しい記事があれば、GROUP、HEAD、BODYのコマンドを使って記事を取得する。

3.2 フィルタリング

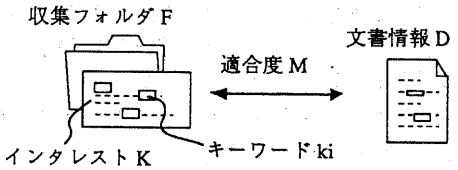
取得した新規文書情報が登録されているインタレストのいずれかに適合しているか検索を行う。

従来の検索システムでは、ユーザが記述した論理式を用いた検索式の真理値を計算することにより検索を行う。効率良く検索する検索式を書くためには、専門的な知識や経験が必要である。また、論理式そのものが非専門家のユーザにはなじみにくい。

本システムでは、ユーザは興味ある事柄に関するキーワードを羅列するだけで検索条件を登録できる。検索は、文書情報とインタレストとの「適合度」を計算することによって行う。

適合度は、インタレストに出現するワードの文書情報における頻度に基づいて図5のように定義する。

各ワードの頻度は全文検索により求めるが、ここで行う全文検索は通常とは異なる特徴を持つ。まず第1に、日々新しく作成され情報サービスに蓄積される大量の文書情報を対象とする点である。たとえば、インタネットニュースでは1日に20000件(約40MB)の文書情報が新しく追加される。特許データベースでは5000件(100MB)／枚



$$M = \sum_{ki\ in\ K} freq(ki\ in\ D) \times weight(ki)$$

図5 適合度の定義

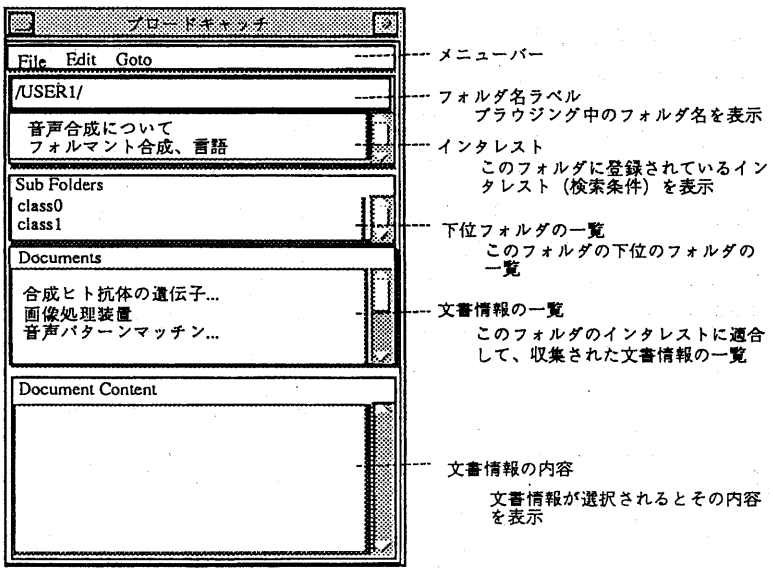


図4 クライアント画面例

のCD-ROMが週に2枚追加される。したがって、高速に検索するためにあらかじめインデックスを作成しておくといった手法は採れない。

第2に、検索条件が多い点である。通常の全文検索では大量のテキストデータベースにおいて一つの検索条件で検索を行う。本システムでは、複数のユーザが複数のフォルダに登録した複数のキーワードを含むインタレストについて同時に検索を行う必要がある。

そこで、本システムでは次のような方法を採った。

まず、全フォルダに登録されたインタレストからワードを切り出し、ハッシュテーブルに登録しておく。このとき各ワードがどのフォルダのインタレストから切り出されたものかわかるようにポインタを張っておく。このテーブルをワード・フォルダテーブルと呼ぶ。

新規文書情報が取得されたら、これからも同様にワードを切り出し、ワードとその頻度をハッシュテーブルに登録する。このテーブルをフォルダ検索テーブルと呼ぶ。ワード・フォルダテーブルとフォルダ検索テーブルを照合することにより全文検索を行い、対象文書情報とインタレストに共通に出現しているワードを調べる。フォルダ検索テーブルに登録されている各ワードの頻度に基づいて、フォルダ毎に対象文書情報との適合度を計算する。

3.3 フォルダへの格納

適合するフォルダがあった文書情報は、インデックス（情報サービス名、ドメイン名、文書情報番号など）をフォルダに格納する。複数の適合フォルダがあった場合、どのフォルダに格納するかは各フォルダとの適合度とフォルダの階層構造から決定する。すなわち、フォルダが下位であるほど分類が細分化されていると考えられるので、適合したフォルダの中でより下位のフォルダに文書情報を格納する。階層構造の枝分れした複数のパス上に該当するフォルダがある場合には複数のフォルダに格納する(図6)。

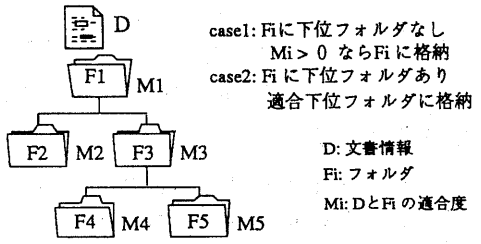


図6 フォルダへの格納

3.4 ワード頻度情報の蓄積

適合した文書情報は、その文書情報にどのようなワードがどのくらいの頻度で出現しているかというワード頻度情報を記録しておく。

ワード頻度情報に現われる文書情報の特徴を用いて後述するフォルダの自動分割を行う。

4. フォルダの自動分割

ネットワークを介して収集される文書情報は、情報フィルタリング機能によってふるいにかけたとしても膨大な量になる。収集した文書情報の整理にも時間と手間がかかり、収集した文書情報を有効に利用できない。

ブロードキャッチサーバは、文書情報の収集状況を監視し、収集状況に応じて、その文書情報群を類似した文書情報群にクラスタリングして、フォルダを分割する。

以下、自動分割の方法について説明する。

4.1 文書情報のモデル化

類似した文書情報を集めてクラスタリングするために、文書情報をワードベクトルで表わす。

ワードベクトルとは文書情報に出現するワードの頻度の並びで、文書情報収集時に記録したワード頻度情報から作成することができる。

文書情報間の類似性をワードベクトルを用いて定義した文書情報間距離で測る。

文書情報Di, Dj のワードベクトルをそれぞれWi, Wjとし、WiとWjのなす角度をθ

とするとDi, Dj間の距離 d(Di, Dj) は次のように書ける。

$$d(Di, Dj) = 1 - \frac{Wi \cdot Wj}{|Wi| \times |Wj|} = 1 - \cos\theta$$

(ただし、・は内積、|Wi|はWiの大きさ)

4.2 クラスタリング

ブロードキャッチサーバは各フォルダに収集される文書情報数を監視している。あらかじめ定めておいた境界値以上の数の文書情報が収集されたフォルダがあると、その文書情報群のクラスタリングを始める。

まず、1文書1クラスタとして文書情報数分のクラスタを作成する。文書情報間の距離をクラスタ間の距離として距離テーブルを作成する。

次に、距離テーブルを参照して、最も近いクラスタを選び統合する。統合されたクラスタはその重心のワードベクトルで表わし、距離テーブルを更新する。

あらかじめ定めておいた距離よりも近いクラスタがなくなるまで、クラスタのマージと距離テーブルの更新を繰り返す。

図7にクラスタリングの様子の具体例を示す。図7(a)は図2に例として挙げたフォルダに格納された文書情報群を適合度順に並べ、インタレストに出現するワード毎の出現頻度を濃度グラフにしたものである。頻度が高いほど白く表現されており、黒は頻度が0であることを示す。

図7(b)はその文書情報群をクラスタリングし、文書情報をクラスタ毎に並べ替えて、同様の濃度グラフにしたものである。図7(b)から元の文書情報群が四つの文書情報群に分れることがわかる。すなわち、「音声」と「処理」の両ワードが出現することを特徴とする文書情報群、「言語」と「処理」の両ワードが出現することを特徴とする文書情報群、「音声」のワードが出現することを特徴とする文書情報群、際だった特徴が現われていないその他の文書情報群である。

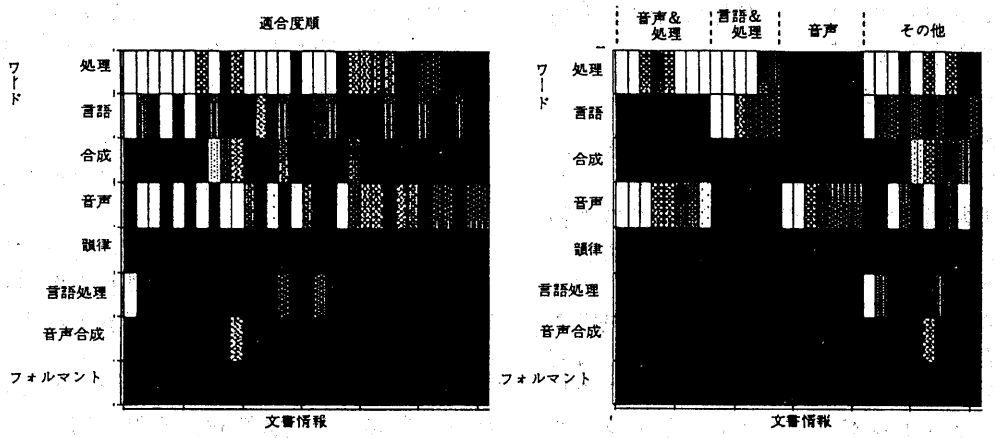
4.3 新フォルダ作成

クラスタに対応して下位フォルダを作成する。

各クラスタに属する文書情報を元のフォルダから対応する下位フォルダに移し、文書情報に共通して出現するワードを各下位フォルダのインタレストとする。

図7の例では顕著な特徴がみとめられる三つの文書情報群にそれぞれ対応するフォルダを元のフォルダの下位フォルダとして作成し、文書情報を移動する。際だった特徴が現われていないその他の文書情報は、元のフォルダにそのまま残す。

元のフォルダに残った文書群もこの後、



(a) クラスタリング前　　(b) クラスタリング後

図7 クラスタリングの例

文書情報の収集を継続して行い、このフォルダに格納される文書情報が増えてくると、特徴が顕著になってくることがある。その場合には、再びフォルダの自動分割を行う。また、生成された下位フォルダにも多くの文書情報が収集されてフォルダの自動分割を行うこともある。

5. 試用評価

平成4年9月より、インタネットニュースサーバへのアクセスと情報フィルタリング機能を組み込んだシステムを当社中央研究所内で募集した25ユーザに試用していただいている。そのユーザの方々に登録していただいたインタレストを用いてフォルダの自動分割の実験を行っている。

5.1 ボランティアユーザによる試用

試用していただいているユーザは、主にコンピュータ関係の研究者であり、普段からインタネットニュースを読んでいる人が多い。したがって、本システムがターゲットとしている非専門家とは言えないが、ニュースから情報収集する時間がなかなか取れない多忙な現代人ではある。このユーザの方々にいただいたご意見を紹介する。

- ふだんは読んでいないニュースグループの興味ある記事を見つけることができた。
- インタネットニュースではいつも読むニュースグループは決まっているし、良いニュースリーダがあるのでブロードキャッチの必要を感じない。
- 無関係の記事が多く収集されてしまうので、情報源とするニュースグループを限定したり、ニュースグループ毎に検索条件を変えるなどのカスタマイズをしたい。
- 従来の検索システムのように自分で論理式を書きたい。
- フィルタリングや分類が自動的に適切になされていて、良いGUIが用意されていても、わざわざクライアントを立ち上げて、フォルダの中身を見に行くのは面倒である。

このほかクライアントの使い勝手に関するご意見も多数いただいた。

5.2 情報収集精度の評価

試用ユーザの方々が登録したインタレストを使って特許文書サーバから情報収集を行い、フォルダ自動分割実験および情報収集精度(precision)の評価を行った。

情報収集精度とは検索結果の中にユーザがその検索条件によって検索しようと意図していた有用な文書情報（以下、有用文書情報と呼ぶ）がどのくらいあるかを現す指標で次の式で表される[4]。

$$\text{情報収集精度} = \frac{\text{有用文書情報数}}{\text{検索された文書情報数}} \times 100$$

情報収集精度の評価には二つの問題がある。まず、有用文書情報かどうかは主観的に評価せざるを得ない。次に、数万件もの膨大な数の文書情報を対象としているため、その中に有用文書情報が何件あるかを検証することは事実上不可能である。そこで、次のような方針で評価を行った。

(1)インタレストの重要と考えられるワードが出現しない文書情報は、有用文書情報ではないとみなす。

(2)上記以外の文書情報については内容を検討して有用文書情報であるかどうかを判断する。

(3)有用文書情報かどうかは、インタレスト毎に内容に応じて基準を設け報告者が評価する。

この方針に基づき、特許データベースの文書情報24000件を対象として情報フィルタリング機能およびフォルダ自動分割の実験を行い、精度を評価した。

図8に図2の例のフォルダにおける文書情報の収集の進行に伴う精度の変化を現すグラフを示す。

一般に全文検索を行った場合、精度は高々10%程度と思われる。実際にこの実験でもフォルダの自動分割機能なしの場合、終始10%前後の精度で推移し最終的には8%であった。

これに対し、フォルダの自動分割機能がある場合には最終的に約80%の精度が得られた。図8の最高の精度の推移をみると精度の向上は段階的に進んでいる。大きく精度が向上しているのは、フォルダの自動分割により意図しない文書情報ばかりが格納されたフォルダが生成されたり、ユーザの意図を良く反映した精度の高いフォルダが生成されたりしたときである。

自動分割により生成されたフォルダのうち、精度の高いフォルダでは、文書情報数が数十件程度に押さえられ、精度の低いフォルダにはユーザが意図していなかった無関係の文書情報が大量に収集された。

6. 終わりに

本稿では、個人の情報収集活動の支援を目的として試作したブロードキャッチシステムについて述べた。本システムの情報フィルタリング機能により、ユーザは興味ある事柄に関するキーワードを羅列するだけの手間で情報サービスの違いを意識せずに情報収集が可能となった。またフォルダ分割機能により、情報収集精度10%程度の収集結果から80%の精度の文書情報を集めたフォルダを生成することができた。

本システムの試用ユーザには、従来の方法では見逃していた情報を検索できたといった本システムの情報フィルタリング機能の効果を認める評価をいただいた一方で、本システムのインタフェースのほかに従来の検索式を記述するインタフェースも欲しいという意見もいただいた。

この意見は情報収集精度の低さへの不満から生じている。この問題はフォルダ自動分割機能といった収集情報を整理する機能により解決できると考えている。今後、フォルダ自動分割機能の改良とフォルダ分割機能を組み込んだシステムの試用評価実験を行う予定である。

今回、インタネットニュースと特許データベースを対象として、文書情報の収集実験を行ったが、インタネットニュースは90%以上が英語文書であり、特許文書には特殊な用語や言い回しが多く用いられている。いずれも一般的な日本語文書とは言えない。今後は、商用データベースやパソコンネットなど一般的な日本語の大規模な情報サービスを対象としたより精密な評価実験が必要である。

[参考文献]

[1] B.Kahle, Wide Area Information Server Concepts, Thinking Machines Technical Memo (1989)

[2] 21世紀の情報化社会, 日経バイト1991. 11 (1991)

[3] P.J.Heyes, S.P.Weinstein, CONSTRUE /TIS: a System for Content-Based Indexing of a Database of News Stories, Proc. of Second Annual Conference on Innovative Applications of Artificial Intelligence (1990)

[4] G.Salton, M.J.McGill, Introduction to Modern Information Retrieval, McGraw-Hill (1985)

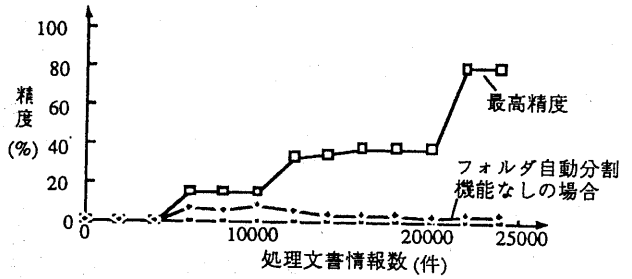


図8 情報収集精度の向上過程