デジタル・ドキュメント　1－3
（1996．5．27）

フロー情報の収集・活用を支援する検索システム

伊藤 史朗，　柴田 昇吾，　上田 隆也，　池田 裕治

キヤノン(株) 情報メディア研究所

ストック情報を利用する従来の電子化文書の利用形態とは異なり，ネットワークを通して流入するフロー情報を収集し，後に活用する形態が一般的になりつつある．こうした形態では，個人により継続的かつ日常的に情報が利用される点に特徴がある．継続的利用が進むと，情報に対する視点が変化する．一方，日常的利用にあたっては，収集にかける時間を短くする必要がある．

我々は以上の点を考慮し，フロー情報の収集と活用を支援するシステムを構築した．このシステムでは，階層のないフォルダで文書を保持し，視点の変化に対応しやすくしている．その上で，収集の時間短縮と活用の精度向上を図るための種々の機能を実現した．また，新聞記事を用いた利用実験により，上記の効果を確認した．

An Information Retrieval System to Collect and Reuse Flowing Information

Fumiaki ITOH,　Shogo SHIBATA,　Takaya UEDA,　Yuji IKEDA

Media Technology Laboratory, Canon Inc.
890-12, Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa, 211, Japan

We have developed a new information retrieval system to collect and reuse flowing information which is used more popularly than stocked information recently. It becomes easy to change points of view to the information by saving documents into folders without hierarchy. Some retrieval functions are implemented
1) to reduce the time for collecting and saving documents and
2) to retrieve the saved documents more precisely.
The evaluation by using the system to read some newspaper, to save articles and to retrieve them shows the effectiveness of these functions.

1 はじめに

パーソナルコンピューターおよびコンピューターネットワークの普及により，電子化文書の利用形態に変化が見られる．

従来の電子化文書の利用形態は，オンラインデータベース，CD-ROM，インハウスデータベースなどの文書データベースから文書を検索して利用することが中心であった[1]．これらの文書は，特定の利用目的が生じた時に利用され，情報としての価値は蓄積されていることにある．そこで，これを**ストック情報**と呼ぶ．ストック情報の利用者は限られていたが，ネットワークの普及でストック情報がオープンになり，利用が広まってきている．

一方，電子メール，ネットニュース，WWWなど，ネットワークを通して流入する電子化文書の利用が一般的になりつつある．これらの文書は，新しい情報を収集するために利用され，情報としての価値は新しい情報が流れてくることにある [5]．そこで，これを**フロー情報**と呼ぶ[2]．さらに，電子化文書は保存が容易なので，フロー情報は後日の活用のために保存されることが多い．こうして，利用者により保存された情報を**セーブ情報**と呼ぶ．

このように，電子化文書の利用形態は，限られた目的でストック情報を利用する形態から，オープンなストック情報・フロー情報・セーブ情報を複合的に使う形態へと変化してきている．それに伴ない，電子化文書の有効利用を図る検索システムも，新しい利用形態に即したものが求められている．

従来の検索システムは，ストック情報の利用を支援することが主目的であった．一方，フロー情報の収集を支援するシステムとして，利用者にとって有益な情報を選択する情報フィルタリングシステムの研究が進められている．しかし，ストック情報の利用，フロー情報の収集，セーブ情報の活用は相互に関連するものであり，この三つのフェーズを統合して支援するシステムが必要であると我々は考えた．

このようなシステムは，次の二点がストック情報の検索システムと異なる．

- 同一人により継続的に利用される．
- 日常的に利用される．

フロー情報の収集もセーブ情報の活用も継続性が重要である．利用者が収集したい情報は，過去に見た情報に依存するし，セーブ情報は収集および活用の中で熟成されていく．この点が，一回の検索で完結するストック情報の検索とは異なる．また，フロー情報の収集は日々行なうものであり，特定の利用目的が生じた時だけ利用されるストック情報の検索とは異なる．

我々は，上記の特徴をまず考慮して，フロー情報の収集とセーブ情報の活用に限定して検討を行なった結果，

- 情報に対する視点を保持し利用すること
- 視点の変化を前提とし，収集時の時間短縮と活用時の精度向上を図ること

が重要であると考えた．そこで，階層のないフォルダにより視点を表現した上で，時間短縮と精度向上を実現する支援システムを作成した．本稿では，このシステムについて，支援目的，支援機能，システム構成，利用実験による評価結果を報告する．

2 支援目的

2.1 視点の変化への対応

情報を利用する場合，ある視点から見て関連する情報を集めて利用することが多い．例えば，電子化文書に関する情報を，表現形式の視点から見たり，利用事例の視点から見たりする．全ての視点を定式化することは難しく，利用者の視点に完全に合う情報を情報検索により得ることは困難である．

そこで，既存のストック情報の利用システムでは，扱う視点を限定し固定している．情報を分類

[1] ここでいうデータベースには，単純なファイルシステムにより実現されるものも含めている．また，単純に文書を取り出すことも含めて検索と言っている．

[2] WWWは基本的にはストック情報であるが，新しい情報を提供することを目的にフロー情報的に用いられることも多い．また，ネットサーフィンのような利用は，利用者にとってのフロー情報を探していると考えられる．

して提供するシステムでは，提供者が体系化した分類に限定されている．また，データベースシステムではスキーマを定義するが，これも限定されたものである．こうしたシステムでは，情報を多数の利用者に提供するために，一般的な視点 (分類やスキーマ) が用いられる．

ところが，個人ごとに異なる視点を利用したい場合もある．特に，セーブ情報に対しては，個人の視点を利用できるように保存しないと意味がない[3]．一般的な視点と同じであればストック情報を利用すればよいからである．

フロー情報の閲覧時に，個人の視点を付与することは可能である．しかし，個人の視点は，継続的な収集・活用を進めるに従って変化する．時間が経つにつれて新しい情報の発生，個人の考えの変化，新しい業務の発生などにより視点が変わる．図1は，時間の推移と視点の変化を模式的に表した図である．A のパスは，活用時と同じ視点を収集時に付与することを示しているが，視点が変化することを考えれば，このパスはあり得ない．従って，視点が変化しないことを前提とした従来のデータベースシステムは，セーブ情報の活用を支援するシステムとしては適さない．視点の変化に対応した支援システムが必要である．

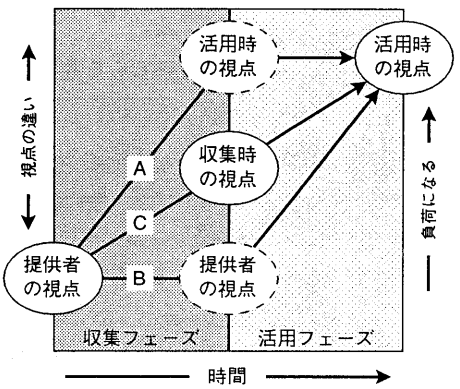

図 1: 視点の変化

[3]コスト面を考えれば，保存により再利用のコストが下がることはある．

2.2 活用時の検索精度の向上

代わって，個人の視点を付与せずに情報を保存し，活用時に情報検索を行ない視点に合う情報を集める方法も考えられる (図1のBのパスに相当)．しかし，前述したように現状の情報検索技術では，情報を完全に集めることはできない．検索で漏れた情報を得ることはできないし，情報が過剰に検索された場合も，件数が多いと，その中から適切な情報を選ぶのは事実上困難である．

ここで，活用時の視点と異なるとはいえ，収集時の視点が情報に付与されていれば，この視点を用いて検索精度を上げることが期待できる．収集時の視点を用いた検索を加えることで，従来の検索では漏れる情報に到達したり，過剰な検索結果を絞ることが可能になる．そこで，図1のCのパスを実現し，活用フェーズで検索精度の向上を図る支援システムが必要となる．

2.3 収集時の時間短縮

フロー情報の収集は日常的に行なわれることを考えると，従来の検索システムに比べて利用時間の重要性は高い．日々の限られた時間内で，多くの情報に目を通せるようにすべきである．そのためには，情報の閲覧にかける時間を多くし，情報の保存に時間をかける時間を短縮することが望ましい．

体系的な分類を作ったり，属性を付与してデータベース化すると，保存にかかる時間が多くなる．従って，これらは図1のCのパスの実現方法としても不適切である．収集フェーズにおける保存のための時間を短縮する支援システムが必要である．

2.4 情報の収集と活用に対する支援

以上の議論をまとめると，フロー情報の収集とセーブ情報の活用では，

- 変化を前提として視点を表現し，
- 収集時に短時間で視点を付与でき，
- 活用時に視点を用いた高精度の検索ができる

ことが重要である．そこで，以上の点を実現する支援システムを考えた．

3 視点の表現方法

容易に付与できる視点の表現方法として，視点に沿って文書を分けて保存する方法がある．実際，多くのメールシステムやWWWブラウザなどは，文書をフォルダに分けて保持している．この方法では，収集時に視点に沿って文書をフォルダに分け，活用時にフォルダを利用する．視点の変化はフォルダの変化ということになる．

フォルダを利用する場合，従来は高々数十のフォルダに分けて保持するか，これを越える場合には階層的に保持するかであった．というのも，フォルダが数十を越えると一覧性が悪くなり，視点の付与も利用も困難になるからである．しかし，この保持方法では次のような問題が生じる．

- フォルダの数を抑えるために，複数のフォルダに分けるべき文書を一つにまとめがちである．
- 階層を作ると，階層を越えた変更が困難になり，視点の変化に対応しづらくなる．

そのため，継続的な利用を進めるうちに，フォルダが視点を適切に反映しなくなる．

そこで，本システムでは，フォルダによる視点表現を採用するが，次の点を実現するようにした．

- フォルダの階層をなくす．
- フォルダの個数を大きく取れるようにする．

この場合，前述したように一覧性の低下による問題が生じる．そこで，後述する支援機能によりこの問題を回避することを目指した．

フォルダによる視点表現では，文書の事例が視点を表している．これ以外にも，視点を表すものとして，次の二点を保持できる．

- 視点を自然言語で表現したラベル
- 視点に沿った文書を集められる条件

ラベルは利用者が視点を直観的に把握するのに有効であり，条件は定式的に表現できる視点に対して有効である．

なお，松永は，名前，キーワード，文書本体の組を保持した情報活用方法を提案しているが[3]，本システムでは，ラベル，条件，文書事例の組で視点を表現している点，視点を用いて情報の収集と活用を行なう点が異なる．

4 本システムの機能

4.1 収集フェーズの機能

収集フェーズでは，多数のフォルダへの保存時間を短縮し，情報の収集を容易に進められるようにするため，以下の二つの機能を用意した．

保存候補のリストアップ機能

文書の保存先として適当なフォルダをリストアップする機能である．

文書の保存にあたって，次のような理由で，保存先を探すのに時間がかかることがある．

- 最後の利用から時間が経っているフォルダは，それが何であったか，あるいは存在そのものを忘れている．
- 多数のフォルダから，望みのフォルダを探すのが困難である．

だが，前述したように収集に時間がかかることは望ましくない．実際，収集に時間がかかると，閲覧する文書を減らしたり，次のようにして無理に時間を短縮することになる．

- フォルダ数を少なく抑える．
- 不適当なフォルダに入れる．
- 複数のフォルダに保存することを減らす．
- 保存を止める．

この場合，フォルダが視点を正しく反映しなくなる．

そこで，保存候補のリストアップ機能を用意した．この機能により，適当なフォルダがリストアップされれば，時間をかけずに保存ができる．このとき，正しい候補だけをリストアップするのではなく，一覧できる候補の中に正解が入れば十分である．従って，現状の検索技術の精度でも実現可能である．

視点に沿った文書提示機能

利用者が指定したフォルダに保存すべき文書を，視点ごとに分けて提示する機能である．

文書が利用者の視点に沿って提示されると，効率的に収集を進められる．文書の集合ごとに，見る時間を調整したり，関連する文書をまとめて見ることができるからである．

なお，視点別に提示される文書については，リストアップ機能において保存先の候補に入る．また，視点別に提示されるのは，利用者が指定した視点だけである．

4.2 活用フェーズの機能

活用フェーズでは，既存の視点[4]を利用した検索を進められるように，多数のフォルダから，活用時の視点に近いフォルダを検索できるようにした．検索機能には，次の二つがある．

条件検索機能

利用者が検索条件を与えて検索を行なう機能である．検索条件に，より合致しているフォルダを検索する．検索条件は，検索語とそのブール演算式として指定する．

類似検索機能

利用者がフォルダを一つ与えて，それに近いフォルダを検索する機能である．活用の視点に近いフォルダを見つけた後，さらにそれに近いフォルダを検索できる．指定するフォルダは，条件検索機能により検索するか，フォルダのリストから探し出す．

両機能とも，文書を検索単位とすることも可能であるが，フォルダを単位とすることで視点が有効に利用される．なぜならば，フォルダ中の文書の共通部分が，特に視点を反映していると考えられるからである．フォルダ中の一部の文書よりも，全体として条件を満足したり類似している方を優先することで，より活用の視点に近いフォルダが得られる．また，検索されたフォルダの視点は，活用の視点に近いので，両者の相関関係は高いと考えられる．検索されたフォルダ内には，単独では条件を満足したり類似していると判定されなくても，活用の視点に合っている文書が含まれる可能性が高い．

[4]後述する活用時に新たに作られる視点も含む．

検索の結果，活用の視点に一致するフォルダが見つかれば，そのフォルダをそのまま利用すればよい．しかし，活用の視点は，それまでの視点と異なることの方が多い．この場合，近いフォルダを利用して，活用の視点に沿った文書を集めることになる．ここで，新たに作成される文書の集合を，そのまま新たなフォルダとして保存できる．このフォルダは，そのまま活用を進める場合にも直ちに利用できる．

5 本システムの構成と動作

前述の機能を実現するため，本システムは図2のような構成になっている．

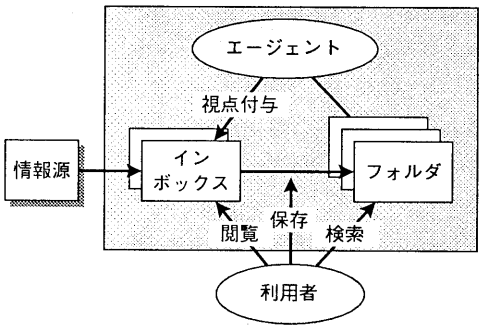

図 2: システムの構成

5.1 本システムの構成要素

文書

文書 d は，次のデータから構成される．

$$d = (t, p, k, v_d)$$

ここで，p と kは，フロー情報の情報源から渡されるテキスト t の大意とキーワードである．v_dはベクトル空間モデル [4] に基づきテキスト t の特徴を表現したベクトルである．いずれも，テキスト t が本システムに到着した時に作成される．ベクトル v_dの作成に用いる語群 [1] は，予め定まっている必要がある．この語群は，保存されている文書を用いて定期的に計算され，同時に全てのベクトル v_dが再計算される．

インボックス

インボックス Iは，ユーザの目に触れる前の文書を保持する．情報源に1対1に対応し，情報源に合わせて文書を提示する情報源別インボックス I_s と，視点に沿って文書を提示する視点別インボックス I_f がある．新たに到着した文書 d は，いずれも情報源別インボックス I_s に入る．

フォルダ

フォルダは，ある視点から見て関連した文書の集合を保持する．フォルダ Fは，次のデータから構成される．

$$F = (l, D, c, v_F)$$

ここで，lはラベルであり，利用者がフォルダを視認するための文字列である．Dは，フォルダに保存される文書の集合である．c は，検索語とそのブール演算式で表現される検索条件である．いずれも利用者により設定される．Dと c は空でもよい．v_Fは，ベクトル $v_d(d \in D)$ の平均をとったベクトルである．v_Fは，Dが変化する度に再計算される．

エージェント

エージェント A は，次のデータを持ち，ユーザの代理となって文書を処理するモジュールである．

$$A = (F, r)$$

ここで，Fはエージェント A が担当するフォルダであり，rはエージェント A の動作を定めるルールである．共に利用者により設定される．

以上の他に，次の関数を実現するモジュールがある．

$$f(c,d) = \begin{cases} 1 & d \text{が} c \text{を満足するとき} \\ 0 & d \text{が} c \text{を満足しないとき} \end{cases}$$

$$g(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}$$

文書 d が条件 c を満足するかどうかは，全文検索 [2] により判定される．フォルダに保存されている文書に対しては，インデックスが作成されており，保存されている文書については，このインデックスが用いられる．それ以外の文書に対してはパターンマッチングにより関数 fが実現される．

5.2 機能の実現

保存候補のリストアップ機能

インボックスの表示部では，利用者により選択された文書 d のテキスト t，大意 p，キーワード k が選択的に表示される．このとき，全てのフォルダ Fに対して，$g(v_d, v_F)$ が計算される．そして，保存候補として，この値の大きい順にフォルダのリストが表示される．

視点に沿った文書提示機能

エージェント A は，情報源別インボックス I_s を定期的に監視しており，未処理の文書 d があるとルール rに従った処理を行なう．ルール rは，担当フォルダ Fに対して，$f(c,d)$ か $g(v_F, v_d)$ の値と閾値との関係を条件として取り，文書 d の移動を動作として取ることができる．$f(c,d)$ か $g(v_F, v_d)$ のうち指定されたいずれかの値が閾値を越えた場合に，文書 d を情報源別インボックス I_sから視点別インボックス I_fに移すルールにより，フォルダ Fの視点に沿った文書は，視点に沿ったインボックスで提示されることになる．

条件検索機能

利用者により与えられた条件 c_tに対して，各フォルダのスコア s_Fを，次のように計算する．

$$s_F = \frac{\sum^{d \in D} f(c_t, d)}{|D|}$$

そして，s_Fの大きい順にフォルダを検索結果として提示する．

類似検索機能

利用者により与えられたフォルダ F_tに対して，各フォルダのスコア S_Fを，次のように計算する．

$$S_F = g(v_F, v_{F_t})$$

そして，S_Fの大きい順にフォルダを検索結果として提示する．

6 本システムの評価

本システムが目指している

- 収集フェーズでの時間短縮
- 活用フェーズでの精度向上

の効果を評価するため，以下の利用実験を行なった．

6.1 収集フェーズでの時間短縮の評価

各被験者が次の二つの方法で収集を行なった．

A 本システムを使って収集を行なう．

B 「保存候補のリストアップ機能」と「視点に沿った文書提示機能」を外したシステムを使って収集を行なう．

5名の被験者を二つのグループP, Qに分け，表1に示すスケジュールで実験を行なった[5]．実験の開始時点は，各被験者ともフォルダがない状態から始めた．

表 1: 収集実験のスケジュール

週	1	2	3
Pグループ	A	A	B
Qグループ	A	B	A

利用時に，次の三つの時間を測定した．

探索時間 閲覧する文書を探している時間．

閲覧時間 文書を読んでいる時間．

保存時間 文書の保存に要する時間．

各時間は，次のようにして推定した．

- 文書が選択され，その本文が表示されている時間から2秒(2秒以下の場合は当該全時間)を引いた時間を閲覧時間とする．
- 保存先を指定するパネルを開いてから，保存操作が完了するまでを保存時間とする．
- それ以外の時間を探索時間とする．

この測定時間に基づき，第2週と第3週のデータを用いて，AとBによる違いを調べた．

図3は，各時間の占める割合を5人の被験者に対して単純平均した値を示している．

[5]次に示す実験とも，日経テレコンBizからダウンロードした1996年4月の日本経済新聞および日経産業新聞の記事を用い，月曜から金曜の毎日収集を行なった．

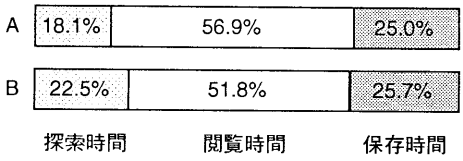

図 3: 所要時間の割合

表2は，1文書あたりの平均保存時間について，Bによる値からAによる値を引いた値を示している．なお，被験者a, b, cがグループPに，他がグループQに属していた．

表 2: 保存時間の短縮状況

被験者	平均	a	b	c	d	e
短縮時間 [秒]	1.0	0.0	2.8	3.9	-3.7	1.8

図3の通り，本システムにより閲覧時間の割合を増やすことができた．これは主に，視点に沿った文書提示機能により探索時間が軽減されたためと思われる．表2に示されるように，保存にかかる時間は若干の改善が見られる．被験者dで時間がかかっているのは，フォルダに保存すべきと誤って判定された文書が多く，保存先から外す動作に時間がかかったためである．参考までに，視点に沿った文書提示機能を使った場合の1文書に付与される視点数の平均値を表3に示す(被験者aは，この機能を用いていない)．

表 3: 1文書あたりの平均視点付与数

被験者	a	b	c	d	e
視点付与数	-	1.6	1.3	4.6	1.6

このように，判定精度が低いと逆効果になる恐れがある．しかし，文書提示機能を使う視点の選択や閾値の設定を適切に行なえば，保存時間を2から3秒短縮できるであろう．

以上の結果から，収集フェーズでの時間短縮に本システムは有効であるといえる．

6.2 活用フェーズでの精度向上の評価

被験者に活用の動機となる業務を課題として与え，その業務を遂行するのに必要な情報を集める

ことを，次の二つの方法を用いて行なった．

C 本システムを用いる．

D 保存された文書に対する文書単位の条件検索機能だけを持ったシステムを用いる．

12個の課題からなる二つのグループX, Yを用意し，5名の被験者を二つのグループS, Tに分けて，表4に示すスケジュールで実験を行なった．表中，xyは，xグループの課題をyの方法で解くことを示している．被験者は実験に先立つ2週間,「情報分野に関連する文書」を収集しており，各自が収集したセーブ情報を用いた．

表 4: 検索実験のスケジュール

Sグループ	XC	YD	YC	XD
Tグループ	YC	XD	XC	YD

表5に，回答ごとの再現率と適合率の方法別平均値を示す．なお，課題に対する正解は，課題ごとの全回答から再度人手により求めた．

表 5: 再現率・適合率の違い

	Cによる回答	Dによる回答
再現率 (%)	61.4	54.9
適合率 (%)	82.8	83.6

同一課題に対する同一被験者の回答において，Cによる正解数からDによる正解数を引いた値を本システムによる改善数とした．改善数別の度数分布を図4に示す．

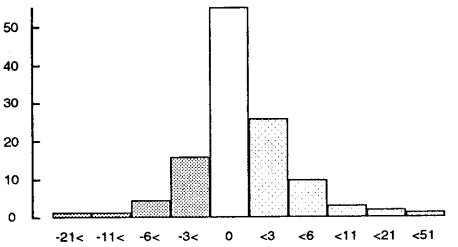

図 4: 改善数別の度数分布

本システムを用いることで，視点を付与しないで文書を保存した場合に比べ，再現率が6%強向上した．また，全体の35%において改善数が正となり，本システムを用いる効果が現れた．一方，改善数が負になるケースも23%見られた．Dの方法でも結果がそれほど多くなく回答を容易に得ることができた一方で，Cでは正解を含むフォルダを調べなかったケースと思われる．実際，これらのケースは他の場合に比べCによる回答時間が短い傾向が見られる．しかし，条件検索の結果のフォルダを調べれば，文書単位の検索と同等の結果は得られるので，時間をかければ改善数が負になることはない．

以上の結果から，活用フェーズでの精度向上に本システムは有効であるといえる．

7 まとめ

フロー情報の収集と活用を支援するため，視点の変化に対応できる文書の保存形態を用い，収集の時間短縮と活用の精度向上を図るシステムを作成した．利用実験により，本システムの支援機能を使った場合に，上記の効果が現れることを確認した．

参考文献

[1] 廣田他, フロー情報を対象にした情報検索システム (4) －文書分類－, 情報処理学会全国大会, 50, 4F-9, 1995.

[2] 菊地他, 全文検索の技術動向とシステム事例, 情処研報, 情報学基礎, 92-FI-25, 1-8, 1992.

[3] 松永, 情報の “三つ組” モデルに基づく情報蓄積・検索ツール:情報箱, 情処研報, 情報学基礎, 95-FI-40, 41-48, 1995.

[4] G.Salton, M.J.McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[5] 上田他, フロー情報を対象にした情報検索システム (1) －概要－, 情報処理学会全国大会, 50, 4F-6, 1995.