自然言語処理 108－11
(1995. 7. 20)

単語集合の自動構造化機能を持つ「情報散策」方式

有田英一、安井照昌、津高新一郎
arita@sys.crl.melco.co.jp
RWCP*新機能三菱研究室
(三菱電機(株)中央研究所内)
*Real Word Computing Partnership

あらまし 「情報散策」という新しい情報アクセス方式を持つ情報ベースシステムについて報告する。対象は新聞などの電子化されたテキストデータである。従来の情報検索システムは、ある程度明確な目的を持っている利用者が、数個のキーワードという抽象化されたレベルで情報ベースへのアクセスを行なっていた。これに対して「情報散策」とは、情報ベースの文書の内容から自動的に構造化され可視化された情報空間を、利用者が興味に従って自由に見て周りながら、しだいに目的を明らかにしてゆくアクセス法である。情報散策のための基本的な機能として、自己組織化マップによる文書の自動分類機能とそれに対応した自動インデキシング機能持つプロトタイプシステムを試作したので、その概要ならびに動作例について説明する。

キーワード 情報検索、シソーラス、連想、全文データベース、発想支援システム、可視化

Information Strolling through Automatically Organized Information Space

Hidekazu ARITA, Terumasa YASUI, and Shin-ichiro TSUDAKA
RWCP* Nobel Functions Mitsubishi Laboratory
*Real Word Computing Partnership

Abstract

This paper describes an information base system with a new access method called "information strolling". The contents of the information base is text data such as newspaper articles. In case of ordinal information retrieval, a user with a clear purpose accesses an information base by several keywords. On the contrary, in case of information strolling, a user's purpose gradually becomes clear after strolling through automatically organized and visualized information space.

As the basic functions of information strolling, automatic classification and indexing of texts by self-organizing map is implemented as a prototype system.

key words information retrieval, thesaurus, association, full-text database, creative thinking support system, information visualization

1. はじめに

テキストデータの情報ベースとして、特許データベースや科学技術文献データベースなどの書誌事項のデータベースが従来から利用されてきた。近年は新聞記事データベースなどの全文データベースが構築され利用できるようになってきた。また最近ではコンピュータネットワークに流れるNetNewsも情報源という意味で一種のデータベースとして考えることができるようになってきている。

一方、それらの情報源に対するアクセスは、従来は情報検索の専門家に依頼していたが、最近では利用者が机上のワークステーションから各種情報源に自由にアクセスできるようになってきた。

利用者が机上のワークステーションで各種情報源に容易にアクセスできるようになったことにより、情報検索をそれ単体で考えるのではなく、知的生産活動の中の一要素として捉えることができるようになる。すなわち情報検索などの情報アクセスの技術と発想支援などの情報利用の技術とを連係させて創造的活動を支援できるようになる。

情報アクセスの技術としては、情報源に容易にアクセスするための技術の開発がなされている。従来のキーワードの論理和／論理積によるBoolean Serachのほかに、テキスト中の任意の単語の有無で検索できる全文検索（フルテキストサーチ）や不要なmail/newsを排除する情報フィルタなどが考えられている。またニューラルネットワーク技術を情報検索に適用した研究も多い[仁木95]。

一方、情報利用の技術としては、利用者がキーワードやフレーズを思い浮かべてそれらを整理するKJ法などの概念レベルの発想支援システムの研究はあるが、大量の事実（文書）を利用するアプローチは、これまではあまり研究がなされていなかったように思われる。

このような問題意識から、我々はこれまで自己組織型情報ベースの研究を実施してきた[津高93][有田93][有田94]。自己組織型情報ベースは文書集合の自動構造化機能、単語集合の自動構造化機能、適応型インタフェースを持つ一種の情報検索システムであるが、特に創造的な活動を支援するツール（発想支援システム）と位置づけている。

本稿では、情報検索と発想支援の両方の機能を持つ「情報散策」の考え方について述べ、その機能を一部実現したプロトタイプシステムの機能／動作について述べる。

2. 情報ベースの分類とアクセス

情報検索は辞書構築、知識ベース構築、電子図書館、自然言語処理などの研究の重要な応用の１つであり、多くの観点から研究されているので、初めに情報ベースの分類とそのアクセス法について整理したあと、本研究の目指すところを述べる。

■情報ベースの分類

情報ベース分類の観点として、長期間にわたって蓄積されるストック型と原則として蓄積しないでネットワークに流通するフロー型という区別と、利用者の個人がファイリングして蓄積した情報と利用者個人以外の人が作成、流通、管理している情報という区別を考える。そうすると次のように分類できる。

(1)外界にあるストック（蓄積）型文書情報源：　特許、文献、新聞、本、WWW　など

(2)外界にあるフロー（流れ）型文書情報源：NetNews, e-mail　など

(3)個人の領域ストック（蓄積）型文書情報源：e-mail, 参考文献、個人資料　など

個人の領域にあるフロー型の情報源に対応するものはない。

■情報ベースのアクセス法の分類

情報ベースのアクセス法としては、様々なアクセス法があるが、大別すると次の４つがある。

(a)検索：文献データベースのようにキーワードを手掛かりに条件にあう情報を一括して得る。主として(1)の情報源に対するアクセス法である。

(b)監視：newsのようにフロー型の情報ベースから特定の目的に合う情報をピックアップする。主として(2)の情報源に対するアクセス法である。

(c)探索：WWWのようにリンクを辿って必要な情報にアクセスする。

(d)ブラウズ：（１つの）情報ベースの内容の全体を見る。主として(1)と(3)の情報源に対するアクセス手法として今後重要になると考えられる。

現在の情報検索システムが使いにくい一番の原因は、情報ベースの全容が見えないことと利用者の判断を情報検索に反映できないことであるので、今後のアクセス手法としては以下の点が重要になると考えられる。

○情報空間の可視化：情報ベースの中身をわかりやすい形で利用者に提示する。

○情報空間のカスタマイズ：利用者の視点や価値観に応じて主観的に情報空間を操作する。

以上の点が、本研究の目指すところであり、本報告では情報空間の可視化について検討している。情報空間を可視化するための第１ステップとして情報空間の自動構造化が必要であり、以下では単語集合の自動構

造化と文書集合の自動構造化について説明する。

3. 単語集合の自動構造化

3.1 上位／下位関係ネットワーク

単語の表わす概念の上位／下位関係を求めることは一般に非常に困難であるが、大規模なネットワークを構成するには、少々間違いがあっても機械的に行なうのが現実的であると考えられる。特に人間の創造的活動の支援システムとして利用する場合は、間違いを簡単に訂正できるインタフェースを用意することで、しだいに正しい上位／下位関係ネットワークになる。

日本語における複合名詞は基本的な単語(概念)にそれを限定する単語(概念)を前方または後方に付加する形で構成されると考えられる[有田91]（英語に対しても同様に考えることができる[頼94]。）ので、第一次近似として、次の関係がある時、単語（概念）ChとClが上位／下位の関係にあるとする。

「文字列Aが文字列Bのsubstringであることを
A⊂Bと記述する。
Ch⊂Clで、Ch⊂Cx⊂ClとなるCxが存在しないとき、ChをClの上位概念、ClをChの下位概念とする。」

このようにして得られる疑似的な上位／下位関係は、JICSTなどの人手を介して作成されたシソーラスと比較して表面的であり誤りもあるが、前処理なしの全文検索の場合にはこの表層的アプローチも有効である。例えば「腎臓移植」について検索したい場合など、この上位／下位ネットワークで「腎移植」「じん臓移植」など異表記も得ることができるので検索の網羅性（Recall Ratio）が改善される。

（シソーラスと文書のインデックス付けは密接な関係があるので、JISCTデータベースのように人間がインデックスを付与する場合は、JICSTシソーラスのように整理された概念体系が良い。）

3.2 連想ネットワーク

■基本チェイン　[有田94]

本研究では文書の構成要素として、少なくとも見出しと本文があるものを対象としている。見出しは多くの場合本文の要約と考えることができるので、見出しに含まれる単語は重要な単語であると考えられる。そこで見出し部に含まれる単語を対象として、本文部と見出し部の共起関係で以下の場合を連想の基本的なチェインと考える。キーワードKWが文書DOCの見出しに含まれることを　KW(DOC) と表わすことにする。一般に単語には複数の意味があるが、KWj(DOCi)は、KWjの意味の中でDOCiという文脈で使用される意味であることを表わしている。

文書DOCmの見出しに含まれるキーワードKWkが文書DOCiの本文に含まれていて、その見出しにキーワードKWjが含まれていることを、

KWk(DOCm) ==>　KWj(DOCi)

と表わすことにする。ここで、

KW1(DOC1) ==> KW2(DOC2)
KW2(DOC2) ==> KW3(DOC3)
KW1(DOC1) ==> KW3(DOC3)

の関係があるとき、KW1, KW2, KW3が連想関係にあると考える。これは２項の偶然の共起関係を排除するためである。上記の３項の共起関係を

KW1(DOC1) → KW2(DOC2) → KW3(DOC3)

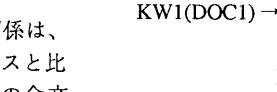
と表わすことにする。これを、連想関係の基本チェインと考える。

■連想ネットワーク

基本チェインに共通部分があるときは、結合してネットワークを構成する。例えば

KW1(DOC1) → KW2(DOC2) → KW3(DOC3)
KW2(DOC2) → KW3(DOC3) → KW4(DOC4)
KW1(DOC1) → KW2(DOC2) → KW5(DOC5)

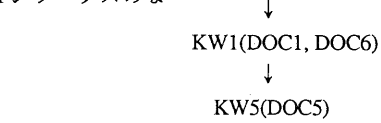
の３つの基本チェインは次のネットワークに結合する。



さらに、DOCi, DOCjが（自動）分類したときに、同じクラスタに分類される文書である場合は、KWk(DOCi), KWk(DOCj)を結合する。例えば、

KW1(DOC1) → KW2(DOC2) → KW3(DOC3)
KW4(DOC4) → KW1(DOC6) → KW5(DOC5)

の２つの基本チェインを結合して下のネットワークを構成する。



このようにして、結合してできる単語のネットワークを連想ネットワークとする。

3.3 単語集合の構造化

上に述べた上位／下位関係ネットワークと連想ネットワークを単語をキーとして結合することにより、情報ベース内の単語集合の構造を見ることができる。図１に例を示す。このような単語集合の自動構造化はJICSTシソーラスやEDR概念辞書のように人手で作成された精巧な構造に及ぶべくもないが、新聞記事や科学技術文献のように急速に新しい概念や言葉が発生す

る情報ベースでは、本研究で述べたような単語集合の自動構造化を併用することにより、情報検索のRecall RatioやPrecision Racioが改善できる。

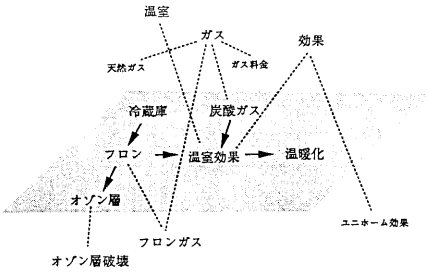


図1　単語集合の構造化の例

4. 文書集合の構造化

4.1.テキストのコーディング

本研究では、文書はそれが含む単語のベクトルとして表現するベクトルモデルを採用している。ベクトルモデルでは単語ベクトルの要素となる単語集合としてどのようなものを選ぶかが重要であるが、計算量の制約から対象となる文書集合に出現する全ての単語を対象とすることはできない。また一般的な単語を採用すると自動分類／自動インデキシングという観点からノイズとなることが予想される。

そこで適切な単語集合を見つけるために、1990年の一般新聞記事データの経済関連の記事（3576記事）について選ばれる以下の3つの場合について、各単語の持つ値の大きさを比較した。（ただし式に含まれるFijは単語word-jが文書DOC-iに出現する頻度、Nは文書集合の文書数、Njは単語word-jを含む文書の数、Ljは単語word-jの文字数である。）

(a)頻度と分布の偏り及び単語長を考慮した場合：
Σ_j(Fij × log(N/Nj) × log(Lj))の値の大きい単語を順に示すと次のようになった。
委員長、大統領、首相、消費税、自民党、日本、リクルート、米国、中国、ソ連、グループ、……

(b)頻度と分布の偏りを考慮した場合：
Σ_j(Fij × log(N/Nj))の値の大きい単語を順に示すと次のようになった。
円、首相、日本、人、委員長、月、大統領、米国、中国、消費税、自民党、ソ連、政府、党、……

(c)頻度のみを考慮した場合：
Σ_jFijの値の大きい単語を順に示すと次のようになった。
円、日本、人、月、首相、米国、委員長、東京、政府、自民党、中国、消費税、大統領、問題、……

単語長を考慮しない場合は、円、人、党、部、金、性、本、家、車、心など一般的な単語が上位に現れるのに対して、単語長を考慮した場合は、リクルート、グループ、政治改革、ポーランド、コンピュータ、パーティー、サミットなどトピックとなるような単語が多く含まれていた。従って単語ベクトルの要素となる単語として、文書集合に出現するすべての単語（異なり）から、単語の出現頻度と分布の偏り及び単語長（文字列長）を考慮した重要度の大きいものから一定数を選ぶのが適当であると判断した。

以上のようにして選んだ単語集合を要素として、文書を単語ベクトルとして表現し、その各単語の値は次式のように設定した。

$$Vij = Fij \times \log(N/Nj) \times \log(Lj)$$

Fij × log(N/Nj)はinverse document frequencyと呼ばれるファクタで情報検索の分野ではよく使われるファクタである。このファクタの意味するところは各単語が文書を分類する能力を表わすものである。log(N/Nj)の意味は、多くの文書に出現する単語はその単語の有無により文書の分類ができないのでその単語の重要度は小さいことを表わしている。全ての文書にある単語が出現する場合は、N=Njなのでlog(N/Nj)=0となり、その単語の重要度Vijは0である。すなわちその単語の有無により文書の分類ができないことを表わしている。Fijの意味は、ある文書に出現する回数の多い単語はその文書の中心的な内容を表わすことが多いことを表わしている。log(Lj)は複合名詞のように単語長が長い単語の方が一般名詞のように単語長が短い単語よりも重要であることを表わしている。

4.2. 文書集合の自動分類／自動インデキシング

文書集合の自動分類／自動インデキシング手法は種々のものが考えられ、目的により最適なものが違うが、ここではプロトタイプシステムで実装している2つの方法について述べる。

4.2.1 SOM法

T.Kohonenの自己組織化マップモデル[Kohonen 90]をベースにした文書の自動分類／自動インデキシング方式[津高93]を1つの方式として採用している。これは単に文書集合をクラスタリングするだけでなく、各クラスタを内容の似たものが近くに来るように2次元平面に配置されるところに特徴がある。（本研究は日本語データを対象としているが、英語データを対象とし

た同様の研究もある[銭94]。）

4.2.2 最大距離－K平均法

クラスタリングアルゴリズムとして知られている最大距離アルゴリズムとK平均アルゴリズムを組み合わせた最大距離－K平均法と呼ぶクラスタリング手法を開発した。これは処理の前半では乱数の影響を受けないように変更した最大距離アルゴリズムを用いて初期クラスタ中心を決定し、処理の後半でK平均アルゴリズムによるクラスタ中心の変更を行なうものである。この手法ではパラメータとしてクラスタの数を指定することができる。シミュレーションによる性能評価によれば、幅広いデータ分布に対して性能が良好でかつ安定していた。

5. 「情報散策」システム

5.1. 情報散策と発想支援

情報散策は検索とブラウズを組み合わせたアクセス法である。利用者は特に明確な目的を持たずに情報ベースの中をブラウズし、興味のある単語（概念）に遭遇した際には、その情報の中身にアクセスすると共に、関連する情報を得るために情報検索を行う。検索結果は自動的に構造化されて可視化される。その可視化された情報空間を見るうちに興味が別の単語に移れば、その単語をキーとして情報検索し、結果の情報空間を眺めるという動作を繰り替えす。情報散策はこのように情報の収集、体系化と検証、シミュレーションなどをスパイラルに繰り返すことにより、しだいに新しい考えをまとめて行く創造的な活動を支援する行為である。

発想支援システムはアイデアをまとめてゆく収束型と様々な観点からのアイデアを出してゆく発散型に分類できる[國藤93]が、情報散策は発散型の発想支援システムに適している。特に情報源としてデータベースを持つことにより網羅性を保証できるところに特徴がある。文書集合の自動分類／自動インデキシングによって全体を把握しながら、関連するキーワードを得ることができ、必要に応じて特定のキーワードで情報検索ができて文書の本文を参照することができる。

5.2 プロトタイプシステム

現在、文書集合の自動構造化（自動分類／自動インデキシング）を中心とする情報散策システムのプロトタイプシステムを開発している。図２に画面のスナップショットを示す。情報ベースとしては一般新聞の経済関係の記事を対象としている。

各ウインドウの説明を以下に示す。

ウインドウ１：「テレビ」というキーワードで全文検索した結果をテレビが出現する頻度が多いものから順にそのタイトル（の一部）を表示したものである。

ウインドウ２：ウインドウ１に検索された文書集合を対象として最大距離－K平均法で自動分類／自動インデキシングした結果を表示している。表示されている単語はそれぞれのクラスタの代表的な単語であり、右側の棒グラフはそれぞれのクラスタに分類された文書の数を表わしている。

ウインドウ３：ウインドウ１に検索された文書集合を対象としてSOM法で自動分類／自動インデキシングした結果を表示している。６角形の各セルの色は、そのセルに分類された文書の数を表わしている。

ウインドウ４：ウインドウ３の１つのセルまたはウインドウ２の１つの分類を指定した時に、そのセルに属する文書のタイトルを表示したものである。この例ではウインドウ３の「粗大ゴミ」というセルを指定した時に、そのセルに分類された３つの文書のタイトルを表示している。

ウインドウ５：ウインドウ１またはウインドウ４の１つの文書を指定したときに、その本文の内容を表示したものである。

通常の情報検索システムであれば、キーワード検索により、適当な文書数まで絞りこんだあと、ウインドウ１のように文書のタイトルを表示し、利用者はタイトルから判断して検索の意図にあっていると思われる文書を選択し、その本文をウインドウ５のような形で参照するという操作を繰り返す。しかしウインドウ１の大きさの制約から検索された文書のタイトルの一部分しか見ることができないので、検索された文書の全容を短時間のうちに把握することが困難である。

このプロトタイプシステムではウインドウ２、３のように検索された文書集合の概要を知ることができる。この例では、「テレビ」という漠然としたキーワードに纏わる話題には、ＮＨＫ、衛星放送といったテレビ放送事業に関する観点や、ＣＭ、コマーシャルといった商業放送の観点や、粗大ゴミといったテレビ受像機の処分の観点や、自動車、輸出、半導体といった製品としてのテレビ受像機の観点や、ロボットアームといった工業製品の一部のモニタとしてのテレビの観点や、ビデオ、ビデオシアターといった放送以外のテレビ受像機の利用の観点などがあることがわかる。

このように情報検索とその検索結果の自動構造化と可視化機能を組み合わせることにより、情報検索の網羅性の特徴を生かした発散型の発想支援システムを構

築することができる。

5. おわりに

本報告では単語集合の自動構造化と文書集合の自動構造化について述べた。また一部プロトタイプシステムを試作して実験し、情報検索とその可視化機能を組み合わせることにより、情報検索の網羅性の特徴を生かした発散型の発想支援システムを構築することができる見通しを得た。

連想ネットワークは現時点では共起情報しか反映していないが、今後は共起頻度も反映して、[岡94][田淵92][堀86]のように単語に位置情報を付加したネットワークの構成を検討している。

なお、本稿はＲＷＣＰ新機能三菱研究室で実施した研究と三菱電機（株）中央研究所で実施した研究をまとめたものである。また本研究の実施に当たり朝日新聞社の紙面データを使用した。

参考文献

津高：「自己組織化マップを用いたテキストの自動分類の試み」、情報処理学会第46回全国大会4-187～188, 1993

有田、山岡、飯田：「問い合わせ対話における名詞句表現の構成」、信学技NLC91-42 (1991)

有田：「AIにおける問題解決技術の新展開－自己組織型情報ベースー」、1993年度電気関係学会東海支部連合大会S5-2, 1993

有田、岡：「新聞記事データからの断片的知識の連鎖の抽出」、信学技報NLC93-66 (1994)

頼、陳、藤原：「概念空間のモデルと専門用語の構造化」、情報処理学会情報学基礎研究会資料FI-35-5 (1994)

仁木、田中：「ニューラルネットワーク技術の情報検索への適用」、人工知能学会誌Vol.10, No.1, pp45-51 (1995)

銭、史、田中：「自己組織化マップと語彙索引を用いたデータベースの抽象化機構」、情報処理学会データベースシステム研究会資料99-22, pp163-170 (1994)

T. Kohonen : "The Self-Organizing Map", Proceedings of the IEEE, Vol.78, No.9, pp1464-1480 (1990)

岡、有田、水谷：「自然言語理解のための記号の空間配置とその結合方式」、信学技報NLC93-65 (1994)

田淵：「記号間の力学に基づく概念マップ生成システムSPRINGS」、情報処理学会論文誌、Vol.33, No.4, pp465-470 (1992)

堀：「単語の意味の学習について」、コンピュータソフトウエア, Vol.3, No.4, pp65-72 (1986)

國藤：「発想支援システムの研究開発動向とその課題」、人工知能学会誌, Vol.8, No.5, pp552-559 (1993)

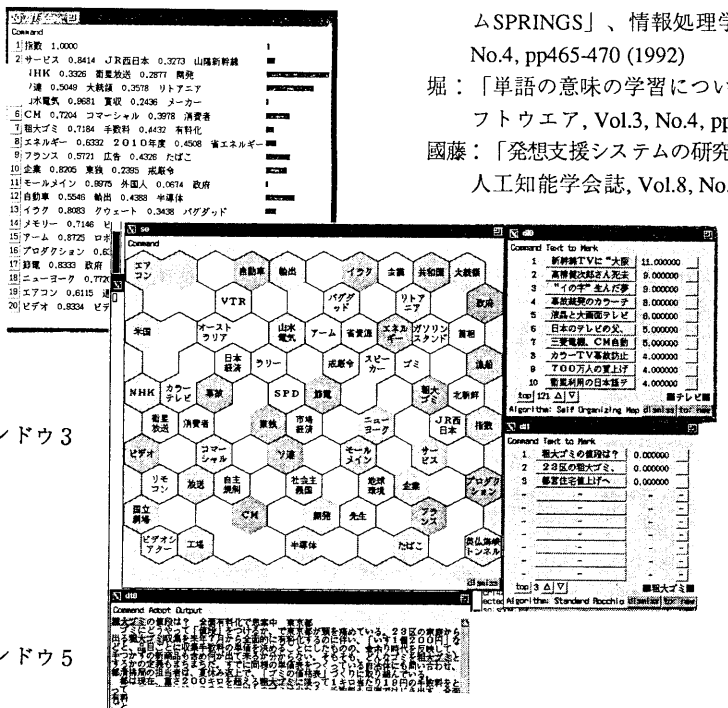


図2　プロトタイプシステムの画面のスナップショット