自然言語処理 112－3
（1996. 3. 14）

文書間の関連性を可視化することによる文献検索システム

杉本 雅則[1] 小山 照夫[1] 堀 浩一[2] 大須賀 節雄[3] 絹川 博之[4] 間瀬 久雄[4]
[1]学術情報センター研究開発部 [2]東京大学工学部 [3]早稲田大学理工学部
[4]日立製作所システム開発研究所

日本語論文を対象とし、文書間の関連性を可視化することによる文献検索システムについて検討する。このシステムの特徴は、論文とそのキーワードを1つの距離空間に同時に可視化することにより、論文間およびキーワード間の意味的な関係を容易に把握することができるという点である。ユーザは、システムに対しインタラクティブな操作を行うことにより、必要な文献を容易に検索することが可能になる。これまでに英語論文を対象としたシステムを構築し、本システムの有効性を示すことができた。

A Retrieval System for Visualizing the Relations between Documents

Masanori Sugimoto[1] Teruo Koyama[1] Koichi Hori[2] Setsuo Ohsuga[3] Hiroshi Kinukawa[4] Hisao Mase[4]
[1]R&D Dept. NACSIS [2]Faculty of Eng. Univ. of Tokyo [3]School of Sci. and Eng. Waseda Univ.
[4]System Development Lab. Hitachi Ltd.

In this paper we discuss a retrieval system which can visualize the relations between Japanese documents. The feature of the system is that it can simultaneously map the documents and their keywords in one metric space. The users can interactively operate it and can easily find documents necessary for them. We have already built a system for English documents and proved its effectiveness through the experiments.

1 はじめに

コンピュータネットワークが発展し、ネットワークニュースや World Wide Web などを利用することにより、世界中の情報を瞬時に獲得できるようになっている。こうした状況においては、大量の情報から必要な情報を容易に検索するための技術が求められていると言えるだろう [1]。本稿では、特に学術論文を対象とし、それらの意味的な関係を可視化することにより、ユーザが必要な文献を容易に検索できるシステムについて検討する。

近年、電子図書館システムの研究開発が世界中で盛んに進められている [2][3]。本研究は、大規模なデータベースを持つ電子図書館システムにおける新しい文献検索手法の一つとして位置付けることもできると考える。

本システムの特徴は、文献とそのキーワードを1つの距離空間に同時に可視化することができる点、およびユーザのインタラクティブな操作が可能である点である。本研究では、まず英語の学術論文を対象とした検索システム (以下「英語版システム」と呼ぶ) を構築し、評価実験を行った [4] [5]。実験の過程で多くのユーザから日本語化への要求を受けたため、これを日本語論文を対象とするシステム (以下「日本語版システム」と呼ぶ) へと拡張した。

本稿の構成は以下の通りである。第2章では日本語版システムの構成について示し、日本語処理に関して本研究でとった手法などについて述べる。第3章では、英語版システムを用いて行った実験および評価について述べる。第4章では、結論と今後の展望について述べる。

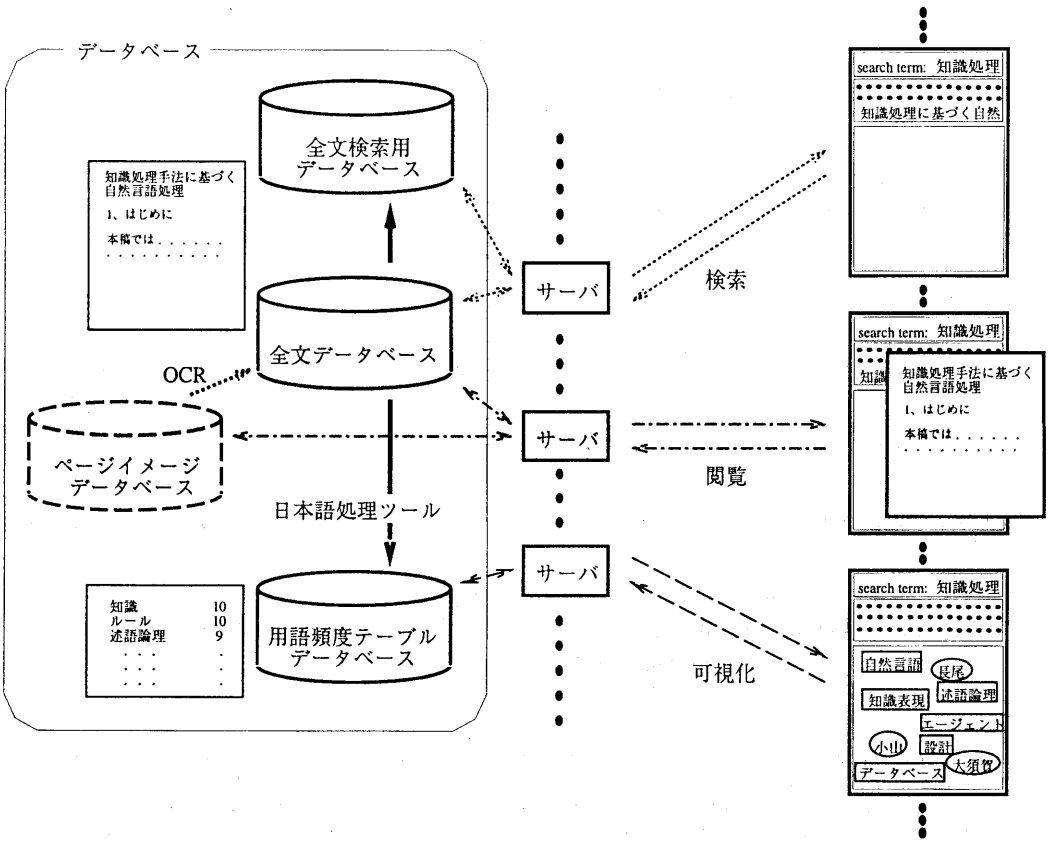

図 1: 本システムの構成

2 本システムの構成

2.1 クライアントサーバモデル

図1は、本システムの構成を示したものである。

本システムは、サーバおよびクライアントより構成される。サーバは以下に示すようなデータベースを管理しており、クライアントからの要求に応じて処理を行い、その結果を送信する。

ユーザはクライアントプログラムのインターフェイス上でインタラクティブに操作することにより、必要な情報を獲得することができる。

2.1.1 本システムのデータベース

本システムは以下の4つのデータベースから構成されている。

- 全文データベース
- ページイメージデータベース
- 全文検索用データベース (インデックスファイル)
- 用語頻度テーブルデータベース

全文データベースには、現在のところ情報処理学会論文誌1994年分の全文テキスト(日本語論文のみ、約250本)がロードされている[1]。

全文検索を高速に行うためには、全文データを検索エンジン用に変換する必要がある。全文検索用データベースとは、高速な検索を可能にするために全文データをインデックスファイルに変換したものを指す。

用語頻度テーブルデータベースは、各文献に出現する用語の頻度を計算し、表の形式にしたものを指す。これは文献間の関係を可視化する際に用いられるものである。用語頻度テーブルの構成については後述する。

2.1.2 サーバの処理モジュール

サーバは、クライアントからの要求に応じて、以下のような処理モジュールを起動する[2]。

検索要求処理モジュール クライアントからの検索要求を受けて全文検索を行う。本研究では、市販の全文検索エンジンを用いることにより、高速な検索を実現する[3]。検索の結果、サーバは検索語にヒットした文献の集合をクライアントに送信する。

閲覧要求処理モジュール クライアントからの閲覧要求を受けて、データベースから必要なデータを取得し、クライアントに送信する[4]。

可視化処理モジュール クライアントからの可視化要求を受けて、処理を行う。指定された文献について、それらの用語頻度テーブルを基に、各キーワードの得点を計算する。このとき以下に示すような自動インデキシングアルゴリズムを用いる[7]。

n 個の文献中で文献 i の用語 k の得点を計算するとして、

1 用語 k が論文 i において占める割合 d_k^i および n 個の論文全体においてに占める割合 $\overline{d_k}$ を計算する。

2 用語 k の文献中での分散 v_k を以下の式で求める。

$$v_k = \frac{\sum_{i=1}^{n}(d_k^i - \overline{d_k})^2}{n-1}$$

(分散の大きい用語は、文献によってその出現の程度が大きく変動することを意味する。)

3 用語 k の得点 s_k を以下の式で求める。

$$s_k^i = \frac{d_k^i \times v_k}{\overline{d_k}^2}$$

(上式では、文献によって出現の程度が異なる用語および出現回数の多い用語ほど高得点が与えられる。)

上記のアルゴリズムにより、各文献に出現する用語には、指定された文献集合における相対的な得点が与えられる。この結果を基に、各文献について得点の高いものから順に一定個数のキーワードを抽出し、これらを次元とするキーワードベクトルを構成する。

次にこのキーワードベクトルを双対尺度法[8]と呼ばれる統計処理手法で解析し、各文献およびそれのキーワードについての関係を表現する多次元空間での座標を計算する。

クライアントでは2次元の可視化空間が構成される。そのためサーバはクライアントに対し、論文およびキーワードの2次元座標を送信する。

2.2 日本語処理

本節では、用語頻度テーブルを構成する際の日本語処理について述べる。

まず全文データベース中の各論文のテキストデータに対し、日本語処理ツールを用いて形態素解析を

[1] 上記のデータベースは、当センターで開発されている電子図書館システム(NACSIS-ELS)[6]のデータを基に構成することを検討している。NACSIS-ELSでは、多数の学会誌のページイメージがデータベースにロードされている。OCRを用いてこれらのページイメージをテキスト化することにより、本システムの全文データベースへロードすることができる。

[2] 以下に示す以外に、サーバはログの管理を行っている。本研究では、学術情報センターの電子図書館プロジェクトで利用している学会誌のデータを用いている。データの利用に当たっては学会側との取り決めに従い、データへのアクセスログ管理を行うことになっている。具体的には、全てのユーザについて、いつどの論文にアクセスしたかに関する情報をサーバで記録している。

[3] 本システムでは、市販の全文検索エンジンをまだインプリメントしていないため、検索終了までに1分程度の時間が必要である。

[4] 現在のところ全文データを送信しているが、将来的にはユーザの要求に応じてページイメージ(GIF, TIFF etc.)を送信することも検討している。

用語	頻度	用語	頻度	用語	頻度	用語	頻度	用語	頻度
ボロノイ	171	点	68	図	77	母	70	領域	56
アルゴリズム	52	添加	34	3次元	28	構造	27	グラフ	24
辺	23	処理	22	性質	22	結果	21	計算	20
数値	19	構成	18	位相	17	算法	16	要素	16
多面体	15	頂点	15	図形	13	点数	12	更新	12
退化	12	決定	11	データ	11	計算機	11	境界	11

表 1: 日本語処理ツールで解析した結果を用いて得られた用語の頻度 (1) (出現頻度数で上位 24 位までのもの)

用語	頻度	用語	頻度	用語	頻度	用語	頻度
母点	44	ボロノイ領域	41	ボロノイ点	29	図	26
アルゴリズム	24	ボロノイ図	22	位相的性質	18	ボロノイ面	17
処理	16	頂点	15	位相構造	14	多面体	14
ボロノイ辺	13	境界	11	要素	11	3次元ボロノイ図	11
点	11	破綻	10	添加	9	母点数	9
部分グラフ	9	結果	9	ボロノイ点集合	8	計算機	8

表 2: 日本語処理ツールで解析した結果を用いて得られた用語の頻度 (2) (出現頻度数で上位 24 位までのもの)

行った[5]。日本語処理ツールへの入力は以下のようなテキストである[6]。

> 3次元ボロノイ図構成のための数値的に安定な逐次添加法
>
> 1．はじめに
>
> 幾何図形を計算機で扱うにあたって，効率のよいアルゴリズムを見つけようとする努力が，計算幾何学と呼ばれる分野において活発になされ，多くの成果を上げている．しかし，そのようなアルゴリズムをそのまま計算機にインプリメントすると，計算誤差が原因となり処理が破綻してしまうことがある．計算機は有限精度の値を扱うものであるために，理論的には正しいアルゴリズムであっても，計算機プログラムに翻訳した際には必ずしも正しく動作するとは限らない．最近になって，このような理論と現実のギャップを埋めようとする試みが，平面ボロノイ図構成算法，ソリッドモデリングなどに対してなされている．
>
> (以下 省略)

日本語処理ツールにより、上記のテキストは分かち書きされ、各語には品詞情報が割り当てられる。まず最初に、この分かち書きされた論文について、普通名詞および未知語のみを対象として頻度計算を行った。その結果は表 1のようになる。

論文などにおいては、著者の独自の判断で、用語を組み合わせて複合語にすることにより、重要な概念を表現する場合があると考えられる。一方その複合語は、読者にとって理解不可能なものではなく、用語を組み合わせることで複合される前の用語よりも表現される概念が明確に絞り込まれるため、むしろ理解を促進するものであると思われる。また、いわゆる専門用語と呼ばれるものには、通常使われる用語を組み合わせることにより、構成されるものも多い。したがって、キーワードを複合語として抽出することは、論文間の関係を可視化して提示する際に重要かつ必要不可欠であると考えられる。

一方あらゆる複合語を辞書に登録しておくことは現実的には不可能である。そこで本研究では名詞および未知語が連続する場合は、これを一つの語と見なし、改めて頻度を計算するという手法を取ることにした。この結果は表 2のようになる。

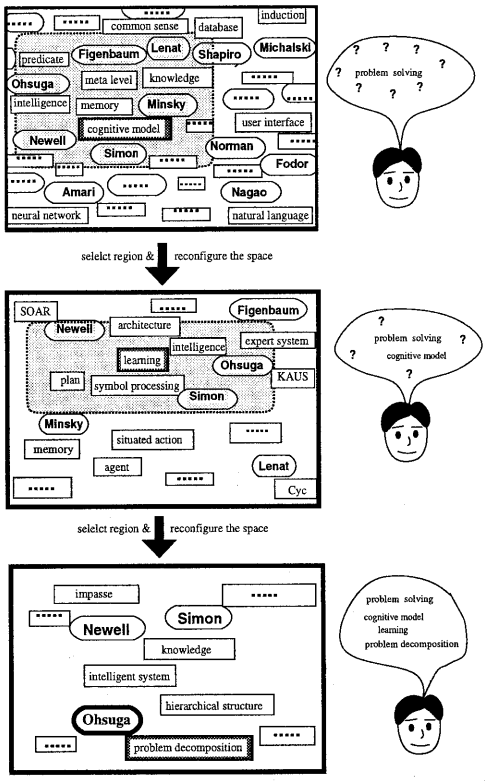

図 2: 本システムの利用法

表 1と表 2を比較すると、表 2 の方が複合語の占める割合が高くなっている。これにより、各論文が研究対象とする分野での特徴語や重要語でありなおかつ複合語であるものをより多く抽出することができると考える。

2.3 本システムの検索手順

本システムでの文献検索の手順は以下のようになる。ユーザはまず、自分の検索要求を表す語を検索語として入力する。その結果、検索語を含む論文のリストが示されることになる。ユーザはこれらのリストからいくつかまたは全ての論文を選ぶことにより、それらの論文間の関係を可視化する。

システムが構成する空間に対して、ユーザはインタラクティブに操作を行う。例えば図 2のように、可視化空間中の論文やキーワード間の関係に着目してある領域を選択し、空間を再構成することにより、論文間のより詳しい関係を可視化することができる。

[5] 日本語処理ツールとしては日立自然言語処理部品ライブラリ [9] を用いた。

[6] 稲垣 宏, 杉原 厚吉: 3次元ボロノイ図構成のための数値的に安定な逐次添加法, 情報処理学会論文誌, Vol. 35, No. 1 (1994) より引用。

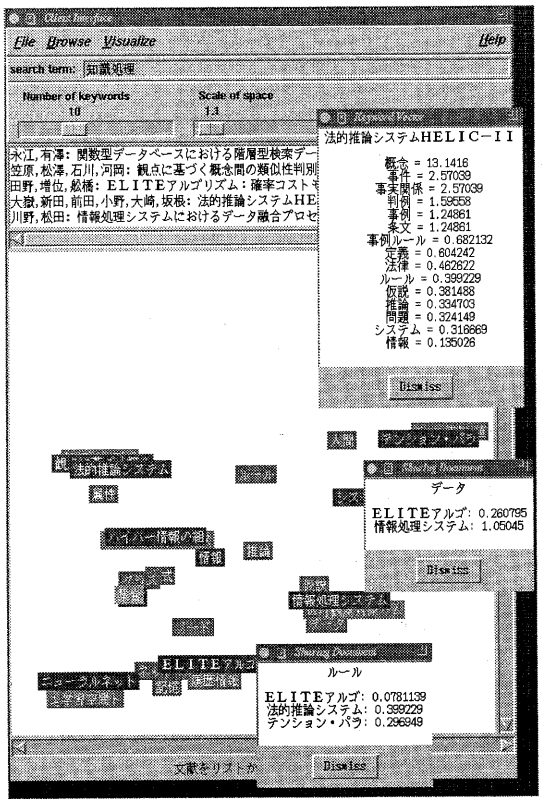

図 3: クライアントのインターフェイスの例

図 3はクライアントのユーザインターフェイスである。この図では、'知識処理' を検索語として検索しヒットした論文 7 本についての可視化空間が構成されている。この図を見ると、例えば'ELITE アルゴリズム...' と '法的推論システム...' と ' テンション・パラメータ...' とは 'ルール' というキーワードで関連付けられていること、'学習空間' は 'ニューラルネット...' のキーワードであることなどが分かる。

システムが作り出す可視化空間では、ユーザがこれまで気づかなかった新たなキーワードが提示されるので、そのキーワードを用いて再度検索したり、あるいは論文間のこれまで気付かなかった関係を発見して、新しい論文を読み進める手掛かりを得たりすることもできると思われる。ユーザは必要な文献を効率良く検索することができるだけでなく、新たに興味を持った論文を読み進めたりすることもできるようになると考えられる。

3 英語版システムでの実験と評価

英語版システム[7]を用いて実験を行い、その有効性を検証した。以下ではその主なものについて簡単に示す。

3.1 実験 (1)

[7]英語版システムでは、人工知能、認知科学、心理学など約 1100 本の英語論文がデータベースに蓄積されている。英語版システムと日本語版システムとの異なる点は、英文ではすでに分かち書きされているので、用語頻度テーブルの構成に際して形態素解析を行う必要が必ずしもないと言う点である。

システムの評価	人数
自分の考えている検索要求がより明確になった	2
気付かなかった新しい検索語を発見した	7
これまでよりも良い検索語を発見した	3
新たによんで見たいと思う文献を発見した	6
特にこれといった効果はなかった	0

表 3: 被験者の本システムに対する評価　(複数回答)

本実験では、被験者 (主に大学院生)9 人に自分の研究テーマに関して自由に文献検索をしてもらった。

表 3は、被験者が本システムを利用した際の評価をまとめたものである。それによると、「検索要求が明確になった」「より良い検索語を発見した」などの意見が得られている。このことは、本システムがユーザの検索要求の形成や修正 [10] [11] を支援する効果があると考えられる。

また、「新たに読んでみたいと思う文献を発見した」り「新しい検索語を発見した」りするということは、研究活動支援や発想支援などの効果を持つと考えることもできるだろう。

3.2 実験 (2)

実験では被験者 6 人に対して、'コンピュータによる生命のシミュレーション' に関する論文の検索をしてもらった。表 4において、'最初の検索語による再現率' および '最初の検索語による適合率' は、通常の Boolean 検索型の情報検索システムの再現率および適合率に対応し、'インタラクション後の再現率' および 'インタラクション後の適合率' は、本システムを用いてインタラクティブに操作した際の再現率および適合率に対応する。表 4では、特に適合率に関して検索効率が改善されていることが示されている。

被験者から得られた意見として多かったのは、「文献間の関係が可視化されることにより、類似した文献が一目で分かる点が従来のシステムよりも使いやすい.」や「可視化空間に浮かぶキーワードを見ながら、インタラクティブに作業を行うことにより、ユーザの意図を検索作業に反映させることがこれまでのシステムよりも容易である.」というものであった。

4 結論と今後の課題

本稿では、文献間の関連性を可視化することによる文献検索システムについての検討を行った。日本語版システムの構成について述べ、日本語処理や検索手法などについて示した。また英語版システムを用いて行った実験について述べ、本システムで用いた手法が文献検索に有効であることを検証した。日本語版システムのプロトタイプが完成し次第、英語版システムと同様その評価実験を行う予定である。

日本語版システムのサーバプログラムは学術情報センターのコンピュータ上で実行されており、クライアントプログラムは各サイトのコンピュータからこのサーバにアクセスすることにより文献検索を行うようになっている。今後の実験により日本語版システムの有効性が検証された後には、クライアントプログラムを一般に公開することを検討している。

謝辞

本研究は日本学術振興会産学共同研究支援事業からの助成を受けました。また、学術情報センター研究開発部 安達教授、大山助教授、高須助教授、片山助手からは本研究に関する多くのコメントを頂きました。ここに感謝致します。

参考文献

[1] 野村浩郷, 井佐原均, 徳永健伸, 中村貞吾. 情報ハイウェイ時代のテキスト情報への知的アクセス. 情報処理, Vol. 37, No. 1, pp. 1 – 9, 1996.

[2] E. A. Fox et al. Digital Libraries. *Communications of the ACM*, No. 4, 1995.

[3] *Digital Libraries '95 (DL'95)*, 1995.

被験者	最初の検索語	最初の検索語による再現率	最初の検索語による適合率	検索過程で見つけたキーワード	インタラクション後の再現率	インタラクション後の適合率	インタラクションの回数
A	artificial, computer, life, simulation	5/8	5/28	environmental	5/8	5/6	3
B	artificial, computer, life, simulation	5/8	5/28	constructive AI	5/8	5/5	2
C	artificial, autonomy, evolution, life	5/8	5/33	animate, genetic information, simulation	5/8	5/5	2
D	computer, life, simulation	6/8	6/48	artifact, evolution, self regulation	6/8	6/6	2
E	computer, life, simulation	6/8	6/48	autonomous system, biological adaptation, biological system, genetic code, intelligence	6/8	6/6	2
F	artificial, computer, life	7/8	7/95	autonomous system, biological system	7/8	7/7	2
平均		34/48	34/280	平均	34/48	34/35	13/6

表 4: 本システムの検索効率に関する実験結果

(http://bush.cs.tamu.edu/dl94/README.html).

[4] M. Sugimoto, K. Hori, and S. Ohsuga. A System for Assisting Creative Research Activity. In Y. Anzai, K. Ogawa, and H. Mori, editors, *Symboiosis of Human and Artifact: Future Computing and Design of Human Computer Interaction*, pp. 685 – 690. 1995.

[5] M. Sugimoto, K. Hori, and S. Ohsuga. A Document Retrieval System for Assisting Creative Research. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, pp. 167 – 170, 1995.

[6] J. Adachi. NACSIS Electronic Library System - Evolution of NACSIS Science Information Services -. Technical report, National Center for Science Information Systems, 1995.

[7] G. Salton. *Dynamic Information and Library Processing*. Prentice-Hall, 1975.

[8] 西里静彦. 質的データの数量化 - 双対尺度法とその応用 -. 朝倉書店, 1982.

[9] 西森裕司, 木山忠博, 絹川博之. 汎用日本語形態素解析ツールの開発. 情報処理学会第 44 回全国大会, pp. 3181 – 3182, 1992.

[10] R.S. Taylor. Question-Negotiation and Information Seeking in Libraries. *Colleges and Research Libraries*, Vol. 29, pp. 178 – 194, 1968.

[11] N.J. Belkin. Anomalous States of Knowledge as a Basis for Information Retrieval. *The Canadian Journal of Information Science*, Vol. 5, pp. 133 – 143, 1980.