デジタル・ドキュメント 22－2
（２０００．３．１７）

# 連想ナビゲーション

蓬莱 尚幸、渡部 勇、齊藤 孝広、三末 和男、松井 くにお
(株)富士通研究所 ドキュメント処理研究部
〒261-8588 千葉市美浜区中瀬 1-9-3

現在、インターネット上の情報アクセス手段としては、キーワード検索とブラウジングが主に利用されている。キーワード検索は検索式作成のための検索知識が必要であり、ブラウジングはコンテンツ製作者の枠組みに束縛されるという問題がある。

そこで、我々は、キーワード検索、ブラウジングに続く第 3 の情報アクセス手段である「連想ナビゲーション」を提案する。連想ナビゲーションでは、ユーザは文書に現れる単語の間の関連を表す連想マップを見て、文書全体に散在する特徴や傾向を把握しながら、漠然とした要求を満たす文書へアクセスできる。

# Association-based Navigation

Hisayuki Horai, Isamu Watanabe, Takahiro Saito, Kazuo Misue, and Kunio Matsui
Document Processing Laboratory, FUJITSU LABORATORIES LIMTED
1-9-3, Mihama-ku, Chiba-shi, Chiba 261-8588 Japan

Keyword search and browsing are the most popular information retrieval methods on the internet. Keyword search requires that users must have high-level knowledge to specify a complicate keyword formula for retrieval. Browsing is very restricted for retrieval because the structure of links completely depends upon the intention of authors.

In order to solve these problems, we suggest a new information retrieval method called "Association-based Navigation" in this report. Association-based Navigation supplies Association Map which represents association among words in target documents. A user recognises characteristics and trends of the target documents from Association Map, and accesses to suitable documents for his/her vague requirements.

## 1. はじめに

本稿では、キーワード検索、ブラウジングとは異なる情報アクセス手段である「連想ナビゲーション」を提案する。連想ナビゲーションでは、ユーザは文書に現れる単語の間の関連を表す連想マップを用いて文書 DB 全体をある着目点に注目してある切り口で概観することで、文書全体に散在する特徴や傾向を把握しながら個々の文書へアクセスできる。

現状の情報アクセス手段としては、キーワード検索とブラウジングが主に利用されている。キーワード検索を用いて大量の情報から欲しいものを得ようとするとき、情報利用者は検索のためのキーワードの選択と組合せ（検索式の作成）を慎重に行わなければならない。そのためには、検索式作成等の検索知識が必要であり、情報利用者の大きな負担となっている。一方、ブラウジングではあらかじめ情報の提供者が用意した情報間のリンクを利用するため、情報提供者の意図を反映したリンク構造が情報利用者の意図に合致していない場合、情報利用者が欲しい情報へアクセスするのは難しい。

たとえば、情報利用者が「競合企業間の事業展開の傾向を知りたい」等の漠然とした要求を持って情報にアクセスしようとした場合、キーワード検索ではユーザが適切なキーワードを指定することが難しい。「競合」「企業」「事業」「展開」等の単語は一般的過ぎて情報の絞り込みに不向きであり、情報の絞り込みに効果的な企業名や事業名を指定することは困難である。また、著者の経験から、このような要求を意図してすでに製作されたリンク構造がある可能性は低く、既存のリンク構造に依存するブラウジングを用いることは困難である。

文書群を対象として上記の問題点を解決するために、我々は連想ナビゲーションを考案した。第 2 節では連想ナビゲーションの概要を例を用いながら説明する。第 3 節では、連想ナビゲーションがユーザに供給する機能およびその実現方法について具体的に述べる。第 4 節では、インターネット上の新聞記事検索サービスへの連想ナビゲーションの適用について述べる。

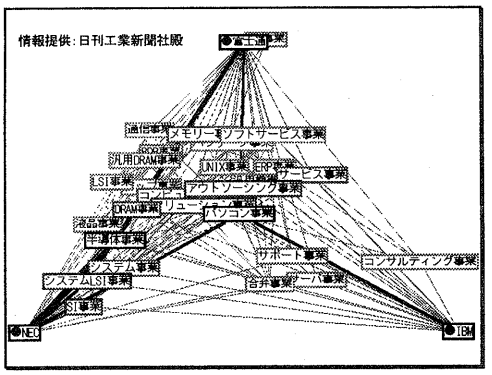


図 1: 連想マップの例

## 2. 連想ナビゲーションの概要

連想ナビゲーションの最大の特徴は、文書に現れる単語の間の関連度を表す連想マップ（図 1）である。連想マップを用いて、ユーザは様々な着目点に注目して様々な切り口で文書全体を概観し、そこに散在する特徴や傾向を把握することができる。たとえば、図 1 の連想マップは、新聞記事データベースにおいて 3 企業（富士通、NEC、IBM）を着目点とし、事業という切り口でそれらの企業がどのように扱われているかを表現している。文書全体に現れる事業名のうち 3 企業との関連の強いものを表示している。その関連の強弱を各単語の枠線の

濃淡で表現している。また、企業名を正三角形の頂点に固定し、事業名を企業名や他の事業名との関連の強弱を反映した位置に配置している。各単語の間の関連の強さは、単語間の関係線の太さで表現されている。このような配置は、ユーザが企業と事業や事業同士が文書全体でどのように関連付けて扱われているかを把握するために役立つ。たとえば、図1からソフトサービス事業はNECやIBMに比べて富士通と関連付けて扱われることが多いことが分かる。このように、図1は「競合企業間の事業展開の傾向」を把握するために有効であると考えられる。

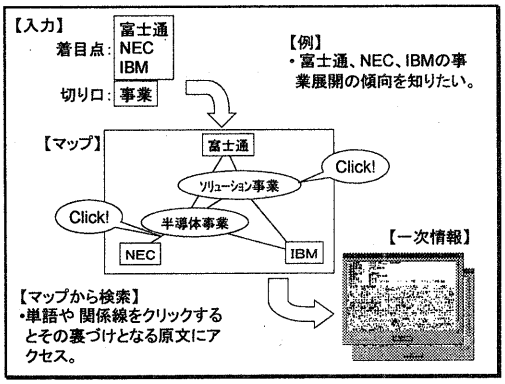


図2: 連想ナビゲーションの概要

図2に連想ナビゲーションの概要を示す。ユーザは漠然とした要求から個々の文書へアクセスするために、ユーザが要求から始めて文書全体の傾向を把握するための連想マップを作成する作業と連想マップから個々の文書へアクセスする作業を行う。前者は、着目点と切り口を入力しマップを表示することで行われる。2.1節では連想マップを作成するための入力について、2.2節では連想マップから一次情報への検索について述べる。

## 2.1 連想マップ作成のための入力

連想マップを作成するために必要な情報は、着目点である連想マップ上で固定する単語（固定単語と呼ぶ）と、切り口である分散して配置する単語（分散単語と呼ぶ）のカテゴリ（たとえば、事業名など）である。すなわち、ユーザが漠然とした要求から連想マップを作成する過程に対する支援とは、ユーザが固定単語と分散単語カテゴリを適切に入力するための支援である。連想ナビゲーションでは、ユーザはカテゴリや単語のリストを用いたシステムとの対話を任意回行い、最終的に適切な固定単語と分散単語カテゴリを選択し連想マップを作成する。本節では、この連想マップ作成支援について、前述の「競合企業間の事業展開の傾向を知りたい」という要求から図1の連想マップを作成する例を用いて説明する。

連想ナビゲーションでは、連想マップの作成に有効な汎用のカテゴリをシステムで用意している。基本的には、ユーザはシステムが頻度を利用して順序付けたカテゴリや単語のリストから選択することでカテゴリや単語の入力を行う。たとえば、図1の分散単語カテゴリはカテゴリのリストからカテゴリ「事業」を選択する。固定単語の指定では、カテゴリ「企業名」の単語リストから競合する企業3社を選択することになる。しかしながら、ユーザがどの3社が競合するかすらわからない場合、ユーザ文書全体に含まれる様々な業種の膨大な数の企業のリストから競合3社を選択できないという問題点が生じる。

このような問題点を解決するために、連想ナビゲーションでは、ユーザが単語や文

書の範囲を指定して、その範囲内で順位付けしたリストを見ることができる。たとえば、コンピュータ業界における競合企業を選択する場合、カテゴリ「業種」から単語「電気／電子機器」を選択し、それに関連するカテゴリ「企業」の単語リストを見ることができる。また、自社（たとえば、富士通）の競業企業を選択する場合、カテゴリ「企業」のリストから単語「富士通株式会社」を選択し、それに関連するカテゴリ「企業」の単語リストを見ることができる。

### 2.2 連想マップから一次情報への検索

連想ナビゲーションでは単語や関係線を指定して、それらに関連する文書にアクセスすることができる。得られた文書は、関連の強弱を利用して順序付けられたリストとして提示される。次に、ユーザはこの文書リストから文書を選択して、文書の原文にアクセスすることができる。

たとえば、図 1 の連想マップに対して、富士通とソフトサービス事業を結ぶ関係線を指定する、あるいは、富士通とソフトサービス事業を指定することで、それらを関連付ける文書にアクセスすることができる。ある事業と各社の連想マップ上の位置を見ながら、関連付ける文書に次々とアクセスすることで、各社のその事業に関する展開の傾向を知ることができる。

## 3. 連想ﾅﾋﾞｹﾞｰｼｮﾝｼｽﾃﾑの主要機能

本節では、2 節で説明した連想ナビゲーションを実現するシステムに必要な機能と方式について述べる。図 3 に連想ナビゲーションシステムの機能概略図を示す。

単語や文書の間の関連度が、連想ナビゲーションにおけるほとんどすべての機能を実現するためのデータとなっている。この関連度は、文書内の単語の共起関係を利用して、あらかじめ文書全体から計算され、単語と文書でインデクス付けられたマトリクスの形式で連想辞書に蓄えられる。3.1 節では、この連想辞書作成機能について述べる。3.2 節では、関連度を利用した連想マップ作成支援のための単語リスト作成機能について述べる。そこでは、2 節で述べた単語や文書の絞り込みについても言及する。3.3 節では連想マップの作成機能について述べる。3.4 節では連想マップから文書にアクセスするための関連文書検索機能について述べる。

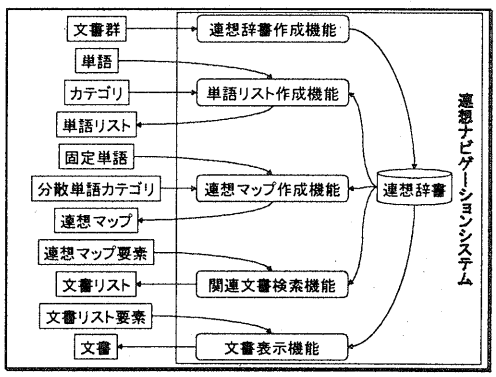


図 3:連想ナビゲーションシステムの機能

### 3.1 連想辞書作成機能

連想ナビゲーションで利用する単語-文書間の関連度は、文書内での単語頻度を基本にあらかじめ計算される。その計算方法は、キーワード検索におけるインデクス作成で利用さているものの応用である（詳細な計算方法については、[1]を参照）。計算結果は単語 w と文書 d でインデクス付けられた関連度マトリクス M(w,d)として連想辞書に格納する。

連想辞書に含まれる単語が、連想ナビゲーションで用いることができる単語である。すなわち、どのような単語を文書から抽出するかによって、連想ナビゲーションのサービス内容が決まる。たとえば、企業間の事業の傾向を知りたいのであれば、企業名や事業名を抽出しなければならない。次に、連想ナビゲーションで抽出する単語の候補を列挙する。

-形態素解析で得られる単語（名詞など）や複数の単語から構成される複合語。
-特定の語尾パターンを持つ語句（「～事業」など）。
-文書に付けられたメタ情報（新聞社が新聞記事データに付けた記事種類など）。

### 3.2 単語リスト作成機能

2 節で述べた連想マップ作成支援のための単語リスト作成機能について説明する。ある単語 X との関連の強弱で順位付けられたあるカテゴリ C の単語リスト(たとえば、自社名や単語「コンピュータ業界」に関連した企業名のリスト）は、連想辞書に格納された関連度マトリクス M から単語 X とカテゴリ C の全単語 w に関する行ベクトル M(X,*)と M(w,*)を求めて、内積 $\Sigma M(X,d)\cdot M(w,d)$ を計算し、その値によってカテゴリ C の単語をソートすることで求められる。

### 3.3 連想マップ作成機能

ユーザが指定した固定単語 Xs と分散単語カテゴリ C をもとに連想マップを作成するためには、まず連想マップに表示する分散単語 Ys を決定し、次に Xs∪Ys の各単語ペア(w1,w2)間の関連度を計算し、最後にこの関連度をもとに各単語の座標を計算する。

単語 Xs とカテゴリ C から Ys を決定する方法は、3.2 節と同様に単語リストを作り上位の単語を選択することで実現する。単語 w1 と単語 w2 の関連度も 3.2 節と同様の関連度マトリクスの行ベクトルの内積を用いる。

単語の座標の計算には、固定ノード付きのスプリングレイアウトを用いる（具体的な計算アルゴリズムは、[2]を参照)。

### 3.4 関連文書検索機能

2 節で述べた文書アクセス支援のための関連文書検索機能について説明する。関連文書検索機能では、まずユーザが指定した複数の単語や関係線を元に検索単語集合 ws を作成し、次に単語集合 ws を元に文書リストを作成する。

関係線の選択は、その両端の単語の選択と等価とみなし、ユーザが選択した単語と関係線から単語集合 ws を作成する。複数の単語の選択は、それらの単語を関連付ける文書の検索であるとみなし、単語集合 ws に関する AND 検索を行う。具体的には、連想マトリクス M の単語集合 ws に関する部分マトリクス M(ws,*)を求め、文書 d のうち列ベクトル M(ws,d)の要素がすべて非 0 であるものを選び、その大きさ |M(ws,d)| でソートする。

## 4. 新聞記事検索サービスへの適用

本節では、連想ナビゲーションの一実装例として、インターネット上の新聞記事検索サービスへの応用について述べる。これまでに我々はテキストマイニングツール ACCENT を開発してきた[3]。連想ナビゲーションシステムの主要な機能は、

ACCENTにおいてすでに実現されている。そこで、ACCENTをベースにして、インターネット新聞記事検索サービスシステムを構築することにした。このとき、解決すべき問題点は以下のとおりである。

-各種Webブラウザへの対応(HTTP対応)。
-連想辞書の短時間での更新(毎日追加される記事への対応)。
-記事検索に適したカテゴリの決定と単語の抽出。

4.1節でACCENTの概要を説明し、4.2節以降で上記の問題点に対する解決策を示す。図4は、本システムの実装方式の概略図である。

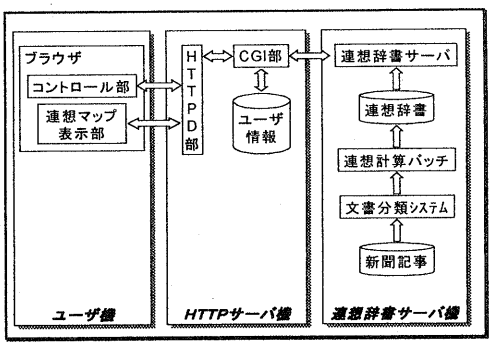


図4：インターネット上の新聞検索ｻｰﾋﾞｽｼｽﾃﾑ

### 4.1 ACCENT

ACCENTは、文書群を対象に様々な分析を行うための汎用テキストマイニングツールである。ACCENTは、文書群から連想辞書を作成する機能、単語-単語／単語-文書／文書-単語／文書-文書間の関連度計算およびそれに基づく検索（リスト作成）の機能、検索やカテゴリ種類や文書情報や単語表記パターンなどを利用した単語や文書の絞り込み機能、スプリングレイアウトや階層レイアウトを用いた様々なマップの作成／描画機能、表形式（csv形式）での出力機能などを持つ。3節で述べた連想ナビゲーションシステムの機能とほぼ同様の機能を持っている。

ACCENTは、連想辞書作成のためのビルダ、連想辞書をもとに様々な関連度計算や検索を行うサーバ、単語入力や単語リスト表示や文書リスト表示などのユーザインタフェース機能を持ったクライアント、クライアントとファイル経由で通信し様々なマップを表示するグラフィカルインタフェースから構成され、各コンポーネントはC言語で作成された専用ソフトである。サーバ・クライアント間の通信は、TCP/IP上で独自のプロトコルを用いて行っており、通常はクライアント起動から終了まで接続を保持する。

図4に示したシステムの構成要素のうち、連想計算バッチと連想辞書サーバは、それぞれ、ACCENTのビルダとサーバを改造して実現した。主な改造点は、大容量化と高速化や、サービス内容に合わせたプロトコルの追加と単語の抽出方式の変更などである。

### 4.2 インターネット対応

インターネット上の新聞検索サービスでは、ユーザインタフェースとしてWebブラウザを用いることにした。そこで、ACCENTのグラフィカルインタフェースとクライアントに対応する連想マップ表示部とコントロール部は、それぞれ、JavaAppletとHTMLで新規作成した。また、ブラウザのページを動的に作成するためにCGI部を新規作成した。

また、サービス提供サイトとユーザサイトはHTTPプロトコルで通信することにし

た。そこで、図4のようにサービス提供サイトではHTTPサーバ機と連想辞書サーバ機を分離し、負荷分散を実現した。連想辞書サーバとCGI部はACCENTの独自のプロトコルを拡張して通信するが、連想辞書サーバの占有を防ぐために連想ナビゲーションの機能ごとに接続を切断することにした。

### 4.3 辞書更新

本新聞検索サービスでは、複数紙の複数年の膨大な新聞記事を対象とし、毎日の短いサービス停止時間を利用して、新たに入手した数千個を含めて連想辞書を更新しなければならない。そこで、連想辞書バッチにおいて、新たに入手した記事から計算する部分と全ての記事から計算する部分を分離した。この際、連想辞書の形式を変更し、記事ごとに独立な情報を漸次蓄積できるようにした。

### 4.4 単語カテゴリの決定と単語の抽出

単語カテゴリとして、新聞記事検索で有効であると考えられる語尾パターン(「～事業」など)と、文書分類システムにより各記事に付与される分類の体系を用いる。

一般的に、単語の抽出には、チューニングが必要である。新聞記事からの単語抽出におけるチューニングには、たとえば、語尾パターン「～事業」について、「三事業」・「大事業」・「水道工事業」(これは、「水道工/事業」ではなく「水道/工事/業」であり、事業名ではない)などの不適切な単語に対する処置などがある。

## 5. 連想ナビゲーションの実例

本節では、様々な要求からの情報アクセスのために有効な連想マップおよびその作成手順を示す。

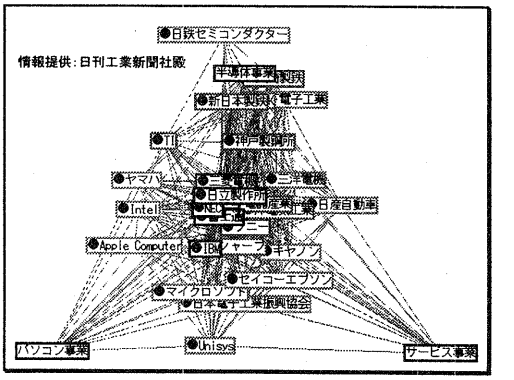


図5:企業間の関連を表す連想マップ

### 5.1 企業間の関連

「ある業界の主要事業に関して競合会社の関連を知りたい」という要求を持って情報アクセスを行う場合を考える。まず、対象業界に関連する文書に範囲を限定し、事業名を検索し、上位の3事業を着目点として指定する。次に、切り口として企業名を指定し、連想マップを作成する。たとえば、図5は、電子/電気機器業界における上位3事業である半導体事業、パソコン事業、サービス事業を着目点とし、企業名を切り口とした連想マップである。図5から、(1)日立製作所、NEC、富士通など中央付近の企業はこれら3事業に関してほぼ均等に扱われている、(2)Unisysやマイクロソフトなど下辺付近の企業は、半導体事業と関連付けられていない、(3)日鉄セミコンダクタなどは半導体事業に特化している企業である、などの事実が読み取れそうである。

### 5.2 事業と市場の関係

「国内市場、米国市場、欧州市場における事業の特性を知りたい」という要求を持って情報アクセスを行う場合を考える。着目点として、国内市場、米国市場、欧州市場を指定し、切り口として事業名を指定し、連想マップを作成する。たとえば、図6は、電子／電気機器業界に文書の範囲を制限して作成した連想マップである。

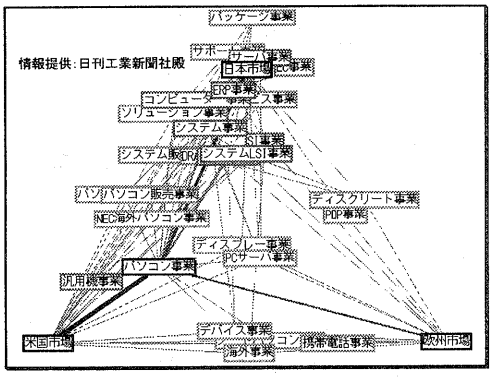


図6：市場と事業の関連を表す連想マップ

## 6. 連想ナビゲーションの将来像

連想ナビゲーションの将来像を図7に示す。そこでは、以下のような機能が提供される。①分類技術により情報のカテゴリをそろえることで、質の高い連想マップを作成する。②情報抽出技術等により分野別の辞書を用意することで、分野ごとに特化した多様な着目点や切り口の連想マップを作成できる。③自動翻訳技術の導入により、異なる言語で記述された文書を統一的に利用するための連想マップが作成できる。④連想マップから文書リスト経由で各文書にアクセスするだけでなく、自動要約技術を利用して関連する文書群を一括してアクセスできるようにする。

## 7. おわりに

本稿では、キーワード検索、ブラウジングに続く第3の情報アクセス手段として連想ナビゲーションを提案した。連想ナビゲーションでは、ユーザは文書に現れる単語の間の関連を表す連想マップを見て、文書全体に散在する特徴や傾向を把握しながら個々の文書へアクセスできる。

また、本稿では、連想ナビゲーションの適用例としてインターネット上の新聞検索サービス について報告した。

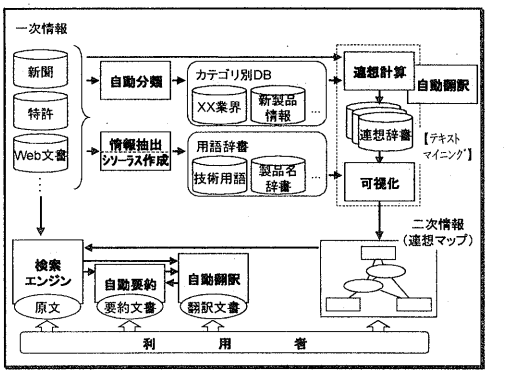


図7：連想ナビゲーションの将来像

## 参考文献

[1] 渡部 勇：発想支援システム「Keyword Associator」第二版，計測事業制御学会第 15 回システム工学部会研究会資料 (1994).

[2] 三末 和男，渡部 勇：テキストマイニングのための可視化技術，情報処理学会第55回 情報学基礎研究会資料(1999).

[3] 渡部 勇，三末 和男：単語の連想関係によるテキストマイング，情報処理学会第55回 情報学基礎研究会資料 (1999).