情　報　学　基　礎　55－9
デジタル・ドキュメント　19－9
（１９９９．７．１６）

テキストマイニングのための連想関係の可視化技術

三末 和男 ・ 渡部 勇
E-mail: {misue,isamu}@flab.fujitsu.co.jp
(株)富士通研究所 コンピュータシステム研究所
〒261-8588 千葉市美浜区中瀬 1-9-3

大量のテキスト情報から知識につながる有用な情報を得るための，連想関係の可視化技術について述べる．ここで言う「連想関係」とは，単語の共起性に基づく統計処理によって数値的に得られる文書や単語間の関係である．本稿は連想関係を利用した「連想分析」のための関係情報の可視化技術に焦点を合わせる．連想分析の概要を説明し，可視化技術としては無向グラフの自動レイアウトのためのスプリング埋め込みとその拡張について述べる．また，分析事例として，新聞記事を利用した車メーカのイメージの分析を紹介する．

Visualization of Keyword Association for Text Mining

Kazuo MISUE and Isamu WATANABE
E-mail: {misue,isamu}@flab.fujitsu.co.jp
Computer Systems Labs, Fujitsu Laboratories Ltd.
1-9-3 Nakase, Mihama-ku, Chiba-shi, Chiba 261-8588 Japan

This paper describes a technique to visualize “association” to mine useful information from large volume of textual data. Here, “association” is relationships among text segments and words, which are statistically calculated by using cooccurrence of words. In an analysis method based on the association, a technique to visualize relational information is focused on. A graph drawing technique – spring embedding – is introduced and extended. An instance of image analysis based on newspapers is also included.

1 はじめに

オフィスのOA化による電子的テキストの増加，ネットワークやWWWの普及による流通・収集の促進，そしてディスク等記憶装置の低価格化による膨大な蓄積の結果，個人や組織の保有するテキスト情報は増加の一途である．それらには貴重な知識が含まれていると推測されるが，その大多数は死蔵であり，有効活用への期待は大きい．

ところが，現状は大量のテキスト情報を「どう」処理すれば「なに」が得られるか，について十分に分っているとは言い難い．そこで，本研究は，大量のテキスト情報から有用な情報を得ることを目指して，上記の「どう」と「なに」について理解を深め，そのための技術を開発することを目的とする．

本稿では，特にテキスト情報を統計的に処理して得られる連想関係(単語間の関連度)から有用な情報を発見しやすくするための，可視化技術の開発に焦点を合わせる．

以下，第2章では，まず本稿で対象とする「連想分析」について概要を説明すると共に，支援ツールを紹介する．続く第3章では連想分析で利用する連想関係の可視化技術について述べる．第4章では連想分析の事例を可視化技術の観点から紹介する．最後に，第5章でまとめと今後の課題を述べる．

2 連想分析

2.1 連想分析の概要

連想分析は次のような手順で行なわれる．

1. 連想辞書作成
2. 関連度計算
3. 連想可視化

以下，それぞれの手順を概説する．ただし，以下の記述は，手順の理解を容易にするためにかなり単純化してある．なお，連想辞書作成および関連度計算については文献[1]を参照されたい．

2.1.1 連想辞書作成

入力テキスト情報を統計処理して連想辞書を作成する．入力は，新聞記事，特許広報，アンケート(自由書式)回答，Webページなどのテキスト情報で，電子化されているものとする．

テキストは，(新聞記事の場合)記事毎，(特許広報の場合)申請毎，(アンケートの場合)質問毎・回答者毎のようにある単位(「単位テキスト」と呼ぶ)に分割される．そして，各単位テキストから単語が抽出され，各単語の頻度が求められる．こして得られた頻度行列をここでは「連想辞書」と呼ぶ．

2.1.2 連想関係抽出

連想辞書を基にして，たとえばTFIDFなどにより，単位テキストと単語の間の関連度，単位テキストと単位テキストの間の関連度，単語と単語の間の関連度を，それぞれ計算する．これらの関連度を総称して「連想関係」と呼ぶ．

関連度は必ずしも頻度行列の全体を用いて計算するわけではなく，しばしば単位テキストや単語の，特定の部分集合についての頻度だけを用いて計算する．このような部分情報による連想関係を利用することで，特定の観点における連想分析が可能になる．

2.1.3 連想可視化

抽出した連想関係から有用な情報を発見しやすいように連想関係を視覚的に表現する．可視化に際しての表現形式は，表現したい関係情報とそこで強調したい特徴に依存して決められるものである．たとえば，リストや表，ネットワーク図(連結図)などが利用される．

本稿で述べる可視化技術は，ネットワーク図を利用する．詳しくは第3章で説明する．

2.2 連想分析支援ツール

筆者らはここで述べたような連想分析を支援するツール「ACCENT」を開発している．その前身は発想支援ツール HIPS[2] であり，HIPS は二つのサブツールを統合したものである．

サブツールの一つは「Keyword Associator(KA)」[3] で，テキスト情報を統計処理して，連想辞書を作成する機能を備える．連想検索ツールとして，ブレインストーミングのような発散的思考を支援する．

もう一つは「D-ABDUCTOR」[4] で，関係情報を図として描き操作する環境を提供する．KA を発散的思考に利用した場合，その後の収束的思考フェーズにおいて KJ 法を支援する．

HIPS は，発想支援の観点では，テキスト情報を入力として KJ 法的な図解を自動的に作成できる．筆者らは HIPS を発想支援ツールとして開発してきたが，様々なテキスト情報を処理しているうちに，分析に有効であると認識した．そこで，連想分析のための機能を強化発展させ，名称を「ACCENT」と改めた．

3 連想関係の可視化

アンケート分析を例に，連想分析における可視化について説明する．その後，関連技術について述べ，可視化技術として無向グラフのレイアウト法について説明する．

3.1 アンケート分析

分析テーマは『ホーム PC 受容調査』であり，その目的は家庭用 PC に関するニーズを把握することにある．

3.1.1 分析対象データ

分析対象データはアンケートの回答である．アンケートは性別や年齢のような選択式の質問とともに，Q14「なぜパソコンを購入したいとおもっているのですか」というような，いくつかの自由記述式の質問から構成されている．回答者数は 229 名で，回答それぞれを一つの単位テキストとして処理した．

3.1.2 キーワードのランキング

表 1 は Q14 に対する回答において重要な単語 (単位テキストの集合に関連の高い単語) の上位 5 個を，回答全体および，性別でランキングしたものである．全体においても男女別セグメントにおいても「インターネット」(表中では紙面の都合で「I-NET」と表記) が首位であることが分る．

表 1: 重要キーワードのランキング・リスト

順位	全体	男性	女性
1	I-NET	I-NET	I-NET
2	仕事	仕事	情報
3	情報	興味	興味
4	興味	情報	子供
5	パソコン	必要	パソコン

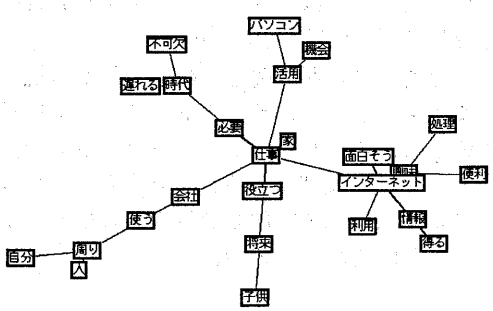


図 1: 男性の重要キーワードの連想関係マップ

3.1.3 キーワードの連想関係のマップ

図 1 は，Q14 の男性による回答から抽出された上位 25 程度の重要単語を関連度の強弱を平面マップとして可視化したものである．表 1 では，たとえば男女それぞれのセグメントに

おいて,「興味」が何に連想付けられているのかは読み取れないが,図1のマップでは,「興味」が「インターネット」に連想付けられていることを読み取ることができる.

3.2 関連技術

情報の可視化は大規模で複雑な情報をうまくすれば効率的に人間に伝達できるため,大量のテキスト情報から有用な情報を発見しようとするテキストマイニングにおいて重視されている [5].

現在主流の可視化技術は,キーワード等なんらかのパタンの頻度をヒストグラムや折れ線グラフなどで表示することで,比較や時間推移を把握しやすくしたものである.筆者らが目指す可視化技術は,これらとは異なり,連想関係を視覚的に提示することを特徴とする.

多変量解析の分野には関連性の可視化に利用できそうな手法がいくつかある.単位テキストに対してそれに含まれる単語を多変量解析的なデータとみなすと,質的データであるため手法としては,コレスポンデンス分析や数量化III類が代表的である.

単語間,文書間の関連性の対話的かつ視覚的な利用としては,テキストマイニング以外の応用もある.たとえば,キーワード間の関連性を多次元尺度構成法により表現したマップにより人の認知空間を計算機上に構成しようとする研究がある [6, 7, 8].

3.3 スプリング埋め込みの利用

コレスポンデンス分析や数量化III類などにより分析を行う場合には,あらかじめ次元を想定した上で変数を設定する必要がある.

本稿で述べる連想分析では,文体や書式が必ずしも統一されていない自由形式の文書を対象とする.形態素解析によりそのような文書から抽出した単語群は,そのままでは分析のために計画された変数とはなりえず,これらの手法で可視化できたとしても何かを発見できるような結果は期待できない.

対策としては,(a) 多変量解析の手法を適用できるような加工をテキストからデータを抽出する段階で行う,あるいは (b) 可視化において分析目的に即した観点を導入できるような操作を可能にする,が考えられる.

本稿では (b) のアプローチをとる.連想関係を,操作が可能な仮想物理モデルによって表現し,分析者がそのモデルを操作することで様々な観点から連想関係を眺め分析を進められるようにする.

3.3.1 Eades のスプリング埋め込み

Eades によって提案された「スプリング埋め込み」[9] は,無向グラフを対象とするグラフレイアウト法の一つで,グラフを仮想的な物理モデルにあてはめ,モデルの安定状態を求めることで適切なレイアウトを得る.Eades のモデルでは,グラフの各頂点を質量0,大きさ0の「リング」とみなし,各辺をリングをつなぐ「スプリング」とみなす.各頂点は,そのスプリングから式 (1) で表される大きさ f_s の力を受けると同時に,辺でつながっていない別の頂点からは式 (2) で表される大きさ f_r の斥力を受ける.

$$(1) \qquad f_s = c_s \cdot \log(d/d_0)$$

$$(2) \qquad f_r = c_r/d^2$$

ここで,d は頂点間の距離,d_0 はスプリングの自然長である.定数 c_s と c_r は力のバランスを制御するための係数である.Eades のモデルでは,定数 c_s, c_r, d_0 はグラフ全体に渡って一様である.

3.3.2 スプリングの自然長と強さ

関連度の可視化のしかたとしては,単語を頂点に割り当て,関連度を頂点間の距離で現わせると都合が良い.そこで Eades のモデルか

ら離れ，それぞれのスプリングの自然長 (d_0) を関連度に応じて変えることにする．

また，関連度の強弱を 2 次元平面上の距離で表現しようとするとどこかに歪みが生じるため，重要な情報を落さないように，そらの歪みをうまく制御する必要がある．そこで，それぞれのスプリングの強さ (c_s) も関連度に応じて変えることにする．

つまり，関連度によって各スプリングの自然長と強さが決定される．以下に関連度 r を自然長と強さに変換するための代表的なパタンを紹介する (実際にはさらに定数倍される)．

パタン 11 関連度が 0 より大きい場合に自然長を 1.0，強さを 1.0 とする．関連性の有無だけが表現される．

パタン RR 自然長を $1/r$ とし，強さを r とする．関連度が大きい単語を近くに配置し，関連度の小さい部分に歪みを集める．

3.4 操作の導入

3.4.1 頂点の固定

いくつかの頂点を 2 次元平面のある位置に固定して動けなくする操作である．たとえば二つの頂点を十分な距離を保って固定すると，その両方につながる頂点は両方が引っ張られ，関連度の高い方に寄って配置される．

固定されていない頂点の配置は他の固定されていない頂点からの影響も受けるため，単純な引き合いとは異なる．つまり，単に固定された頂点に対する直接的な関連度の比較ではなく，頂点間の間接的な関連性を反映した比較が可能になる．

3.4.2 辺の間引き

関連度の大きい辺 (スプリング) だけを残し，関連度の小さい辺を削除する操作である．関連度の高い辺だけが残ることで，関連性の主要な骨格が顕在化される．

削除する辺の選び方としては，単純に関連度の小さい方から必要な数だけ削除するという方法の他，全体の連結性を保持しつつ関連度の小さい方から削除する方法もある．

4 分析事例

新聞記事を利用した自動車メーカのイメージ分析を紹介する．新聞記事は日経産業新聞 3 年分 (1994 年～1996 年) で，記事数は 157,756，抽出された異り単語数は 667,670 であった．

ここではイメージを抽出しやすくするために「～性」というパタンの属性名詞に着目して分析を進める [1]．

4.1 キーワード・ランキング

表 2 は「トヨタ」,「日産」,「ホンダ」というメーカ名に対して「～性」という属性名詞を関連度によってランキングしたものである．表中「(3 社)」は「トヨタ」,「日産」,「ホンダ」それぞれからの関連度の和によるランキングである．どの列においても「安全性」は 6 位以内であり，どのメーカも安全性と関連付けられていることが分る．

4.2 「頂点固定」マップ

表 2 から各社とも「～性」という属性名詞については「安全性」との関連が強いことが分った．しかしながら，自動車業界において (特に 3 社に対して)「安全性」がどういう位置にあるのかは読み取れない．

そのような位置を知るためには，関連度の可視化が効果的である．一つの単語についてだけなら，たとえばヒストグラムの利用が簡単である．しかしながら，通常イメージは単独に形成されるのではなく，他のイメージと影響し合って形成されるため，一つの言葉の位置よりも，他の言葉との関係を同時に見るべきである．そのためには図 2 のような単語

表 2: 車メーカ名に対する属性名詞「～性」の関連度によるランキング

順位	'94～'96 (3社)	1995年			1996年		
		トヨタ	日産	ホンダ	トヨタ	日産	ホンダ
1	安全性	安全性	安全性	安定性	安全性	生産性	剛性
2	生産性	生産性	生産性	走行安定性	安定性	安全性	安全性
3	安定性	走行安定性	方向性	耐久性	居住性	剛性	耐久性
4	居住性	安定性	安定性	出力特性	生産性	耐久性	トルク特性
5	耐久性	快適性	操縦安定性	実用性	必要性	静粛性	経済性
6	走行安定性	視認性	居住性	安全性	快適性	安定性	静粛性
7	剛性	耐衝撃性	採算性	作業性	走行安定性	優位性	発がん性
8	実用性	成形流動性	走行性	居住性	独立性	双方向性	実用性
9	静粛性	閉鎖性	操縦性	走破性	独自性	方向性	ファッション性
10	方向性	自主性	耐久性	機作業性	チャネル個性	地域性	安定性

のマップが有効である.

図 2 は三つのメーカ名を三角形の頂点に固定したマップである (パターン RR を利用). このマップにより,「～性」という属性名詞の 3 社に対する位置関係が直感的に捉えられる. たとえば, 1995 年には「安全性」は「トヨタ」と「日産」に強く引き寄せられている. 1996 年には, さらに「トヨタ」だけに極端に引き寄せられている. ということが分る.

4.3 「辺間引き」マップ

次に 1995 年から 1996 年にかけて, 自動車業界において (特に 3 社に対して)「安全性」に関して何が起ったかを探ってみる.

図 3 は,「トヨタ」,「日産」,「ホンダ」のいずれかに関連する記事[1]を対象に,「安全」に関連する単語間の関連性を辺を最大限間引いて可視化したものである (パターン 11 を利用).

図 3 では紙面の都合で上位 30 単語程度に留めてあるものの, 1995 年に安全と言うと「エアバック」が一つの中核を成していることが分る. それに対して, 1996 年にはエアバックとは別に「衝突」という中核が出現していることが分る. 1996 年頃から衝突に関するボディの安全基準が問題視され始めて, 各社ともその対応を進めたと推測される.

これより先は, マップの情報を越えた分析になるが, 図 3(b) 中の「衝突」の上にある「GOA」に着目して, KA の検索機能により記事を検索すると, トヨタが提唱する衝突安全基準 GOA に関する多くの記事により「トヨタ」と「安全性」が関連付けられていることが分る.

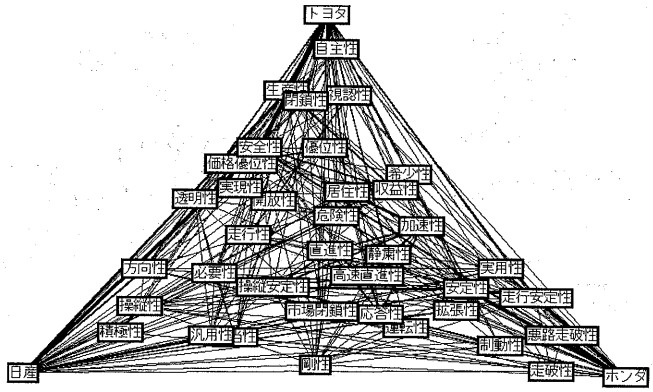
4.4 マップから読み取れない情報

連想関係を表現したマップからいくつかの情報を読み取ったが, 次の情報は現在のマップからは読み取れない. これらをいかに抽出し, どう提示するかは今後の研究課題である.

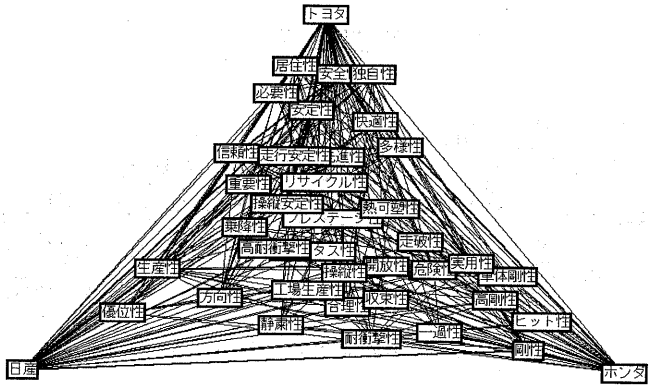
肯定・否定: 「トヨタ」と「安全性」の関連度が高いということは,「トヨタ」と「安全性」の共起性に基づく統計量が大きいだけで, トヨタの安全性が「高い」か「低い」かについては情報を与えない.

絶対量: 1995 年のマップにおいて「安全性」の「日産」からの距離は「トヨタ」からの約 2 倍である. しかし,「トヨタ」の安全性のイメージが「日産」より「2 倍」強いわけではない.

[1]「トヨタ」,「日産」,「ホンダ」のいずれか, または「ニッサン」や「本田技研工業」などの別名を含む記事.



(a) 1995 年の記事による分析結果



(b) 1996 年の記事による分析結果

図 2: 車メーカーの 3 社 (トヨタ，日産，ホンダ) の頂点固定マップ

5 まとめと今後の課題

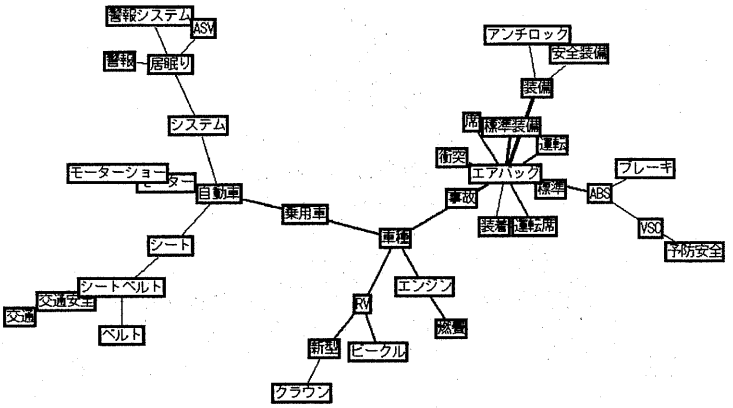
大量テキスト情報から統計処理に基づいて抽出した連想関係を可視化する技術について，関連性の見せ方を工夫することで，個々のテキスト情報を個別に読むだけでは把握が困難な情報を容易に得られることを，実例を通して示した．

評価に代えて，ある利用者のコメントをあげると，ある企業のイメージに関するアンケート調査の分析結果に対して，利用者は，マップが役に立ったと述べた．さらに，その理由として，うすうすそうではないかと思っていたことが，客観的に明らかになった，そしてさらに詳細になった，と述べている．

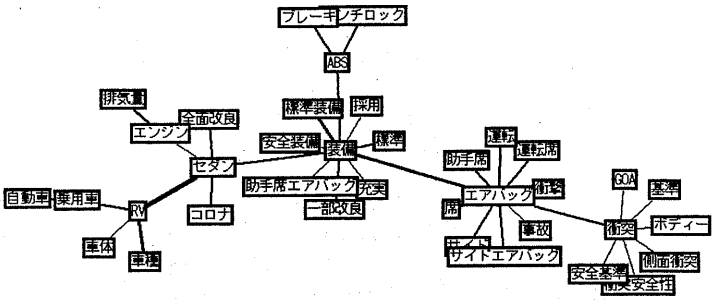
今後の課題としては，第 4 章であげた他に，「どのような」データに対して「どう」処理すれば「なに」が得られるかを体系化すること，そしてそれらが「なぜ」得られるのかを数学的に裏付けることが重要と考えている．

参考文献

[1] 渡部 勇，三末和男: 単語の連想関係によるテ



(a) 1995 年の記事による分析結果



(b) 1996 年の記事による分析結果

図 3: 車メーカーの辺間引きマップ

キストマイニング, 情報処理学会第 55 回 情報学基礎研究会資料 (1999).

[2] 渡部 勇, 三末和男, 新田 清, 杉山公造: ハイブリッド発想支援システム「HIPS」, 計測自動制御学会 第 17 回システム工学部会研究会「発想支援ツール」資料, pp. 77–84 (1995).

[3] 渡部 勇: 発散的思考支援システム「Keyword Associator」第二版, 計測自動制御学会 第 15 回システム工学部会研究会資料, pp. 9–16 (1994).

[4] 三末 和男, 杉山 公造: 図的発想支援システム D-ABDUCTOR の開発について, 情報処理学会論文誌, Vol. 35, No. 9, pp. 1739–1749 (1994).

[5] 那須川 哲哉, 諸橋 正幸, 長野 徹: テキストマイニング – 膨大な文書データの自動分析による知識発見 –, 情報処理, Vol. 40, No. 4, pp. 358–364 (1999).

[6] Hori, K.: A System for Aiding Creative Concept Formation, *IEEE Trans. SMC*, Vol. 24, No. 6, pp. 882–894 (1994).

[7] 杉本 雅則, 堀 浩一, 大須賀 節雄: 設計問題への発想支援システムの応用と発想過程のモデル化の試み, 人工知能学会誌, Vol. 8, No. 5, pp. 575–582 (1993).

[8] 角 康之, 堀 浩一, 大須賀 節雄: テキストオブジェクトを空間配置することによる思考支援システム, 人工知能学会誌, Vol. 9, No. 1, pp. 139–147 (1994).

[9] Eades, P.: A Heuristics for Graph Drawing, *Congressus Numerantium*, Vol. 42, pp. 146–160 (1984).