情　報　学　基　礎　55－8
デジタル・ドキュメント　19－8
（１９９９．７．１６）

単語の連想関係によるテキストマイニング

渡部 勇　三末 和男

E-Mail: {isamu,misue}@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有用な情報を発見するためのテキストマイニングツール「ACCENT」を開発した。ACCENT は、文書群から抽出された単語の間の「連想関係」を、単語の共起性に基づいて計算し、マップ（ネットワーク図）として可視化する。文書を個別に調べてもわからない、文書群全体が持つ特徴・傾向を、この単語の連想マップを通して読み取ることが可能となる。本稿では、まずACCENTによる連想分析の流れを説明したあと、連想関係の計算方法・マップを構成する単語群の選択方法について詳細に述べる。また、新聞データを用いた分析実験例を紹介しながら、分析の目的・視点をマップに反映させるための方法およびその効果について報告する。

Text Mining Based on Keyword Association

Isamu WATANABE and Kazuo MISUE

E-Mail: {isamu,misue}@flab.fujitsu.co.jp

Computer Systems Labs, Fujitsu Laboratories Ltd.
1-9-3, Mihama-ku, Chiba-shi, Chiba 261-8588 Japan

We developed a text mining tool called ACCENT that facilitates us to mine useful information from large volume of textual data. The tool calculates associations among text segments and words in documents using coocurrence of words, and then visualizes them as a keyword map. Through the map, we can easily grasp traits and trends in the whole documents, which cannot be captured from the individual documents. In this paper, first, a flow of association analysis with ACCENT is described. Next, methods to calculate associations and to select words in a map are described. Finally, techniques to make a map according to the purpose of analysis are illustrated with examples.

1. はじめに

計算機ネットワーク環境の中に蓄積された大量のテキスト情報を有効に活用していくための道具として、今日では情報検索システムが広く利用されるようになってきている。近年、情報活用のためのもう一つの道具として、テキストマイニングが注目を集めている[1]。

本稿では、筆者らが開発をしたテキストマイニングツールACCENTについて述べる。ACCENTは、文書群から抽出された単語の間の「連想関係」を、単語の共起性に基づいて計算し、マップ（ネットワーク図）として可視化する。文書を個別に調べてもわからない、文書群全体が持つ特徴・傾向を、この単語の連想マップを通して読み取ることが可能となる。

以下、第２章ではACCENTによる「連想分析」の流れを説明する。第３章では連想分析のために使用される連想辞書の作成方法、第４章では関連度（連想関係の強さ）の計算方法についてそれぞれ詳細に述べる。第５章では新聞データを用いた分析実験例を紹介しながら、分析の目的・視点をマップに反映させるための方法およびその効果について報告する。最後に、第６章でまとめと今後の課題を述べる。

2. 連想分析

2.1. 連想分析支援ツール

筆者らはこれまでに、発想支援ツールHIPS[2]の開発を行なってきた。HIPSは、連想検索ツールKeyword Associator (KA) [3]と関係情報の可視化ツールD-ABDUCTOR (DA) [4]の要素技術を統合したものであり、当初は、テキスト情報を構造化・整理するためのツールとして開発した。その後、要求会議の分析[5]、あるいは自由回答形式のアンケート分析などに応用していく中で、テキスト情報を分析するツールとしても使えるようにさまざなま機能拡張を行なってきた。

本稿で紹介するACCENTは、HIPSの分析機能をさらに強化することにより、単語間の連想関係の可視化機能を中心とする連想分析支援ツールとして発展させたものである。

2.2. 目的指向分析

HIPSにおける分析処理の流れは、文書群を入力すると、分析結果が出力されるというバッチ的なものであった。

たとえばアンケート分析などに関しては、分析対象となる文書（アンケート回答）に書かれている内容が、すでに分析目的を反映したものになっているため、このようなバッチ的な方法でもそれなりの分析結果を出すことができた。

しかし、さまざまなタイプの文書を分析していく中で、分析結果に対し、分析者の意図・目的・視点を反映するための仕組みが不可欠であるとの認識に達した。

ACCENTには、分析結果である単語の連想マップを作成する際の、

(1) マップを構成する単語の選択方法
(2) 単語間の関連度の計算方法
(3) マップのレイアウト方法

に対し、それぞれ分析目的・視点を反映するためのさまざまな機能が用意されており、これがACCENTの分析ツールとしての大きな特徴になっている。

以下、次節では、(1)～(3)をコントロールしながらインタラクティブに進めていく「連想分析」の流れを、第3章、第4章では、(1)および(2)について説明する。なお、(3)については文献[6]を参照されたい。

2.3. 連想分析の流れ

図1にACCENTを用いた連想分析の流れを、図2にACCENTの画面例を示す。図1において、上の段はシステムが行なう処理を、下の段は分析者が行なう作業の流れを、真中の段は情報の流れをそれぞれ表している。

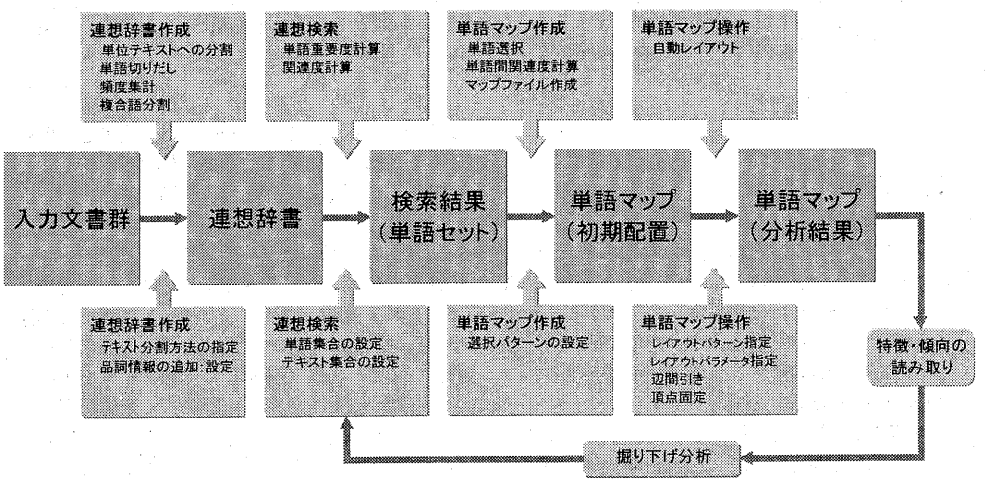

図1　ACCENTによる連想分析の流れ

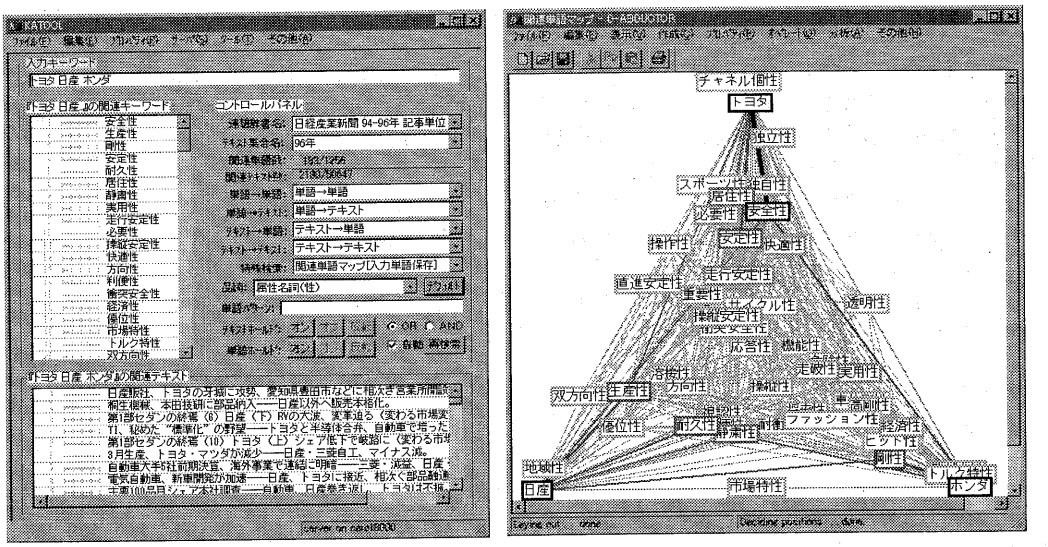

図2　ACCENT の画面例

■　連想辞書作成

システムにより、まず、分析対象となる文書群が「単位テキスト」と呼ぶ部分テキスト情報の集合に分割される。単位テキストは、単語の重要度および単語やテキスト間の関連度を計算する際の基本単位となる。

次に、各単位テキストからの単語切り出し・頻度集計・複合語の分割処理などが行なわれる。

上記の過程で得られる、各単位テキスト内での単語の出現頻度行列を「連想辞書」と呼んでいる。

分析者は、単位テキストへの分割方法、単語の品詞情報などを指定することにより、作成される連想辞書の性質をコントロールすることができる。

■　連想検索

単語群・連想関係の絞り込みは、連想検索ツール（図2左側の画面）上で、以下の連想検索機能を用いることによって行なう。

A)　重要単語を検索
B)　単語から関連単語を検索
C)　単語から関連テキストを検索
D)　テキストから関連単語を検索
E)　テキストから関連テキストを検索

連想検索には、連想辞書中の頻度情報から計算される、単語の重要度および単語やテキスト間の関連度が使用される。このとき、頻度行列の行および列を指定することによって、部分行列を用いて重要度・関連度を計算することも可能である。

行を指定する（計算に使用する単位テキスト集合を指定する）方法としては、

- 単位テキストの属性情報を用いる方法
- テキスト検索（C,E）の結果を用いる方法

列を指定する（計算に使用する単語集合を指定する）方法としては、

- 単語の品詞情報を用いる方法
- 単語の文字列パターンを指定する方法
- 単語検索（A,B,D）の結果を用いる方法

を利用することができる。

■　単語マップ作成

マップを構成する単語を選択するパターンとしては、以下の3種類がある。

①　重要単語検索（A）の結果の上位単語
②　関連単語検索（B）の結果の上位単語
③　単語検索の結果のうち指定された単語

このうち、①のパターンで作成されるマップを「重要単語マップ」、②のパターンで作成されるマップを「関連単語マップ」と呼ぶ。

システムは、分析者によって指定されたパターンにしたがってマップを構成する単語を選択し、単語間の関連度行列の計算を行ない、単語マップを作成する。作成された単語マップは、可視化ツール（図2右画面）上に表示される。

■ 単語マップ操作

可視化ツール上には、単語がグリッド状に配置される。単語を囲む枠の色は、重要単語マップの場合は重要度の大きさ、関連単語マップの場合は関連度の大きさを反映したものになる[1]。

単語間には、連想関係を表す辺（線）が引かれる。辺の太さは単語間の関連度の大きさを反映したいる。

システムは、単語間に引かれた辺をスプリングと見立てて、関連の強い単語が近くに配置されるように自動レイアウトを行なう。

分析者は、

- レイアウトパターンの指定
- レイアウトパラメータの指定
- 辺の間引き
- 頂点（単語）の固定

を行なうことによって、レイアウトをコントロールしながら、さまざまな観点から連想関係を眺め、単語マップから特徴・傾向などを読み取っていく。

■ 掘り下げ分析

単語マップから読み取った情報を元に、さらに掘り下げて分析を行なう場合には、連想検索機能を利用する。たとえば、ある単語とある単語の連想関係の背後にあるものを調べたい場合であれば、それらを入力としたAND検索を行なうことによって、二つの単語を結びつける単語やテキスト情報を見つけ出すことができる。あるいは、単語群・連想関係を絞り込むことによって再度単語マップを作成し、そこから背後にある要因を捉えることも可能である[6]。

3. 連想辞書作成

3.1. 単位テキストへの分割

連想検索の際に利用される単語の重要度および単語やテキスト間の関連度は、すべて単位テキスト内での単語の頻度情報から計算される。

特に、関連単語検索（B）および単語マップ作成の際に利用される単語間関連度に関しては、単位テキスト内での単語の共起性に基づいて計算されるため、同一単位テキスト内に現れない単語ペアの間の連想関係は抽出されない。したがって、単位テキストをどのように取るかが、単語の連想関係抽出のスコープを決めることになる。

3.2. 単語抽出

単語マップの可読性を高めるためには、「単語」をある程度まとまった意味を持つ単位として切り出す必要がある。一方、単語間の関連度を抽出するという観点からは、「単語」が細かく分かれていた方が都合よい。

そこで、ACCENTでは、未登録語に対して以下のような処理を行なうことにより、「まとまった単位」と「細かい単位」の両方を抽出するようにしている。

まず、単語を切り出す際、漢字とカタカナのみで構成される文字列を、一単語として抽出し、頻度集計を行なう。たとえば、「連想辞書作成方法」は一単語として抽出される。

全単位テキストについての頻度集計が終わったら、辞書情報・頻度情報を用いて、未登録語の分割処理を行なう。たとえば、「連想辞書」「作成方法」という単語が高頻度で抽出されていれば、「連想辞書作成方法」は、まず「連想辞書＋作成方法」に分割される。「連想辞書」「作成方法」が未登録語の場合には、それぞれ「連想＋辞書」「作成＋方法」に分割される。

上記の分割の過程で生成される部分文字列のうち、指定された長さより短いものはすべて「単語」として抽出される。たとえば、6文字以下の部分文字列を登録するような設定であれば、「連想辞書作成方法」からは、「連想辞書」「作成方法」「連想」「辞書」「作成」「方法」が単語として抽出されることになる。

3.3. 単語の品詞情報

ACCENTには、特定の品詞の単語のみを用いてマップを作成する機能がある。

分析の目的・視点を反映した単語の選択を行なうためには、名詞、固有名詞といった「文法的なカテゴリー」だけでは不充分であるため、分析者が自由に品詞情報を追加・設定し、「意味的なカテ

[1] 図2では、重要度・関連度が大きいものほど、枠の色が濃くなるように設定されている。

ゴリー」として利用することができるようになっている。

品詞情報は、ユーザ辞書に登録することで、追加・設定することができる。たとえば、自動車メーカー名を「自動車メーカー名」という品詞で登録すれば、自動車メーカー名のみを含む単語マップを作成することが可能となる。

特定の接尾文字列を持つ単語に対し、自動的に特別の品詞を付与することも可能である。たとえば、企業イメージ分析・製品イメージ分析などを行なう場合には、「的」「性」といった接尾文字列を設定する。こうすることで、イメージ表現によく利用される「～性」「～的」というパターンの単語に対し、それぞれ「属性名詞（性）」「属性名詞(的)」という品詞情報が付与され、他の品詞の単語と区別して扱うことができるようになる。

4. 関連度計算

4.1. 単語重要度

単位テキスト t における単語 w の重要度 $S(t,w)$ は、t における w の出現確率 $p(t,w)$（t における w の出現頻度を t における全単語の総頻度で割った値）と全単位テキスト集合における w の出現確率 $q(w)$（w の総頻度を全単語の総頻度で割った値）を用いて以下のように計算される（ただし、計算値が負になる場合は重要度を0とする）[1]。

$$S(t,w) = p(t,w)\log\frac{p(t,w)}{q(w)}$$

なお、以降の計算では、単位テキスト中での単語の重要度の二乗和が1になるように正規化した値 $S'(t,w)$ を使用する[2]。

単位テキスト集合 T における単語の重要度 $S'(T,w)$ は、以下の式で示すように、T に含まれる単位テキストにおける単語重要度の和を取った値として計算される。

$$S'(T,w) = \sum_{t\in T} S(t,w)$$

重要単語の検索では、この $S'(T,w)$ の値によって単語がソートされて出力される[3]。

4.2. 関連度

単語とテキストの間の関連度は、単語重要度を用いて、以下の式により計算される。

$$R_{wt}(w,t) = R_{tw}(t,w) = S'(t,w)$$

$$R_{ww}(w_1,w_2) = \sum_{t\in T} S'(t,w_1)S'(t,w_2)$$

$$R_{tt}(t_1,t_2) = \sum_{t\in T} S'(t_1,w)S'(t_2,w)$$

ここで、Rwt、Rtw、Rww、Rtt は、それぞれ、単語とテキスト、テキストと単語、単語と単語、テキストとテキストの間の関連度を、T、W は、それぞれ、検索時に設定されている単位テキスト集合、単語集合を表す。

単語間の関連度は、同一単位テキスト内での重要度の積の和として計算される。したがって、共起する回数が多く、かつ重要度が大きい単語間の関連度が大きくなる。

単位テキスト t における単語 w の出現頻度が0の場合、その単語の重要度 $S'(t,w)$ は0になる。したがって、単位テキスト集合 T に含まれる単位テキストにおいて、一度も共起しない単語間の関連度は0になる。

4.3. 単位テキスト集合の設定

ACCENT では、重要度・関連度の計算に利用される単位テキスト集合 T として、特定のパターンの名前を持つ単位テキスト群を指定することが可能になっている。

---

[1] 設定により、情報検索の分野でよく使われている TFIDF、あるいは TF（Term Frequency）、DF（Document Frequency）を重要度の値として用いることも可能である。

[2] 設定により、ノーマライズを行なわないようにすることも可能である。

[3] 設定により、重要度の値とは独立に、TF や DF の値によってランク付けを行なうことも可能である。

単位テキストの名前は、文書群を単位テキストに分割する際に、システムによって付与される。このときに使用される命名規則は分析者が指定可能であり、文書群の中の情報を名前の一部に取り込むこともできる。

したがって、たとえば新聞記事の場合、日付や掲載ページといった記事の属性情報を名前に埋め込んでおくことにより、特定の期間の記事、特定のページの記事（たとえば一面）における重要単語、連想関係を抽出することが可能となる。

テキスト検索の結果として得られる単位テキスト群を、単位テキスト集合として指定することも可能である。たとえば、単語から関連テキスト検索を行なうことによって、特定の話題の単位テキスト集合を選び出し、その中での重要単語、連想関係を抽出する場合などに用いる。

4.4. 単語集合の設定

ACCENT では、単語検索の結果として出力される単語を規定する単語集合 W として、特定の品詞情報を持つ単語群を指定することが可能になっている。

特定の文字列パターンの単語群を、単語集合として指定することも可能である。接尾文字列により特別な品詞を付与する機能は、連想辞書作成時に接尾文字列を指定しておかないといけない、文字列パターンとして接尾文字列しか指定できないといった制限があった。文字列パターンによる単語集合の指定は、こういった制限をなくし、検索実行時にパターンを変更しながら単語を絞り込むことができるようにするための機能である。

単位テキスト集合の場合と同様に、単語集合の指定に関しても、検索結果を利用する方法が用意されている。

5. 分析実験

以下，新聞データを用いた分析実験例を紹介しながら、分析の目的・視点をマップに反映させるための方法およびその効果について報告する。

実験に用いたのは、日経産業新聞 3 年分（1994 年～1996 年）の記事（記事数：194,714）から、人事、訃報記事を除いたもの（記事数：157,756）で、1 記事を 1 単位テキストとする方法と、1 段落を 1 単位テキストとする方法の両方でそれぞれ連想辞書を作成した。

単位テキスト	記事単位	段落単位
単位テキスト数	157,756	930,086
総単語数	686,485	686,485
平均テキスト長	585 文字	99 文字
連想辞書の非零要素数	25,792,628	36,885,378

5.1. 分析実験例1

図 3 は、自動車メーカーの提携関係を分析することを目的とし、記事単位で処理した連想辞書を利用し、以下の手順で作成したマップである。

【step1】 日米欧の自動車メーカー名を「自動車メーカー名」という品詞で登録

【step2】 「提携」「合弁」「出資」「資本参加」「供与」「供給」「共同」「協力」「買収」

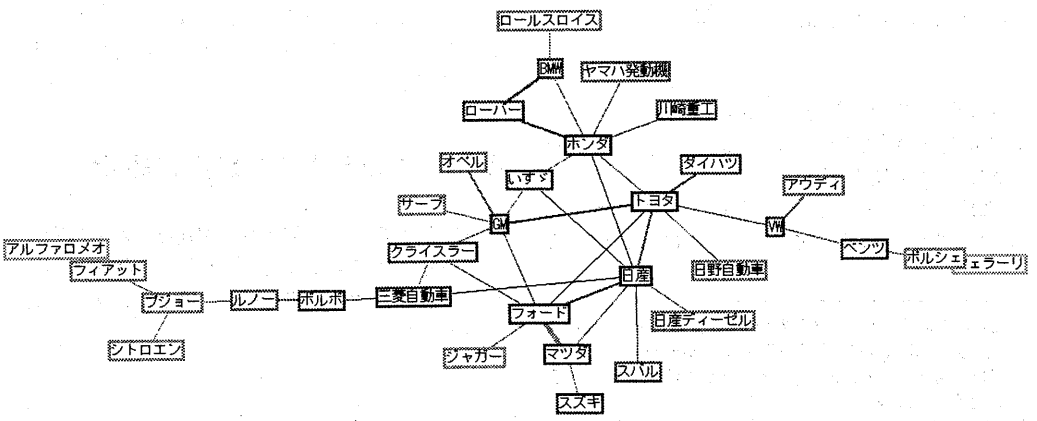

図3　自動車メーカーの提携マップ（記事単位で処理した連想辞書を利用）

を入力として、関連テキストを検索

【step3】 step1 の検索結果を単位テキスト集合として設定

【step4】 単語検索の結果として出力される品詞を「自動車メーカー名」に設定

【step5】 重要単語検索の結果全単語を利用し、重要単語マップを作成

【step6】 辺間引きを実施

【step7】 パターン 11 を利用しレイアウト

step1 および step4 は、自動車メーカー名のマップを作成するために行なった処理であり、品詞を追加・設定する機能を利用している。

step2 および step3 は、提携関係について報じた記事の中での連想関係を抽出するために行なった処理である。まず提携関係を表現するいくつかの単語を入力とし、関連テキスト検索を行なうことによって、提携関係を報じた記事が、検索結果として得られる。この検索結果を単位テキスト集合として設定することにより、マップとして表現される単語間関連度を計算するために利用する単位テキスト群を提携関係に関する記事に限定することができる。

出力結果を見ると、実際に提携関係にある自動車メーカー名の間に強い連想関係の辺が引かれており、連想関係の抽出に利用するテキスト情報を絞り込むことによって、提携関係を分析するという分析の目的・視点を反映したマップを作成できていることが分かる。

一方、直接の提携関係にはない、たとえば米自動車メーカーのビッグスリー「GM」「フォード」「クライスラー」の間にも、連想関係の辺が引かれている。このような連想関係が抽出された原因を、掘り下げ分析を行なうことによって調べた結果、以下のようなことがわかった。

94 年から 95 年の記事から、日米欧の各自動車メーカーによる中国進出の話題が数多く見つかった。この中で、中国最大の自動車メーカーである「上海汽車」の合弁相手として、「トヨタ」「GM」「フォード」が競合していることが何度も報じられている。このことから、直接の提携関係にない自動メーカー名の間に抽出された連想関係の一部は、提携・合弁の競合関係を表しているものであることが分かる。

また、提携関係を報じている記事には、解説記事的なものが多く、提携の報道とともに、他社の動向も合わせて解説されているケースが多いことも分かった。ACCENT では単語の共起により連想関係を抽出しているため、このような記事に表れる、「A 社と X 社が提携」「B 社と Y 社が提携」という記述から、提携関係にない A 社と B 社の連想関係が抽出されることになる。特に、ビッグスリー、あるいはトヨタ・日産・ホンダに関しては、比較して解説されることが多いため、上記のようなパターンで抽出された連想関係が強く働いたいるものと考えられる

5.2. 分析実験例2

図 4 は、より強い連想関係のみを抽出することを意図して、段落単位で処理した連想辞書を利用して作成した提携マップである。

図 4 作成の手順は、基本的に分析実験例 1 の場合と同様であるが、step2 の前で、まず単位テキスト集合を、記事の見出し部分だけに限定してから検索を行なっている。また、step6 の辺間引きは実行していない。

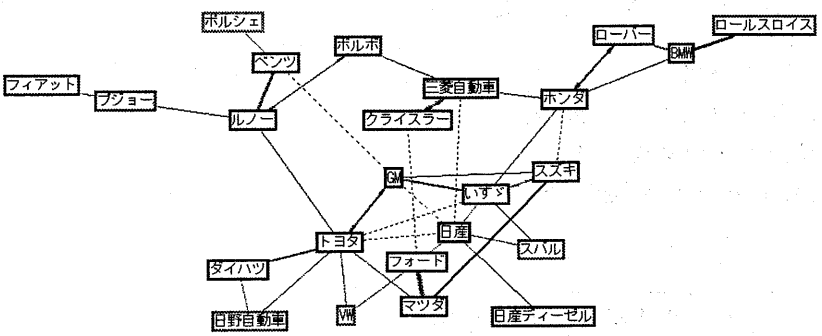

図 4 自動車メーカーの提携マップ（段落単位に処理した連想辞書を利用）

図4に関しては、得られた連想関係の数も少なかったため、辺で結ばれたすべての単語ペアについて掘り下げ分析を行なった。結果は線種の違いとして図に表現してある。

実線で結ばれた自動車メーカーのペアに関しては、「ホンダ」「BMW」のケース[1]を除き、すべて直接的な提携あるいは提携解消を報じる記事が見つかっている。

点線で結ばれている自動車メーカーのペアに関しては、直接的な提携関係は見つかっていない。ただし、分析実験例1のケースとは異なり、いずれも、『トヨタ系アイシン・日産系ユニシアが中国合弁合意』『日産系の土屋製作所、フィルター部品、英トヨタに供給』といったような、部品メーカーを媒介とした系列企業としての提携の記事が見つかっている。

このように、分析実験例2では、連想関係が抽出されるスコープを小さくしたこと、見出し部分のみを利用したことにより、抽出される連想関係の数は減るものの、より実際の提携関係を反映したマップを作成することができている。

6. まとめと今後の課題

以上、連想検索の機能を使いながら、連想関係が抽出されるスコープをコントロールすることにより、分析の目的・視点を反映した結果（単語マップ）が容易に得られることを、実例を通して示した。

なお、3.3 節で説明した、単語の接尾文字列パターンにより分析目的に合った単語を選択する方法を利用した実験例が文献[6]にまとめてあるので、そちらも合わせて参照されたい。

今後の課題に関しては、以下の点が重要であると考えている。

- 評価方法の確立
- より深い分析のための技術開発
- 比較分析のための技術開発
- 時系列分析のための技術開発

[1] 「ホンダ」と「BMW」の間の辺は、「ホンダ」と提携関係にあった「ローバー」を「BMW」が買収したことを報じる記事から抽出されたものであり、分析実験例1における「トヨタ」「GM」「フォード」のケースと同様に競合関係を表している。

参考文献

[1] 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).

[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 (1999).

[3] 渡部 勇: 発想支援システム「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] 渡部 勇: Keyword Associator による情報の構造化支援－要求会議分析への応用－, 人工知能学会 第7回AIシンポジウム資料 (1996).

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