データベースシステム　124－18
情　報　学　基　礎　62－18
（２００１．５．２２）

テキストマイニングにおける概念関係視覚化方式の検討

城塚　音也

株式会社ＮＴＴデータ　開発本部

大量の自由記述されたテキストデータからの知識発見を支援するテキストマイニングでは、テキストデータの分析結果を視覚的に提示することが重要となる。本論文では、我々が試作したテキストマイニングシステムに実装した、複数の概念関係の視覚化方式を組み合わせたトップダウンの知識発見アプローチを実現する概念関係視覚化方式について報告する。実際の自由記述型のアンケートデータを使用し、本概念関係視覚化方式の、注目すべき概念の発見や重要な概念関係への絞込みといった作業における有効性を確認した。また、概念関係抽出性能の検証を行った結果、従来手法に比べて、高い精度で概念関係を抽出できることを確認した

A Study on Visualization Method of Concept Relationship for Text Mining

Otoya SHIROTSUKA

Research and Development Headquarters
NTT DATA CORPORATION

Visualization of relationship between concepts is important in text mining as the process of knowledge discovery from large amount of unstructured text data. In this paper, we proposed an unique visualization method of concept relationship in top-down knowledge discovery approach that is implemented in a system prototyped by authours.

Text data analysis is performed to real Q&A data in free form, which is retrieved from information providing service in a commercial web site and we found that our visualization method helps analyst's work to select concepts to be focused and to organize concepts relationship for reporting. Experimental results show that the prototyped system can extract concepts relationship with higher precision compared to conventional method.

1 はじめに

インターネット、イントラネットの普及により、電子情報としてのテキストデータが大量にネットワーク上を流通し、企業等の組織内に蓄積されるようになってきた。消費者と企業とのコミュニケーションチャネルも、郵便や電話に加えて、電子メール、Web が活用されてきており、コールセンタに蓄積された消費者からの質問、クレームメールや、インターネットマーケティングの結果得られた消費者のコメント等のテキストデータを分析し活用することが求められている。テキスト処理技術へのニーズも、従来のテキスト検索では対応が困難な、分析的なアプローチを可能とする技術が求められており、従来のテキスト処理技術とデータマイニングに用いられる各種データ分析技術を組み合わせたテキストマイニングが脚光を浴びている。その結果、近年、テキストマイニングに関する研究が盛んに行われており[1][2]、製品化も行われている[3][4][5]。しかし、データマイニングツールと同様、知識発見に至る作業の流れをサポートする機能を備えていないこと、言語情報を分析しているにもかかわらず、分析結果が分かりにくい等の問題から、現状ではユーザのニーズを充分に満たしているとはいえない。

本論文は、トップダウンの知識発見アプローチを実現する概念関係視覚化方式を提案する。本アプローチで使用する概念関係視覚化は、俯瞰的な概念関係、注目すべき概念を絞り込んだ視覚化、絞り込んだ概念の文脈情報を提示する視覚化および注目した概念同士の比較点を明確にした視覚化の、四段階の視覚化手順によって実現される。

まず、2章で試作したテキストマイニングシステム KnowledgeOcean について述べ、3章において、KnowledgeOcean で実現される概念関係視覚化方式と、その実装方式を説明する。4章では、自由記述型のアンケートデータを使用して、概念、概念共起、概念意味関係のそれぞれを使用した分析結果について比較評価する。5章では、本論文のまとめと、今後の課題を述べる。

2 テキストマイニングシステム KnowledgeOcean

KnowledgeOcean は、我々が試作した Web ベースのクライアントサーバー型テキストマイニングシステムであり各種分析処理を NT サーバー上で行い、Web ブラウザを通じて、複数のユーザが、それぞれ分析用辞書および分析対象データを用いてマイニングを行うことができる。

本システムの特徴は、マイニング作業のインタラクティブ性が良好である点にある。インタラクティブ性を高めるために、分析、視覚化処理の高速化を行っている。また、分析目的ごとに複数の視覚化機能を提供することにより、分析結果を多種類の表現でフィードバックしている。

2.1 テキストデータからの概念抽出

テキストデータから抽出する特徴単位としては、単語、名詞句、動詞句および「何が＋どうした」といったメタデータを用いている。本論文ではこれらの抽出単位を総称して「概念」と呼ぶこととする。

単語、名詞句の抽出には、形態素解析を用いる方法と、単語辞書および文字種類の変化点情報を用いたロバストな方法の二種類を用いており、分析対象データの性質やドメイン辞書のチューニング状況に応じて使い分けることが出来る。動詞句の抽出では、記述意図を区別するため、「疑問（～か？）」「要望（～たい）」「否定（～ない）」「可能（～できる）」の四タイプを区別して扱うようになっている。

概念の抽出には、係り受け解析に基づく方式が提案されているが[6]、本システムではユーザによるカスタマイズ性、処理の高速性を重視し、情報抽出におけるパタンマッチング手法を用いている。ドメインの性質、分析目的に応じて、必要な概念のタイプが異なってくる問題に対しては、抽出する概念の種類を事前に指定しておくことによって対応している。

2.2 分析手法

KnowledgeOcean は概念の共起分析と 抽出された概念に基づいたクラスター分析を行うことが出来る。共起分析は文書内で規定された距離内に現れる２つの概念の関係を共起関係とみなし、多数の共起関係を総合的に捉えることによって行う。

クラスター分析には一般的な K-means 法に基づいたトップダウンのクラスタリング手法を用いており、分析のインタラクティブ性を損なわないように、次元数の圧縮による高速化の工夫を行っている。

3 情報視覚化方式

上述した共起分析結果を視覚化する方式として我々は以下のトップダウン型視覚方式を提案する。

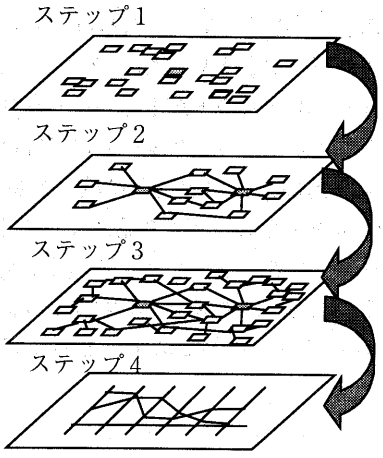

図 1. トップダウン型視覚化方式

トップダウン型視覚化方式は、図１に示す四つの視覚化ステップに分かれている。

ステップ１では、分析対象データ中で、どのような「モノ」、「コト」が述べられているかどうかを把握するための支援として、多量の概念間の関係を俯瞰表示する。ステップ２では特定の「モノ」、「コト」に注目して、それについて「どうである」と述べられているかを把握するための支援として、より表示対象を絞った形で概念間の関係を視覚化する。ステップ３では、特定の「モノ」、「コト」が「どうである」という記述がどのような文脈で述べられているかを把握するための支援として、ステップ２の内容に二次共起情報を加えた表示を行う。

ステップ４では、ステップまでで把握した定性的傾向について定量的に把握するための統計表示を行う。各ステップ間の移動は、マウスによる概念アイコンの選択とマウスメニューによりスムーズに行うことが出来る。各ステップの詳細について以下に説明する。

3.1 ステップ１（俯瞰）

図２にステップ１での俯瞰表示の例を示す。二次元上に配置された概念は、共起度が強いほど距離的に近くになるように配置されている。一般にお互いに関係度のような相関のある多数のデータを二次元上に配置するアルゴリズムは多数提案されているが[7][8]、一般に精度と計算速度がトレードオフの関係になっている。マイニング作業では分析のインタラクティブ性を重視するため、データ配置の繰り返しによる配置最適化を行わない、速度重視のアルゴリズムを採用している。

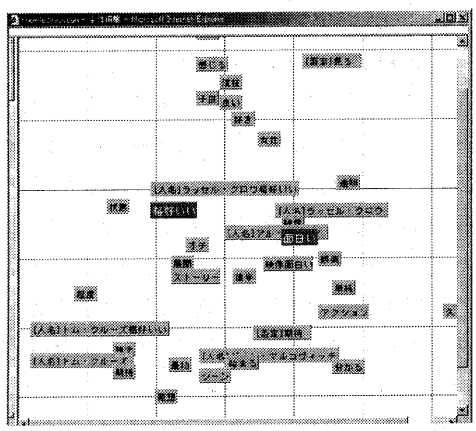

図 2. 俯瞰表示

図 2 では映画の評論データ 500 件を対象に、概念を抽出し、「格好いい」「面白い」という二つの概念と関連のある概念を表示している。たとえば図中の「トム・クルーズ格好いい」という概念は、「トム・クルーズは、やっぱり格好いいですね」といったコメントから抽出されたメタデータである。

3.2 ステップ 2（注目）

図 3 にステップ 2 における注目した事象についての記述内容把握のための視覚化の例を示す。例では、ステップ 1 で着目した 3 人の俳優「アル・パチーノ」「トム・クルーズ」および「ラッセル・クロウ」を核にして、共起する概念が表示されている。各概念を結ぶ線の太さはステップ 1 における概念間距離と同様に概念間の共起度の強さを表している。このようなコンセプトをアイコンで示し、その関係を線で結ぶ視覚化方式は非常に分かりやすく、色々なテキストマイニングシステムで使用されている[9][10]。

ステップ 2 の視覚化方法では、二つ以上の核概念と共起する概念は他の概念と区別できるように表示色を変えて中心に配置される。一方、ひとつの核概念のみとしか共起しない概念は、その核概念独特の性質をあらわす概念として、その核概念の周りに配置される。

この視覚化は複数の事象を比較する際に、事象共通の性質と各事象固有の性質を明確に視覚化することが出来るところに特徴がある。

3.3 ステップ 3（文脈参照）

図 4 にステップ 3 における視覚化例を示す。ステップ 3 では、ステップ 2 の視覚化結果中の核概念と共起する子概念に対して共起するような概念（孫概念）を追加した表示を行う。たとえば、図 4 では子概念「マキシマス」に対して「歓声を上げたくなる」（表示は「歓声[要求]上げる」）という孫概念が表示されており、子概念の文脈情報となっている。

3.4 ステップ 4（比較）

図 5 にステップ 4 における視覚化例を示す。ステップ 4 では、核概念と比較対象の子概念を指定することにより、定量的な比較グラフの自動作成が可能である。図 5 では 3. 2 節で説明した 3 つの核概念に対して共起する 5 つの子概念「格好いい」「頑張る」「楽しめる」「面白い」「いつも通り」の共起回数の比較が行える形でグラフ化されている。

核概念と子概念の共起関係は、同一の意味的な関係を表しておらず、たとえば、「トム・クルーズ」と「格好いい」の共起は「トム・クルーズが格好いい」という記述を連想させるが、「トム・クルーズの車が格好いい」等の場合も考えられる。「トム・クルーズが格好いい」という意味的関係を示す概念に絞って抽出を行うことも考えられるが、文脈によって「トム・クルーズ」が省略されたり、照応語で表されたりする場合も多々あるため、すべての記述をカバー出来ているわけではないという問題がある。

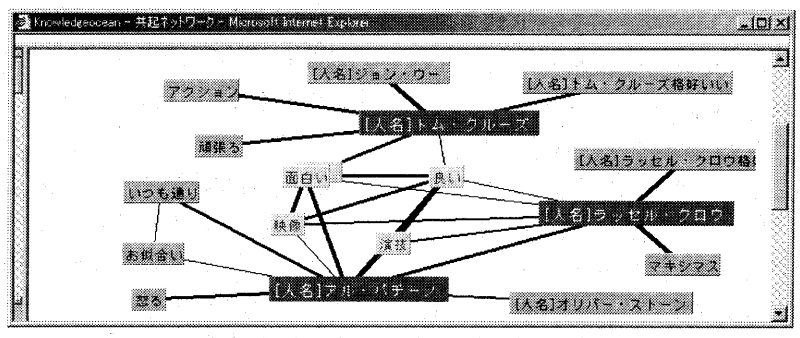

図 3. 概念共起関係表示（一次共起）

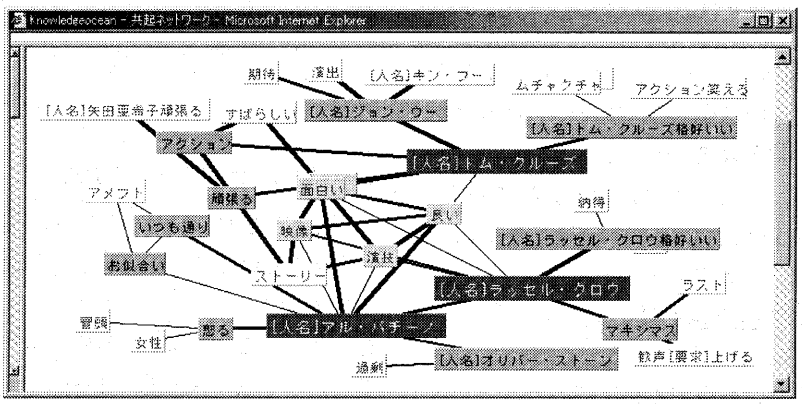

図 4. 概念共起関係表示（二次共起）

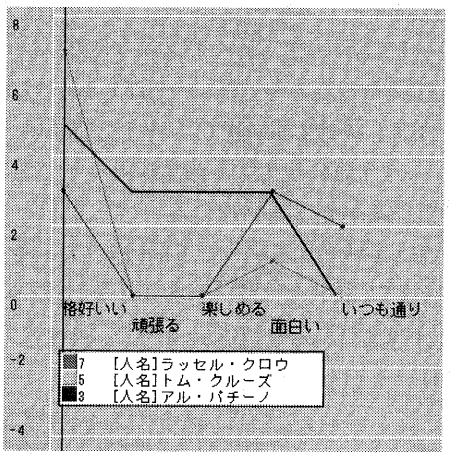

図 5.共起回数比較表示

4 性能評価実験

3 章で説明した概念関係方式を実際のアンケートデータの分析に適用し、知識発見を行った。評価に使用したデータについて 4.1 節で説明し、4.2 節において、本視覚化方式を使用したことによって得られた知見について考察する。

また、3.4 節で指摘した、核概念と子概念の共起関係が同一の意味関係を表していない問題を定量的に把握することを目的に、概念共起と情報抽出によるメタデータによる分析が、実際の分析対象データの意味内容をどの程度正確に抽出できるか評価実験を行った。4.3 節において評価実験の内容について説明する。

4.1 評価データ

評価データとして用いたものは、ヘルスケア関連の情報提供 Web サイト上での、2 種類の健康食品に対するアンケート結果の自由記述コメント 6468 件である。内訳は商品 X に対するコメントは 5012 件、商品 Y に対するコメントは 1456 件であり、一件あたりのコメントの分量は 1 文～3 文程度の記述がほとんどである。

4.2 知識発見の支援性能

4.1 節の評価データを使用して知識発見作業を行った。作業は商品 X、Y それぞれのアンケートデータに対する分析および、それぞれのアンケートデータを合わせた比較分析である。作業を通じて提案する概念視覚化方式に以下の特徴があることが分かった。

トータルとしての作業時間が短くなる

ドメイン辞書を整備し、重要な概念を洗い出して、興味深い概念同士の関係を選択し、定量的な把握を行い、レポートとしてまとめ上げる一般的な分析作業は、時間がかかる仕事であり、分析対象データが多くなればなるほど、作業が困難になる。本視覚化方式を使用した結果、レポート作成までの作業は、通常半日程度かかる分析が 1 ～2 時間程度で実行できた。

重要な概念への絞込みが簡単

個別の作業について述べると、商品 X、Y のうち、分析者が事前に商品知識をもっていなかった商品 Y の分析では 3.1 節で述べた俯瞰機能が有効であった。これは、俯瞰機能が、ドメイン全体の概念関係を全体的に把握するのに有効であるためと考えられる。また、3.2 節のの注目機能において、なぜ、核概念と関係があるか不明な子概念に対して文脈情報を得ることが出来る 3.3 節の文脈参照機能も非常に役に立つことがわかった。

逆にある程度事前に、分析者が事前に商品知識のあった商品 X に関しては、3.2 節の注目機能が有効であった。

複数の商品の比較が迅速に行える

商品 X、Y 両方のアンケートデータを合わせて分析する場合は、3.2 節の注目機能と 3.4 節の比較機能が有効であった。特に 3.2 節における複数の核概念と関係がある子概念と、単独の核概念とのみ関係がある子概念を区別した視覚化は、核概念に比較対象の商品関連の概念を選んだとき非常に有効であった。

4.3 概念抽出性能の評価実験

商品 X、Y について、それぞれ「おいしい」、「好き」という意見記述が、何件あるかという正解データと、下表の 4 手法による分析結果とを比較することにより、各分析手法の精度を調査した。正解データは、分析手法 A の抽出結果を人手で確認することによって作成した。そのため、「おいしい」「好き」という表層的表現を含まないが同義であるような文については正解としていない。

分析手法 A は「おいしい」「好き」といった述語的概念をキーにした分析である。分析手法 B は、概念の共起分析である分析手法 C から、共起判断に使用する距離制限を除いた分析手法と捉えることが出来る。

各分析を行う前には、商品 X および商品 Y に特有の用語（商品名等）の形態素辞書への登録、同義語の定義（「おいしい」「美味い」等）を行っている。

表1．実験条件

分析手法	条件
A	商品A、商品Bそれぞれ、「おいしい」または「好き」という概念を含む記述を取り出す。
B	商品名と「おいしい」、商品名と「好き」の二つの概念を含む記述を取り出す。
C	商品名と「おいしい」、商品名と「好き」の共起関係を含む記述を取り出す。
D	「商品Aがおいしい」「商品Bが好き」というメタデータを含む記述を取り出す。

4.4 共起出現の定義

本システムでは、概念aと概念bの共起を次のように定義している。

共起出現：テキストが概念抽出により概念列に変換された状態で、概念aの前後n個の概念以内に概念bが存在するとき、「概念aと概念bは共起している」と定義する。

本実験においては近傍とする概念距離 n=3 とした。

4.5 メタデータの抽出

本実験に用いたメタデータの抽出は、形態素解析の出力結果に対して、あらかじめ作成したメタデータ抽出用のパタンマッチングルールを適用することによって実現される。形態素解析には奈良先端科学技術大学院大学の「茶筌」を使用している。

4.6 評価結果

表2に評価結果を示す。評価尺度である抽出精度と抽出率は以下の式により計算される。

抽出精度：

抽出された正解概念数／抽出概念数・100（%）

抽出率：

抽出された正解概念数／正解概念数・100（%）

各分析手法を比較すると、手法Aでは平均50%程度の抽出精度であるのに対して、手法Bでは平均86%、手法Cでは平均89%の精度が得られている。逆に抽出率については、手法B、手法C、手法Dの順に低下しており、抽出精度の向上の代わりに、抽出率が低下することが分かる。

「商品Xがおいしい」に対する手法Aの抽出精度が他に比べて大幅に低い原因は、アンケート対象者に商品を食べたことがない人間が多く、「おいしければ買いたい」という内容のデータが多かったことにある。これは、「おいしい」という概念に「おいしい」の仮定形を含めない等の工夫による対処が必要となる。

表2．評価結果

「商品X（Y）がおいしい」

分析手法	商品	正解	誤り	抽出精度	抽出率
A	X	54	95	36%	100%
	Y	31	24	56%	100%
B	X	15	1	94%	28%
	Y	24	5	77%	83%
C	X	9	1	90%	17%
	Y	18	3	86%	58%
D	X	4	0	100%	7%
	Y	10	0	100%	32%

「商品X（Y）が好き」

分析手法	商品	正解	誤り	抽出精度	抽出率
A	X	33	45	42%	100%
	Y	38	21	64%	100%
B	X	14	3	82%	42%
	Y	29	3	91%	76%
C	X	8	1	89%	24%
	Y	17	2	89%	45%
D	X	4	0	100%	12%
	Y	12	0	100%	32%

4.7 考察

実験結果から、3.4節で説明した定量的評価に手法Bや手法Cの結果を用いることは、抽出率が悪いことから、分析対象データの内容傾向を定量的に正しく反映できているとはいえず、概念抽出性能の向上が必要であることが判明した。そのためには、パタンマッチングルールの精緻化や統語解析の導入による手法Dの抽出率の向上が必要と考える。

5 おわりに

本論文では、テキストマイニングにおける知識発見作業支援のための分析結果視覚化方式として、トップダウン型の概念関係視覚化方式を検討した。本手法を用いることにより、ユーザはインタラクティブな視覚化を通じて、漸進的な知識発見に対する支援を受けることが出来る。

商用の情報提供ウェブサイト上で収集された、健康食品に関するアンケートデータを分析した結果、本視覚化方式がユーザの知識発見に有益であることを確認した。また、複数の概念抽出方式を比較した結果、共起や文法的関係に基づく概念関係抽出は、テキストデータ内の意味内容を精度よく抽出することができるが、定量的な分析を行うために、漏れなく概念関係を抽出するためには抽出性能が不足していることを確認した。

今後の課題は、現在対応していない、概念の属性情報を含めた概念関係の視覚化方式および、高い抽出精度および抽出率をもった概念抽出機能の実現である。

参考文献

[1] M.A Hearst: Untangling Text Data Mining, Proceedings of ACL '99 the 37th Annual Meeting of the Association for Computational Linguistics, 1999

[2] 長野徹, 武田浩一, 那須川哲哉: テキストマイニングのための情報抽出, ディジタル図書館 No. 18, pp79-86, 2000.

[3] コマツソフト :Vext Search, http://www.komatsusoft.co.jp/develp/vxtsc/index.html

[4] 東芝 :KnowledgeMeister, http://www.toshiba.co.jp/product/cn/filling/kmeister/

[5] 富士通 :Symphoware Text Mining Server, http://software.fujitsu.com/jp/symfoware/products/tmining-v4/index.html

[6] 松澤裕史: テキストデータからの頻出パターンのマイニング, 電子情報通信学会言語理解とコミュニケーション研究会・対話システム研究会 共催, 「知識発見のための自然言語処理」シンポジウム 1999.

[7] T. Kohonen, S. Kaski, K. Lagus et al.: Self organization of a massive document collection, IEEE Transactions on Neural Networks, Vol. 11, No.3, pp574-585, 2000.

[8] Aurigin Corporation: Themescape , http://www.aurigin.com/aureka.html#themescape

[9] 渡辺勇他:単語の連想関係によるテキストマイニング, 情報処理学会研究会報告 FI-55-8, DD-19-8, pp.57-64, 1999.

[10] Semio Corporation: Semio Map, http://www.semio.com/products/semiomap.html