グループウェアと
ネットワークサービス 46－8
（２００３．１．１６）

新聞の知識を利用したテキストマイニング支援システムの提案

井前 吾郎* 奈須 庄健* 重野 寛* 岡田 謙一* 松下 温†

本論文では，テキスト情報として大量で，かつ一般的な情報源として重要度の高い新聞の記事データを対象としてテキストマイニングを行い，得られた結果から有用な情報を発見しやすくするための視覚化手法について提案する．本システムでは企業を魚の群れの中心とし，企業と関連のある単語を，群れを構成している魚として表現している．本システムの視覚化により，テキストマイニングの結果を時間軸に沿った情報として表現するとともに，単語のクラスタリングも表現し，複雑なテキストマイニングの解析結果を理解することが容易となった．

The proposal using newspaper of a text mining support system

Goro Inomae* Shoken Nasu* Hiroshi Shigeno* Kenichi Okada* Yutaka Matsushita†

In this paper, we proposed about the visualization technique for making useful information easy to perform text mining and to discover from the obtained result. We minig newspapers which are general sources of information in large quantities as text information and important. As a fish which constitutes the group for the word which sets a company as the center of the group of a fish. This system expressing the result of text mining as information in alignment with the time-axis, and became easy to also express clustering of a word and to understand the analysis result of complicated text mining by visualization.

1 はじめに

近年，過去に蓄積されたノウハウを体系化し，知識を共有活用するナレッジマネージメントなどが注目されるとともに，オフィスの OA 化による電子テキストの増加，ネットワークの普及によるデータの流通と収集の促進，そしてハードウェアの高性能化と低価格により電子化された文書を大量に保有することが可能となった．そして，それらの大量データの中から，新たな知識を発見するマイニングとよばれる技術が注目されるようになった [1]．一般に，構造化された数値データを対象にする場合はデータマイニングと呼ばれるが，自然言語のような非構造データをも対象にする場合はテキストマイニングと呼んで区別している [2]．

しかし，マイニングの結果として得られる情報は複雑であり解析結果を理解することは我々ユーザにとって大きな負担となる．そのため，マイニング手法だけでなくマイニング結果の視覚化手法についても研究が進められている [3]．

そこで本論文では，テキスト情報として大量で，かつ重要度の高い新聞の記事データを対象としてテキストマイニングを行い，得られた結果から有用な情報を発見しやすくするための視覚化手法を実装した．なお，分析対象の新聞記事については日経四紙 (日本経済新聞, 日経産業新聞，日経流通新聞ＭＪ, 日経金融新聞) の 1 年分 (2000 年) のデータを使用した．

以下，第 2 章では，本研究で用いたテキストマイニングの手法について説明を行い，第 3 章ではテキストマイニングの結果の視覚化手法について説明をする．第 4 章では本研究のシステムついて説明を行い，第 5 章でシステムの評価と考察したことについて述べる．最後に，第 6 章でまとめと今後の課題について述べる．

* 慶應義塾大学大学院 理工学研究科
Faculty fo Science and Technology, Keio University

† 東京工科大学
Tokyo University of Technology

2　新聞記事のテキストマイニング

2.1　新聞記事の特徴分析

我々はまず, 新聞記事には他のテキストデータ (web ページやアンケートデータなど) にはないいくつかの特徴があり，それらの特徴を利用すれば記事データのマイニングの精度が向上するのではないかと考えた.

新聞は毎日発行されおり，記事の内容は社会を反映したものとなっている．つまり, 新聞記事は時系列に沿ったテキストデータであり，1 月 1 日の記事と，12 月 31 日の記事は同じまとまりとしてテキストマイニングを行うべきではないといえる．このことから，新聞記事をマイニングするにあたり対象データを全体でまとめて取り扱うのではなく，ある長さの時間で区切って時間単位で分析したほうがよいと我々は考えた.

また，新聞は印刷物であるので紙面の大きさ，枚数などの急な変更は困難である.

そのため，購読者に伝えたい情報を伝えるためには簡潔に表現しなければならないといえる．加えて，購買意欲を高めるために，多くの記事ではタイトルにその記事の最も重要なキーワードが記されており，一方，タイトルに含まれていないキーワードはその新聞記事においては，タイトルに含まれているキーワードと比較して重要度が低いということがいえる．さらに，紙面が限られているにもかかわらず，多くの面積を占めている記事も重要であるということがいえる.

そこで本研究では，新聞記事のテキストマイニングを行う際にキーワードがタイトルに含まれているのか否かで記事を絞り込むこととした．そして，1 年間の新聞記事データをある時間単位ごとに分割してテキストマイニングを行い，また，文字数については視覚化の部分で表現することとした.

なお，時間の単位についてであるが，分析対象とした新聞記事が日経四紙であることから，時間的に意味を持つ単位は日曜日から土曜日までの 1 週間であると我々は考え，本研究では 1 週間単位で取り扱うこことした.

2.2　関連技術

現在，テキストマイニングでよく用いられる手法は単語間, 文書間の関連性を算出する手法や [4], 文書をクラスタリングすることで文書の検索精度を向上させる手法 [5]，そして，単語間の相関ルールを用いた手法である [6]．これらの手法は，大量のテキストデータ全体の分析を行い色々な特徴や知識を発見する．つまり，例えば新聞記事については，解析対象すべてをまとめてマイニングを行っている.

本研究ではこれらの手法とは異なり，ある時点での新聞記事のデータマイニングの結果が次の時点でどのように変化したのかを示すことを特徴としている．そのため，得られるマイニングの結果は 1 年間のトータルとしての結果ではなく，時系列に沿った 1 年間の変化の様子についての新たな知識である．例えば単語の出現頻度を算出した場合では，あるキーワードは年の前半のほうに多く出現していたが，年の後半のほうでは出現しなくなっていたといった情報が表現できるということである．我々はこのように，1 年間全体のマイニングの結果を示しただけでは発見できなかった別の知識を発見することができるのではないかと考えた.

2.3　単語の重みの算出

本研究では，新聞記事に対し，TFIDF 法を用いることで各新聞記事における単語の出現頻度を算出し，単語と新聞記事との関連度をもとめることにした.

TFIDF 法はある文書を特徴付けるようなキーワードとして

1. ある文書に高い頻度で現れる
2. 少ない数の文書にしか現れない

という 2 通りのキーワードが存在するという考え方をしており，1 は tf (term frequency) と呼ばれ，次式で表される.

$$tf{=}W_i/W$$

ここで，W は 1 つの文書に含まれている全キーワード数，W_i は 1 つの文書に含まれているあるキーワードの数を表している．つまり，tf は文書に含まれるキーワードの出現回数を算出している．また，2 は idf (inverse document frequency) と呼ばれ，次式で表される.

$$idf = log(N/n) + 1$$

ここで，N は全文書数，n はキーワードが含まれる文書の数を表している．また，log は稀にしか出現しない単語の「重み」を重くし，頻繁に出現する単語の重みを少なくする役割を担っており，また，キー

ワードが全文書に含まれる場合 (i.e. $N = n$) は $idf = 0$ となるので，これを避けるため,「$+1$」している．そして，TFIDF 法は tf および idf で求めた値をもとに次式を用いて単語の出現頻度を算出する．

$$\text{重み} = tf * idf$$

つまり，TFIDF 法は単語の重要度に関する 2 つの異なる考えを組み合わせて，単語の重要度をバランスよく算出する方法である．

ここで TFIDF 法の問題として，対象の文書が例えば法律の条文やサポートセンターの質問事項などのように 1 つの文書の文字数が少ない場合，TF 法の結果がどれも同じになってしまうということがある．そのため，IDF 法の単語の出現文書数だけで単語重要度を決めることになり，TFIDF 法では精密な出現頻度の算出は期待できないということがある．しかし，新聞記事のように長い文書ならば重要な単語が繰り返し出現するため TF 法の仮定が有効となるといえる．また，TFIDF 法は情報検索の分野で最もよく用いられていることから，本研究では使用することにした．

3 視覚化

テキストマイニングで利用されている視覚化の関連技術について述べ，本研究の視覚化の手法の概要について述べる．

3.1 関連技術

情報の視覚化は効果的に使用すれば，大規模で複雑な情報を効率的に人間に伝達することができるので，テキストマイニングでも結果を表示するにあたり重視されている [7]．現在主流となっているテキストマイニングの結果を表示する視覚化手法は，キーワードなどの出現頻度を棒グラフや折れ線グラフで表示することで，比較や時間的な推移を把握しやすくするものや，単語間の連想関係を抽出し視覚化するといった手法がある [8]．

また，情報視覚化の分野では，3 次元を用いて視覚化をすることも主流になりつつある．3 次元空間を利用することで表示できる情報量を多くすることができるが，一方で 3 次元で視覚化した場合にはいくつかの問題が起きることも指摘されている [9]．例えば，コンピュータ画面は 2 次元であるから，3 次元の構造をひとつの図だけで完全に理解するのは不可能である．そのために，3 次元空間内で視点の移動を行わなければならないが，大部分のコンピュータに採用されているマウスのような 2 次元的コントロールでは 3 次元空間の操作を行うにはある程度の熟練度が必要となる．そのため，3 次元空間に慣れていないユーザには 2 次元空間より理解するのが困難となり，3 次元空間内で自分の位置を把握できなくなる場合がある．

3.2 動きを用いた視覚化

本研究では折れ線グラフなどで表示すると表示できない情報があることと，ノードを固定してしまうと, 単語と文書との関係が理解しづらくなること，また，新聞記事が時系列データであることを考慮して，動きのある視覚化を用いることにした．そして，我々は動きのある視覚化を行うことで以下のような効果をもたらすことができると考えた．

- 推移の連続性を示すことができる．つまり，あるオブジェクトに 2 つ以上の状態がある場合，その推移がアニメーションになっていれば，静止画よりも状態間の変化が理解しやすくなり，各オブジェクト間の対応関係を直感的に理解できる．
- 複合的な表示ができる．アニメーションは複数の情報オブジェクトを同じ場所に表示するのに利用できる．
- 時間による変化を図解できる．アニメーションは経時変化の表示だから，時間と共に変化する現象に一対一の対応をつけることができる．
- グラフィック的な表示を豊かにできる．ある種の情報は，静止画を使うより，動きを与えた方が視覚化しやすい場合がある．

また，3 次元空間を用いた場合逆に理解しづらくなる場合があることから本研究では 2 次元空間を用いた．

3.3 ノードの表示

動きを用いた視覚化をするにあたり，我々は表示を複雑にしてしまうと理解しづらくなりユーザの負担が増加してしまうということと，四角い箱といった無機質なもので描画しても動きを持つことに違和

感を感じてしまうことから，結果として直感的な把握ができなくなると考えた．そこで，身近なものでかつノードが動きを持っていたとしても違和感がないようにするために，魚の群れを用いて描画することにした．ここで，魚の群れを用いて描画することにより以下のようなメリットも考えられる．

- 群れで行動するのでノードのクラスタリングを表示しやすい
- 子供を産むのでノードの分割, つまり同じ属性でもノードは異なっていることが表現できる
- 成長するので時間による重みの変化を表現できる

また，色を利用することによりユーザは情報認識がしやすくなることから，キーワードごとに色分けすることにした．

4　システムの説明

4.1　インターフェース

本システムのインタフェースについて説明する (図 1).

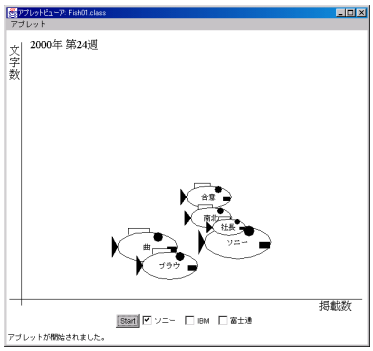

図 1: インターフェース

まず，企業を群れのリーダーにし，企業と関係している単語を小魚として出現させ時系列に沿って動きを変化させた．ここで，時系列で表示するので 1 度にすべての単語，つまり小魚を表示する必要はない．小魚の体には単語が記されており，小魚の大きさは単語の重みに比例して，重みの重い単語ほど大きさが大きくなるように表示した．また，同じ単語でも関係している企業が複数ある場合は，小魚を複数表示するようにした．これにより，ある単語を 1 つのノードで表現した場合，複数の企業と関連があったときに, ノードの重みの内訳が分からないという問題が解決できると考えた．また，図 1 のように群れとして表示することで，企業による単語のクラスタリングも表現するとともに，群れごとに色分けすることでたとえ同じ単語でもどの企業と関係している単語であるのかを区別できるようにした．そして，x 軸は掲載数，y 軸は文字数であり，例えば x 座標が大きく y 座標が小さい場合は 1 つ 1 つの記事の文字数がさほど多くはないということが理解できる．

このように，ある企業について，どの程度の記事が掲載されて，また，それらの記事にどのような単語が含まれているのかといったことや，それらの単語がどの企業と関係を持っているのかが視覚的にしめされているので，ユーザは直感的に企業と新聞記事とそして単語の重みの関係を知ることができる．

4.2　システムの操作

次に本システムの操作方法を説明する (図 2). 本システムはマイニング結果を時系列に沿って表示することから, 現在表示されている画面が第何週 (又は第何 Term 目) であるのかを左上に表示し, 企業を選択するためのチェックボックスと視覚化を開始するボタンのみを配置した. ここで、操作部を単純にした理由は, なるべく, マイニング結果の視覚化に多くの画面領域を使用したかったからである.

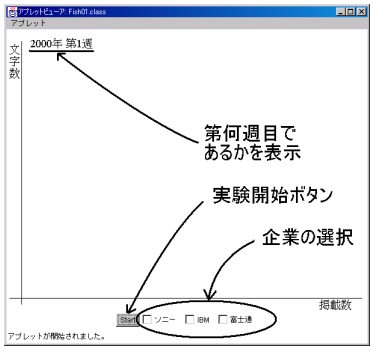

図 2: システムの操作部

ユーザはマイニング結果の見たい企業名を選択する．企業は複数選択可能である．選択後「Start」ボタンを押すことで，マイニング結果としてあらかじめ求められている単語の重みをもとに，企業と単語の関係の深さ (実際にはどれだけその単語の重みが重いのか) が時系列に沿って示されていく．

4.3 システムの流れ

本システムでは 1 年分の新聞記事データを 1 週間単位で分割してテキストマイニングを行った．そこで，視覚化としてはマイニング結果を時系列に沿って表示した（図 3，図 4）．ここで，図 4 は図 3 の次の週のマイニング結果である．

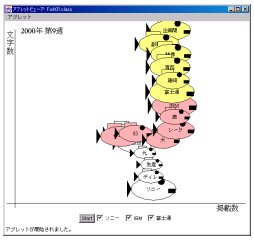

図 3: 第 9 週目

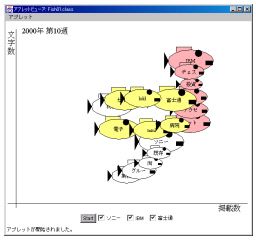

図 4: 第 10 週目

図 3 および図 4 のように時系列で表示することで，新聞記事の掲載数や文字数の変化の様子や，各週における単語の重みの変化の様子が理解できる．例えば，図 3 から図 4 に変化したことで，SONY は文字数および掲載数が増加しており，また，重みの重い単語を見ることでどのようなことについて注目をされたのかが理解できる．このように，1 年間全体でのマイニング結果のみからでは発見できなかった知識が新たに発見できる．

5 システムの評価

5.1 被験者の回答

本研究の有用性を示すために，新聞記事データを 1 年間分まとめてテキストマイニングをした結果から発見できた情報や知識と，本システムを用いて 1 年の間においてどのように変化していったのかを表示することで発見した情報や知識について，各々比較することとした．

具体的には，3 つの企業について，既存の手法のように 1 年間分をまとめて取り扱ったマイニング結果について参照してもらい，次に本システムを使用した後に，システムの感想を述べてもらった．なお，表示した情報は，3 つの企業の新聞記事の掲載数，文字数，および企業と関連している単語とその重みである．

システムについて述べてもらったことを以下に示す．

1. 全体概要の把握はしやすかった
2. 解析過程で全体の結果とは異なる結果があることが理解できた
3. 魚が重なってしまい見にくくなることがあった
4. 企業間の比較と同時に，同じ企業に属している単語間の比較や，別な企業に属している単語間の関係が把握しやすかった

5.2 本システムに関する考察

(1) より，本研究の視覚化手法は，他の研究と同様に全体概要の把握ができることが示された．ただし，他の研究における全体概要の把握とは，対象としているデータの傾向や単語の相関といったものであり，本研究の全体概要とは対象データが時系列データであることから，時間的な推移を示している．

また，(2)(4) より，全体のマイニング結果を示すのみでは気づかなかった知識が発見できていることが示されているといえる．つまり，全体のマイニングでは一つの企業に関する複数の単語の重みの時間推移や，ある単語についての複数の企業の推移といった表示になってしまうが，本研究のマイニング手法と視覚化手法では，複数の企業について複数の単語の重みがどのように時間推移していったのかを示すことができるということである．このことから，本システムのは，テキストマイニングの時系列を考慮した解析結果の表示に適しているといえる．

最後に (3) についてであるが，これは，描画する上で避けられない問題であるといえる．ディスプレイが有限な広さである以上，描画している物体が重なっていしまうことは避けられないことであると考えられる．しかし，操作性を向上させることである程度の問題は解決できるものと考えている．

これらのことから，本研究の提案したシステムについて以下のようにまとめた．

- 時系列表示を行うことで時間軸に沿った概要の把握が容易になる
- 全体のマイニング結果では気づかない知識にも注目することができる
- 企業間と単語間の比較を同時に行うことができる
- ユーザが注目したいデータを，容易に取得できるような操作方法を考察する必要がある

6 まとめ

本研究では，新聞記事 1 年間分をまとめてマイニングするのでなく，時間的に意味を持つ単位に分解してマイニングを行い，また，時系列を考慮した動きのある視覚化を用いて，結果を表示する手法を提案した．

これにより，時間軸に沿ったマイニング結果の全体概要の把握ができるとともに，解析過程から解析結果とは異なった知識も発見できることを述べた．

ここで，今後の課題として新聞記事の他の特徴 (レイアウト，何ページ目に掲載されていたかなど) について考慮した重みの算出方法と，単語が多くなった場合の視覚化手法の検討，そして評価方法の確立といったことが考えられる．今後はこれらの問題点やシステムの操作性について検討していくつもりである．

参考文献

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth: Knowledge Discovery and Data Mining: Towards a Unifying Framework, Proceedings of Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996.

[2] Marti A. Hearst: Untangling Text Data Mining, Proceedings of ACL'99, June 20-26, 1999.

[3] 高田 哲司, 小池 英樹: 見えログ:情報視覚化とテキストマイニングを用いたログ情報ブラウザ, 情報処理学会論文誌 Vol.41, No.12, pp.3265-3275, 2000.

[4] 渡辺 勇，三末 和男: 単語の連想関係によるテキストマイニング，情報学基礎，Vol.55，No8，pp.57-64，1999.

[5] 吉田 尚史, 清木 康, 北川 高嗣: 意味的連想処理機構を用いた大量データ分析のための動的クラスタリング方式， 情報処理学会研究報告, 98-DBS-116(1), pp.143-150, 1998.

[6] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir: Text mining at the term level, In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, pages 65-73, September 1998.

[7] 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，No4，pp.358-364，1999.

[8] 渡辺 勇，三末 和男: テキストマイニングのための連想関係の可視化技術，情報学基礎，Vol.55，No8，pp.65-72，1999.

[9] Marc M. Sebrechts, Joanna Vasilakis, Michael S. Miller, John V. Cugini, Sharon J. Laskowski: Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces, 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, August 1999.