

新聞記事のマイニング結果についての視覚化手法の提案*

3C-05

井前 吾郎 奈須 庄健 淡谷 浩平 重野 寛 松下 温†

慶應義塾大学理工学部‡

1 はじめに

近年、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]。

そこで本研究では、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 M J、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について説明を行い、最後に、第 5 章でまとめと今後の課題について述べる。

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ており、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 (=文字数) 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が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ととした。

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、1 年間といったまとまっ

*Proposition of visualization technique concerning the mining result of the newspaper article

†Goro Inomae, Shoken Nasu, Kouhei Awaya, Hiroshi Shigeno, Yutaka Matsushita

‡Faculty of Science and Technology, Keio University

た期間で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法はある文書の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 $idf = 0$ になるので、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、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定めることになり、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う手法がある [8]。

また、情報視覚化の分野では、3次元空間を用いて視覚化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はできるが、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、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で採用されているマウスのような2次元的控制では、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もある。

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对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次元空間を用いることとした。

3.3 ノードの表示の仕方

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、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ることができた。

- 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

新聞記事において、企業名をもとにした分析例について説明する。

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)。

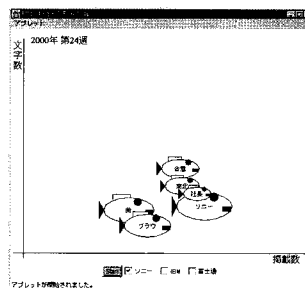


図 1: インターフェース

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ことにした。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を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ことを示すことができる。

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を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).

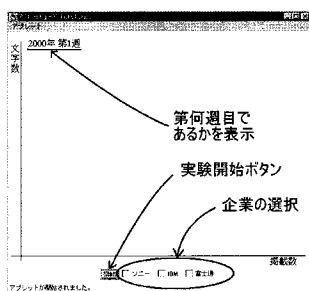


図2: システムの操作部

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。

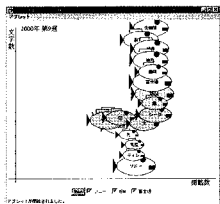


図3: 第9週目

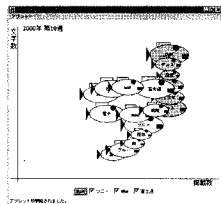


図4: 第10週目

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 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のではなく、時間的に意味を持つ単位に分割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ということを述べた。

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るかなど)について考慮した重みの算出方法と、視覚化手法において単語が多くなった場合の検討、そして評価方法の確立といったことが考えられる。本研究ではこれらの課題について十分な研究を進めたいので本システムの評価を行う予定である。

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. Sebrecths, 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.