頻度・鮮度の多面分析に基づく動向分析の試行

Trial of Trend Analysis Method by OLAP based on Frequency and Freshness

寺地　雅弘†，佐賀　亮介†，辻　洋‡

Masahiro Terachi, Ryosuke Saga, Hiroshi Tsuji
大阪府立大学　経営情報システムグループ
Management Information System Group, Osaka Prefecture University
E-mail：†{terachi, saga}@mis.cs.osakafu-u.ac.jp　‡tsuji@cs.osakafu-u.ac.jp
URL：http://www.cs.osakafu-u.ac.jp/mis/

概要　本研究では，各記事カテゴリを構成するキーワードを抽出し，その出現頻度の変遷や出現期間といった情報を広義の動向情報として捉え，それらをもとに，個々の記事からは把握できない動向を多数の記事から把握するための一手法を提案する．手法は記事のカテゴリ，その中に現れるキーワード，キーワードの当初出現時，最新出現時，出現記事数，出現頻度を軸として，OLAP を利用するものである．OLAP のダイシング，スライシング，ドリルダウン，ドリルアップによって各キーワードのポジショニングを行い，キーワード出現の栄枯盛衰を可視化するためのデータベースの構築や多次元データ構造の設計手順及び分析手法について提案する．

キーワード：OLAP，データベース，テキストマイニング，RFM 分析

1　はじめに

企業活動において，計算機上に蓄積されたデータに対して様々な角度から分析を行い，意思決定を行うことが求められていた．また大量のデータの蓄積と，そのデータに対しての高速な処理も合わせて必要とされていた．

近年では，計算機性能やハードウェアのコストパフォーマンスの向上，電子化された情報の増加に伴い，データウェアハウスの構築についての関心が高まり [1]，ユーザの関心や興味，傾向を分析する研究も求められている．そのための方法として，出現キーワードの頻度変動や，他のキーワードとの関係の情報可視化などが行われている [2] – [7]．ただ，複数のキーワードの動向を調査するには，グラフなどでは情報量が多くなりすぎてしまい，動向の変化が見にくいことが問題であった．また，データベースに基づく分析については，SQL の理解などの技術が求められることも少なくなく，データ管理者がデータベースを構築しても，全てのエンドユーザが，それらのデータに基づいた意思決定をできるわけではなかった．

本論文では，マーケティングで用いられている RFM分析 [8], [9] の概念を基に，キーワードの変動の可視化手法を提案する．これは，単語の出現頻度といった頻度情報と，出現時期から得られる鮮度情報とを行列上にマッピングすることで，キーワードのトレンド動向についての分析を試みるもので，前者の問題を解決するものである．後者の問題については，そうした複雑な SQL を知らない一般ユーザでも簡単に多次元データの分析を行うためのシステムとして OLAP（On-line Analytical Processing）が提案され，企業の販売分析などに利用されている．本論文では，前述の分析手法に OLAP の応用も考慮し，そのためのデータベース構築と結果の可視化とを試みる．

2　提案システム

2．1　提案システムの概略

加藤らは，動向情報の要約と可視化に関するワークショップ（MuST）を運営している [10]．本論文では，ここで提供されているデータセットを基に，キーワードのトレンドの変化をマクロな動向情報とみなし，これを可視化しエンドユーザの知識獲得に役立たせることを目指す．

提案システムの概略を図 1 に示す．まず，コーパスで与えられた記事群から記事番号を抽出し，対応する原文に対して ChaSen*1を用いて形態素解析を行う．その結果から名詞，未知語を抽出し，記事カテゴリごとに TF-IDF アルゴリズムによってウェイト計算を行い [11]，その上位語をキーワードとした分野別データベースを構築する．エンドユーザは，これらを統合したデータベースを基に，OLAP 用分析キューブ，RFM クロス集計表の操作を行う．以下の節では，これらの分析操作についての詳細を説明する．

＊ http://chasen.naist.jp/hiki/ChaSen/

・ 奈良先端科学技術大学院大学松本研究室

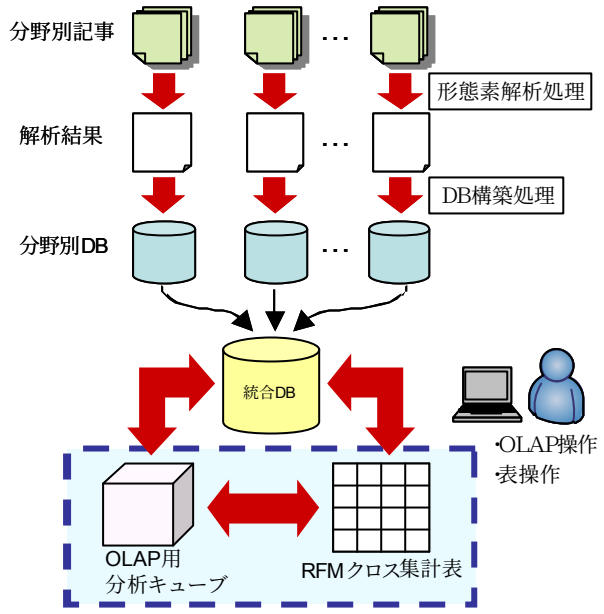
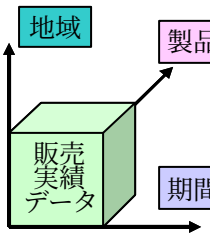


図1．　提案手法の概略

2．2　OLAP（オンライン分析処理）の概念

OLAP とは，企業内の販売管理や顧客管理といった業務アプリケーションからデータを抽出し，それらの特徴やパターンを複数の次元から分析するものである．OLAP ではキューブと呼ばれる複数の次元で構成されているデータ集合をもとに分析する．図２は，販売実績のデータベースを例に OLAP について模式的に示した図で，地域，製品，期間の三軸を持つ販売実績データがキューブ内に格納されている．このキューブから，エンドユーザは，ダイシング，スライシング，ドリルダウンと呼ばれる操作で時々の状況に応じた分析レポートを自分自身で動的に生成する事が可能となる．

分析操作は，キューブの手前の面を操作することで行われ，ダイシングとは手前の面の軸を変更する操作（例：製品軸と期間軸を入れ替える），スライシングとは，奥行きになっている次元について対象を絞り込む操作（例：特定の製品のみについてのレポートを見る），ドリルダウンとは次元の階層を掘り下げる操作（例：2004 年 1 月分のデータを参照する）である．またドリルダウンの逆向き，上の階層に進む操作をドリルアップと呼ぶ．



全製品	2004年	2005年	2006年
東日本	20	50	70
西日本	80	90	100
海外	0	10	15

図2．　OLAP 操作キューブの例

これらの操作によって得られたレポートに基づいて指針，方針を決定する事は有効だと思われるが，本体や動作に必要な環境が高価だったため，大企業以外ではなかなか導入されなかった．今回利用する OpenOLAP*2は情報処理機構の支援を受け IAF コンサルティング㈱が開発したオープンソースで，無償で利用可能なソフト類で動作可能である．

2．3　RFM分析とその応用

RFM 分析とは，マーケティングにおける分析の一つで，Recency：ある顧客が最後に来店したのはいつかといった時系列の情報と，Frequency：どのぐらいの頻度で来店しているのか，Monetary：どのぐらいの額を購入しているのかといった頻度情報とを組み合わせ，企業が顧客との間に，より良いリレーションシップを築くための判断に役立てるものである．例えば，表１のような顧客分析を行い，通算の来店回数は少ないが最近来店して興味を示している顧客を優良顧客へと引き上げるといった戦略である．

表1．　R-F による顧客分析の例

	F: 多い	F: 普通	F: 少ない
R: 最近	優良顧客	優良顧客候補	試み客
R: 普通	不満客	事情のある客	
R: 昔	不信客	他社の客	

本論文では，キーワード分析に，この考えを応用することを考える．つまり，顧客の購買行動を各単語がキーワードと見なされる事象，即ち各記事内において一定の閾値以上のウェイトを記録する事象に置き換えた分析を行う．使用する指標としては TF，DF といった頻度情報の他に，RFM 分析の R に該当する頻度情報として，単語が最後にキーワードと見なされた日付 LA（Last Appearance），初めてキーワードと見なされた日付 FA（First Appearance）を導入し，表２のような分析例を得る．

表2．　キーワード分析への応用例

	TF: 多い	TF: 普通	TF: 少ない
LA: 最近	旬の話題	成長可能性のある話題	新興の話題
LA: 普通	成熟期の話題	グレーゾーン	
LA: 昔	衰退期の話題	無視できる話題	

＊2　http://sourceforge.jp/projects/openolap/

・　SourceForge.jp

2．4　OLAP を用いたクロス集計分析

提案手法では，最終的には図3に示されるような，より柔軟なキーワード分析を可能とすることを目的としている．例えば，クロス集計による分析結果を見ながら，ダイシング操作によって軸を変更してみたり，スライシング操作によって条件を絞り込んだりと分析結果を動的に得る事ができる．また特定の条件や期間に条件を絞ったドリルダウン操作によって，細かいトレンド動向にも注意を向け，着目した箇所の詳細を知る事ができる．

提案手法は，自動的に注目語，高い価値を持ったキーワードが抽出されるといったものではない．そうした判断は人間に依るところであるが，柔軟な分析をサポートすることで，ユーザの意思決定に役立つものと思われる．

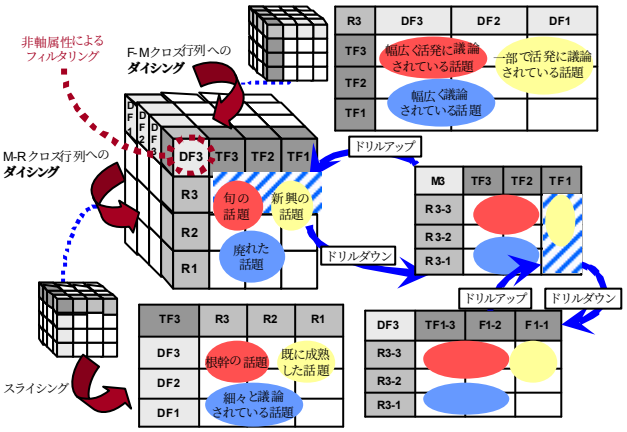


図3　OLAP 操作とキーワード分析の統合

3　データベースの設計と構築

3．1　データセット

今回の実験の対象データとしては，MuST で提供されているデータセット，毎日新聞の 1998-9 年のテキストデータから，コーパスを基に記事数 30 以上の記事カテゴリを対象とし，7つの記事カテゴリの計 249 件の記事を分析対象とした．これらについて ChaSen を用い形態素解析を行い，品詞によるフィルタリングを行った結果，延べ 21,612 語，2,974 種の単語を抽出した．さらに各記事カテゴリの単語についてウェイトに対して仮の閾値を設け，延べ 1,404 語，835 種のキーワードを抽出した．

3．2　OLAP 用データベースの設計・構築

OLAP で扱うキューブはディメンションと呼ばれる次元の組み合わせで構成されている．今回はキーワード，品詞，登場した記事カテゴリ，及び登場時期（時間軸）をディメンションとして定義した．また，OpenOLAP にはセグメントディメンションと呼ばれる，ディメンションのメンバーを値によって分類するものが備わっている．TF や DF といった頻度情報，本論文で導入している LA や FA といった鮮度情報については，これを用いて表2で示されているようなクラス分けを行い次元として取り扱った．

また，キューブ内には，メジャーと呼ばれる数値データが格納されている．ここでは，TF や DF といった頻度情報，及びそれらに基づいて計算されるウェイトをメジャーとして定義した．

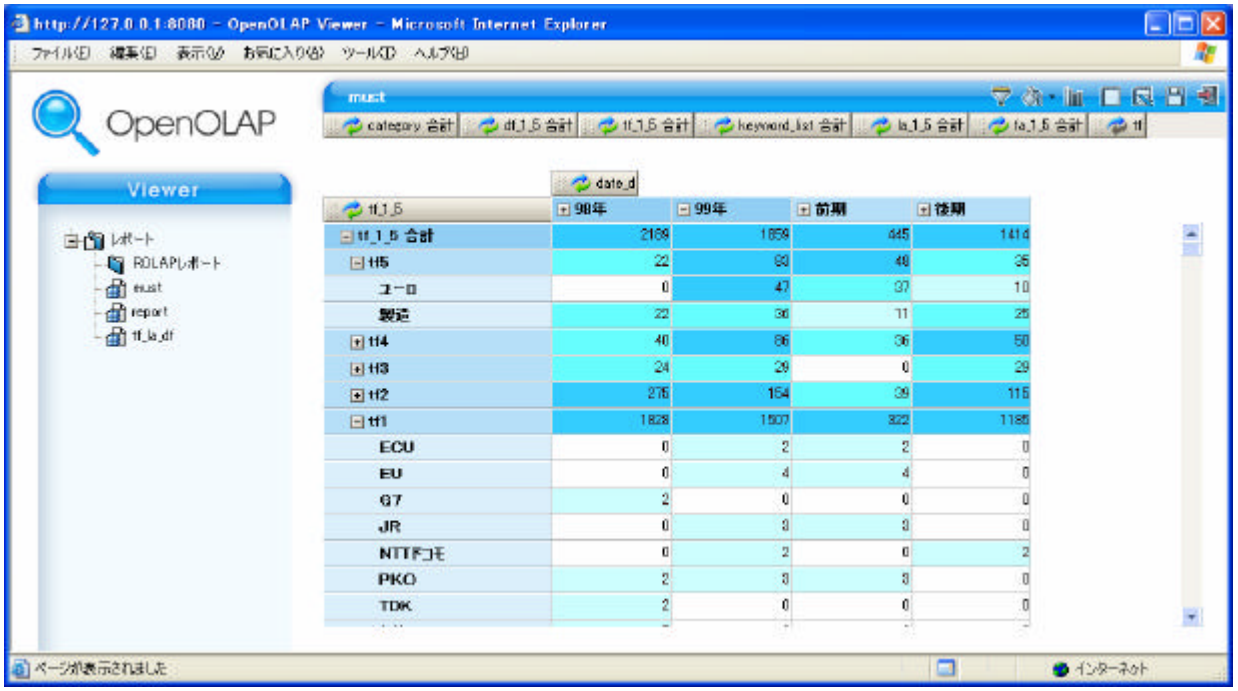


図4　OpenOLAP 実行画面例

以上のような次元と尺度の設定で，OLAP キューブを構築した．また，このキューブの基となるデータテーブルは，データベースサーバである PostgreSQL*3に格納されている．データを追加，変更したモデルで実行したい場合，キューブ設計に変更の無い限りは，ここに格納されているテーブルを更新後，OpenOLAP 上でリロードするだけで，キューブを再構築する事ができる．また閾値を変更したモデルについては，OpenOLAP 上で閾値の調整ができ，キューブを再構築する事で既存のレポートにも結果が反映される．

4　OLAPを利用した分析

OpenOLAP の実行画面例が図４である．これは縦軸にキーワードを TF の値でクラス分けしたセグメントディメンションを，横軸に時間軸が来るようにダイシングしたものである．セル内には対象期間でのメジャーのうち TF が表示されており，その値に応じて各セルの背景が色分けされている（ユーザ側で閾値を設定可能．ここでは４段階で，大きいものほど濃い）．

表２のキーワード分析を参考に，まず TF の多いキーワードについてドリルダウンを行い，詳細を見てみると「ユーロ」というキーワードが，特徴的な TF の変動を見せていることが分かる（図４上部）．この単語は 1998 年中には出現が無く，FA が 1999 年 1 月となっており，LA も新しい．すなわち「旬の話題」に該当するものである．（実際，ユーロが銀行取引などの通貨として導入され始めたのがこの 1999 年からである．）

一方で，TF の少ないキーワードにも焦点をあてると，TF は小さいが LA の新しいもの，即ち「新興の話題」として「NTT ドコモ」が挙げられる（図４下部）．このキーワードについても，横軸に「記事カテゴリ」が来るようにダイシングすると，この単語は「日経平均株価」に登場していることが分かる．さらに記事カテゴリの階層にドリルダウンを行うことで，出展記事を得ることができる．その出展記事を参照してみると，「NTT ドコモ」が時価総額で前年の２位から１位に躍進したという記事（記事番号：991231047）に辿り着いた．

このような分析は一例であり，また結果の判断自体はユーザ依存のものである．しかし，OpenOLAP を利用した可視化によって，本論文で提案したキーワード分析を基に，ユーザが状況に応じてキーワードのトレンド動向を多角的に分析することが可能と考えられる．

5　おわりに

本論文では，複数の次元で構成されるデータベースに対して，頻度情報と鮮度情報とのクロス集計によるマッピングで分析を行う手法を提案した．また，OLAP の考えと組み合わせる事によって，この分析に基づいた柔軟な分析と可視化を実現した．OpenOLAP を用いた実装部分では GUI ベースの操作によるインターフェースを利用し，専門的な知識を持たないエンドユーザでも，簡単に分析を行うことができる環境を設計した．

この提案手法は，ユーザが指定した条件に対して，それにマッチする答えが自動で導き出される類のものではない．しかし．エンドユーザがその柔軟性を活かして試行錯誤的な分析を重ねる事で，価値ある知見を得て，意思決定に役立てることができるものと考えている．

今後の課題としては，キーワード分析で着目した話題について共起関係にある語を視覚的に示すことで，ユーザの探索活動にさらなる幅を持たせたいと考えている．また，データセットを拡張することで，長期視点でのキーワードの動向変動などを鳥瞰できるようになればとも考えている．

文　献

(1) William H. Inmon:Building the Data Warehouse, Wiley Publishing, Inc. (2005)

(2) Koichi Yamada, Hsashi Komine, Hroshi Kinukawa, Hiroshi Nakagawa：“Abstract of Abstract: A New Summarizing Method based on Document Frequency and Clause Length”, SCI2004 (The 8th World Multi-Conference on Systemics, Cybernetics and Informatics), Orland, July. (2004)

(3) Hiroshi Yamamoto, Seishiro Ohmi, Hiroshi Tsuji : “Entropy-based Indexing Term Selection for N-gram Text Search System”, IEEE International Conference on Systems, Man & Cybernetics (IEEE/SMC 2003), pp.4852-4857 (2003)

(4) Hisao Mase and Hiroshi Tsuji：“Experiments on Automatic Web Page Categorization for Information Retrieval System”, IPSJ Journal, Vol. 42, No. 2, pp. 334-348 (2001)

(5) Kenji. Yamanishi, Hang Li : “Mining Open Answers in Questionnaire Data”, IEEE Intelligent Systems, pp58-64 (2002)

(6) Gerald Salton:Automatic Text Processing, Addison-Wesley Publishing Company (1989)

(7) Yutaka Matsuo, Mitsuru Ishizuka：“Keyword Extraction from a Document using Word Co-occurrence Statistical Information”, JASI Journal, Vol.17, No.3, pp.217-223 (2002)

(8) Donald R. Libey : Libey on RFM value, e-Book, http://www.e-rfm.com.

(9) Masahiro Terachi, Ryosuke Saga, Hiroshi Tsuji : Trends Recognition in Journal Papers by Text Mining, IEEE International Conference on Systems, Man & Cybernetics (IEEE/SMC 2006), pp4784-4789 (2006)

(10) Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando：“MuST: A Workshop on Multimodal Summarization for Trend Information”, Proc. NTCIR-5 Workshop Meeting, pp. 556-563 (2005)

(11) Akinori Kageyama, Hiroshi Tsuji：“Web-based Characteristics Analysis for Industrial Departments in Universities”, The 8th World Multiconference on Systemics, Cybernetics and Informatics, Vol.10, pp.23-28(2004)

＊3　http://www.postgresql.jp/

・ 日本PostgreSQLユーザ会