GUI による非構造データ統合分析方式

桐村 綾子† 高山 茂伸† 菅野 幹人†

†三菱電機株式会社 情報技術総合研究所 〒247-8501 神奈川県鎌倉市大船 5-1-1

E-mail: †(Kirimura.Ayako@cw, Takayama.Shigenobu@db, Kanno.Mikihito@bc).MitsubishiElectric.co.jp

あらまし 情報システムの発達に伴って発生するデータの蓄積量の増大に伴い、それを有効に活用するための技術の必要性が増している。文書などの非構造データから単語の共起関係を分析するだけでなく、それに関連する構造データを共通な分析軸で統合することにより、新たな知見の発見や、分析結果の意味的な補強の研究を行っている。本稿では、これらのデータを同一画面上に表示し、試行錯誤しながら分析するためのユーザインタフェースの方式を提案する。

キーワード 非構造データ，データ統合

Integrated analysis of structured and unstructured data on graphical user interface

Ayako KIRIMURA† Shigenobu TAKAYAMA† and Mikihito KANNO†

† Information Technology R&D Center, Mitsubishi Electric Corporation

5-1-1 Ohuna, Kamakura city, Kanagawa, 247-8501 Japan

E-mail: †(Kirimura.Ayako@cw, Takayama.Shigenobu@db, Kanno.Mikihito@bc).MitsubishiElectric.co.jp

Abstract The Amount of data generated and stored continues to grow as the development of the information system technology. Especially such unstructured data as text, e-mail, www is increasing rapidly. The necessity of the technology to use it effectively also increases. We are doing the research that the discovery of the new insight to analyze not only the co-occurrence relation of the word of unstructured data but also structured data with common analysis axes. In this paper, we propose the method of the User Interface to analyze unstructured data and structured data at once which supports user interaction and try and error analysis.

Keyword Unstructured data, Data integration

1. はじめに

デジタルデータにおいて、数値データのように構造化されていない、メールや文書ファイル・音声・画像などの、非構造データの割合が著しく増加している。ストレージの大容量化・低価格化に伴い、それらはより気軽に蓄積されるようになっている。企業内でも、企業機密漏えい防止や金融商品取引法への対策などの観点から、メールや各種ログの蓄積は進んでいる。しかし、これらは有事に備える性質のものであり、蓄積することそのものが目的となっていることが多い。

一方、リレーショナルデータベースなどの発達に伴い、企業では内外のデータを系統的に分類・加工して構造化することにより、ビジネスインテリジェンス(BI)として経営に利用してきた。非構造データについても､コンテンツとして BI に取り込み企業活動に生かしたいというニーズがある。しかし、非構造データは加工や画面表示が容易でないことや、企業内に分散するシステム毎に蓄積されていることから、活用が難しいという課題があった。

この課題に対し、我々は非構造データとそれに関連するデータベースを統合した分析を行なう手法について提案してきた[1]。この手法は、両データに共通な分析軸を用いることで非構造データを構造化して OLAP 分析を行なうものである（図 1)。両データの統合により、データウェアハウスなどを用いた分析結果の精度向上や、統合データから得られる数値データによる分析の裏づけの補強などの効果が期待できる。

この手法では、非構造データに含まれる単語や付随するメタデータ、構造化データのカラム情報から、共通に分析が可能な項目を軸として抽出して使用する。例えば、ユーザ情報（年齢、性別、職業等)、製品情報（名称等)、店舗情報などである。これらはマスターデータテーブルとして構造化データの次元テーブルの形式で格納されている場合が多い。さらに、非構造データを分析するためには、分析したいキーワードをテキストから抽出することで分析軸として用いることが可

能である。例えば、機能の名称(スイッチ、画面、電源、通信等)、要望事項としての形容詞（大きい、小さい、強力、簡単）などの項目である。これらの分析項目を階層化しておくことにより、グループ集計やドリルダウンといった OLAP 処理を行うことができる。

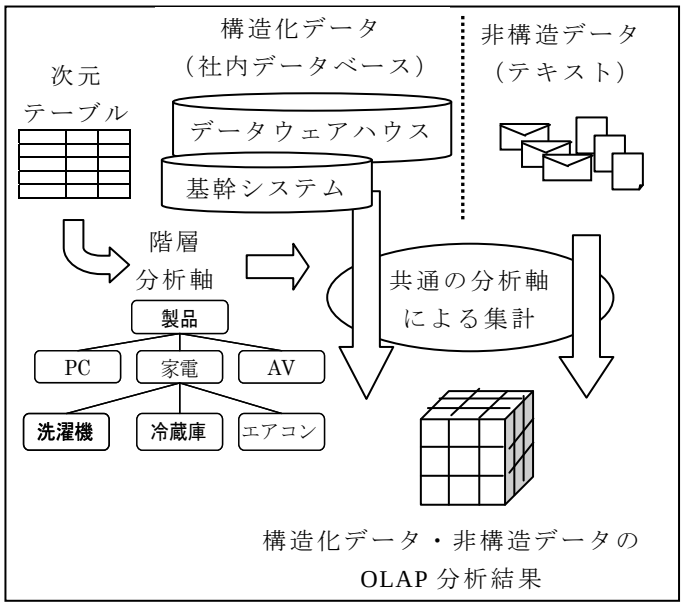

図 1　統合分析の手法

OLAP 分析は、高速な問合せ処理によりその場限りの問合せの繰り返しを可能にして、可視化されたデータ分析過程から必要な情報を見つけさせる手法である。我々の提案手法において、より直感的に分析操作を行うため、分析軸を画面上で操作するための可視化インタフェースを開発することとした。本稿では、企業システムのような大規模データを対象に、非構造データをマップグラフにて可視化する分析方式を提案する。

本論文は以下のように構成される。２章ではマップによる文書情報の可視化についての課題と本方式による解決策について示す。３章では本方式を実装したプロトタイプについて説明する。４章はまとめである。

2. 課題と解決策

2.1. マップを用いた文書情報の可視化

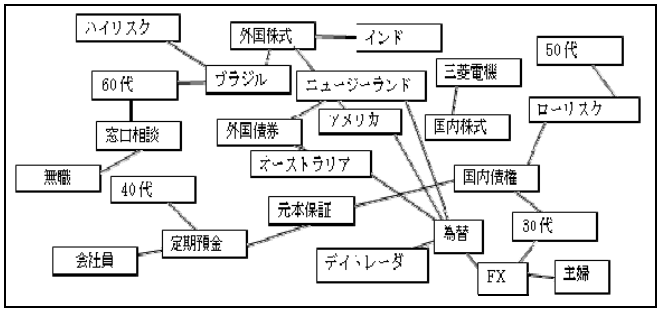

図 2　マップによる単語間関係の可視化

テキストマイニング[2][3]では、抽出された単語同士の関連を、集計値のグラフや図 2 のようなネットワーク化したマップによって可視化することでデータの傾向を表現する手法が取られている。

マップによる可視化には以下のようなメリットがある。

- 全体を俯瞰的に表示することで直感的な理解を助ける
- 要素を柔軟に配置できるため、属性の異なる要素同士を比べるのに向いている。（表形式のように同じ属性の要素で比べなくてもよい）
- ノードの類似度合いなどを視覚的に把握できる

企業内での用途を想定すると、取引先リストなどのマスターデータが 1 万件を超えることも稀ではない。人間が画面上で一見して把握できる情報量は限られていることから、マップ表現によって全体把握ができるデータ量は限られており、マスターデータの全ての項目を画面上に表示することはできない。視認性を確保するためには、マッピングするデータを絞り込むことでデータ量を減らさなければならない。

2.2. データの絞込みの課題

絞込みの際にユーザの視点が入ることによって、有用な情報を切り捨ててしまう可能性があるため、できるだけデータ全体の特徴を失わないように表示する必要がある。マップデータの絞込みの方法としては、閾値処理による方法と、階層化やクラスタリングといった抽象化による方法の二つがある[4]。

大量データに対する閾値処理では、下記のような課題がある。

課題1　わずかな閾値の差に大量のデータが含まれるため、目的のデータが表示されるまでに画面上に表示可能なデータ数に達してしまう。

課題2　属性値の分布度が異なるデータ同士の関係を分析するような場合は、ひとつの閾値を全体に適用してもデータの特徴を的確に表すことができない(例えば、小規模店と大規模店を売上げ額で比較することは困難である)。

データ構造を抽象化する手法を用いた場合でも、データ量が多い場合は抽象化の度合いが大きくなりデータの特徴を表現しにくい。また、階層構造の場合、分析の目的によってはひとつの親に数百の子が属するようなデータもある。このような場合に子ノードを表示するには、課題 1 と同じ問題が発生するため、閾値処理などを組み合わせることによってさらに絞り込む必要が在る。

2.3. 階層構造を用いた共起関係の可視化

本手法では、分析軸の各項目同士の関係を画面上にノード・エッジで構成されたネットワークグラフとして表示する。

共起分析を行う分析軸を二つ選択し、階層構造を木構造として画面上に対応するように配置、階層間の共起関係をノード間にエッジとして表示する。階層を用いることによって属性の異なる要素同士の共起関係を整理して提示することが可能になり、分析作業を効率的に行うことができるようになる。また初期画面はノードを縮退した形状で表示することでデータを抽象化し、各ノードを展開していくことにより、OLAP 分析におけるドリルスルーと同様の操作が可能になる

さらに、階層に属するノード数が多い場合に表示するノードを絞り込む。絞込みの基準となる値には、非構造データからは分析軸の項目である文書から抽出可能な統計値属性（各単語の共起回数など）と、構造化データからは項目に対応する数値属性(売上げなど)と、複数の属性値を用いることによって閾値処理に使用する属性値を変更可能にする。また、ノード毎に下階層に適用する閾値の値や、絞込みに用いる属性を変更可能にすることで課題 2 に対応する。また、これにより、注目箇所以外の場所のデータの表示量を増やさずに閾値処理が可能になり、課題 1 に対応する。

このように、階層化構造によるマップグラフの抽象化と柔軟な閾値処理により、データの特性に合った情報を表示することができるようになり、適切な分析結果を選択可能にする。

3. 統合分析プロトタイプシステム

本分析方式を評価するために非構造データ統合分析可視化プロトタイプを実装した。

3.1. システム構成

図 3 にシステム構成を示す。システムは、グラフデータ作成部と可視化部から成る。

3.1.1. グラフデータ作成部

グラフデータ作成部は、元データからグラフデータを作成するまでの下記の機能を持つ。

① 単語共起関係の抽出

非構造データ（メール、テキストデータなど）から分析軸の項目となる単語を抽出する。

② 分析軸情報の作成

顧客情報などのマスターデータを階層化設計（次元テーブルの作成）し、非構造データから抽出された上記以外の単語情報を分析軸として階層化設計し、分析軸となる次元テーブルを作成する。

③ 分析軸による集計

分析軸の項目ごとに分析対象データを集計する。システムは、集計までを事前に行う。

- 非構造データ
  各文書内の単語の出現状況から、単語同士の共起関係を集計する。
- 構造化データ
  売上げや顧客属性値(契約年数や従業員数など)などを分析軸項目ごとに集計する。

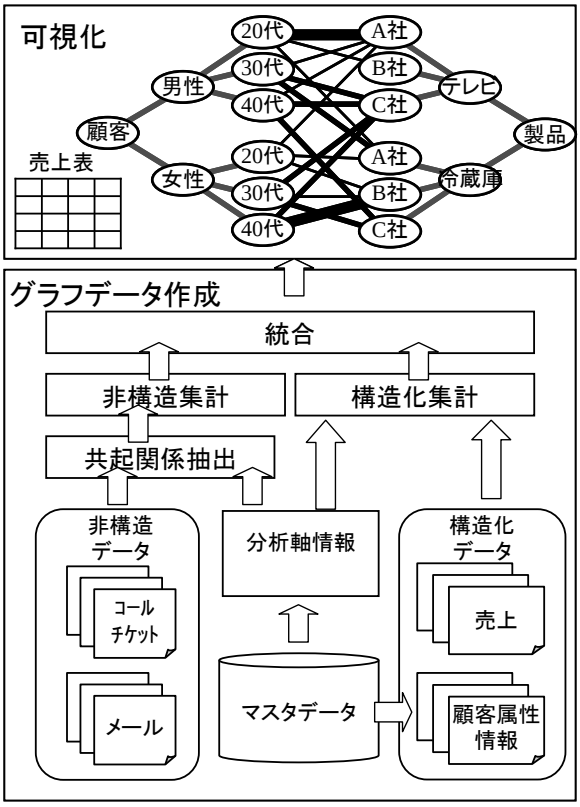

図 3 統合分析プロトタイプシステム構成

④ 統合

可視化の設定（マッピングする分析軸や分析対象の期間など）に応じて各集計結果から必要部分を抽出する。設定によって、期間毎の中間集計値などをさらに集計する。

3.1.2. 可視化部

統合ざれたデータをマップグラフにて可視化し操作するユーザインタフェースツール。階層関係を用いたマッピンググラフに対し、インタラクティブな分析を行うため下記のような機能を持つ。

① グラフ表示

分析軸の各項目をノード、共起関係をエッジとして表示する。ノードは子ノードの縮退表示を可能とし、縮退している場合は親ノード項目としての集計値でノード、接続するエッジを表示する。

② 属性値によるフィルタリング・表示

各ノード、エッジに複数の属性値を付与する(打ち上げ値、文書数など)。表示の際にこれらの属性値のうちのひとつを選択し、閾値や表示色を設定可能とする。また属性値を表などの形式でマップとは別に表示する。

③ 非表示化

ユーザが分析に不要と判断したノードを非表示化可能とする。表示するノード・エッジを減らすことでグラフの視認性を向上する。

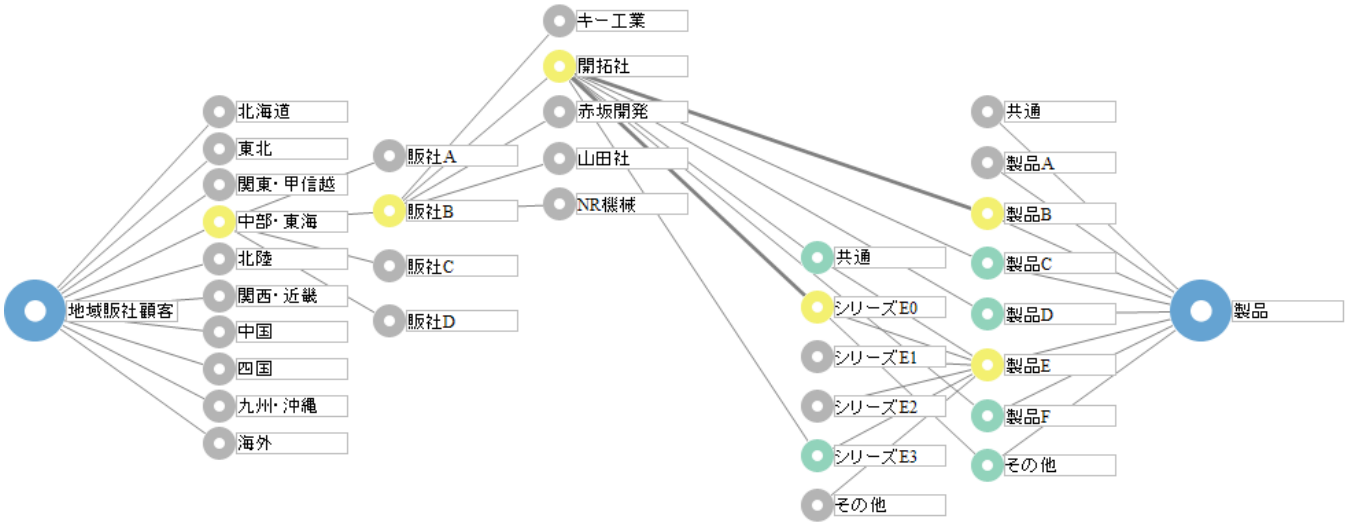

図 4　プロトタイプ画面

④　3 重共起関係の表示

3 つ目の分析軸を設定し､エッジに属する文書に対して多く出現する分析項目を表示する。(テキストから抽出したキーワードなどを表示する)

⑤　ドリルダウン

縮退したノードを展開する操作を行うことで分析項目を詳細化する。また、ドリルダウンする際に子ノード数が多い場合は、視認性を確保するため、上限表示数とソート属性を設定した上で、属性値が上位の子ノードから順に表示し、スクロールバーを用いて下位子ノードと入れ替え表示を可能とする。

⑥　ドリルスルー

ノード、エッジから対応する分析軸項目が出現する文書へアクセス可能とする。

3.2. 実装画面例

プロトタイプでは､グラフ表示と共起数を元にしたフィルタリング・表示を行う機能を実装した。図 4 にプロトタイプの画面の例を示す。製品に対する問合せ内容の分析用途を想定し、問合せ回数を問合せ元と製品間の共起関係としてマッピングした。問合せ元は地域→販社→顧客と階層化した分析軸とすることで、問合せ回数の多い販社の抽出とその担当顧客の中の要注意顧客を同一画面上でトレースが可能になる。

3.3. 考察

表示ノード数が 100 程度であってもエッジの数は対応する階層の末端ノード数に相当するため煩雑になる。注目部分を強調するためのその他の部分の省略表示や、階層が深い場合の表示方法についても同様で、上位階層の圧縮表示などの可視化方法の洗練が必要であることがわかった。

4. まとめと今後の課題

テキストデータと構造化データを、それらの共通項目を軸として分析する統合分析方式に対し、グラフを用いた可視化によって直感的な分析を行うための可視化方式を提案した。今後はプロトタイプへの機能追加を行った上で、実システムデータを用いた実証実験を行い、効率性等について評価していく。

今後の課題として以下を挙げる。

- 単語の表記の揺れ

  構造化データではマスターデータを用いることによって入力内容を一意にすることができるが、メール等の場合は表示ゆれが避けられない。単語抽出時に意味的な統合を行う必要がある。

- カテゴリ設計

  日々文書が蓄積される中で、新しく出現するキーワードを分析軸に追加していく必要がある。

- データ管理部集計方法

  大規模コールセンターなどでは、顧客数は万単位、製品数も 1000 件を超えることは珍しくなく､日々 1000 件以上の問合せが発生する。また、大量文書を分析軸に応じて集計する必要がある。分析軸の組み合わせが多くなるに従い計算量が増し、事前の集計が困難になる。

文　献

[1] 高山，平田，“非構造データと構造化データを統合した分析手法の提案”,情報処理学会全国大会講演論文集．巻：71st　号：1 頁:1.491-1.492

[2] 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 19990415

[3] 渡辺 ”テキストマイニングの技術と応用”情報の科学と技術　Vol.53, No.1,　pp.28-33 20030101

[4] 井上，吉廣，中川“大規模クラスタリング結果のグラフによるインタラクティブな可視化手法”情報処理学会研究報告．情報学基礎研究会報告 2006(118)　pp.21-28 20061116