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

高山　茂伸　　平田　飛仙
三菱電機株式会社　情報技術総合研究所

1. はじめに

人類によって創出される情報量が爆発的に増加している。その中でも特に、メールや文書ファイル、画像などの非構造データの増加が著しく、その活用が企業における課題の一つである。また、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を蓄積するニーズが高まっている一方で、それらの情報は有事に備えて蓄積することそのものが目的となっており、コンテンツは十分に活用されていない場合が多い。

非構造データに関連する研究としては、大量の文書データの集合を分析して傾向をつかむために、多次元データベース[1]を用いたオンライン分析処理(Online Analytical Processing: OLAP)の手法が提案されている[2]。しかしながら、メールや文書などの非構造データの分析と、構造化データである社内の情報システムなどの分析は別々に実行されている場合が多く、企業においてはそれらを統合した活用がなされていないという課題がある。

本稿では、メールや文書ファイルなどの非構造データと構造化データである社内データベース・データウェアハウスを共通の分析軸で OLAP の手法を用いて分析することで、非構造データの分析結果によりデータウェアハウスなどの構造化データの分析結果の精度向上や分析の裏づけの補強をするなどの効果を目指して、両データを統合した分析手法の提案を行う。

2. 非構造データ分析の現状

企業で所有するアンケート結果やユーザからの要望のメールなどのテキストデータを活用する場合においては、非構造データ独自で分析を行なう、もしくは関連する企業情報システムと紐付ける場合でも、テキストから抽出したキーワードを参考にそれと関連するレコードの分析を行なうなど、異なる視点の分析結果の紐付けである。そもそも企業においては、基幹システムや情報システムなどのデータベースもしくは分析系のデータウェアハウスと、メールやログなどの非構造データは別々に管理されているという現状もある。また、非構造データの分析をそのまま構造化データの分析と紐付ける場合には、非構造データから抽出したキーワードの数が多い、キーワードの共起関係が複雑であるなど、非構造データの分析と構造化データの分析を統合する際の課題もあり、両データに共通の課題を設定し統合した分析を行なうことにより、分析結果の裏づけを補強する、原因を深堀するなどは実現されていない。

3. 統合分析手法の提案

本稿では、非構造データに含まれる単語や構造化データに含まれるカラム情報から、共通に分析が可能な項目（軸）を抽出し、共通の軸を用いて非構造データを OLAP 分析し、その分析結果を構造化データの OLAP 分析にて活用する手法を提案する。

3.1 分析軸の抽出

構造化データと共通の分析軸とは例えば、ユーザ情報（年齢、性別、職業など）、製品・サービス情報（製品名、製品カテゴリなど）、支店・店舗情報、担当者情報などである。これらは、構造化データの次元テーブルとして格納されている場合が多い。非構造データと構造化データの分析結果を有意なものとするためには、分析軸の抽出が重要である。抽出の手法としては、テキストマイニングなどの手法を用いる。例えば、あらかじめ非構造データに出現する単語のランキングを作成し、それとデータウェアハウスなどの構造化データのカラム情報、特に多次元データベースの次元テーブルのカラムの情報と照らし合わせ、分析軸を抽出するなどが考えられる。また、このように、分析軸の項目をあらかじめ構造化データに出現するキーワードと

An Integrated Analysis of Structured and Unstructured Data
Shigenobu Takayama, Takahisa Hirata
MITSUBISHI ELECTRIC CORPORATION INFORMATION TECHNOLOGY R&D CENTER

照らし合わせて制限することで、非構造データから抽出したキーワードの数が多い、キーワードの共起関係が複雑であるなどの課題も解決可能である。

3.2 非構造データの分析

非構造データを OLAP 分析するために、分析軸となるキーワードおよびそれらの分析軸で分析したいキーワードをテキストマイニングなどの手法を用いて抽出する。例えば、ユーザ情報（年齢、性別、職業など）と製品情報（ＰＣ、デジカメ、ＴＶなど）および、製品に対する要望事項として、おしゃれ、かわいい、小さい、などのキーワードが考えられる。ユーザからの問い合わせもしくは要望メールから、メール発信者のユーザ情報および要望の対象となる製品情報、さらに要望の内容に含まれるキーワードを抽出し、キーワード間の共起関係を利用することで、表１のような OLAP 分析が可能である。

表 1 ユーザ別、製品別の要望事項一覧の例

	ＰＣ	出現回数	デジカメ	出現回数
20 代女性	おしゃれ	50 回	かわいい	51 回
30 代女性	安い	46 回	小さい	62 回
40 代女性	簡単	23 回	簡単	45 回

3.3 構造化データの分析

非構造データの分析軸をあらかじめ構造化データの項目と照らし合わせて決めているため、非構造データと関連する社内データベース・データウェアハウスを上記と同じ分析軸で OLAP 分析することが可能である。例えば、製品の売上げデータを格納したデータウェアハウスの分析を行なうことで、表２のような結果を取得可能である。

表 2 ユーザ別、製品別の売上げ金額一覧例

	ＰＣ	デジカメ
20 代女性	3,800,000	800,000
30 代女性	2,300,000	1,200,000
40 代女性	1,400,000	500,000

3.4 統合した分析

上記で示したとおり、非構造データと構造化データを共通の分析軸で分析することで、図１に示すとおり非構造データの分析結果を構造化データの分析に活用することが可能である。

例えば、20 代の女性ユーザであれば、“おしゃれ”という理由で当社のＰＣを購入しているなどということがわかり、それとＰＣの売上分析結果を統合することが可能となる。さらには、グループ集計やドリルダウン[1]などの処理も可能であり、分析者がインタラクティブな操作で非構造データと構造化データを統合して分析することで、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分析の裏づけの補強をするなどの効果が期待できる。

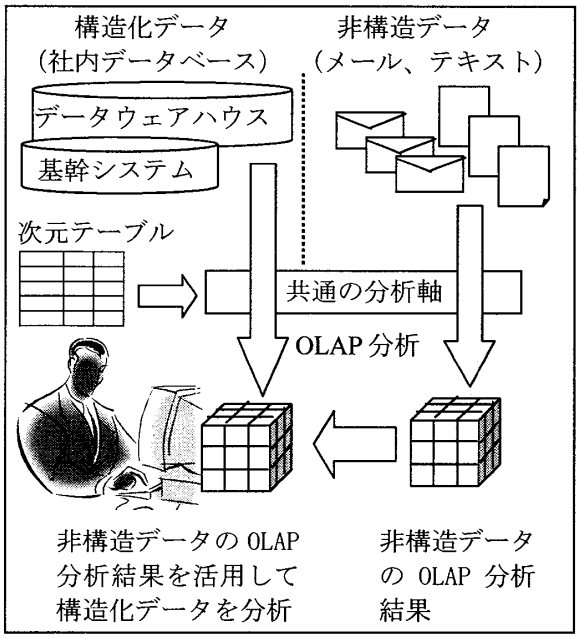

図 1 統合分析の手法

4. おわりに

本稿では、非構造データとそれに関連するデータベースもしくはデータウェアハウスを共通な分析軸を用いて OLAP 分析を行なうことで統合した分析を行なう手法を提案した。今後は、分析軸や集計項目の選択および分析結果の活用方法の検討を進め、提案モデルを実装し評価を行なう予定である。

5. 参考文献

[1] Pedesrsen, Jensen,” Multidimensional Database Technology” IEEE Computer, Vol34

[2] 猪口 明博, 武田 浩一, “テキスト分析のための OLAP システム”, 情報処理学会論文誌. Vol.48, No.SIG_11(TOD_34) pp. 58-68