E-003

時系列データと文章データベースからの知識抽出に関する研究

Research on knowledge extraction from temporal data and text data base

西村　康成†　　平山　正治†

Yasunari Nishimura　　Masaharu Hirayama

１．まえがき

本研究は，時間的に変化する時系列データと文章データ間の規則性や因果関係を抽出し，利益拡大・機会損失の防止・リスク回避に役立つ意思決定支援を行うことを目的とした研究である。対象データとして時間的に変化する時系列データである株価データ[1]と，文章データである新聞記事[2]を用い，これらから株価変動とその原因となった新聞記事との関係を，キーワード集合と記事掲載から株価変動に至る遅延時間の組み合わせで表現する知識の抽出方法について述べる。

２．知識発見のプロセス

1990年代の後半から，データウェアハウスの構築が進み，この中からの知識発見の手法として KDD(Knowledge Discovery in Databases)が注目されている。KDD の処理手順[3]は図１であり，詳細を以下に述べる。

１）データウェアハウス

蓄積データに検索可能な状態を構築する。

２）項目選択

分析・解析対象を選択する。

３）マイニング

データマイニング手法を用いて大量データの中に潜む価値ある情報（宝物）を発見する。例としては，相関分析，重回帰分析，ニューラルネットワークがある。

４）抽出・解釈された特徴

抽出された項目を検証し，不十分であれば前ステップに処理を戻して作業を反復する。

５）知識

蓄積データから抽出された規則性や因果関係が「知識」となる。

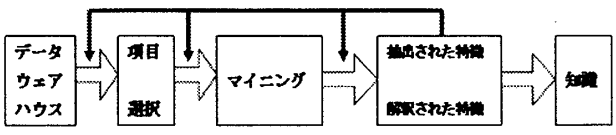

図１　ＫＤＤの処理手順

上記で述べたようにデータウェアハウスから知識抽出の手法として，データマイニングが大きな役割を担っている。データマイニングは数値データ間の関連性を見つけることを念頭とした技術である。また最近では，文章データ間の関連性を見つける手法としてテキストマイニングが考案され，顧客分析やマーケティングに活用されている。どちらの手法も，数値または文章といった単一データ構造を対象としており，顧客の苦情や要望といった文章データと，その結果の販売数や解約数といった時系列データとの関連性を分析することが困難である。本研究では，このような知識抽出の問題を解決し，時系列データと文章データ間に存在する価値ある知識を自動的に抽出しようとする研究である。

３．時系列データと文章データからの知識抽出

３．１　概要

著者は，上記の目的に沿った実証システムとして時系列データに株価データを，文章データに新聞記事を用い，時系列データと文章データ間の規則性や因果関係を抽出し，「合理化に関する記事が掲載された直後に株価が上がる」といった規則を「早期退職・合理化・リストラ・記者会見」というようなキーワード集合と，記事掲載から株価変動に至る遅延時間とその変動値の組み合わせで表現し，知識として提示する。抽出された知識の例を図２に示す。

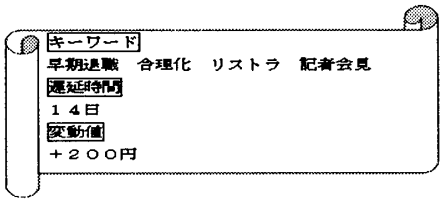

図２　知識

３．２　知識獲得の手順

時系列データと文章データベースからの知識獲得の手順は図３および図４であり，詳細を以下に述べる。

１）株価変動調査

特定企業の株価変動を調べ，大きく変動した日付（トリガ日）を取得する。

２）記事検索

企業名や検索期間を入力して，関連のありそうな新聞記事を検索する。

３）原因記事表示＆選択

ユーザは前ステップで検索・提示された新聞記事の中から，株価変動の原因と思われる新聞記事を複数件選択

†大阪工業大学大学院　情報科学研究科

する。(ユーザ選択は必ず正しいとする) このユーザ選択時におけるトリガ日と新聞記事の関係を「ケース知識」と呼ぶ。１）２）３）を繰返すことで多くのケース知識を取得し，ケース知識ＤＢに蓄積する。

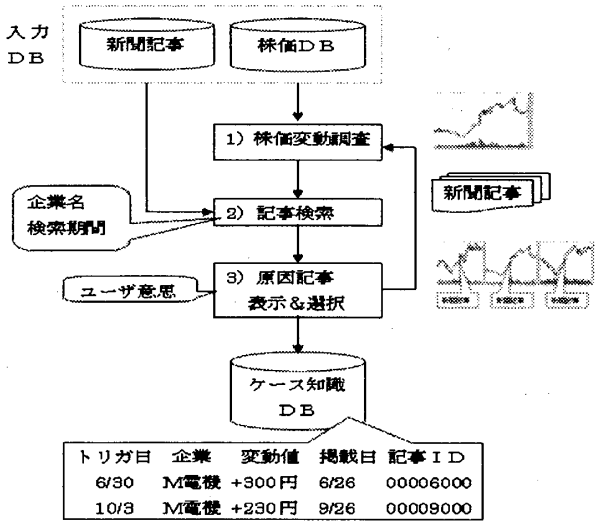

図３　知識化手順１

４）キーワード重要度算出

ケース知識毎に単語の重要度を計算する。ケース知識の新聞記事から得られた単語リストを「ケース知識キーワード」と呼ぶ。また，ユーザが選択した全ての新聞記事に含まれる単語の重要度を計算する。ユーザが選択した全ての新聞記事から得られた単語リストを「全知識キーワード」と呼ぶ。キーワードの重要度算出手順[3.3 節参照]を以下に示す。

・形態素解析システム「茶筌」[4]を用いて新聞記事の分かち書きを行う。
・tf・idf 法を用いて記事内の単語の重要度を求める。

５）重要キーワード毎のケース知識抽出

全知識キーワードにおいて重要度がｎ位以上のキーワードに対して，そのキーワードがケース知識キーワードにおいて重要度１位となっているケース知識を知識抽出の対象として選択する。

６）知識抽出

選択されたケース知識全体に対する単語の重要度を計算し，このうちの上位のキーワードおよび遅延時間，変動値の平均を知識として提示する。

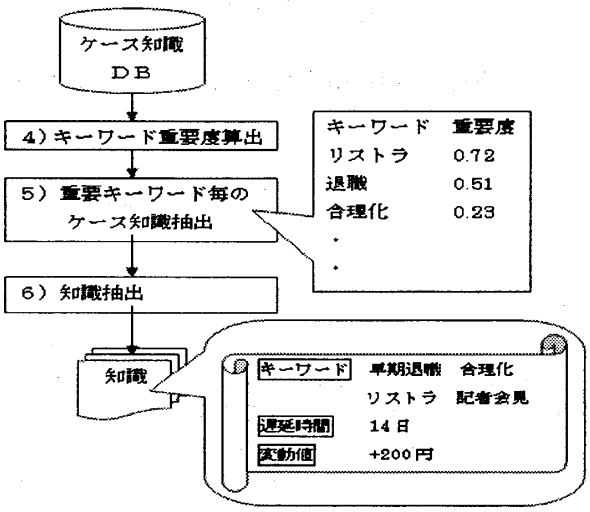

図４　知識化手順２

３．３　キーワード重要度算出

文章内の単語の特徴量を算出する tf･idf 法[5]を用いて各キーワードの重要度を計算する。tf 値と idf 値の積が文書中の単語の重要度となる。

tf（Term frequency）値: 同一文書中の単語の出現頻度

idf (Inverse document frequency) 値: 全文書数Ｎの内，該当する単語を含む文書がｎ個ある時，その比の対数 log(N/n)となる

tf・idf 値は以下の特徴を持つ。

・ 同一文書中に何回も出現する単語は重要単語である。
・ 出現する文書数が少ない単語は文書の絞込みに役立つ重要単語である。

４．まとめ

本論では，時系列データと文章データ間の規則性や因果関係から，知識を抽出する手法について述べた。本研究は，売上げデータ・在庫データ・生産データなどの時系列データと，営業日誌・イベント報告・生産記録などの文書データの相互の関連性を知識として抽出できると考える。

謝辞

本研究に際して，株価データをご提供下さったデータ・ゲット株式会社，新聞記事データをご提供下さった日外アソシエーツ社，形態素解析システム「茶筌」をご提供下さった奈良先端科学技術大学院大学の松本研究室の方々に感謝致します。

参考文献

[1]データ・ゲット株式会社，10 年株価データ CD，2004.

[2]毎日新聞社，CD-毎日新聞 2003 データ集，日外アソシエーツ，2004.

[3]石川慎也，データマイニングの宝箱，
http://www5.ocn.ne.jp/~shinya91/.

[4]松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：形態素解析システム「茶筌」Version2.3.3 使用説明書，松本研究室，2003.

[5]増井俊之，インターフェイスの街角—類似検索，
UNIXMAGAZINE 2002.12.