様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １５ 日現在

機関番号：３２７１４
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１８２

研究課題名（和文）データ融合のモデル化と不確定性を扱うデータマイニング

研究課題名（英文）Modeling of Data Integration and Datamining with Ambiguity

研究代表者
松本　一教（Matsumoto, Kazunori）

神奈川工科大学・情報学部・教授

研究者番号：４０３５０６７３

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：データマイニングのために真に有効なデータを準備する手法を開発する．単一情報源として与えられるデータだけでは不十分である．そこで本研究では，3通りの観点からの手法を開発した．まず，元データに背景知識が与えられているとして，それによりWeb検索で関連するデータを検索し付加する手法である．第2には，複数データベースを単一データベースに融合する方法であり，複数の可能性を見逃さずに扱えるようにする手法である．第3には，元データの信頼性を自動的に推測して付加する方法である．これら3つの方法を組み込んだデータマイニングを実現して，その有効性を実証することができた．

研究成果の概要（英文）：This study develops effective methods to prepare useful databases for datamining. It is in general not sufficient to use a database from a single source. We develop three types of methods that depend on different viewpoints. First, we assume a set of background knowledge is attached to the original database. Then a developed method automatically discovers useful data using the Web search, and attaches it to the database. Second, we develop a method to integrate multiple sources of databases into the one of a single source. The point here is to maintain many possibilities of the integration. Third, we develop an automatic acquisition of data with which we estimate the degree of reliability for the original database. These three types of methods and new datamining algorithms with them are implemented, and then the effectiveness is verified through experiments.

研究分野：知識工学

キーワード：データマイニング　データ融合　多重世界　優先度　センサーデータ

１．研究開始当初の背景

膨大なデータベースを対象として，データの中に潜んでいる知識を発見するためのデータマイニング技術の重要性が認識されるようになってから久しい．データベースの形式や性質に応じて，あるいは発見する知識の性質に応じて，様々なデータマイニング技術が開発されてきた．このような研究と並行して，知識発見の出発点となるデータベースの質が最終的な知識の質に大きく影響するという指摘も以前からなされている．欠損値や外れ値の処理問題としては，十分な研究実績があり，豊富な成果も得られている．このような従来からの研究成果にもとづいて，データクリーニングとよばれる技術が開発されるようになってきている．しかし，データマイングが多くの領域で実用的に用いられるにつれて，本来は必要なデータが単一のデータベースとして都合良く揃っていることはむしろ稀であることが明確になってきた．複数の情報源（データベース）に分散したデータを融合して，データマイニングのためのデータベースを作成する作業が重要な意味を持つことも次第に明らかになっていた．または，不足しているデータを自動的に探索して融合する手法の重要性も明らかになりつつあった．本研究の開始時点では，このように問題の重要性が認識され始めていたという状況であったが，そのための技術開発は不十分であり，研究も十分になされているとは言い難い状況であった．

２．研究の目的

上記の背景で述べたように，データを自動的に探索して融合すること，あるいは複数データベースを融合することで単一データベースにする技術が必要となる．本研究では，この問題に対して，３つの全く異なる観点からの目標と目的を設定して取り組んだ．

第１の観点は，不足しているデータを探索して自動的に付加するという技術である．本観点での研究では，元のデータに対して，背景知識が与えられていると仮定し，Web 空間を背景知識から導出できるキーワードにより検索することで，新たな知識を見出すことを目的として設定した．

第２の観点は，融合できる候補のデータベースが与えられているときに，合理的な方法により融合を実行する方法の開発である．従来技術では，まずデータ間に距離あるいはそれと類似の尺度を導入して，データ間の類似度を判定できるようにしておき，類似度を利用して，データベースの結合演算に似た手法を適用することで１つのデータベースに融合する．この方法は，類似度が自然に導入できる場合に限っては，効果的に用いることができる．しかし，一般の場合には，融合が一意的に定まるようにすることは困難である．そこで，データ融合の可能性を唯一に決定するのではなく，複数の可能性があるとして扱う手法を開発することにした．このとき，複数の可能性を維持したままでは，データマイニングの処理コストが増大し，現実的に適用できなくなる．そこで，データ融合をその後に続くデータマイニング手法と一体化して扱い，処理コストを考慮した段階的な手法とすることで，現実的な処理ができる手法を開発することを目的とした．

第３の観点は，データマイニングの候補となるデータに対する信頼性や曖昧性を考慮した手法の開発であり，そのために用いるデータを自動的に得ることである．本研究では，センサーにより自動的に収集できるデータを取得した上で，知識にもとづく前処理を行い，その結果を込めて元のデータに埋め込む手法について開発することを目的とした．

３．研究の方法

第１の観点に関しては，対象を経済データに絞り込んで研究を進めることにした．その上で，元データに当初から背景知識が与えられているものと想定し，それを用いた探索結果を活用したデータ融合手法を研究した．

第３の観点に関しては，授業中に学生からアンケート形式で取得するデータに絞込んだ上で，その信頼性の推定とデータマイニングへの組み込みについて焦点を絞り研究を進めることにした．

第２の観点での方法を説明する（図 1）．まず，リレーショナルデータベースにおける１組のデータ（タプル）を（A　α）と表現する．ここに，A は任意のデータの並びを表し，α はあるデータ項目の値の並びを表すものとする．データベース X 中のデータ(A　α)とデータベース Y 中のデータ(β　B)に対して，ただしβは事前に定めた類似度に関して，αと最大の類似度を持つものとすると，新たなデータとして(A　α β　B)を得る．これをデータベース中の全てのタプルに対して適用して，拡大したデータベースを得ることがデータ融合となる．このβが一意に定まらない場合の処理方法を本研究では開発する．

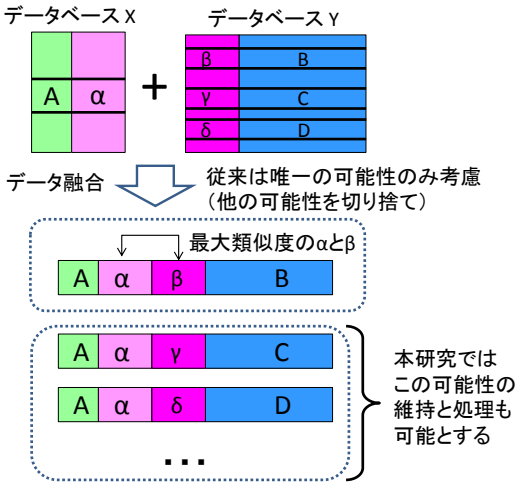

図1　本研究でのデータ融合イメージ

従来手法では，このようなβを唯一に定めることが前提となっていたが，厳しい条件であり実用化の障害となっていた．そこで本研究では，複数のβが同等にαと等しいという可能性を維持したまま処理が進めるようにする．そのためには，複数の可能性を簡潔に，また利用者にとっても分かりやすく表現する手法が必要である．通常のリレーショナルデータベースでの表現方法を自然に拡張して実現できる方法の開発を目指す．

４．研究成果

本研究により得られた成果を，上述の研究目的で述べたように，3 通りに分類して説明する．

(1) 第 1 の目的達成のために，対象範囲を経済データ，特に株価データに絞り込んだ．株価データに対するデータマイニングは社会的に極めて高い需要がある領域である．そして，最初に与えられる株価データに対して，その銘柄，基本的な企業名や企業情報をテキストデータとしての背景知識として与える．この条件の下，テキストデータから抽出するキーワードにより Web 検索を行い（既存の検索エンジンを利用），Web 上のテキストデータを関連知識の候補として抽出する（図 2）．関連知識の抽出の際に，その情報に含まれる時刻に関する表現を分析することにより，情報に対する時刻情報を決定する．これは本研究で開発した推論方式により判定するため，誤りの可能性のある判定である．株価データは時系列データであるから，時系列上の各時点に対して，先に得られた関連知識中から，時刻が近い情報を関連付ける．どの程度の範囲を近いと見なすかについての最大許容範囲をパラメータとして設定できる．

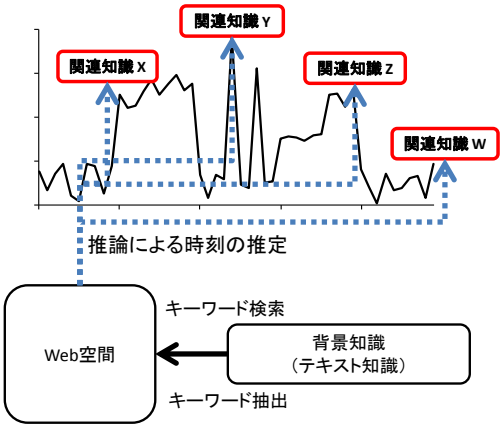

図2 時系列データと関連知識の融合イメージ

このように，時系列データに対して，テキストデータを融合したデータが得られる．これに対して，以下のデータマイニング手法を開発して実験により評価した：

① 相関ルールを拡張した手法によるデータマイニングを行う手法を開発した．すなわち，時系列データ上の特徴的な部分パターン $P=\{P_1, \cdots, P_n\}$ と，テキストとしての関連知識 $T=\{T_1, \cdots, T_m\}$ に対して，X→Y なる形式のルールが抽出される．ここに，X⊆P，Y⊆T あるいは X⊆T，Y⊆P としている．通常の相関ルールでの支持度と確信度を拡張して用いることで，ルール発見の制御ができるようにしている．

② 決定木学習を用いる方式．テキストとしての関連知識のうちのいずれかをクラス属性として用いる．他の属性部分については，上述の①で述べたパターン集合 P をクラスタ分割し，その代表元（セントロイド）との距離を属性値として用いる．クラスタリングには k-means などの従来手法を用いるが，そのときの距離としてはダイナミックタイムワープにもとづく方法など，いくつかの手法を利用できるようにしている．また，クラスタ数の決定については，実験的な方法による．また，パターンの汎用性を高めることで知識の精度を高めるために，遺伝的アルゴリズム GA と組み合わせる方法も開発している．この手法を従来からのデータマイニング手法（決定木，サポートベクターマシン SVM，ニューラルネットワーク）との評価実験を行い，本研究で開発した手法が有効であることを実証した．すなわち，本研究で開発したデータ融合手法と新たなデータマイニング手法の組合せが高品質の知識発見につながることが明らかになった．

手法		精度	
従来マイニング手法の場合	Decision tree learning (C4.5)	30.89 %	
	Support Vector Machine	27.91 %	
	Neural Network	30.04 %	
本研究で開発したマイニング手法の場合	5 clusters	34.60 %	65.23 %
	10 clusters	46.13 %	85.28 %
	20 clusters	69.96 %	89.87 %

GAなし　GA 100世代

図3 本研究で開発したマイニング手法の性能

(2) 第 2 の目的達成のために，先に図 1 で示したような，複数の可能性の全てを同時に扱いながらデータマイニングを実行するシステムを開発した．ここでは，データマイニング方式として，相関ルールの抽出に限定した．すなわち，全ての融合可能なデータベースを同時に対象として，その中から事前に与えた支持度（サポート値，support 値）を超える頻出アイテム集合を抽出し，それらに対して確信度（コンフィデンス値，confidence 値）による選択を行って相関ルールを発見することになる．この流れの基本的な考え方は，従来の手法と同一である．最も計算コストを

要する処理は，頻出アイテム集合の抽出であり，不要な候補をできるだけ早い段階で除去することが有効である．そこで本研究では，データ間に一種の順序関係（選好順序）が与えられるという,応用上妥当な仮定を置くことにした．利用者の手間を考慮して，与えられるのは単一のデータ項目内における選好順序関係と仮定するので，これをデータ項目集合間の順序に拡張する方法を開発し，それに基づいて以下の手続きを開発することができた．なお，この拡張された順序関係において極大な元をレベル 1 の極大データとよび,極大データを取り除いた集合中での極大データをレベル 2 の極大データとよぶ．これを繰り返すことでレベル k の極大データが定義できる．これに関して以下の手続きが開発できる：

入力 INPUT：
①各データ項目 D_i 上の選好順序 $\leqq_i$ $(1 \leqq i \leqq n)$
②選好順序のレベル k
手続き ORDER_HANTEI:
レベル k の極大データ $x \in D_1 \times \cdots \times D_n$ を求め出力

上記の手続き ORDER_HANTEI を頻出アイテム集合の抽出アルゴリズムに組み込むことにより，与えられた選好順序関係の下で，優先度の低い対象を速やかに除去できるようになる．その骨格は以下のようになり，レベルに応じて可能なデータ融合を生成して，それにもとづくデータマイニングを実行するようになっている：

入力 INPUT：
① 各データ項目 D_i 上の選好順序 $\leqq_i$ $(1 \leqq i \leqq n)$
② 上述の手続き ORDER_HANTEI
手続き 段階的なデータ融合とデータマイニング:
k := 1 /* レベルを 1 に初期設定 */
repeat
レベル k で ORDER_HANTEI 実行；
レベル k の極大データ $x \in D_1 \times \cdots \times D_n$ に対してデータマイニングを実行；
k := k+1;
until 現在のマイニング結果に満足する

(3) 第 3 のアプローチとしては，センサーデータから自動的に取得できるデータを利用する技術を開発した．様々なセンサーが利用可能であるが，本研究ではスマートフォンにも搭載されており，容易に利用できる加速度センサーおよび視線移動データを取得できるアイトラッカーを用いた．これらを用いて,実験的に以下の 2 つのシステムを構築して有効性を実証した．

① 加速度センサーを胸ポケットの位置に装着しデータを収集する（図 4）．このセンサーデータを大量に取得し，その時点での行動状況および行動に対する集中度をデータマイニングにより推定できるようにする．

図4 身体に装着して加速度データの取得

学生を対象として授業中に実施するアンケートデータに対して，このように推定した行動状況と集中度を融合したデータベースの詳細な検討を行った．その結果，アンケート回答の信頼度が行動状況や集中度と密接に関係していることが判明した．そこで相関ルールのデータマイニングにおいて，支持度と確信度に加えて信頼度という新たな指標を導入し，データの不確実性を反映したデータマイニングが可能となる方式の開発と実証に成功した．

② 学生に対して実施する，授業理解度に対するアンケート回答を想定したシステム開発を行った．授業中に閲覧させるプログラムに対して，読解時のアイトラッカーによる視線パターンデータを収集し，先の加速度センサーデータの場合と同様に，データマイニング技術を開発して適用し，プログラムの理解度と視線パターンとの関連性の知識を発見するようにした．この場合にもアンケートの信頼性と理解度との間に深い関係があることが判明したので，支持度と確信度に加えて信頼度という新たな指標を導入し，データの不確実性を反映したデータマイニングが可能となる方式の開発と実証に成功した．

本章で研究の成果を述べたように，研究の目的として設定した 3 通りの観点に対しての方式開発を行うことができ，その有効性も実験により実証することができた．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 ２ 件）

(1) 杉村博，松本一教，アノテーション付き時系列データからの知識発見システムの開発，電気学会論文誌 C 電子・情報・システム部門誌，査読有，Vol.132，No.4，2012，pp. 592-597.
(2) 杉村博，松本一教，Web 情報からの自動タギングのデータマイニングによる応用，電気学会論文誌 C 電子・情報・システム部門誌，査読有，Vol.132，No.4，2012，pp. 623-624.

〔学会発表〕（計 １６ 件）

(1) Hiroto HOSHINO, Daisuke YAGI, Kazunori MATSUMOTO, Towards Automatic Evaluation of Program Understanding Degree using Eye Tracking Patterns,

30th Int'l Conf. on Computers and Their Applications (CATA2015), 2015 March 9 - 11, Honolulu, Hawaii, USA.
(2) Daisuke YAGI, Hiroto HOSHINO, Tetsuo TANAKA, Kazunori MATSUMOTO, Fundamental Study on Effective Design of Digital Documents, Asian Conf. on Information Systems (ASIS2015),2014 December 1-3, Nha Trang, Vietnam.
(3) 八木大介, 松本一教, 視線データを用いた文章強調の効果分析, 人工知能学会知識流通ネットワーク研究会, 2014 年 9 月 26 日, 名古屋大学, 愛知県名古屋市.
(4) Hiroyuki NAKAGAWA, Takumitsu KUDO, Yuichi SEI, Yasuyuki YAHARA, Akihiko OHSUGA, Towards Software Evolution for Embedded Systems Based on MAPE Loop Encapsulation, IEEE 8th Int'l Conf. on Self Adaptive and Self Organizing Systems (SASO2014), 2014 September 8-12,London, UK.
(5) Yuichi SEI, Akihiko OHSUGA, Randomized Addition of Sensitive Attributes for I-diverstity, 11th Int'l Conf. on Security and Cryptography (SECRYPT 2014), 2014 August 28-30, Vienna, Austria.
(6) 星野寛登, 松本一教, 視線移動パターンデータの文書理解への応用, 人工知能学会全国大会, 2014 年 5 月 12 日―15 日, ひめぎんホール, 愛媛県松山市.
(7) Hiroshi SUGIMURA, Masao ISSIKI, Takeaki MORI, Kazunori MATSUMOTO, Building Open HEMS Lifelog over Social Networking Service with Sensor Data, 7th IADIS Int'l Conf. on Information Systems (IADIS IS2014), 2014 February 28 - March 2, Madrid, Spain.
(8) Tomoaki UEDA, Hiroshi SUGIMURA, Tetsuo TANAKA, Kazunori MATSUMOTO, Estimating Degree oc Concentration through Activity Recognition: Use in a Classroom with SNS, 2nd Asian Conf. on Information Systems (ACIS2013), 2013 October 31 - November 2 , Phuket, Thailand.
(9) 植田智明, 杉村博, 松本一教, スマートフォンによる行動推定を用いるライフログシステム設計, 電気学会情報システム研究会, 2013 年 6 月 21 日－22 日, 広島工業大学, 広島県広島市.
(10) 植田智明, 杉村博, 松本一教, 一色正男, センサーデータからの人間の行動推定, 人工知能学会全国大会, 2013 年 6 月 4 日－7 日, 富山国際会議場, 富山県富山市.
(11) Masaki FUJISAWA, Hiroshi SUGIMURA, Kazunori MATSUMOTO, Intelligent Classroom Information System using Sensing Data and Personal Records, 6th IADIS Int'l Conf. on Information Systems (IADIS IS2013), 2013 March 13 - 15, Lisbon, Portugal.
(12) Hiroto HOSHINO, Hiroshi SUGIMURA, Kazunori MATSUMOTO, Time-Series Datamining System with Annotated Information, 6th IADIS Int'l Conf. on Information Systems (IADIS IS2013), 2013 March 13 - 15, Lisbon, Portugal.
(13) Kazunori MATSUMOTO, Hiroshi SUGIMURA, Cooperative Datamining with Numerical Time-Series and Textual Data, Int'l Conf. on Computers and Their Applications (CATA 2013), 2013 March 4-6, Honolulu, Hawaii, USA.
(14) Seiko TAKAMIZAWA, Hiroshi SUGIMURA, Masaki FUJISAWA, Kazunori MATSUMOTO, Towards Authorship Detection based on Datamining, 1st Asian Conf. on Information Systems (ACIS2012), 2012 December 6-8, Siem Rap, Cambodia.
(15) 杉村博, 高見澤聖子, 松本一教, 時系列データマイニングに行動推定技術の開発, 電気学会 情報システム研究会, 2012 年 7 月 19 日－20 日, はこだて未来大学, 北海道函館市.
(16) 杉村博, 松本一教, 背景知識を利用したデータマイニング, 人工知能学会全国大会, 2012 年 6 月 12 日－15 日, 山口県教育会館, 山口県山口市.

〔図書〕(計　1　件)
月本 洋, 松本 一教, やさしい確率・情報・データマイニング 第２版, 森北出版, 2013 年, 167 ページ.

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等
① 神奈川工科大学 情報学部 情報工学科 松本研究室のホームページ
http://laurel.ic.kanagawa-it.ac.jp/
② 電気通信大学 大学院情報システム学研究科 大須賀研究室のホームページ
http://www.ohsuga.is.uec.ac.jp/

６．研究組織
(1)研究代表者
松本 一教 (MATSUMOTO, Kazunori)
神奈川工科大学・情報学部・教授
研究者番号：40350673

(2)研究分担者
大須賀 昭彦 (OHSUGA, Akihiko)
電気通信大学・大学院情報システム学研究科・教授
研究者番号：90393842