2M2-OS-34a-3

コンテクスト検索エンジンを利用した時系列データマイニングの提案

Proposal of Temporal Data Mining Using Context Search Engine

佐藤 宏貴　　高間 康史
Hiroki Sato　　Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

This paper tries to apply data mining methods to temporal data obtained using a context search engine. A context search engine has proposed to access temporal trend information collected from the Web. While its effectiveness as the means for accessing temporal data has been shown, this paper uses it as a resource for temporal data mining. This paper shows some results of temporal data mining, such as clustering of items in terms of cyclic nature, and discovery of items affected by the same event.

1. はじめに

本稿では、「動向に対する問い」を対象タスクとしたコンテクスト検索エンジンを利用した，時系列データのマイニングについて提案する．現在，Web 検索エンジンを用いることで多種多様な情報を発見することができるが，これに伴いにユーザが Web から入手することを期待する情報も多様化している。その結果，既存の Web 検索エンジンが提供する基本検索機能と，ユーザの情報要求との乖離が大きくなっていることが指摘されている．この問題についての解決策の一つとして，コンテクスト検索エンジンが開発されている [高間 15a] ．コンテクスト検索エンジンとは,「動向に対する問い」という特定だが一般にみられる情報要求に検索対象を限定することにより，ドメインを限定せずに既存 Web 検索エンジンよりも高度な検索機能を提供する検索エンジンである．ユーザ実験により，時系列データへの効率的なアクセスが可能であることが示されている．従来の Web 検索エンジンでは，その検索結果として得られるヒット件数や検索回数，スニペットなどを解析することで，有益な情報を取得する研究が行われている [Goel 10]．コンテクスト検索エンジンでも同様のデータマイニングを行うことで，時系列データからの有益な情報・知識の発見が期待できると考える．そこで本稿では，特徴的変動が発生した時期や回数，特徴的変動の組合せなどといった，コンテクスト検索エンジンの検索結果から計算可能な特徴量を新たに提案し，これを用いて時系列データのマイニングを行うことを試みる．

2. 関連研究

2.1 コンテクスト検索エンジン

Web 検索エンジンはインターネット上の膨大な情報から情報を見つける手段として普及している．検索クエリとして指定されたキーワードが含まれるページの中で，関連性の高いものを検索結果として返す，わかりやすい機能により誰でも気軽に利用可能である．しかし，既存の Web 検索エンジンが提供する基本検索機能とユーザの情報要求との乖離が大きく，ユーザの負担が大きくなっている．この問題点の解決策として，検索エンジンの基本機能を見直すアプローチが研究されている．

連絡先: 高間 康史，首都大学東京大学院システムデザイン研究科，〒 191-0065，東京都日野市旭が丘 6-6，ytakama@tmu.ac.jp

ある商品の価格や売上，会社の業績状況等の時系列データを基として，その変化を通時的にとらえつつ総合的にまとめ上げた情報は動向情報と呼ばれ [加藤 04]，世の中の社会活動に深く関わっている．高間らは，この「動向に対する問い」を対象タスクとした検索をコンテクスト検索と定義し，Web 上でコンテンツとして公開されている動向情報と Web 利用に関する動向情報の 2 つを Web から収集し，検索エンジンを構成している．この検索エンジンでは以下の 3 つの基本検索機能を実装している [高間 15b].

1. 指定したアイテムに関する動向が特徴的変動を示した期間の検索
2. 指定した期間に特徴的変動を示したアイテム・動向の検索
3. 指定したアイテムに関する動向が特徴的変動を示した期間に同様の変動を示したアイテム・動向の検索

2.2 検索エンジンを利用したデータマイニング

従来の Web 検索エンジンの検索結果や検索回数を解析することで，新たな情報を発見したり，予測を行う手法が研究されている．Web 検索エンジンの検索結果の情報を用いて予測を行う研究として，Yahoo!における検索回数を素性として映画の初週興行収入やテレビゲームの初週売上，音楽の週間ランキングといった，コンテンツの人気を表すデータの予測を行う手法が提案されている [Goel 10]．Web 上にコンテンツとして公開されている動向情報を用いた研究として，株価や決算等のデータを用いて株価の予測を行う研究がある [植田 07].

3. 提案手法

本研究では，既存 Web 検索エンジンで行われている研究と同様に，コンテクスト検索エンジンの検索結果から得られる情報を用いて情報抽出や予測などを行うことを目的とする．しかし，Web 検索エンジンを利用したマイニングで用いられる素性は検索結果や検索回数，スニペットなどであるが，コンテクスト検索エンジンの API から得ることのできる情報は，アイテム名やリソース名，地域・対象，データの期間，そして変動の種類であり従来の Web 検索エンジンとは大きく異なるため，従来と同じ素性・手法を用いることはできない．また，新たに発見できる情報も従来とは異なると考えられる．そこで本稿ではコンテクスト検索エンジンの検索結果から発見が期待できる情報を検討し，それに適した新たな素性とマイニング手法として以下の 4 点を提案する．

3.1 特徴的変動の組み合わせによる短期的変動の発見

コンテクスト検索エンジンでクエリに指定可能な特徴的変動に，山形 (PEAK)，谷形 (BOTTOM) があり，同一期間に PEAK，BOTTOM があったアイテムは，共通の要因に影響を受けた可能性がある [高間 15b]．しかし，この変動のみでは変動が長期的か，短期的か区別することができないため，急上昇 (SI)，急下降 (SD) という変動を組み合わせることで，短期的な変動を表す新たな特徴的変動 SPEAK，SBOTTOM を提案する．この変動を用いることで，ある事象に短期的な影響を受けた可能性のあるアイテムを発見することが期待できる．

3.2 クラスタリングによる周期性の分析

コンテクスト検索エンジンの検索結果として得られるアイテムには，周期性のあるアイテム，ないアイテムが存在する．周期性のあるアイテムの特徴としては季節や定期的なイベントに影響を受けていることが想定される．共通する周期性を持つアイテムを検索することで，例えば旬が同じ野菜を発見することが可能となる．そこで，周期性のあるアイテムを発見 red するために，各アイテムの各季節における変動の回数を素性としてクラスタリングを行う．

3.3 相関係数による変動要因アイテムの発見

コンテクスト検索エンジンで検索可能なアイテムには，他の動向情報の変動要因となっているアイテムが存在する．例えば，台風の被害によって野菜の価格が高騰する事例は数多く観測されており，台風が農作物の価格に影響を与える一要因となっている．そこで，PEAK，BOTTOM の値を用いて相関係数を求め，他のアイテムの変動要因となるアイテムを発見することを試みる．

3.4 SVM を用いた動向予測

Web 上でコンテンツとして公開されている動向情報には農産物や水産物の価格の動向情報が含まれる．これらについて，過去の動向から将来の動向情報の上昇・下降を，SVM を用いて予測することを試みる．

4. 実験

3 節で述べた 4 つの手法について実験を行った．コンテクスト検索エンジンの検索結果として得られた，2006 年から 2012 年までの 7 年間の PEAK，BOTTOM，SI，SD の特徴的変動をデータセットとして用いる．

4.1 短期的変動の発見

提案手法である SPEAK，SBOTTOM を計算し，他のアイテムに大きな影響を与えそうな事象として，リーマンショック (2008 年 9 月)，東日本大震災 (2011 年 3 月)，ロンドンオリンピック (2012 年 7-8 月) の時期の SPEAK，SBOTTOM について分析した．結果として，東日本大震災に影響を受けたアイテムとして，東京電力やミネラルウォーター，自転車などが当該時期に SPEAK，SBOTTOM を持つことを確認した．

4.2 周期性に基づくアイテムクラスタリング

1 年を 1-3 月，4-6 月，7-9 月，10-12 月の各四半期に分割し，それぞれにおける SI，SD の頻度の合計 8 種類を素性として k-means によりクラスタリングを行った．リソースを「価格」と「Google Trends」に限定してクラスタリングを行った結果，旬が同じアイテムのクラスタが形成された．表 1 は同一クラスタに属するアイテムの一例であり，夏から秋にかけて旬を迎えるアイテムがクラスタを形成していることがわかる．

表 1: 夏から秋にかけて旬を迎えるアイテム (Q:四半期)

アイテム名	リソース	SI1Q	SI2Q	SI3Q	SI4Q	SD1Q	SD2Q	SD3Q	SD4Q
ぶどう	価格	0	10	0	0	0	4	5	0
みかん	価格	0	7	0	0	0	6	7	0
もも	価格	0	2	0	2	0	14	0	0

4.3 変動要因アイテムの発見

台風の検索数に関する動向情報の PEAK の値，および 15 種類の農作物，海産物の旬の時期における BOTTOM の値を取り出したデータセットを作成し，これら 16 アイテム間で相関係数を求めた．その結果，秋に旬を迎えるくりの価格については台風の検索回数との相関係数が高くなった（0.706）．しかし，くりと同様に台風の検索数が多くなる 9 月以降に旬を迎える，ぶどう，みかん，なし，柿の相関係数は低く，高い相関は得られなかった．今回台風の動向情報として検索数を用いたが，実際の被害と検索回数が必ずしも比例関係にないことが原因として考えられる．

4.4 SVM を用いた予測

各アイテムについて，PEAK，BOTTOM の際の値を用いて予測を行うが，コンテクスト検索エンジンの検索結果から得られる情報は特徴的変動が起きた時期のみであり，時系列データとしては欠損値の多い点過程データであるため，欠損値の線形補完を行った．補完可能な 2007 年から 2011 年までの 5 年間のデータを用いて，アイテム数 47 個のデータセットを作成した．前月に比べて価格が上昇したか，下降したかを目的変数として，前 4 年分を訓練データとし残りの 1 年分の予測を行う．入力を予測月の前月，前 2ヶ月，前 3ヶ月の三種類で予測を行ったが，正解率はそれぞれ約 69.0 %，67.7 %，67.9 %となり精度の向上は見られなかった．

5. おわりに

本稿では、「動向に対する問い」を対象タスクとしたコンテクスト検索エンジンを利用した，時系列データのマイニングを行った．様々なマイニング手法を利用した結果，同時期に旬を迎えるアイテムの発見や，ある事象に対して影響を受けたアイテムの発見が可能であることを示したが，精度の点ではまだ課題が残る結果となった．今後は，精度の高い予測やアイテム間の新たな関係性発見を可能とするために，より有効な素性や分析に適したマイニング手法の検討が課題としてあげられる．

参考文献

[高間 15a] 高間 康史, 加藤 優, 桑折 章吾, 石川 博, 動向に関する問いを対象とした検索エンジンの提案, 人工知能学会論文誌, Vol. 30, No. 1, pp. 138-147, 2015.

[Goel 10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J., Predicting consumer behavior with Web Search, in Proceedings of the National Academy of Sciences of the United States of America, Vol.107, No.41, pp.17486-17490, 2010.

[加藤 04] 加藤恒昭, 松下光範, 平尾努, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会研究報告／自然言語処理研究会報告, Vol. 2004, No. 108, pp.88-94, 2004.

[高間 15b] 高間康史, Yanjun Zhu, 桑折章吾, 山口晃一, 瀧口慈勇, 動向に関する問いに答えるコンテクスト検索エンジンの開発, 情報アクセスと可視化マイニング研究会 (第 10 回), SIG-AM-10-02, pp.9-15, 2015.

[植田 07] 植田英三郎, 時系列解析による株価予測, 大阪府立大學經濟研究 53(3), pp.95-111, 2007.