様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　６月　６日現在

機関番号：　２５４０３
研究種目：　若手研究（B）
研究期間：　２０１１ ～ ２０１２
課題番号：　２３７００１２４
研究課題名（和文）
時空間文書ストリーム上における文書データからの知識発見に関する研究
研究課題名（英文） Research on Knowledge Discovery on Documents in Spatiotemporal Document Streams
研究代表者
田村　慶一（TAMURA KEIICHI）
広島市立大学・情報科学研究科・准教授
研究者番号：　８０３４７６１６

研究成果の概要（和文）：
本研究では，時空間文書ストリームから時間と位置に関連した社会的なイベントやホットな話題を抽出するための手法を開発した．時空間文書ストリームの数理モデルを作成し，位置に基づくバースト検出アルゴリズムを用いることで，地域的なイベントや話題を取り出すことができるようなった．また，トピック単位でバースト検出できるようにするために，クラスタリングに基づくバースト検出アルゴリズムを開発した．あわせて，大規模な時空間文書ストリーム上のバースト検出について並列処理による高速化を行った．

研究成果の概要（英文）：
This study has developed a novel method for extracting spatiotemporal social events and hot topics in a spatiotemporal document stream. In this study, a mathematical model for spatiotemporal document stream is defined. Spatiotemporal social events and hot topics can be extracted by using the location-based burst detection algorithm. To extract bursts for topics, the clustering-based burst detection algorithm is proposed. Moreover, the parallelization method for burst detection algorithm, which is developed in this study, archives the speed-up of extracting bursts on the large-scale spatiotemporal document streams.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,900,000	870,000	3,770,000

研究分野：データマイニング，並列処理
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：テキストマイニング，文書ストリーム，情報検索，ソーシャルメディア，並列分散処理

１．研究開始当初の背景

インターネット上でリアルタイムに生成される文書データは爆発的に増加しており，ソーシャルメディアの急速な発達とともに，生成された文書データは集団的な知識（集合知）を有するようになってきている．そこで，代表的なソーシャルメディアである電子掲示板，ブログデータ，wikipedia，twitter に代表されるマイクロブログやカスタマ・レビューなどで生成される文書データを時間とともに到着するストリームデータ（以下，文書ストリームと呼ぶ．）として扱い（図 1），文書ストリームから有益な知識を発見する研究が盛んに行われている．特に，文書ストリーム上の文書データには時々刻々と変化する社会的な事象や話題が記述されており，

文書ストリームから様々な社会的なイベントやホットな話題を抽出する手法が研究されてきた．

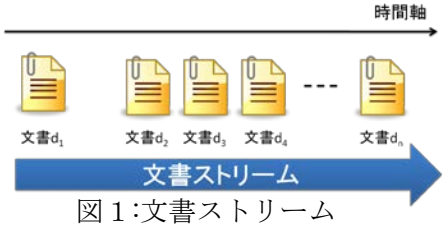

図１:文書ストリーム

図２:時空間文書ストリーム

また，近年，GPS 付き携帯情報端末やスマートフォンなどの普及とともに，インターネット上でリアルタイムに生成される文書データには，文書データが生成された時間だけではなく，文書データが生成された位置に関する情報（位置情報）が付与されるようになってきている．位置情報が付与された文書データの内容は，様々な時間に様々な位置で人々が目にしたことなど，位置に関連する情報と結びついている可能性が高い．そこで，文書データの内容や生成された時間だけでなく，文書データが生成された位置情報も考慮して文書ストリームから有益な知識を発見することが重要となる．

本研究では，時間と位置情報が付与された文書データから構成される文書ストリームのことを時空間文書ストリームと呼ぶこととする．時空間文書ストリーム上では文書データを時間軸だけではなく，座標空間も考慮した時空間上に配置して（図 2），分析をする必要がある．特に，時空間文書ストリーム上の文書データは生成された位置と結びついた事象や話題と関連している可能性が高く，そこに現れてくる時間と位置に関連した社会的なイベントやホットな話題を抽出するための新しい抽出手法が必要となる．

２．研究の目的

本研究では，時空間文書ストリームから時間と位置に関連した社会的なイベントやホットな話題を抽出するための手法を開発することを研究目的とする．時空間文書ストリームから時間と位置に関連した社会的なイベントやホットな話題を抽出することができれば，例えば，GPS 付き携帯情報端末やスマートフォンを持っているユーザに，現在地周辺の社会的なイベントやホットな話題をリアルタイムに提示することが可能となる．

３．研究の方法

研究期間内に，（１）時空間文書ストリームの数理モデルの作成，（２）時空間文書ストリームから社会的なイベントとホットな話題を抽出する手法，（３）効率的な情報検索と情報提示手法，（４）オンラインアルゴリズムと高速化に取り組む．また，数理モデルや手法を検討するだけではなく，実データによる評価を行い，数理モデルや提案手法の優位性を明らかにする．

４．研究成果

（１）時空間文書ストリームの数理モデルの作成，（２）時空間文書ストリームから社会的なイベントとホットな話題を抽出する手法，（３）効率的な情報検索と情報提示手法，（４）オンラインアルゴリズムと高速化に取り組んだ．

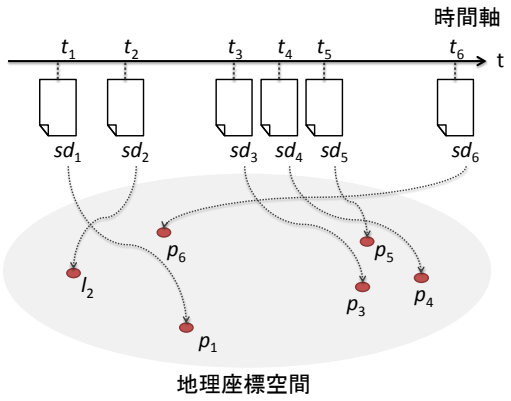

図３： 時空間文書ストリーム

（１） 時空間文書ストリームの数理モデルの作成

文書ストリームのデータモデルを拡張し，時空間文書ストリームのデータモデルを作成した．時空間文書ストリーム上の文書データ sd_i は 3 つのデータ要素 $sd_i = <id_i, text_i, t_i, p_i>$ から構成される．ここで，id_i は当該文書データの識別子であり，$text_i$ は当該文書データの内容（タイトルやテキストデータなど），t_i は当該文書データの生成時刻，p_i は位置情報（経度・緯度）である（図３）．

また，文書データに影響度を規定し，ユーザからの距離が大きくなるほど，指数関数的に影響度が小さくなるモデルを作成した．文書データの影響度は，ユーザの周辺のイベントやトピックを取り出すための基準とすることができる．さらに，ユーザの進行方向も考慮した文書データの影響度を考案した．この影響度は，ユーザの進行方向を考慮して，ユーザの周辺のイベントやトピックを取り出すための基準とすることができる．

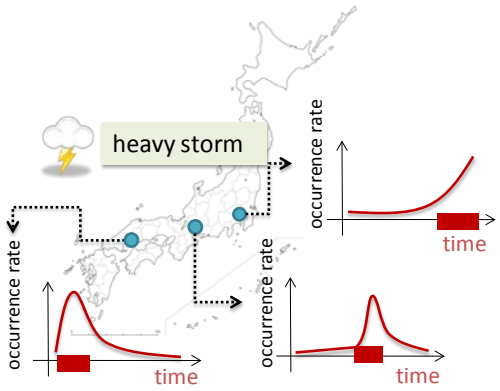

図４： 位置に基づくバースト検出

（２） 時空間文書ストリームから社会的なイベントとホットな話題を抽出する手法

時空間文書ストリームからユーザ周辺に存在するイベントやホットな話題を抽出するために位置に基づくバースト検出アルゴリズムを開発した．バーストとは，ある事象の出現頻度が通常の出現頻度と比較して多く，また，急激に増加している現象のことである．文書ストリーム上においてバーストを検出することで，インターネット上でユーザの関心の高い事象を検出することができる．しかしながら，既存のバースト検出手法はユーザと文書データとの距離関係を考慮していない．位置に基づくバースト検出アルゴリズムでは，（１）で導入した文書データの影響度をバースト検出アルゴリズムに反映させ，ユーザの現在位置により，バーストを変化させることができる．例えば，日本で大雨が降り，各地で大雨が降った時間が異なった場合を考える．図４に示すように，各地で大雨というキーワードがバーストする時間を変化させることができれば，各地で話題となっているキーワードを的確に提示することが可能となる．

また，時空間文書ストリームにおいて，イベントが発生した場所や，話題となっている地域を取り出す手法を開発した．密度に基づく空間クラスタリング手法を（ε， τ）密度に基づく時空間クラスタリング手法として拡張し，時空間文書ストリーム上の文書データをクラスタリングする手法となっている．

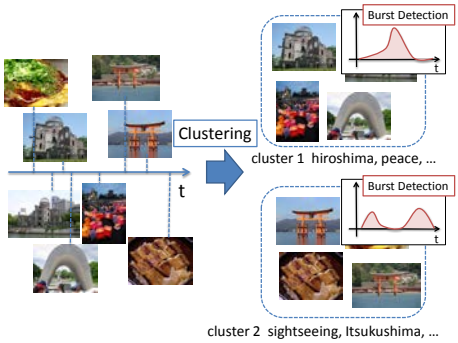

図５：クラスタリングに基づくバースト検出

（３） 効率的な情報検索と情報提示手法

時空間文書ストリーム上において，ユーザが文書データを情報検索しやすくするために，クラスタリングに基づくバースト検出手法，また，クラスタリングの精度を向上するための進化計算手法，検索結果をまとめるための汎化手法の高速化を行った．

クラスタリングに基づくバースト検出手法では，時空間文書ストリーム上の文書データに含まれる語句を用いて，文書データをクラスタリングし，クラスタ毎にバーストを検出することで（図５），トピックやイベントを検出することができる．

クラスタリングに基づくバースト検出手法では，クラスタリングの精度が重要となる．そこで，進化的計算のひとつである改良版の Extremal Optimization(EO)を作成し，島モデルを用い性能向上を図った．研究期間内にクラスタリング手法に組み込むことができなかったが，現在，組み込むための研究を進めている．また，時空間文書ストリーム上の文書デー

タに対して，キーワード検索を行うと大量の類似した文書データが検索結果として得られる．文書データをまとめて分かりやすい形で提示する汎化処理について，その高速化手法を開発した．

（４） オンラインアルゴリズムと高速化

（２）と（３）とで開発を行ったアルゴリズムにおいてバーストを検出するアルゴリズムとして Kleinberg のバースト検出アルゴリズムを使用している．ソーシャルメディアへの関心の高まりとともにインターネット上で生成される文書データは指数関数的に増加している．時空間文書ストリーム上に現れる様々な語句に対して，Kleinberg のバースト検出アルゴリズムを適用すると非常に処理時間を必要とする．文書データ数が増加すると文書ストリーム上に現れる語句数が多くなり，すべての語句に対してバーストを検出するのに処理時間がかかる．また，Kleinberg のバースト検出アルゴリズムでは語句の発生回数×状態数の領域が必要となり，メインメモリ上で検出アルゴリズムを実行できないという課題が分かった．

そこで，Kleinberg のバースト検出アルゴリズムに焦点をあて，大規模時空間文書ストリームを対象としたバースト検出アルゴリズムのマルチコア CPU 上における並列化手法を開発した．具体的には，大規模な大規模時空間文書ストリームにおいて，タスク間並列化にタスク内並列化を併用することで，負荷の偏りを最小限に抑え，大規模なタスクをオンメモリで処理可能なアルゴリズムとなっている．実際のデータを用いて実機上で性能評価を行ったところ，効果的な並列化手法となっていることを確認できた．

５．主な発表論文等

〔雑誌論文〕（計４件）

[1] 中田章宏，田村 慶一，北上 始，高橋誉文：CMO問題に対する改良版EOを用いた発見的解法，情報処理学会論文誌 数理モデル化と応用，査読有，2013 年（掲載決定済）．

[2] Keiichi Tamura，Hajime Kitakami，and Akihiro Nakada：Distributed Modified Extremal Optimization using Island Model for Reducing Crossovers in Reconciliation Graph，Engineering Letters，International Association of Engineers, 査読有, Vol.21，Issue.2，pp.81-88，May 2013.
URL:http://www.engineeringletters.com/issues_v21/issue_2/EL_21_2_05.pdf

[3] Kaishi Hirahara, Keiichi Tamura, Hajime Kitakami, and Shingo Tamura: Parallel Processing of Burst Detection in Large-Scale Document Streams , GSTF Journal on Computing (JoC), 査読有, Vol.2, No.4, 7pages，2013.
URL:
http://dl4.globalstf.org/?wpsc-product=parallel-processing-of-burst-detection-in-large%E2%80%90scale-document-streams-and-its-performance-evaluation

[4] Yagi Shinpei，Keiichi Tamura, and Hajime Kitakami: Parallel processing for stepwise generalisation method on multi-core PC cluster，Special Issue on "Advanced Soft Computing Methodologies and Applications in Web Intelligences," International Journal of Knowledge and Web Intelligence (IJKWI)，Inderscience Publishers，査読有，Vol. 3，No. 2，pp.88-109，2012.
DOI: 10.1504/IJKWI.2012.050282

〔学会発表〕（計１２件）

[1] Yosuke Watanuki, Keiichi Tamura，Hajime Kitakami，Yoshifumi Takahashi: Parallel Processing of Approximate Sequence matching using Disk-based Suffix Tree on Multi-core CPU ，2013 IEEE 6th International Workshop on Computational Intelligence and Applications (IWCIA)，査読有，Hiroshima City University，July 13 2013 (発表決定済).

[2] Tomoki Matsui，Keiichi Tamura，and Hajime Kitakami: Location-based Burst Detection Algorithm for Georeferenced Document Streams based on User's Moving Direction，2013 IEEE 6th International Workshop on Computational Intelligence and Applications (IWCIA)，査読有，Hiroshima City University，July 13 2013 (掲載決定済).

[3] Keiichi Tamura，Hajime Kitakami，and Nakada Akihiro: Distributed Modified Extremal Optimization for Reducing Crossovers in Reconciliation Graph，The 2013 IAENG International Conference on Artificial Intelligence and Applications，査読有，pp.1-6，The Royal Garden Hotel Kowloon，Hong Kong，13-15 March，

2013.
[4] Shingo Tamura, Keiichi Tamura, Hajime Kitakami , and Kaishi Hirahara:Clustering-based Burst-detection Algorithm for Web-image Document Stream on Social Media, The 2012 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2012), 査読有, Seoul in Korea , pp.703-708 , 14-17 October 2012.
[5] Akihiko Nakada, Keiichi Tamura, and Hajime Kitakami:Optimal Protein Structure alignment using Modified ExtremalOptimization , The 2012 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2012) , 査読有, Seoul in Korea, pp.697-702, 14-17 October 2012.
[6] Kaishi Hirahara, Keiichi Tamura,, Hajime Kitakami, and Shingo Tamura: Parallel Processing of Burst Detection in Large-Scale Document Streams, , 3rd Annual International Conference on Advances in Distributed and Parallel Computing, 査読有, pp.60-65, Bali in Indonesia, 17-18 September, 2012.
[7] Keiichi Tamura , and Hajime Kitakami: Location-Based Burst Detection Algorithm in Spatiotemporal Document Stream,The 2012 International Conference on Data Mining (DMIN12), 査読有, Las Vegas, NV, USA, pp.195-201, July 16-19, 2012.
[8] 田村 真吾, 田村 慶一, 北上 始 , 平原 海詞 : 画像付き文書データストリームにおけるバースト検出手法, 2012 IEEE SMC Hiroshima Chapter若手研究会, 査読無, pp.47-50, 広島市立大学, 2012 年 7 月 14 日.
[9] 平原 海詞, 田村 慶一, 北上 始, 田村 真吾 : マルチコアCPU上における文書ストリーム上のバースト検出手法, 2012 IEEE SMC Hiroshima Chapter若手研究会, 査読無, pp.59-62, 広島市立大学, 2012 年 7 月 14 日.
[10] 綿貫 陽介, 田村 慶一, 北上 始, Nguyen Phuong Bac, 高橋 誉文:マルチコアCPU上におけるサフィックス木を用いた文字列検索の並列処理, 2012 IEEE SMC Hiroshima Chapter若手研究会, 査読無, 広島市立大学, 2012 年 7 月 14 日.
[11] 平原 海詞, 田村 慶一, 北上 始 : マルチコアCPU上でのマルチプルアラインメントの並列処理, 2011 IEEE SMC Hiroshima Chapter若手研究会, 査読無, pp.79-82, 広島市立大学, 2011 年 7 月 9 日.
[12] 八木 真平, 田村 慶一, 北上 始 : マルチコア CPU 上での段階的一般化法の並列処理, 2011 IEEE SMC Hiroshima Chapter 若手研究会, 査読無, pp.75-78, 広島市立大学, 2011 年 7 月 9 日.

６．研究組織

(1)研究代表者

田村　慶一（TAMURA KEIICHI）
広島市立大学・情報科学研究科・准教授
研究者番号：80347616

(2)研究分担者

なし

(3)連携研究者

なし