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１０日現在

機関番号：１０１０３
研究種目：若手研究（Ｂ）
研究期間：2011～2012
課題番号：２３７００１２９
研究課題名（和文）　ウェブから時空間依存データを抽出するウェブセンサに関する研究

研究課題名（英文）　A Research on Web Sensors to Extract Spatio-Temporal Data from the Web

研究代表者
服部　峻（HATTORI SHUN）
室蘭工業大学・工学研究科・助教
研究者番号：４０５５５２２３

研究成果の概要（和文）：実世界でスマート空間を実現するには、状況を監視し続け、サービスや構造を最適化する必要がある。しかし、従来の物理的なリアルセンサだけでは、ある場所（空間）ある時間に起きた現象に関して、人々がどのように認識しているか、評判や印象までセンスするのは困難である。そこで本研究では、様々な現象に関して大量のウェブ文書から時空間依存データを抽出するウェブセンサ技術を開発し、その可能性や信頼性を多角的に検証した。

研究成果の概要（英文）：To build Smart Spaces in the real world, they need to keep on sensing their situation and optimize their services and structure. But it is difficult for existing real sensors to physically sense what cognition, reputation, and impression people have on a physical phenomenon's occurring in a place (space) and a time. Therefore, this research has developed Web Sensors to extract spatiotemporal data from a huge amount of Web documents about various physical phenomena, and validated the potential and reliability of the Web-sensed spatiotemporal data multidirectionally.

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	2,600,000	780,000	3,380,000

研究分野：総合領域
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：ウェブマイニング、ウェブセンサ、時空間解析、スマート空間、特異画像検索

１．研究開始当初の背景

かつてのウェブ世界は実世界と互いに疎な関係で独立した存在と言っても過言ではなかったが、ウェブの利用が広く一般の老若男女に普及し、加えてブログや口コミサイト、SNS などが盛んになって来ており、少数のプロの書き手や編集者から成る組織だけでなく、大多数の一般消費者個々人によって、実世界で起きた（今後起こるであろう）様々な現象やイベントに関して、ウェブ世界でもウェブ文書として記述されることが非常に多くなり、ウェブ世界は実世界と互いにより密な関係になって来ていた。

このようなウェブ世界と実世界との関係がより密になるに伴って、ウェブ世界は、ウェブ独自のサービスや活動だけでなく、実世界がどのように変化しているかをモニタするための情報源（センサ）としての側面にも注目されて来ており、実世界での様々な現象に関する知識をウェブから抽出し、可視化するための手法、その知識の活用方法などが盛んに研究されていた。例えば、実世界で提供されている製品やサービスなどの評判抽出、実世界のある場所である期間に味わえる体験（イベント）の抽出およびマッピングなどが提案されていた。ウェブから自動抽出された情報の一部は既にウェブサービスとして一般に提供され始めており、実世界で製品や

サービス、行動を選択する際に多くの一般ユーザが参考にするようになって来ていた。

本研究の代表者も、実世界オブジェクト（語概念）間の上位下位関係や部分全体関係といった階層構造、実世界オブジェクトの外観記述や典型画像などの五感情報などをウェブからマイニングする研究を行っていた。一方で､実空間に様々なIT機器を埋め込み、実世界でサービス配置や構造を最適化し、スマート（セキュア）空間を実現するための研究も行っていた。しかし、従来の物理的なリアルセンサだけでは、ある場所ある期間に起きた様々な現象に関して、その現象を人々がどのように認識しているか、評判や印象までセンスすることは困難である。そこで本研究では、実世界での様々な現象を人々が体験した上で記述されたウェブ文書を活用して、自動的に時空間依存データを高精度に抽出するウェブセンサ技術の開発を行った。

マイニング技術の進歩によりウェブから何らかの知識らしいデータを抽出することは難しくなく、これらのデータを単に見て楽しむだけであるならば特に問題は無いかもしれないが、実世界のセンサとしてのウェブの利用可能性や信頼性などが保証されていないままでは、鵜呑みにしてしまい易い一般ユーザへの提供や、よりクリティカルな情報システムでの活用にも大きな問題がある。そこで本研究では、実世界で起きた現象やイベントについてどの程度正確にウェブ世界で記述されているのかなど、広範で詳細な調査を行うことで、実世界の状況を監視するセンサとしてのウェブの利用可能性や信頼性を科学的に保証することを試みた。

２．研究の目的

本研究の主な目的は、実世界のある場所ある期間に起きた様々な現象に関して人々がどのように認識しているか、その評判や印象など、従来の物理的なリアルセンサでは取得が困難な時空間依存データを、多種多様、玉石混交なウェブ文書から高精度に抽出するウェブセンサ技術を開発することであった。また、ウェブセンサで抽出した時空間依存データの利用可能性や信頼性について詳細な検証を行うことで、これらを担保することも目的の一つであった。

ウェブから抽出された時空間依存データの検証内容としては、実世界で起きた（今後起こるであろう）様々な現象を取り上げ、総じて、ウェブ世界でどのくらい（量的）、どのように（質的）記述されているのか、その対応関係、一致性などについて調査を行い明らかにすることであった。また、実世界での現象の種類に依ってウェブでの記述され易さ、タイムラグに違いがあるのか、その現象が起きた時空間に依って異なるのかなど、より詳細な分析を行うこと、ウェブ世界は実世界でのどのような種類の現象をモニタするのが得意なのか、一方、どのような種類の現象についてはデータが不足したり遅れ過ぎるなど信頼性に欠けるのかを明らかにすることであった。さらに、ウェブ世界には多様なメディアがあり、従来のウェブ文書、ニュース記事、ブログ、マイクロブログ、SNS など、情報源（コーパス）の種類に依って、抽出された時空間依存データと実世界との対応関係に違いがあるのかなど、様々な側面からウェブのリアルセンサ代替可能性や信頼性を多角的に調査する必要があった。

３．研究の方法

爆発的に増大して行くウェブから大量のウェブ文書を自動的に収集・格納するクローリング機能、これらのウェブ文書群をリアルタイム解析して実世界上のある地理的位置である期間に起きた現象に関するデータ（時空間依存データ）を抽出する機能、リアルセンサデータとの相関を計算する機能などを備える、ウェブから時空間依存データを抽出するウェブセンサ基盤システムを構築した。その上で、実世界での様々な現象に関する統計情報を幅広く収集し、各現象についてウェブセンサで抽出した時空間依存データとリアルセンサデータとの相関などを計算することで、どのような種類の現象については相関が高いか、どのような抽出手法や情報源（コーパス）が適しているかなど、利用可能性や信頼性に関して多角的に検証した。また、ウェブから時空間依存データを高精度に抽出する手法の改善、実世界の現象やオブジェクトに関してテキストデータだけでなく画像データをウェブセンスする技術の改善、ウェブセンサデータの実空間（セキュア空間）への活用方法などについても検討を行った。

(1) ウェブセンサのプロトタイプ開発

日々爆発的に増大して行くウェブ世界、大量のウェブ文書群から時空間依存データを自動的に抽出するウェブセンサのプロトタイプシステムを構築した。大量のウェブ文書を自動的に収集・格納するクローリング機能、これらのウェブ文書群をリアルタイム解析して実世界上のある地理的位置である期間に起きた現象に関するデータ（時空間依存データ）を抽出する機能などが必要であった。まず最も単純な語共起頻度に基づく抽出手法を実装した後、空間的シフト、空間的伝播、線形結合などによる改善を考案した。

(2) リアルセンサとの相関調査

地震、降雨、降雪、気温、インフルエンザ、交通事故など、実世界における様々な現象について、多種多様な抽出手法を実装したウェ

ブセンサによってウェブから抽出した時空間依存データと、実世界のリアルセンサによる時空間依存データ（気象庁などの統計）との相関を計算し、ウェブセンサの利用可能性や信頼性を多角的に検証した。ウェブ世界でどのくらい（量的）、どのように（質的）記述されているのか、その対応関係、また、実世界での現象の種類に依ってウェブでの記述され易さ、タイムラグに違いがあるのか、その現象が起きた時空間に依って異なるのか、現象の種類への依存性、従来のウェブ文書、ニュース記事、ブログ記事、SNS など、ウェブセンサの情報源（コーパス）の種類への依存性などについて詳細な調査を行った。

(3)ウェブセンサによる画像データ抽出

ウェブセンサによってウェブから時空間依存のテキスト（数値）データを抽出するだけでなく、時空間依存の画像データを高精度に、効率的に収集する技術の開発に取り組んだ。まず、実世界オブジェクトの典型画像ではなく特異画像をウェブ検索する手法を考案し、クロス言語（機械翻訳）や語概念階層による改善も行った。また、ある地域（空間）、ある時間において典型的な、あるいは、特異な画像を検索する手法の検討も行った。

(4)ウェブセンサデータの実空間への活用

信頼性（相関）の高い現象に関するウェブセンサの時空間依存データはどのようにユーザに呈示すればより良いか、信頼性（相関）の低い現象に関する時空間依存データはどのようにユーザに呈示すれば誤解の問題などを軽減できるかなど、ウェブセンサデータを実世界サービスや実空間へ活用する方法について検討した。また、ウェブセンサで抽出された知識データ（潜在的なニーズなど）に応じて、リアルタイムに実世界に反映し、実空間（セキュア空間）でのサービス配置や構造を適応させる研究にも取り組んだ。

４．研究成果

実世界で様々な現象を人々が体験した上で記述された大量のウェブ文書から、時空間依存データ（実世界上のある地理的位置である期間に起きた現象に関するデータ）を高精度に抽出するウェブセンサシステムの基盤技術を構築し、ウェブセンサによって抽出した時空間依存データの利用可能性や信頼性を担保するため、様々な現象に関してリアルセンサとの相関など多角的な検証を行った。

(1)ウェブセンサの多角的検証

実世界の様々な現象に関して、多種多様な抽出手法を実装したウェブセンサと、実世界のリアルセンサとの相関などを計算し、ウェブセンサでウェブから抽出した時空間依存データの利用可能性や信頼性について多角的な検証を行った。

①現象や情報源の種類への依存性

時空間依存データの抽出対象である実世界の現象の種類や、ウェブセンサの情報源（コーパス）の種類に依って、ウェブセンサによってウェブから抽出した時空間依存データと、実世界のリアルセンサによる時空間依存データとの相関の強弱の違いを調査した。表１は、実世界の現象の種類として有感地震回数、降雨量、降雪量の３種類、ウェブセンサの情報源（コーパス）の種類として一般ウェブ文書、ニュース記事、ブログ記事、マイクロブログである Twitter のツイート、ウェブ文書を検索する条件である検索クエリのログの５種類についてクロス分析した結果である。但し、空間は 47 都道府県、時間は 2011 年の 52 週間である。

ウェブ世界でウェブ文書を投稿（作成）するアクションよりも、ウェブ世界でウェブ文書を検索するアクションをウェブセンサの情報源（コーパス）とする方が概ね強い相関が得られることが明らかになった。また、ニュース記事や検索クエリのログではどの現象に対しても概ね一様な相関が得られるが、一般ウェブ文書やブログ記事では地震が不得意、Twitter のツイートでは降雨が不得意など、情報源（コーパス）の種類に依って、得意・不得意な現象に大きな違いも見られる。従って、ウェブセンサで抽出した時空間依存データを実世界サービスや時空間へ活用する場合、目的の現象の種類に依って適切な情報源（コーパス）を選択する必要がある。

表１　現象や情報源の種類への依存性

	地震数	降雨量	降雪量
一般ウェブ	0.08377	0.28348	0.24496
ニュース	0.30187	0.27939	0.31399
ブログ記事	0.36803	0.52155	0.54296
Twitter	0.40375	0.19275	0.35212
検索クエリ	0.54341	0.45692	0.60143

②タイムラグ（時間的なズレ）の差異

抽出対象の現象を表すキーワードの語共起頻度に基づくだけでなく、時間的シフトも組み込んだウェブセンサによって、ある現象が実世界で発生してから、その現象に関してウェブ世界でウェブ文書の投稿や検索などのアクションが行われるまでの時間的なズレを調査した。その結果、降雨に対しては時間的なズレはないが、予報が充実している降雪に対しては実世界での発生よりもウェブ世界でのアクションの方が早い。一方、図 1 のように、予報が困難な地震に対しては実世界での発生よりもウェブ世界でのアクションの方が遅いという結果が得られた。

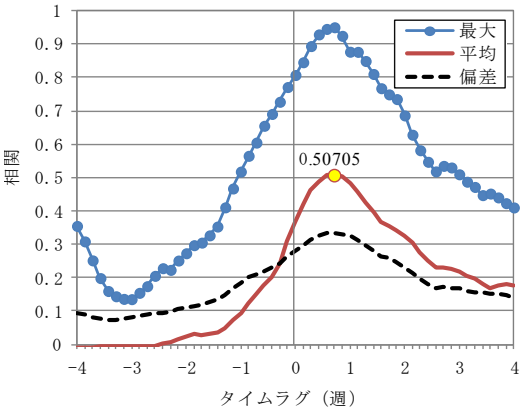

図１　地震に対するタイムラグ（遅延）

③ウェブセンサの信頼性の空間依存性

ウェブセンサによってウェブから抽出した時空間依存データの信頼性（リアルセンサとの相関の強弱）が、空間（47 都道府県）に依ってどのように異なるかを調査した。その結果、台風やゲリラ豪雨を除いて広域で概ね一様に共有されることが多い現象である降雨に対しては一様な空間的分布が得られたが、より局所的な現象である降雪や、ある地震源で発生した後、減衰しながら空間的に伝播して行く地震に対しては、その現象（のインパクト）がより大きかった空間ほど、リアルセンサとの間に強い相関が得られるという傾向が見られた。図 2 は、2011 年に起きた地震現象（日毎）に対して、ウェブセンサの時空間依存データの相関の空間的濃淡分布、及び、2011 年 3 月 11 日に発生した東日本大震災の地震源を示している。

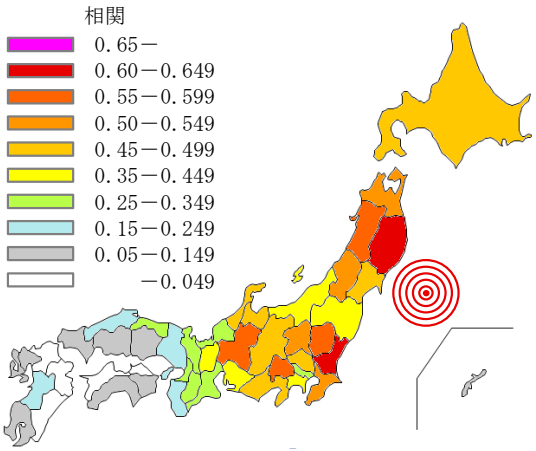

図２　地震に対する空間依存性

④ウェブセンサ手法の線形結合による改善

地震という現象に対して、「地震」という単一のキーワードと時空間との語共起頻度に基づくウェブセンサよりも、より粒度の細かい「大地震」というキーワードの語共起頻度と線形結合することで改善することが出来た。一方で、ウェブ世界に投稿されたブログ記事などの単一の情報源（コーパス）だけに基づくウェブセンサよりも、ウェブ世界でウェブ文書を検索する条件である検索クエリのログに基づくウェブセンサと線形結合させることで改善させられることが明らかになった。今後の研究課題としては、線形結合パラメータの最適化などが挙げられる。

(2)実世界オブジェクトの特異画像検索

ウェブセンサによって時空間依存のテキスト（数値）データを抽出するだけでなく、時空間依存の画像データを高精度に、効率的に収集する技術の一つとして、実世界オブジェクトの典型画像ではなく特異画像をウェブ検索する手法を開発した。また、クロス言語（機械翻訳）や語概念階層による改善も行った。図 3 は、赤白の東京タワーという一般的な画像ではなく、ピンク色や緑色、青色などでライトアップされた特異な画像を検索できている。今後の研究課題としては、ある地域（空間）、ある時間における典型画像や特異画像を検索する手法が挙げられる。

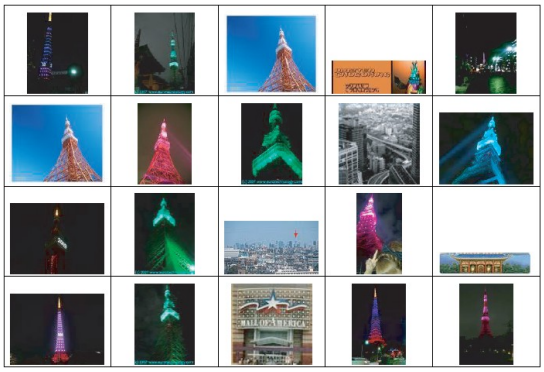

図３　「東京タワー」の特異画像検索

(3) ウェブセンサとセキュア空間の連携

図 4 のように、IT 機器やリアルセンサが埋め込まれた実空間（セキュア空間）との連携として、ウェブセンサで抽出した知識データを活用したコンテキスト・アウェアな検索クエリ制御や代替クエリ発見を考案した。

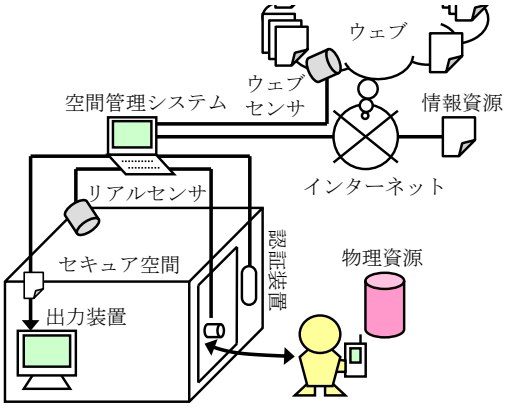

図４　セキュア空間におけるウェブセンサ

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５件）
(1) Shun Hattori、Spatio-Temporal Web Sensors Using Web Queries vs. Documents、Journal of Automation and Control Engineering (JOACE)、査読有、Vol.1、No.3、2013、pp.192-197
DOI: 10.12720/joace.1.3.192-197

(2) Shun Hattori、Hyponymy-Based Peculiar Image Retrieval、International Journal of Computer Information Systems and Industrial Management (IJCISIM)、査読有、Vol.5、2013、pp.79-88
http://www.mirlabs.org/ijcisim/regular_papers_2013/Paper83.pdf

(3) Shun Hattori、Context-Aware Query Control for Secure Spaces、Journal of Computer Technology and Application (JCTA)、査読有、Vol.3、No.2、2012、pp.130-139
http://www.davidpublishing.org/DownLoad/?id=4314

(4) Shun Hattori、Peculiar Image Retrieval by Cross-Language Web-extracted Appearance Descriptions、International Journal of Computer Information Systems and Industrial Management (IJCISIM)、査読有、Vol.4、2012、pp.486-495
http://www.mirlabs.org/ijcisim/regular_papers_2012/Paper53.pdf

(5) Shun Hattori、Secure Spaces and Spatio-temporal Weblog Sensors with Temporal Shift and Propagation、Recent Progress in Data Engineering and Internet Technology、査読有、LNEE Vol.157、2012、pp.343-349
DOI: 10.1007/978-3-642-28798-5_46

〔学会発表〕（計10件）
(1) Shun Hattori、Granularity Analysis for Spatio-Temporal Web Sensors、the WASET International Conference on Knowledge Management (ICKM'13), 2013年2月15日、クアラルンプール（マレーシア）

(2) Shun Hattori、Ability-Based Expression Control for Secure Spaces、the Joint 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on advanced Intelligent Systems (SCIS&ISIS'12)、2012年11月23日、神戸

(3) Shun Hattori、Hyponym Extraction from the Web based on Property Inheritance of Text and Image Features、the Sixth IARIA International Conference on Advances in Semantic Processing (SEMAPRO'12)、2012年9月27日、バルセロナ（スペイン）

(4) Shun Hattori、Spatio-Temporal Web Sensors by Social Network Analysis、the 3rd International Workshop on Business Applications of Social Network Analysis (BASNA'12)、2012年8月26日、イスタンブール（トルコ）

(5) Shun Hattori、Linearly-Combined Web Sensors for Spatio-Temporal Data Extraction from the Web、the 6th International Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'11)、2011年12月11日、バンクーバー（カナダ）

(6) Shun Hattori、Query Expansion for Peculiar Images by Web-extracted Hyponyms、the Fifth IARIA International Conference on Advances in Semantic Processing (SEMAPRO'11)、2011年11月23日、リスボン（ポルトガル）

(7) Shun Hattori、Searching the Web for Peculiar Images based on Hand-made Concept Hierarchies、the Seventh International Conference on Next Generation Web Services Practices (NWeSP'11)、2011年10月20日、サラマンカ（スペイン）

(8) Shun Hattori、Alternative Query Discovery from the Web for Daily Mobile Decision Support、the 5th IADIS International Conference on Wireless Applications and Computing (WAC'11)、2011年7月20日、ローマ（イタリア）

(9) 服部峻、地震統計との相関に見るウェブセンサの可能性、第1回 テキストマイニング・シンポジウム、2011年7月8日、東京

６．研究組織
(1)研究代表者
服部　峻（HATTORI SHUN）
室蘭工業大学・工学研究科・助教
研究者番号：４０５５５２２３