Vol.2010-DBS-151 No.4
2010/11/12

気象統計との相関に見るWebセンサの可能性

服部　峻[†1]

実世界で起きた（る）様々な現象や事象に関する知識データを日々情報爆発し続ける Web から，特に Web ブログなどの CGM からマイニングする研究が盛んに行われ，一般向けへのサービス化も試行され始めている．しかしながら，大量の Web 文書からマイニングされたデータが，実世界でのリアルなデータをどの程度正確に反映しているかの詳細な調査は見当たらず，精度が十分に保証されていないまま盲目的に利用するのは問題があると考える．そこで本論文では，実世界で起きた現象のリアルなデータとして気象庁の気象統計情報から気温および降水量を用い，Web から抽出した時空間依存データとの相関を評価することで，Web センサによってマイニングされたデータの利用可能性や信頼性を検証する．

The Potential of Web Sensors in Correlation with Weather Statistics

Shun Hattori[†1]

Many researches on mining the explosively-growing Web, especially CGM (Consumer Generated Media) such as Weblogs, for knowledge about various phenomena and events in the real world have been done actively, and Web services with the Web-mined knowledge have begun to be developed for the public. However, there is no detailed investigation on how accurately Web-mined data reflect real-world data. It must be problematic to idolatrously utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently. Therefore, this paper tries to validate the potential and reliability of Web sensor's spatio-temporal data by measuring the correlation with weather (temperature and precipitation) statistics of Japan Meteorological Agency as real-world data.

†1 東京工科大学コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology

1. はじめに

創世期の Web 世界は実世界とは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，Web の利用が広く一般の老若男女に普及し，加えて Web ブログや口コミサイト，ソーシャルネットワーキングサービスといった CGM (Consumer Generated Media) が非常に盛んになって来ており，少数精鋭のプロの書き手や編集者から成る新聞社などの組織だけでなく，大多数の一般消費者個々人によって，実世界で実際に起きた，または，今後起こるであろう様々な現象やイベントに関して，Web 文書として記述されることが多くなり，今日の Web 世界は実世界と互いにより密な関係になって来ている．

このように Web 世界と実世界との関係がより密になるに伴って，今日の Web 世界は Web 独自のサービスや活動の場としてだけでなく，実世界でどのような現象やイベントが起きているのか，どのように変化しているのかを監視するための情報源（センサ）としての側面も注目されており，実世界での様々な現象に関する知識を Web 世界から抽出・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る．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[1)]，実世界のある場所である期間に味わうことができる体験（イベント）のマイニング[2)] などが提案されている．他にも，語概念の階層構造 (is-a/has-a 関係など）の抽出[3)]，実世界オブジェクトの外観などの五感情報の抽出[4),5)] など，実世界の様々な事象に関する知識を日々情報爆発し続ける Web から，特に Web ブログなどの CGM からマイニングする研究が盛んに行われている．これらの Web から抽出された情報の一部は既に Web サービスとして一般にも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もなって来ている．

しかしながら，実世界で実際に起きた，または，今後起こるであろう現象や事象（イベント）について，どの程度正確に Web 世界に Web 文書として記述されているのか，実世界をモニターするための情報源（センサ）としての Web の利用可能性や信頼性などについて，詳細な調査は未だ不十分であると考える．テキストマイニング技術の進歩により Web から何らかの知識らしいデータを抽出することは全く難しくなく，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 Web 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，実世界で製品やサービス，行動を選択する際に，これらのデータを本当に参考にしても良いのかは非常に怪しく，よりクリティカルなシステムへの活用には大きな問題があると考える．Web 世界の実態に詳しく「そもそも Web は疑わしいもの」と認識している（研究）者であれば Web から抽出されたデー

タを過信（盲信）することは無いかもしれないが，鵜呑みにしてしまう一般ユーザは少なくないであろう．

Web から何らかの知識データを抽出する手法の研究やサービスの開発は既に多種多様に行われているが，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や事象に関するデータ）を Web から抽出する手法（本論文では Web センサと呼ぶ）の研究は未だ少なく，Web センサによって Web 抽出された時空間依存データの利用可能性や信頼性についての詳細な検証は見たらない．そこで，実世界の様々な現象や事象に関して，実世界に設置されたリアルセンサによって収集された統計データと多角的に照合および考察することで，Web センサによって Web 抽出された時空間依存データの利用可能性や信頼性を検証することは（社会的にも）非常に重要であると考える．本調査により，Web センサによってマイニングされたデータの利用可能性や信頼性が保証されれば，一般ユーザが日常生活において製品やサービス，行動を選択する際に誰もがより安心して気軽に参考にすることができるようになり，同時に，Web センサのデータを利活用した応用システムもより拡充されて行くと考える．

本論文では，実世界で実際に起きた現象のリアルなデータとして，全国各地に観測所があり，長期間に亘り，公式データとして公開されている気象庁の気象統計情報[6] から気温および降水量を用いる．従って，まず，実世界の気温や降水量に関する時空間依存データを Web 抽出する Web センサを構成する必要がある．Web から何らかの知識データを抽出する手法は様々に提案されているが，Web 検索エンジンのインデックス情報（検索件数やスニペットなど）を用い，特定の時空間で制限した Web 文書の中で「暑い」や「雨」といった言葉を含む頻度に基づいて Web センサを構成する．その上で，実世界の気温や降水量に関して Web 世界から抽出した時空間依存データと気象庁の気象統計との相関を多角的に評価することで，Web センサによってマイニングされたデータの利用可能性や信頼性を検証する．一般の Web 文書と Web ブログとで，実世界の現象（気温変化や降水量）に依って，実世界の現象が起きた空間（都道府県）や時間（月毎）の違いに依って，気象庁の気象統計との相関に差が見られるかについても考察を行う．

本論文の以下の構成を示す．まず，2 章では，実世界の気温および降水量に関する時空間依存データを Web 世界から抽出する Web センサの構成方法を示す．次に，3 章では，2 章で構成した Web センサによって Web 抽出した時空間依存データと，気象庁の気象統計情報の気温および降水量との相関を多角的に評価・考察する．最後に，4 章で本論文をまとめ，今後の課題についても述べる．

2. 気温および降水量に関する時空間依存データを抽出する Web センサ

次章では，Web センサによって Web 抽出した時空間依存データと，気象庁の気象統計情報の気温および降水量との相関を多角的に評価・考察する．従って，実世界の現象として気温および降水量に関する時空間依存データを抽出する Web センサを構成する必要がある．Web から何らかの知識らしいデータを抽出する手法は様々に提案されているが，本論文では，Web 検索エンジンのインデックス情報の一つである検索件数だけを用い，特定の時空間で制限した Web 文書の中で，対象の現象（気温や降水量）を連想させる「暑い」や「雨」といった言葉を含む文書頻度に基づいて Web センサを構成する．

まず，検索対象の Web 文書を特定の時間および空間で制限する必要がある．時間制約に関しては，Google ウェブ検索や Google ブログ検索がサポートしている期間指定オプションを利用する．但し，この期間指定オプションは Web 文書の最終更新日（クロールされた日）に基づいて制限するため，必ずしも真に必要な特定の時間に起きた実世界の現象について記述された Web 文書であるとは限らない．このような Web 文書のメタデータとしての日付ではなく，Web 文書の内容に含まれる日付表現を切り出す方法も考えられるが，日付表現は多様であり，検索件数を求めるために検索クエリを構成するのが容易ではないため，本論文では Google の期間指定オプションを利用している．例えば，「2000 年 1 月」という期間だけに Web 文書を限定するためには，「2000/1/1」から「2000/1/31」までという期間指定オプションを設定する．空間制約に関しては，Google の期間指定オプションのように Web 文書のメタデータを用いて制限することができないため，特定の空間（地域）表現を内容に含むか否かで Web 文書を制限する．例えば，「東京」という地域だけに Web 文書を限定するためには，「東京」という言葉を検索クエリ自体に含める．

次に，時間 t および空間 s における実世界の現象「気温」に関する Web センサが出力する時空間依存データ（数値）を，「暑い」という言葉の Web 文書の頻度で定義する．

$$\text{WebSensor-Temperature}(t,s) := \text{wf}_t(\texttt{[“暑い” AND “}s\texttt{”]})$$

但し，$\text{wf}_t(\texttt{[}q\texttt{]})$ は，時間 t で期間指定オプションした上で検索クエリ q で Google ウェブ検索した結果の Web 文書の検索件数を表す．同様に，一般の Web 文書ではなく，Web ブログだけを用いた Weblog センサを以下のように定義する．

$$\text{WeblogSensor-Temperature}(t,s) := \text{bf}_t(\texttt{[“暑い” AND “}s\texttt{”]})$$

但し，$\text{bf}_t(\texttt{[}q\texttt{]})$ は，時間 t で期間指定オプションした上で検索クエリ q で Google ブログ検索した結果の Web ブログの検索件数を表す．

3. 気象統計との相関の評価

本章では，前章で構成したWebセンサによってWeb抽出した時空間依存データと，気象庁の気象統計情報の気温および降水量との相関を多角的に評価することで，Webセンサによってマイニングされたデータの利用可能性や信頼性を検証する．一般のWeb文書とWebブログとの違いに依って，実世界の現象（気温や降水量）の違いに依って，実世界の現象が起きた空間（都道府県）や時間（月毎）の違いに依って，気象庁の気象統計との相関に差が見られるかについて考察して行く．

時間制約としては,「2000年1月」から「2010年5月」までの月毎，全部で125区間を用いている．空間制約としては，47都道府県の気象台がある場所（大部分は県庁所在地）の名称を用いている．つまり，実世界で実際に起きた現象のリアルなデータとしては，各気象台で観測された月毎の気温および降水量データを用い，一方，Webセンサの数値データとしては，各期間でオプション指定した上で「暑さ」や「雨」などを各気象台の名称で拡張した検索クエリで検索したWeb文書やWebブログの頻度を用いている．

図 1 は,「暑い」という言葉のWeb文書頻度 wf_t(["暑い" AND "s"]) に基づくWebセンサの数値データと気象庁の気温統計データとの各年月から「2010年5月」までの大局的な相関係数について，47都道府県の気象台の最大値，平均値，最小値の推移を示している．より過去の年月から「2010年5月」までの相関係数は正の相関が多少見られる程度であるが，より直近の年月から「2010年5月」までだけの相関係数を見るとより強い正の相関があり,「暑い」という言葉のWeb文書頻度に基づいてWeb抽出されたWebセンサの数値データは，実世界の現象「気温」のリアルセンサの統計データをある程度反映していることが分かる．逆に，**図 2** は,「暑い」の反対語である「寒い」という言葉のWeb文書頻度 wf_t(["寒い" AND "s"]) に基づくWebセンサの数値データと気象庁の気温統計データとの各年月から「2010年5月」までの大局的な相関係数について，47都道府県の気象台の最大値，平均値，最小値の推移を示している．より過去の年月から「2010年5月」までの相関係数は負の相関が多少見られる程度であるが，より直近の年月から「2010年5月」までだけの相関係数を見るとより強い負の相関があり,「寒い」という言葉のWeb文書頻度に基づいてWeb抽出されたWebセンサの数値データも逆に使えば，実世界の現象「気温」のリアルセンサの統計データをある程度反映していることが分かる．

図 3 および**図 4** は,「暑い」および「寒い」という言葉のWeb文書頻度ではなくWebブログ頻度 bf_t(["暑い" AND "s"]) および bf_t(["寒い" AND "s"]) に基づくWeblogセンサの数値データと気象庁の気温統計データとの各年月から「2010年5月」までの大局的な相関係数について，47都道府県の気象台の最大値，平均値，最小値の推移をそれぞれ示している．Web文書頻度に基づくWebセンサの図1および図2と比べて，標準偏差を悪化させることなく，Webブログ頻度に基づくWebセンサの図3および図4の方がより強い正の相関および負の相関をそれぞれ示していることが分かる．従って，少なくとも実世界の様々な現象の中で「気温」に関しては，一般のWeb文書よりもWebブログに制限して抽出した数値データの方が，実世界のリアルな統計データとより強い相関があると言える．また，標準偏差があまり大きくないので，実世界の現象「気温」が起きた空間（都道府県）の違いに対する依存性は小さいとも言える．

図 5 から**図 8** は,「晴れ」や「雨」という言葉のWeb文書頻度に基づくWebセンサやWebブログ頻度に基づくWeblogセンサの数値データと気象庁の気温統計データとの各年月から「2010年5月」までの大局的な相関係数について，47都道府県の気象台の最大値，平均値，最小値の推移をそれぞれ示している.「気温」を連想させる「暑い」や「寒い」といった言葉と異なり,「気温」を必ずしも連想させない「雨」や「晴れ」といった言葉を用いて，実世界の現象「気温」に関する時空間依存データを抽出するWebセンサを構成した場合，やはり，実世界のリアルな統計データとほとんど相関を示さないことが分かる．

一方，**図 9** から**図 16** は,「暑い」「寒い」「晴れ」「雨」という言葉のWeb文書頻度に基づくWebセンサやWebブログ頻度に基づくWeblogセンサの数値データと気象庁の気温統計データとの各年月から1年（12ヶ月）分の局所的な相関係数について，47都道府県の気象台の最大値，平均値，最小値の推移をそれぞれ示している．図1や図3のように大域的に見ると，実世界のリアルな統計データとやや強い正の相関を示していた「暑い」という言葉のWeb文書頻度に基づくWebセンサやWebブログ頻度に基づくWeblogセンサであっても，図9や**図 11** のように局所的に見ると，逆に負の相関を示してしまっている期間が存在する．図2や図4のように大域的に見ると，実世界のリアルな統計データとやや強い負の相関を示していた「寒い」という言葉のWeb文書頻度に基づくWebセンサやWebブログ頻度に基づくWeblogセンサであっても，**図 10** や**図 12** のように局所的に見ると，かなり激しく振動し，逆に正の相関を示してしまっている期間も存在する．従って，実世界の現象「気温」が起きた時間（月毎）に対する依存性は小さくないと言える．時間依存性を生んだ一つの要因としては，Webセンサを構成するために用いたGoogle検索エンジンの期間指定オプションが必ずしも正確に機能していないと考えられる．特に，より過去の期間において，Web文書のメタデータ（最終更新日）の信頼性が乏しいように考えられる．

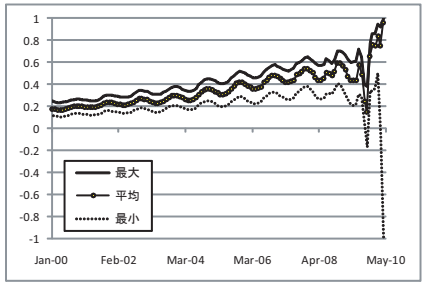

図 1 wf_t(["暑い" AND "s"]) に基づく Web センサと気温統計との大域的な相関係数
Fig. 1 Global Correlation between Temperature and Web Sensors by wf_t(["atsui" AND "s"])

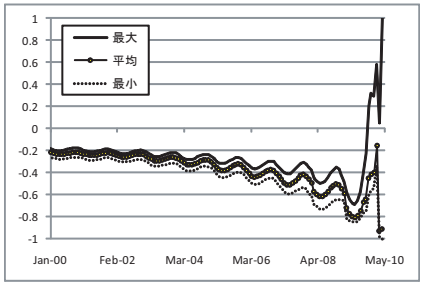

図 2 wf_t(["寒い" AND "s"]) に基づく Web センサと気温統計との大域的な相関係数
Fig. 2 Global Correlation between Temperature and Web Sensors by wf_t(["samui" AND "s"])

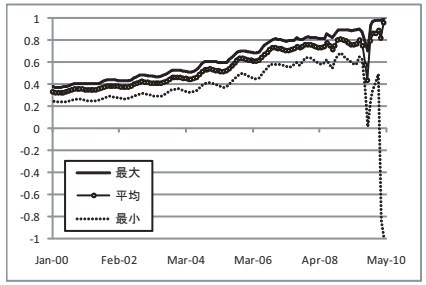

図 3 bf_t(["暑い" AND "s"]) に基づく Weblog センサと気温統計との大域的な相関係数
Fig. 3 Global Correlation between Temperature and Weblog Sensors by bf_t(["atsui" AND "s"])

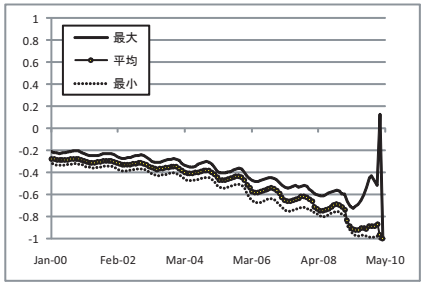

図 4 bf_t(["寒い" AND "s"]) に基づく Weblog センサと気温統計との大域的な相関係数
Fig. 4 Global Correlation between Temperature and Weblog Sensors by bf_t(["samui" AND "s"])

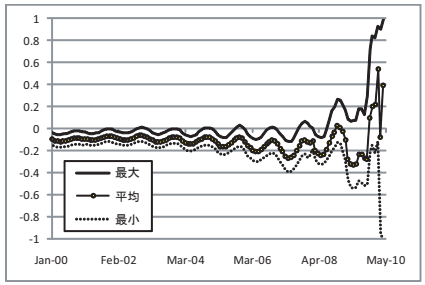

図 5 wf_t(["晴れ" AND "s"]) に基づく Web センサと気温統計との大域的な相関係数
Fig. 5 Global Correlation between Temperature and Web Sensors by wf_t(["hare" AND "s"])

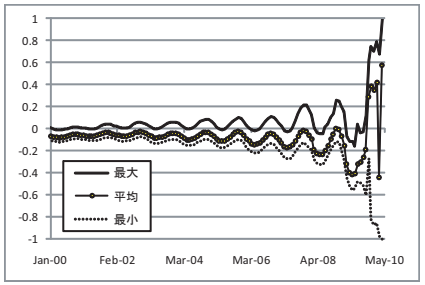

図 6 wf_t(["雨" AND "s"]) に基づく Web センサと気温統計との大域的な相関係数
Fig. 6 Global Correlation between Temperature and Web Sensors by wf_t(["ame" AND "s"])

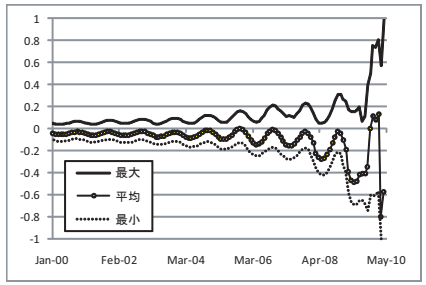

図 7 bf_t(["晴れ" AND "s"]) に基づく Weblog センサと気温統計との大域的な相関係数
Fig. 7 Global Correlation between Temperature and Weblog Sensors by bf_t(["hare" AND "s"])

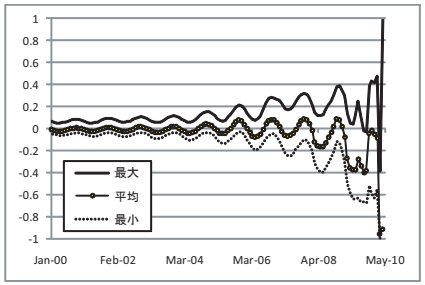

図 8 bf_t(["雨" AND "s"]) に基づく Weblog センサと気温統計との大域的な相関係数
Fig. 8 Global Correlation between Temperature and Weblog Sensors by bf_t(["ame" AND "s"])

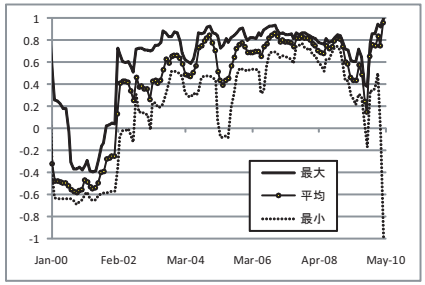

図 9 wf_t([“暑い” AND “s”]) に基づく Web センサと気温統計との局所的な相関係数

Fig. 9 Local Correlation between Temperature and Web Sensors by wf_t([“atsui” AND “s”])

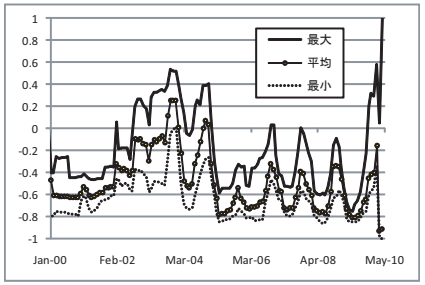

図 10 wf_t([“寒い” AND “s”]) に基づく Web センサと気温統計との局所的な相関係数

Fig. 10 Local Correlation between Temperature and Web Sensors by wf_t([“samui” AND “s”])

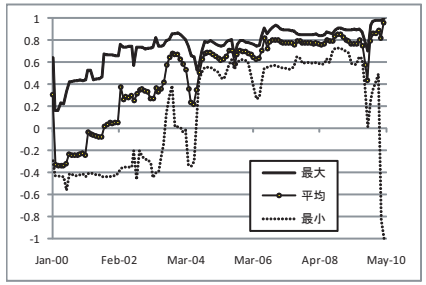

図 11 bf_t([“暑い” AND “s”]) に基づく Weblog センサと気温統計との局所的な相関係数

Fig. 11 Local Correlation between Temperature and Weblog Sensors by bf_t([“atsui” AND “s”])

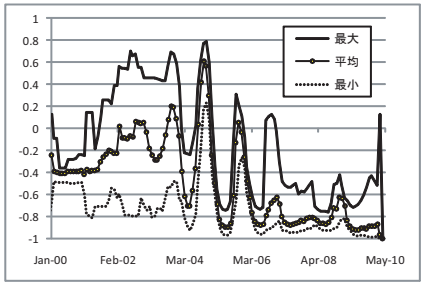

図 12 bf_t([“寒い” AND “s”]) に基づく Weblog センサと気温統計との局所的な相関係数

Fig. 12 Local Correlation between Temperature and Weblog Sensors by bf_t([“samui” AND “s”])

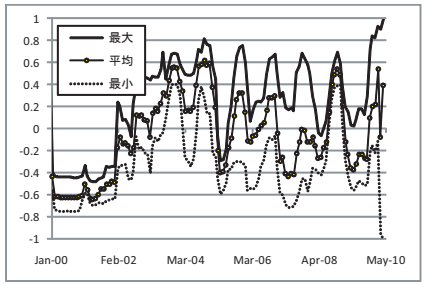

図 13 wf_t([“晴れ” AND “s”]) に基づく Web センサと気温統計との局所的な相関係数

Fig. 13 Local Correlation between Temperature and Web Sensors by wf_t([“hare” AND “s”])

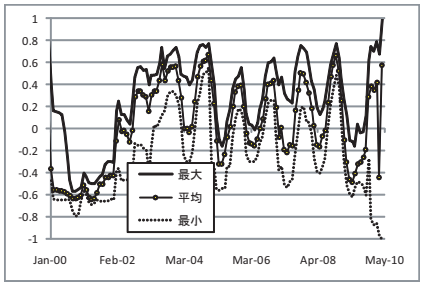

図 14 wf_t([“雨” AND “s”]) に基づく Web センサと気温統計との局所的な相関係数

Fig. 14 Local Correlation between Temperature and Web Sensors by wf_t([“ame” AND “s”])

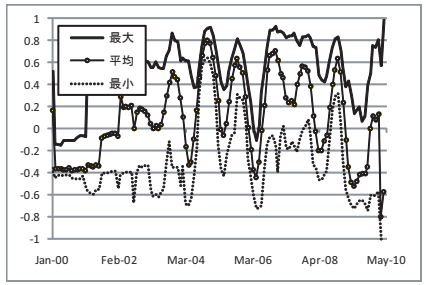

図 15 bf_t([“晴れ” AND “s”]) に基づく Weblog センサと気温統計との局所的な相関係数

Fig. 15 Local Correlation between Temperature and Weblog Sensors by bf_t([“hare” AND “s”])

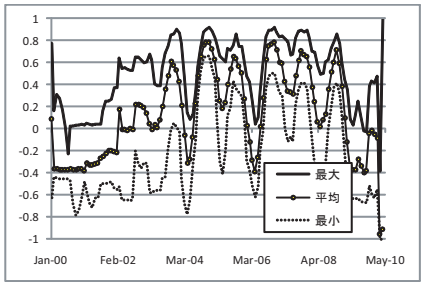

図 16 bf_t([“雨” AND “s”]) に基づく Weblog センサと気温統計との局所的な相関係数

Fig. 16 Local Correlation between Temperature and Weblog Sensors by bf_t([“ame” AND “s”])

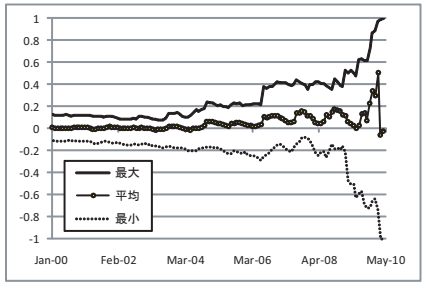

図 17 wf_t(["雨" AND "s"]) に基づく Web センサと降水量統計との大域的な相関係数

Fig. 17 Global Correlation between Precipitation and Web Sensors by wf_t(["ame" AND "s"])

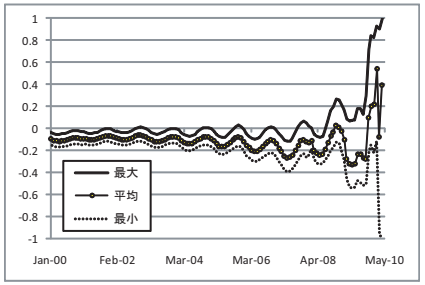

図 18 wf_t(["晴れ" AND "s"]) に基づく Web センサと降水量統計との大域的な相関係数

Fig. 18 Global Correlation between Precipitation and Web Sensors by wf_t(["hare" AND "s"])

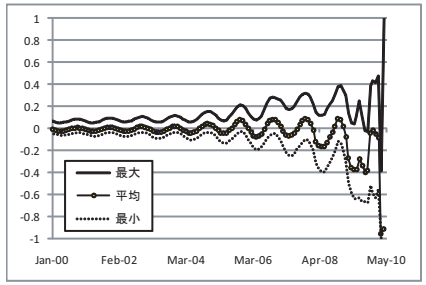

図 19 bf_t(["雨" AND "s"]) に基づく Weblog センサと降水量統計との大域的な相関係数

Fig. 19 Global Correlation between Precipitation and Weblog Sensors by bf_t(["ame" AND "s"])

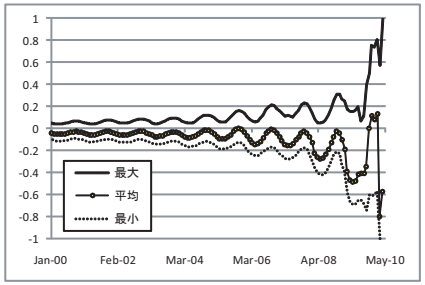

図 20 bf_t(["晴れ" AND "s"]) に基づく Weblog センサと降水量統計との大域的な相関係数

Fig. 20 Global Correlation between Precipitation and Weblog Sensors by bf_t(["hare" AND "s"])

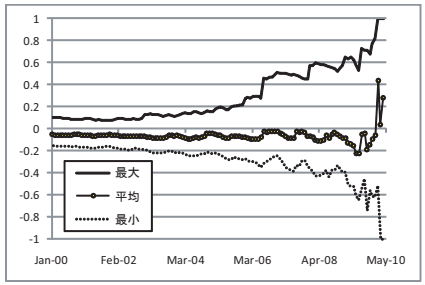

図 21 wf_t(["寒い" AND "s"]) に基づく Web センサと降水量統計との大域的な相関係数

Fig. 21 Global Correlation between Precipitation and Web Sensors by wf_t(["samui" AND "s"])

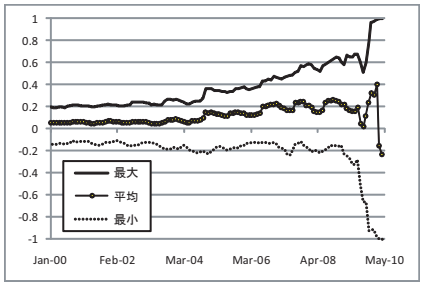

図 22 wf_t(["暑い" AND "s"]) に基づく Web センサと降水量統計との大域的な相関係数

Fig. 22 Global Correlation between Precipitation and Web Sensors by wf_t(["atsui" AND "s"])

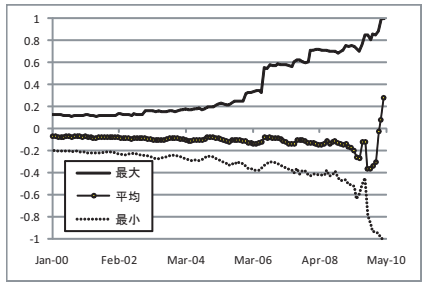

図 23 bf_t(["寒い" AND "s"]) に基づく Weblog センサと降水量統計との大域的な相関係数

Fig. 23 Global Correlation between Precipitation and Weblog Sensors by bf_t(["samui" AND "s"])

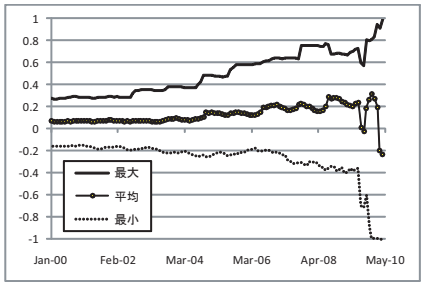

図 24 bf_t(["暑い" AND "s"]) に基づく Weblog センサと降水量統計との大域的な相関係数

Fig. 24 Global Correlation between Precipitation and Weblog Sensors by bf_t(["atsui" AND "s"])

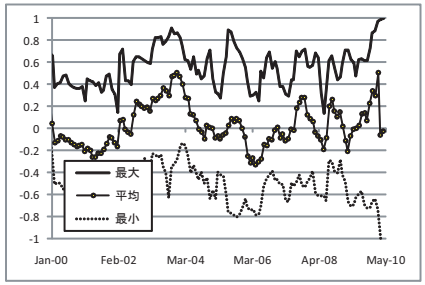

図 25 wf_t(["雨" AND "s"]) に基づく Web センサと降水量統計との局所的な相関係数

Fig. 25 Local Correlation between Precipitation and Web Sensors by wf_t(["ame" AND "s"])

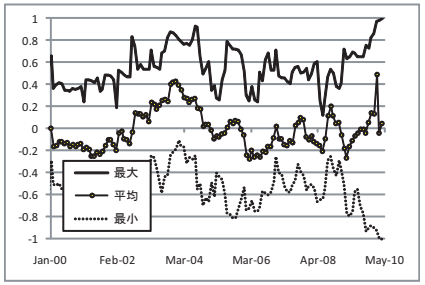

図 26 wf_t(["晴れ" AND "s"]) に基づく Web センサと降水量統計との局所的な相関係数

Fig. 26 Local Correlation between Precipitation and Web Sensors by wf_t(["hare" AND "s"])

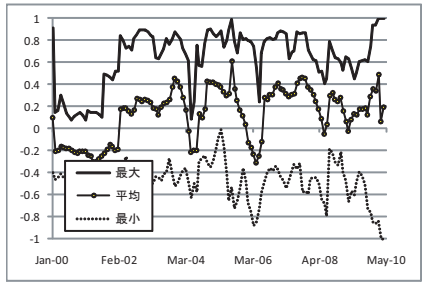

図 27 bf_t(["暑い" AND "s"]) に基づく Weblog センサと降水量統計との局所的な相関係数

Fig. 27 Local Correlation between Precipitation and Weblog Sensors by bf_t(["ame" AND "s"])

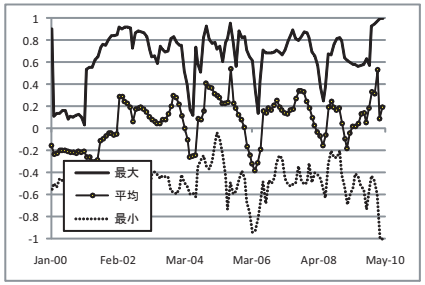

図 28 bf_t(["晴れ" AND "s"]) に基づく Weblog センサと降水量統計との局所的な相関係数

Fig. 28 Local Correlation between Precipitation and Weblog Sensors by bf_t(["hare" AND "s"])

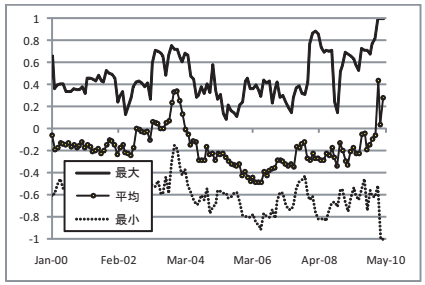

図 29 wf_t(["寒い" AND "s"]) に基づく Web センサと降水量統計との局所的な相関係数

Fig. 29 Local Correlation between Precipitation and Web Sensors by wf_t(["samui" AND "s"])

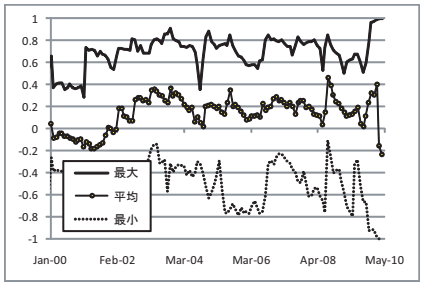

図 30 wf_t(["暑い" AND "s"]) に基づく Web センサと降水量統計との局所的な相関係数

Fig. 30 Local Correlation between Precipitation and Web Sensors by wf_t(["atsui" AND "s"])

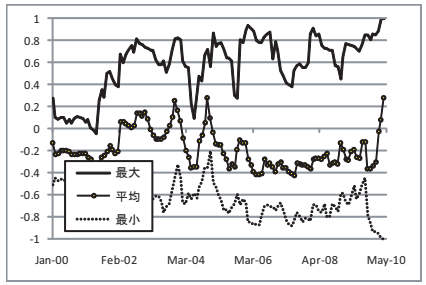

図 31 bf_t(["寒い" AND "s"]) に基づく Weblog センサと降水量統計との局所的な相関係数

Fig. 31 Local Correlation between Precipitation and Weblog Sensors by bf_t(["samui" AND "s"])

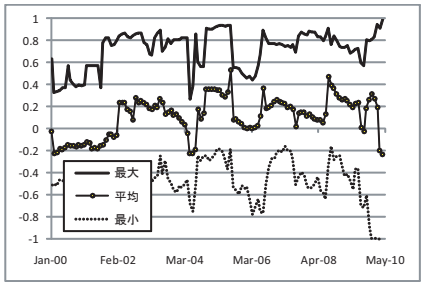

図 32 bf_t(["暑い" AND "s"]) に基づく Weblog センサと降水量統計との局所的な相関係数

Fig. 32 Local Correlation between Precipitation and Weblog Sensors by bf_t(["samui" AND "s"])

次に，**図 17** から**図 24** は気象庁の降水量統計データとの各年月から「2010 年 5 月」までの大局的な相関係数について，**図 25** から**図 32** は気象庁の降水量統計データとの各年月から 1 年（12ヶ月）分の局所的な相関係数について，47 都道府県の気象台の最大値，平均値，最小値の推移をそれぞれ示している．実世界の現象「気温」に関しては,「暑い」という言葉の文書頻度を用いると正の相関が,「寒い」という言葉の文書頻度を用いると負の相関が得られたが，一方，実世界の現象「降水量」に関しては,「降水量」を連想させる「雨」や「晴れ」といった言葉の文書頻度を用いても,「降水量」を必ずしも連想させない「寒い」や「暑い」といった言葉の文書頻度を用いた場合と同様に，実世界のリアルな統計データとほとんど相関を示していないことが分かる．従って，実世界の現象（気温や降水量など）の違いに依存して，Web センサによってマイニングされたデータと実世界に設置されたリアルセンサの統計データとの相関に大きな差があると言える．

4. まとめと今後の課題

本論文では，実世界で起きた現象のリアルなデータとして気象庁の気象統計情報から気温および降水量を用い，Web から抽出した時空間依存データとの相関を評価することで，Web センサによってマイニングされたデータの利用可能性や信頼性を検証を試みた．Web から何らかの知識データを抽出する手法は様々に提案されているが，本論文では Web 検索エンジンのインデックス情報の一つである検索件数だけを用い，特定の時空間で制限した Web 文書や Web ブログの中で「暑い」や「雨」といった言葉を含む頻度に基づいて Web センサを構成し，実世界の気温や降水量に関する時空間依存データを Web から抽出した．その上で，実世界の気温や降水量に関して Web 世界から抽出した時空間依存データと気象庁の気象統計との相関を多角的に評価・考察を行った．

その結果，実世界の現象「気温」に関しては,「気温」を連想させる「暑い」や「寒い」といった言葉の文書頻度に基づいて Web センサを構成することで，実世界に設置されたリアルセンサの統計データとの相関を得ることができた．また，一般の Web 文書よりも Web ブログから抽出した方が相関がより強いこと，実世界の現象が起きた空間（都道府県）の違いにはあまり依存しないが，実世界の現象（気温や降水量など）や実世界の現象が起きた時間（月毎）の違いに依存して相関に大きな差が見られることが分かった．つまり，Web センサには，時空間依存データを Web 抽出する対象の時空間（特に時間），実世界の現象や事象に依って得意・不得意が存在するため，Web センサによってマイニングされたデータを安易に盲信することは出来ないと言える．少なくとも，Google の期間指定オプションそのままでは，Web センサの時間軸の信頼性が乏しいと考えられる．Web 文書の更新日時といったメタデータに依存した時間指定に基づく文書頻度ではなく，Web 文書の内容から，例えば，文単位で見て時間表現，空間表現，そして，現象表現を含む文を抜き出し，その文頻度を用いて Web センサを構成し，本論文と同様に気象統計との相関を評価することを今後検討して行く予定である．

今後の研究課題としては，気象統計以外の実世界のリアルなデータを収集し，自然現象以外のより多くの種類の実世界の現象についても Web センサとの相関を評価する必要がある．また，本論文では，最も単純な文書頻度を用いて Web センサを構成したが，関数を工夫するなど，抽出手法を高度化することで，より強い相関が得られるかどうかについても評価する必要があると考える．さらには，Web ニュース記事や Twitter のつぶやきなど，他の Web メディアからの時空間依存データの抽出実験も行う予定である．Twitter のつぶやきを用いれば，月毎や日毎といった粒度よりも細かく，よりリアルタイムに数値データを出力する Web センサを構成できる可能性がある．

最後に，実世界に物理的に設置する必要があるリアルセンサとは異なり，あらゆる場所の時空間依存データを仮想的に抽出できる可能性がある Web センサを活用して，実世界に設置されたリアルセンサの代替あるいは補完センサと位置付け，Web 抽出された時空間依存データ（潜在的なニーズなど）をリアルタイムに実世界にフィードバックし，実空間の構造や実世界でのサービスの配置などを自動的に変えて行く新しい仕組みの研究にも取り組んで行きたいと考えている．

参 考 文 献

1) 藤村 滋，豊田 正史，喜連川 優：“文の構造を考慮した評判抽出手法,” 電子情報通信学会 第 16 回データ工学ワークショップ (DEWS2005), 6C-i8 (2005).
2) 倉島 健，藤村 考，奥田 英範：“大規模テキストからの経験マイニング,” 電子情報通信学会 第 20 回データ工学ワークショップ (DEWS2008), A1-4 (2008).
3) 服部 峻，田中 克己：“性質継承と概念の再帰的適用に基づく Web からの概念階層抽出,” 情報処理学会論文誌：データベース，Vol.1, No.3 (TOD40), pp.60–81 (2008).
4) 服部 峻，手塚 太郎，田中 克己：“文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出,” 情報処理学会論文誌：データベース，Vol.48, No.SIG11 (TOD34), pp.69–82 (2007).
5) 服部 峻，田中 克己：“Web 抽出した特異な色名と色特徴量変換に基づく特異画像の Web 検索,” 情報処理学会論文誌：データベース，Vol.3, No.1 (TOD45), pp.49–63 (2010).
6) 気象庁 - 気象統計情報, `http://www.jma.go.jp/jma/menu/report.html` (2010).