ニュースサイトと履歴ウェブによるトピックセンサー

Topic Sensor Utilizing News Sites and Web Usage Data

吉岡 由智* 平野 真太郎†
成 凱‡ 上林 弥彦

Yoshitomo YOSHIOKA Shintaro HIRANO
Kai CHENG Yahiko KAMBAYASHI

現在マスコミはウェブの利用に強い影響力があると考えられる．例えば，テレビで選挙が報道されると，ウェブで選挙に関するデータの利用が増えるといった具合である．本稿では，利用者の興味が集まっていると考えられるニュースサイトのデータ内容とその利用履歴を考慮することで，利用者の興味を反映したデータの重要度決定を行うトピックセンサーを考案し，開発のための予備実験を行い，有用性を検証した．そして重要トピックを利用者へ推薦することによりウェブデータを効率的に扱えることが期待できる．

We believe that usage of web is influenced significantly by mass media today. For example, when news about an election is being transmitted by television, usage of the data on the web about this election will increase. The increase of web usage is reasonably regarded as an indicator of emerging interests. In this study, we develop Topic Sensor, a middleware that determines data priority in terms of users' interest by taking into consideration the contents of web data and its usage. In our experiments, we focused on the news sites of most interest. Topic Sensor is expected to apply in content recommendation, so that users can keep informed of important topics.

1. はじめに

現在のインターネットの世界では，ウェブデータは膨大な量になっている．そしてこれからも増加の一途をたどっていくと考えられるが，このような環境下では利用者が自分の必要とするデータだけを取得するのは困難である．従って，利用者の WWW 上での活動を支援するために，利用者から必要とされているデータを検出することが重要になる．そこで本稿ではこの状況を解決するために，ウェブデータの内容やその利用状況を考慮して，ウェブデータに対して重要度決定を行うことで，利用者の興味が集中しているデータを検出する方法を考案した．

実社会において利用者の欲するデータ，つまり利用者の興味が集中しているデータは，マスコミの発信するデータ（例．ニュースサイト）に左右されるところが大きい．例えばテレビで選挙が報道されると，ウェブで選挙に関するデータの利用が増えるといった具合である．しかし，マスコミの発信するデータはそれ自体で膨大な量になっている．このようなマスコミの発信するデータ量の膨大化に着目し，そのデータを効率的に扱い利用者にわかりやすい形で提供することを目的とする研究は多数存在する([1], [2])．本稿でも，マスコミの発信するニュースサイトのデータを用いた．そして，データの内容を端的に表したものをトピックと定義し，サイト内の各データをトピックにより分類する．

次にデータの利用状況を考慮する方法であるが，本稿ではウェブ上でのデータの利用状況を表したデータを履歴ウェブと名づけた．履歴ウェブの例としては，ウェブ上での利用者の活動を時間順に保持したものであるプロキシログがある．分類したトピックごとにプロキシログに現れるデータへのアクセス回数を集計して，トピックごとの重要度決定を行う．これにより，ウェブ上で利用頻度の高いトピックのデータが高い重要度を持つことになる．利用頻度の高いデータに対しては利用者の興味が集中していると判断できるので，トピックの重要度決定の結果に利用者の興味という情報が反映され，利用者の興味が集中しているトピックを検出できる．

そして，トピックセンサーによって決定されるトピックごとの重要度の結果を用いて利用者に重要トピックを推薦することにより，利用者はその時点で話題になっているトピックを知ることができる．このことは膨大なウェブデータの中から利用者が求める内容のデータへたどり着くことの支援になると期待できる．さらに，利用者だけでなくウェブ上でデータを配信する側に推薦することにより，配信側は利用者にもっと見てもらうために，重要度の高いトピックに関連するデータを増やすことが期待される．また，このことはウェブデータを見る利用側にとっても，自分たちの興味を集めている話題が多く提供されるということを意味し，配信者という供給側と，利用者という需要側の双方にとって意味をもたらすと考えられる．この様子を図１に示した．

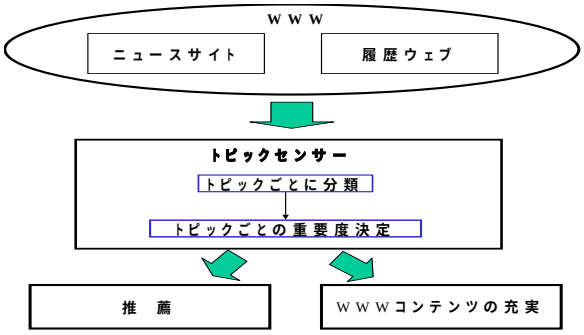

図１ トピックセンサーの利用例

Fig.1 An Example of the Utility of Topic Sensor

以下 2 章では関連研究を述べ，3 章ではトピックセンサーの詳細，4 章では開発したトピックセンサーの実装及び評価，最後に 5 章で本稿のまとめと将来研究について述べる．

2. 関連研究

TDT(Topic Detection and Tracking)[1]は，マスコミの発

* 学生会員 京都大学大学院情報学研究科修士課程
yoshitomo@dl.kuis.kyoto-u.ac.jp
† 京都大学大学院情報学研究科修士課程
shin@db.soc.i.kyoto-u.ac.jp
‡ 正会員 九州産業大学情報科学部
chengk@is.kyusan-u.ac.jp

信するニュースデータに着目し，そのデータを効率的に扱い利用者にわかりやすい形で提供することを目的とする研究である．具体的にはニュースデータを個々のニュース記事に分割した後，個々のニュース記事をトピックごとに分類する手法を用いている．そして TDT の発展研究である Trend Analysis[2]では，TDT を実行した後に世間で注目を集めているトピックを検出することを目的としている．対象とするニュースは本稿と同じテキスト形式のニュースサイトである CNET[6]などに掲載されているニュース記事である．それぞれのサイトに掲載されている各ニュース記事を，ART(Adaptive Resonance Theory)[8]を用いてトピックごとに分類する．さらに，Trend Analysis によって分類されたトピックごとに重要度を決定する．Trend Analysis では，各トピックに含まれる 1 週間あたりの記事の数を比較してトピックの重要度決定を行っている．

このように[2]はニュース記事をトピックごとに分類し，トピックごとの重要度を決定するという，本稿のトピックセンサーと機能的に非常に類似するものである．しかし，トピックの重要度決定の結果において[2]ではトピック内の記事数で決定されるため，ニュースの制作者（供給側）が重要とみなしたトピックが高い重要度を持つことになる．しかしこれはニュースの供給側の主観的な評価であるといえる．これに対して，トピックセンサーでは一般の ISP(インターネットサービスプロバイダ)の履歴ウェブを用いて ISP の会員という利用者集合の利用状況を考慮することで，利用者の興味が反映されるため，その結果はより信頼性の高い客観的な評価になることが期待できるところが優位点であるといえる．

さらに，本稿で利用しているプロキシログである京都市の ASTEM(京都高度技術研究所)の運営する Kyoto I-net のデータは，京都に住んでいる人のデータが多く含まれるので，京都に住む利用者集合の特徴が反映されると期待できる．これにより，京都の地域性も反映されると考えられる．例えば，野球というトピックに関して京都の利用者は阪神タイガースに関連するニュースデータをよく見るが，福岡の利用者は福岡ダイエーホークスに関連するニュースデータをよく見るというように，利用者の居住している地域によって利用者の興味が集中しているデータには違いがあるということが予想できる．このような地域による利用者の興味の相違がわかれば，地域ごとの利用者の興味に合わせてデータを個別に提供することが可能になる．従って，このような地域性を考慮したトピックの重要度決定は重要であり，[2]にはないトピックセンサーの優位点であるといえる．

3. トピックセンサー

本稿で開発したトピックセンサーは 2 つの機能から構成されており，1 つは利用者の興味が集中しているデータをトピックごとに分類する機能，もう 1 つはそれを用いて履歴ウェブを考慮することにより，トピックごとの重要度決定を行うものである．これらの機能について具体的に説明する．

3.1 トピックごとに分類する機能

本稿において，トピックごとに分類するために用いるデータは，ニュースサイトを用いる．

ニュースサイト内の各データをトピックごとに分類するためには，まず各データからトピックを抽出しなければならない．アサヒ・コム[3]を始めとする一般的なニュースサイトの構造はトップページを中心とした有向グラフと見ることができる．図 2 にこの様子を示す．

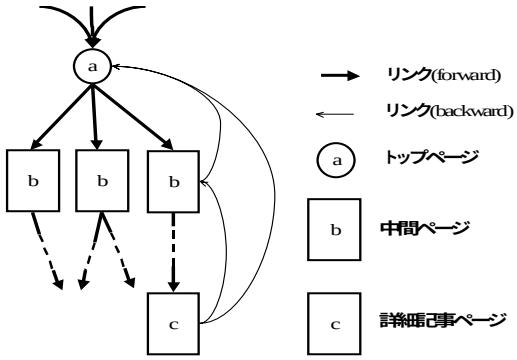

図 2　ニュースサイトの構造図
Fig.2　Structure Diagram of a News Site

図 2 の有向グラフから a への枝と，a や b に戻るような backward の枝を取り除くと，a を根とする有向木と見ることができる．a のような木の根となるページはニュースサイトのトップページである．c のような木の葉となる節点は 1 つのニュースについて詳細に記述したページである．これを詳細記事ページ(c)と名づけた．また，木の根と葉以外の節点はトップページと詳細記事ページとの間のページである．これを中間ページ(b)と名づけた．この(b)と(c)の各ページからトピックを抽出する手法をそれぞれ説明する．

まず(b)のページからトピックを抽出する方法であるが，ニュースサイトの中にはそのページの内容を端的に表したキーワード集合（トピック）がページのタイトルタグ部分に出現するものがある．このようなページに対しては，そのタイトルタグ部分に含まれるキーワードを抽出して，そのページのトピックとした．また，あるページのタイトルタグ部分に現れるキーワードは，そのページの親となっているページのタイトルタグ部分に現れるキーワードに 1 つのキーワードを追加した形式になっている．このため，葉節点に近いページほどトピックを構成するキーワードの数が多くなり，より詳細にトピックを定義できる．このように中間ページに対しては，根となるトップページからの木の深さによって，トピックの詳細度を定義できるというような， トピックの階層構造を考慮した定義が可能になる．本稿では根からの深さが n のトピックを詳細度 n のトピックとして定義した．

次に(c)のページからトピックを抽出する方法について述べる．詳細記事の内容を端的に表すデータは，その記事のタイトル部に出現している名詞のキーワードであると考えられる．従って，記事のタイトル部を *tf * idf* [5]の考え方を用いて解析し，出現している名詞を抽出してそのページのトピックとする．このようにトピックの抽出の際には，トピックの詳細度という考えを用いて行う．以下の図 3 に抽出したトピックの例を示す．

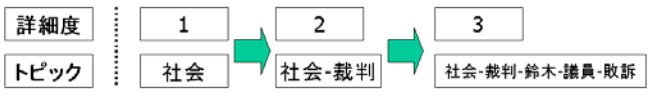

図 3　各詳細度での抽出したトピックの例
Fig.3　Example of Extracted Topics in each Degree

上記の手法によってニュースサイト内の各データからトピックを抽出した後，同じトピックの記事をまとめていくことにより，トピックごとに分類することができる．具体的には，まず中間ページから抽出されるトピックを含むページ同

士を同じトピックとして分類する．このようにして分類された詳細記事をさらに細かく分類するために，抽出した詳細記事のキーワード間の類似度をコサイン類似度を用いて計算し，類似度が閾値を超えるもの同士を同じトピックにまとめる．以上により，中間ページに対しては，その詳細度にあわせてトピックの階層構造を考慮した分類が可能になる．更には詳細記事ページのタイトルタグ部分と記事のタイトルを解析することにより，詳細な分類が可能になる．これによりトピックごとに分類する精度の向上が期待できる．

3.2 トピックごとの重要度決定の機能

まず，トピックごとの重要度を決定するために用いるデータについて説明する．次に実際に重要度決定を行う手法について説明する．

本稿では，分類されたトピックに対して履歴ウェブを考慮することにより，トピックごとの重要度を決定する．履歴ウェブの例としては，ウェブ上での利用者の活動を時間順に保持したプロキシログがある．プロキシログの主要な情報のみを記したテーブルは以下のとおりである．

- **Proxy(Time, IP, URL)**

Time：リクエストされた時刻
IP ：URL をリクエストした IP アドレス
URL：リクエストされた URL

上記のテーブルから，利用者がいつどのウェブページを見たのかが分かる．本稿ではこのようにプロキシログのデータを用いてトピックごとの重要度決定を行った．

分類された各トピックに含まれるそれぞれのニュース記事(ウェブページ)の，プロキシログに現れる回数を集計し，そのアクセス頻度を比較する．そして頻度の高いトピックには大きな重みを，頻度の低いトピックには小さな重みを割り当てることにより，利用者の興味を反映したトピックの重要度決定が可能になると考えられる．しかし，トピック自体が詳細度ごとに分類されているので，アクセス頻度をただ単に集計するだけでは，異なる詳細度間でのトピックの重要度の比較をする場合に，詳細度の小さいトピックが必ず高い重要度を有してしまうことになる．しかし，利用者にとっては詳細度の大きいトピックの方が，利用者の興味をより具体的に表した意味のある情報だと思われる．従って，トピックの重要度を決定する際には，詳細度の大きいトピックになるほど高い重要度を割り当てるべきである．

このため，トピックの重要度決定の際にはただ単にトピック毎にアクセス回数を集計して比較するのではなく，詳細度の大きいトピックに対応するページへのアクセスがあれば，詳細度の小さいトピックでのアクセスよりも重みをつけることとする．これを計算する式は以下のとおりである．

記事を i，トピックを T とし，トピック T に分類される記事数を k とする．

$$i \in T(1 \leq i \leq k)$$

となる記事 i に対するアクセス回数を A_i とする．

次に詳細記事ページのトピックの詳細度を n とすると，ページ i に対する重み w_i をそのページ i のトピックの詳細度を m_i として，次の式で定義する．

$$w_i = \frac{m_i}{n}$$

これらを用いてトピック T の重要度 $Weight(T)$ を定義する．

$$Weight(T) = \sum_{i=1}^{k} A_i * w_i$$

これにより，利用者の興味をより詳細に表したトピックが，より高い重要度を持つことになる．

4. 予備実験及び評価

4.1 実験の目的と利用したデータ

本稿ではトピックセンサーによって決定されるトピックごとの重要度決定の結果が，本当に利用者の興味を集めているかどうかを検証した．具体的にはトピックセンサーと Trend Analysis との比較を行った．トピックの重要度決定の結果において Trend Analysis ではトピック内の記事数で決定されるため，ニュースの制作者（供給側）が重要とみなしたトピックが高い重要度を持つことになる．しかしこれはニュースの供給側の主観的な評価であるといえる．これに対して，トピックセンサーでは一般の ISP の履歴ウェブを用いて ISP の会員という利用者集合の利用状況を考慮することで，利用者(需要側)の興味が反映されるため，その結果はより信頼性の高い客観的な評価になることが期待できる．両者を比較した場合，利用者にとっては利用者の興味が反映されるトピックセンサーの方が有用であると考えた．

実験では 03/12/23 から 03/12/30 の 8 日間に渡って，トピックセンサーと Trend Analysis の双方でトピックの重要度決定を行い，その結果の比較を行ってトピックセンサーの有効性を検証した．実験の際に用いたニュースサイトはアサヒ・コムであり，プロキシログは Kyoto I-net のデータを利用した．なおプロキシログに現れる IP アドレスは利用者を特定できないように暗号化されたものを用いることにより、Kyoto I-net の利用者のプライバシーの点には十分注意した．

4.2 トピックセンサーの有用性の検証

開発したトピックセンサーの実験評価について述べる．実験方法については扱ったアサヒ・コムのニュースサイトがトップページから詳細記事ページまでの深さが 3 であったため，抽出したトピックは詳細度 1 から 3 のトピックであった．従ってまず詳細度が 1，つまりトップページからの深さが 1 のページから抽出したトピックの重要度を計算した．そして詳細度が 2，3，でのトピックの重要度計算を順に行った．そして，それぞれの詳細度でのトピックごとの重要度決定をトピックセンサーと Trend Analysis の 2 つの方法で行い，両者の結果の比較をそれぞれの詳細度で行った．最後にトピックセンサーに関する評価のまとめについて述べる．

- **詳細度 1 と 2 によるトピックの比較**

詳細度 1 と 2 のトピックはページのタイトルタグから抽出したキーワードであり，”national”や”incident”など意味が漠然とした普通名詞であった．実際にトピックセンサーと Trend Analysis の両方でトピックの重要度を比較した結果，両者の間に大きな差異は見られなかった．差異がない以上，計算コストの小さい Trend Analysis の方がこの場合は有用であるといえる．

- **詳細度 3 によるトピックの比較**

詳細度 3 のトピックはタイトルタグの他に，詳細記事ページの記事のタイトルから抽出したキーワードから構成されている．よって，トピックの数は一定ではなく，その種類も多くなる．また，記事のタイトルから抽出したキーワードは具体的な固有名詞が多く，詳細度 3 のトピックは利用者にとってはわかりやすい．ここでは最も特徴的なトピックごとの

重要度の結果を示した 03/12/25 における各トピックの重要度決定について述べる．以下の表 1 は 03/12/25 でのトピックセンサーと Trend Analysis でのトピックの重要度決定の結果を表したものである．

表 1　詳細度 3 でのトピックの比較
Table 1　Comparison of Topics in Detail-Degree 3

トピックセンサー

トピック	重要度
社会 事件 栃木 佐野 トラック	10
サイエンス 永久磁石 鉄球 浮遊 160年 社会 裁判 出生届 最高裁 曽	7
社会 裁判 鈴木 議員 敗訴	5
サイエンス 火星探査 周回軌道 政治 国政 小泉 アルジャジーラ 社会 裁判 医療ミス 滋恵 青戸 文化 芸能 漢剣慶太 四季 婦関届 経済 産業 任天堂 ゲーム機 来年	4
5件	3
6件	2
17件	1

Trend Analysis

トピック	重要度
政治 国政 小泉	2
社会 事件 千葉 国際 イラク バグダッド 経済 市況 経済指 ダウ 経済 市況 経済指 NY 経済 市況 経済指 東証	
74件	1

表から Trend Analysis ではトピックの重要度にほとんど差は無くほとんどのトピックの重要度が 1 となっており，どのトピックが重要なのかはわからない．一方トピックセンサーでは各トピック間で重要度の差は大きく，多くのトピックの中から少数のトピックを明確に検出できているといえる．この結果は利用者にとってわかりやすいものであり，トピックセンサーの方が有用な結果が出ているといえる．また表 1 からトピックセンサーで重要度が高いと判断されたトピックを見てみると，1 位のトピックは 03/12/24 に栃木県佐野市でトラックにはねられて 15 歳の少女が死亡した事件に関するトピックである．このトピックは少女がクリスマスプレゼントを同級生に渡した後の帰り道での悲劇としてマスコミでも多く報じられた．これは実験を行ったクリスマス(12/25)という日を特徴付けるトピックであるといえる．しかし，Trend Analysis ではこのトピックを重要と判断していないことから，トピックセンサーの方が有用なトピックの重要度決定を行ったといえる．また，2 位以下のトピックも鈴木宗男議員のトピックや火星探査，医療ミスなど世間で話題のあると思われるトピックである．これらはいずれも Trend Analysis での結果では重要度が高いトピックとして検出されていなかったトピックであり，Trend Analysis よりもトピックセンサーによるトピックの重要度決定の結果に利用者の興味が顕著に反映されているということがわかる．

以上より詳細度の大きい，つまり利用者にとってより理解しやすいトピックの重要度決定においてはトピックセンサーの方が Trend Analysis よりも有用であるといえる．

5. まとめと今後の課題

本稿では利用者の興味が集中しているデータをトピックごとに分類し，履歴ウェブを考慮することにより，トピックごとの重要度を決定する機能であるトピックセンサーを提案した．トピックセンサーの新規点はプロキシログを用いて履歴ウェブを考慮することにより利用者の興味を反映させたトピックの重要度決定を行った点である．これにより詳細なトピックや，1 日ごとなどの短い期間でのトピックの重要度決定の結果に優れた結果をもたらしたことを検証した．

今回の実験結果からは京都という地域性を反映させたものは得られなかった．従って，今後はより長期間に渡った実験，さらには京都新聞[4]のような地方の新聞社の運営するニュースサイトを含む複数のニュースサイトを用いた実験を行って地域性の特色を反映できているかの検証を行うとともにトピックセンサーの信頼性の検証も行う予定である．

[謝辞]

本稿の一部は、文部科学省科学研究費基盤研究(A)(2)「高水準ウェブデータウェアハウスとそれを基準とする教育システムの研究開発」と科学技術振興機構(JST)戦略的創造研究推進事業・CREST における「デジタルシティのユニバーサルデザイン」による支援を受けている。

[文献]

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang: “Topic Detection and Tracking Pilot Study Final Report,” In Proceedings of the Broadcast News Transcription and Understanding Workshop, pp194-218, 1998

[2] K. Rajaraman and Ah-Hwee Tan:”Topic Detection,Tracking and Trend Analysis Using Self-organizing Neural Networks,”Pacific-Asia Conference on Knowledge Discovery and Data Mining , pp102-107, 2001

[3] http://www.asahi.com

[4] http://www.kyoto-np.co.jp

[5] G. Salton, A. Wang, and C. Yang: “A vector space model for information retrieval,” In Journal of the American Society for information Science. Vol 18, pp 613-620, 1975

[6] http://www.cnet.com

[7] Y. Kambayashi and K. Cheng: “Capacity Bound-free Web Warehouse,” In Proceedings of the First Biennial Conference on Innovative Data Systems Research 2003 pp47-57

[8] G. A. Carpenter, S. Grossberg, and D. B. Rosen.: “ Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, ” Neural Networks, 4:759-771, 1991

吉岡 由智　Yoshitomo YOSHIOKA

京都大学大学院情報学研究科社会情報学専攻修士課程在学中．2004 年 京都大学工学部情報学科卒業．ウェブウェアハウスの研究に従事．日本データベース学会学生会員．

平野 真太郎　Shintaro HIRANO

京都大学大学院情報学研究科社会情報学専攻修士課程在学中．2003 年 京都大学工学部情報学科卒業．利用者の興味を反映した情報検索，WWW 活用の研究・開発に従事．情報処理学会学生会員．

成 凱　Kai CHENG

九州産業大学情報科学部社会情報システム学科助教授。2002 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．ウェブウェアハウス、ウェブデータベース、ウェブ情報検索の研究に従事．ACM SIGMOD，情報処理学会，日本データベース学会各会員

上林 弥彦　Yahiko KAMBAYASHI

1970 年京都大学大学院工学研究科博士課程修了．京都大学大学院情報学研究科教授および同研究科長，電子情報通信学会情報システムソサイエティ会長，日本データベース学会会長などを務める．2004 年 2 月 6 日逝去．