ニュースサイトと履歴ウェブによるトピックセンサー

吉岡　由智[†]　平野　真太郎[‡]　　成　凱[§]　　上林　弥彦[‡*]

†京都大学工学部情報学科　〒606-8501 京都市左京区吉田本町

‡京都大学情報学研究科社会情報学専攻　〒606-8501 京都市左京区吉田本町

§九州産業大学情報科学部社会情報システム学科　〒813-8503 福岡市東区松香台 2-3-1

E-mail: †‡{yoshitomo,shin,yahiko}@db.soc.i.kyoto-u.ac.jp,　§chengk@is.kyusan.ac.jp

あらまし　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，さらに利用期間の予測できるデータを考慮することで，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．さらにウェブキャッシュにおいては，LRU ではうまく扱えない未利用のデータについても，そのトピックによって正しく評価できる．

キーワード　トピックセンサー，ウェブとインターネット，データウェアハウス，キャッシュ，データマイニング

Topic Sensor Utilizing News Sites and Web Usage Data

Yoshitomo YOSHIOKA[†]　Shintaro HIRANO[‡]　Kai CHENG[§]　and　Yahiko KAMBAYASHI[*‡]

†Department of Information Science, Kyoto University

‡Department of Social Informatics, Kyoto University

§Department of Social Information Systems Faculty of Information Science, Kyushu Sangyo University

†‡Yoshidahonmachi, Sakyo-ku, Kyoto-shi, 606-8501, Japan

§3-1 Matsukadai 2-Chome, Higashi-ku, Fukuoka 813-8503, Japan

E-mail:†‡{yoshitomo,shin,yahiko}@db.soc.i.kyoto-u.ac.jp, chengk@is.kyusan-u.ac.jp

Abstract　We believe that usage of web is influenced significantly by mass media today. For example, when an election is transmitted by television, usage of the data about the election increases in the web. In this study, we develop Topic Sensor, which makes decisions of data priority reflecting users' interest, by taking into consideration the contents of the data and its web usage of the news site considered that a users have interest, and taking into consideration the data which can predict use periods further. If Topic Sensor is applied to recommendation, users can know important topics. Furthermore, if Topic Sensor is applied to web cache, Topic Sensor estimate correctly the unused data which cannot be carried well at LRU in web cash.

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

＊ 2004 年 2 月 6 日逝去

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．

また，トピックの中にはあらかじめ利用期間の予測できるものも存在する．例えば，ワールドカップサッカーなどは，その開催前や開催期間中は利用者の興味が集中し，ウェブ上でもワールドカップサッカーに関連するデータの利用が増大すると予測できる．しかし，開催期間を過ぎると利用者の興味は無くなり，ウェブ上でも利用は激減すると予測できる．本稿ではこのようなトピックの寿命を考慮し，あらかじめ利用者の興味が集中する期間の予測できるトピックのデータに対し，その期間中は高い重要度をつけるが，期間外には重要度を低くするということを考案した．本稿では，このように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，履歴ウェブとトピックの寿命を考慮することによって，トピックごとの重要度決定を行う機能を有するシステムをトピックセンサーと名付けた．

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．また，推薦の他にトピックセンサーにより決定されたトピックごとの重要度の結果をウェブキャッシュに適用することも考えられる．キャッシュにおける従来のデータ重要度決定にはLRU[3]が用いられている．しかしLRUではデータの使われ方で重要度が決まる．そのため最低限１度は使用されないとキャッシュの中には入れられないという問題点がある．一方トピックセンサーでは利用者の興味が集中しているトピックのデータが高い重要度を持つ．従って１度も使用されていないデータであってもウェブキャッシュに入れておくことができ，LRUでの問題点は解消できると期待している．このことはさらに，全体としてのデータ取得のための通信時間の軽減を期待できる．このようにトピックセンサーはウェブデータの推薦やウェブキャッシュの効率化に有用であると考えた．以下の図１はトピックセンサーの機能とその結果の利用例を示した概念図である．

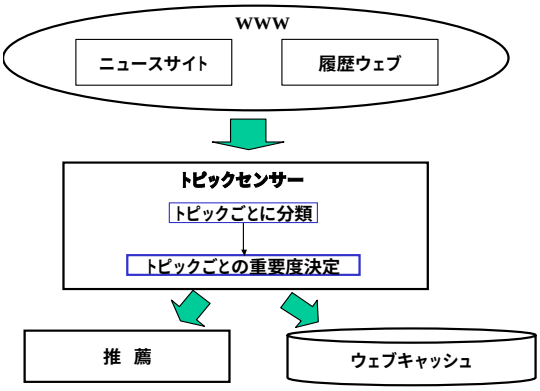

図 1. トピックセンサーの機能と利用例

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．TDT で対象となるデータはウェブ上のニュースサイトだけではなく，テレビやラジオなどで放送されるニュースデータも扱っている．TDT の機能は，ニュースデータを個々のニュースに分割する Segmentation，分割された個々のニュースをトピックごとに分類する Topic Detection，新たに配信されたニュース記事に対してトピックごとに分類を行う Topic Tracking の３つから構成されている．トピックセンサーはニュースサイト内のデータをトピックごとに分類する機能と，履歴ウェブとトピックの寿命を考慮してトピックご

との重要度を決定する機能から構成されている．このトピックごとに分類する機能は TDT における Topic Detection と Topic Tracking と機能を等しくするものである．しかし，トピックセンサーはトピックごとの分類を行った後に，更にトピックごとの重要度決定を行う機能を有している．その結果には利用者の興味が反映されるため，利用者が現在利用者の間で話題になっているトピックを知ることができるという優位点がある．

TDT の発展研究には，トピックごとの重要度を決定する研究も存在する．Trend Analysis[2]では，TDT を実行した後，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[8]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に Topic Detection と Topic Tracking を適用し，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．Topic Detection と Topic Tracking，つまり各ニュース記事をトピックごとに分類する手法は ART(Adaptive Resonance Theory)[11]を用いている．

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 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．以下の図 2 にトピックセンサーと Trend Analysis の比較を示した．

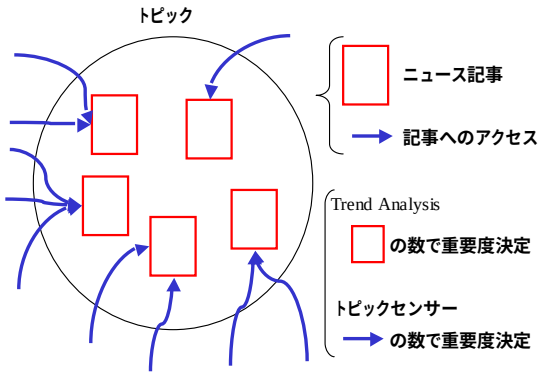

図 2. トピックの重要度決定手法の比較

さらに，本稿で利用している 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とトピックの寿命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，ウェブ上のマスコミの発信するデータである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[4]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．以下の図 3 にこの様子を示す．

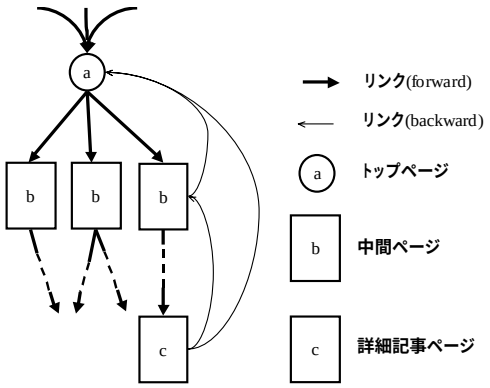

図 3．ニュースサイトの構造図

図 3 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

詳細記事ページへのリンクを集めた，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．

以下では(1)中間ページからトピックを抽出する手法と，(2)詳細記事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．

(1)中間ページのトピック抽出

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．以下の図 4 にこの様子を示した．

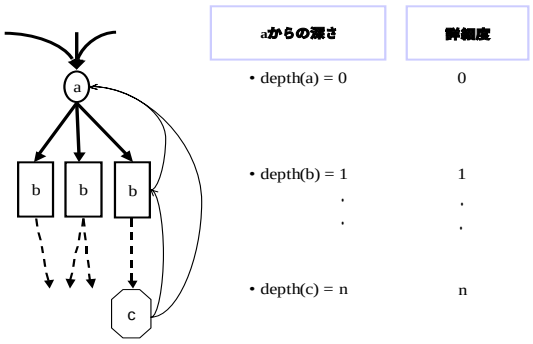

図 4. トピックの詳細度とサイト構造の関連

(2)詳細記事ページのトピック抽出

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．さらに詳細記事の中から記事のタイトル部に現れるキーワード以外の重要なキーワードを抽出して，トピックを表すキーワードとする．具体的には *tf *idf* 法を用いる．*tf *idf* 法はデータに含まれるキーワードの出現状況を計算に用いる手法であり，次の 2 つの要素を用いる．

- Term Frequency

  データ中に多く含まれるキーワードはそのデータを良く特徴付ける．式は次のように定義する．

$$tf(t,d) = freq(t,d)$$

$freq(t,d)$ はキーワード t のデータ d における出現頻度である．

- Inverse Document Frequency

  キーワードは出現するデータの数が少ないほどその出現元のデータを特徴付ける．キーワード t の重要度を次の式で定義する．

$$idf(t) = \log(\frac{D}{freq(t,D)}) + 1$$

D は全データの集合を表す．

これらの 2 つの値を利用することでデータ d におけるキーワード t の重要度 $Weight(t,d)$ を次のように計算することができる

$$Weight(t,d) = tf(t,d) * idf(t) \quad <式1>$$

各詳細記事を *tf *idf* 法を用いてその記事に出現するキーワードに重みをつける．記事 d 中のキーワード t の重み $Weight(t,d)$ は<式 1>の値を用いる．$Weight(t,d)$ が最大となるキーワードから順に抽出し，その記事のトピックとする．抽出するキーワードの数は，詳細記事の内容を端的に表すのに十分な個数とする．

これら記事のタイトル部と記事全文から抽出されるキーワードをつなげることにより，その記事のトピックをさらに詳細に定義することができ，分類の精度を向上させることができると考えた．これら詳細記事と中間ページから抽出されるキーワードをあわせることにより，詳細記事ページのトピックをさらに詳細に定義する．

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．しかし，本稿では複数のニュースサイトを用いるが，詳細記事ページのタイトルタグ部分に現れるキーワードの数は同じではない．このため，サイト間で詳細記事ページでのトピックの詳細度を等しくするために，キーワードの数の不足しているものは補う必要がある．従って，キーワードの数が不足しているページに対しては，記事のタイトル部と記事全文から抽出した複数個のキーワードの中で，$Weight(t,d)$ の値が最大となるキーワードを補う．これにより詳細記事のページでのトピックの詳細度を均一化できる．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をタイトルタグ部分と記事のタイトル，記事の全文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．

３．２．１．プロキシログ

本稿では，分類されたトピックに対して履歴ウェブとトピックの寿命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．本稿ではこのようなプロキシログのデータに現れる履歴ウェブを用いてトピックごとの重要度決定を行った．

３．２．２．履歴ウェブを考慮

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，プロキシログに表れる回数を集計し，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，異なる詳細度間のトピックの重要度の比較をする場合に，小さい詳細度の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，利用者にとっては大きい詳細度のトピックの方が，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，大きい詳細度の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，大きい詳細度のトピックに対応するページへのアクセスがあれば，小さい詳細度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．

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

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 として，次の式で定義する．

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

このようなトピックの詳細度によるアクセス回数への重み付けの概念図を示したものが図 5 である．

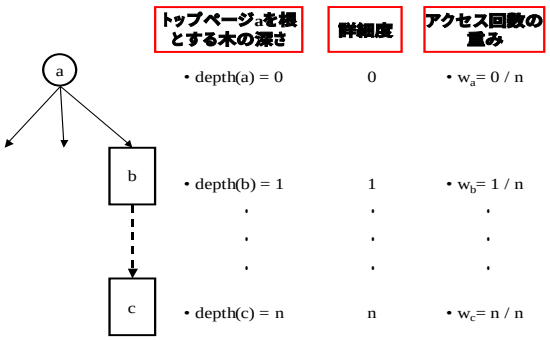

図 5.　異なる階層に渡るトピックの重み付け

これらを用いてトピック T の重要度 *Tweight* を次の式で定義する．

$$Tweight = \sum_{i=1}^{k} A_i * w_i \qquad \text{<式 2>}$$

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．

３．２．３．トピックの寿命を考慮

トピックを寿命という観点から見た場合，定期的なトピックと不定期的なトピックの 2 種類のものがある．定期的なトピックとは毎年や毎月の決まった時期に開催されるようなトピックであり，あらかじめその開催期間がわかっているようなものである．その例としては相撲などがある．これに対し不定期的なトピックは，事前に開催期間がわかっていないようなトピックである．例としては地震などの事故に関するトピックがあげられる．

- 定期的なトピック

定期的なトピックに対しては，過去のデータからその寿命を予測することができる．本稿では，このように過去のデータを考慮してトピックの重要度を決定する手法として，λ-aging[9]を利用した．λ-aging は利用頻度の計算に過去の利用頻度を考慮して集計する手法であり，以下の式で計算する．

トピックを T ,過去のデータの中に存在する T をトピック T' とする．そして，トピック T が開催されてから i 日目のトピック T の重要度 $NewWeight(T,i)$ を

$$NewWeight(T,i) = wWeight(T,i) + (1-w)Weight(T',i) \qquad \text{<式 3>}$$

$Weight(T,i)$ は本日の利用頻度であり，この値には<式 2>を用いる．そして，$Weight(T',i)$ は過去のデータの中のトピック T' の i 日目の利用頻度である． w は定数であり,本日と過去のデータに対する重みを表す． w を適切に設定することによって，本日の利用頻度に過去の利用頻度を考慮することができる．これにより，過去のトピックの寿命を考慮したトピックの重要度決定が可能になると考えた．

- 不定期的なトピック

次に不定期のトピックに対してであるが，これについては過去のデータの中に一致するトピックが存在しない．しかし，過去のデータの中から類似度の高いトピックを過去のデータとすることによって<式 3>を扱うことができる．

４．予備実験および評価

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 のそれぞれ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する．

４．１．利用したデータ

４．１．１．プロキシログ

本稿では京都市の ASTEM(京都高度技術研究所)の運営する ISP である Kyoto I-net のプロキシログを利用した．利用したデータは 03/12/23 から 03/12/31 までのものである. 28 台のプロキシサーバーが稼動しており，各々が独立にプロキシログを記録している．実際に利用したテーブル総数は 13,653,359 にも及ぶ．このうち，プロキシログに現れるファイルタイプが HTML 形式のものだけを集計すると, テーブル数は 2,037,616(14.9%)であった. また，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イシーの点には十分注意した.

４．１．２．ニュースサイト

本稿で扱うニュースサイトは，利用者の興味が集中しているものである必要がある．従って，知名度の高い代表的なニュースサイトとしてアサヒ・コム[4]と YOMIURI ON-LINE[5]が重要であると考えた．また，利用したプロキシログは京都市の ASTEM(京都高度技術研究所)の運営する ISP である Kyoto I-net のデータである．従ってその利用者には京都市民が多く含まれていると考えられる．京都市民にとってはアサヒ・コムや YOMIURI ON-LINE などのニュースサイトの他にも，京都という地域に密着した情報を発信するニュースサイトである京都新聞[6]も知名度が高いと思われる．従って本稿では，この京都新聞のニュースサイトも重要であると考えた.

実際にこれら 3 つのニュースサイトへのアクセス回数の推移を 03/12/23 から 03/12/30 に渡って比較したところ，アサヒ・コムへのアクセスが平均的に最も多いことがわかった. 従って, 本稿ではアサヒ・コムのニュースを利用して予備実験を行った．また 03/12/23 から 03/12/30 までのアサヒ･コムへのアクセスを示すテーブルの総数は 3,433 であった．これは上記で示した HTML 形式のページへのアクセス数 2,037,616 のうち 0.17%であった.

また，各ページからのトピックの抽出では，中間ページに対してはタイトルタグ解析を行ってページのタイトルタグに含まれるキーワードを自動的に抽出してトピックとした．詳細記事に対しては，記事のタイトルに出現するキーワードを手動で抽出して, タイトルタグから抽出したキーワードとあわせて 5 個のキーワードを用いてトピックとした.

次に各ページをトピックごとに分類する手法であるが，タイトルタグから抽出したキーワードのトピックを含むページに対しては，そのキーワードを含むものを同じトピックに分類した．記事のタイトルから抽出したキーワードのトピックに対しては，5 個のキーワードのうち 3 個以上が同じものになるページ同士を手動で同じトピックに分類した.

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 によるトピックでの比較**

まず，タイトルタグから抽出されるキーワードを解析してトピックごとの重要度決定を行った.

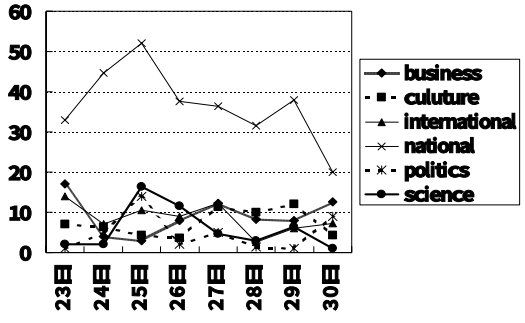

図 6. 詳細度 1 でのトピックセンサーの結果

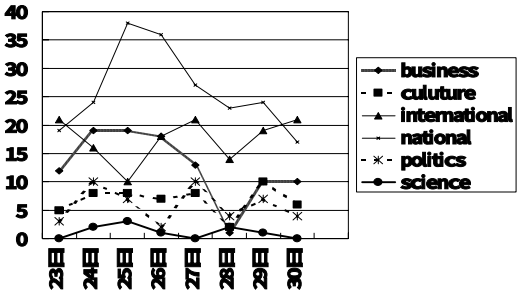

図 7.　詳細度 1 での Trend Analysis の結果

上の図 6 と図 7 は詳細度 1 でのトピックセンサーと Trend Analysis の結果である. 抽出したトピックは"national", "business", "culture", "international", "politics", "science", "sports"の 7 つのトピックであり，収集したアサヒ・コムの各ページのタイトルタグ部分を解析し，上記のトピックを含むものをそのトピックに分類した．そして各トピックへのアクセス回数でトピックごとの重要度決定を行ったものである．上記の図 6 と 7 から"national"の重要度がどちらも高いことがわかる．全体的には両者とも類似した重要度の推移をしているといえるが，トピックセンサーの結果の方が"national"の重要度が高いことが明確にわかる．また 03/12/25 は"national"の重要度が最も高く，特徴的な日であるといえる．

- **詳細度 2 によるトピックでの比較**

次にトピックの詳細度を 2 に増やしてトピックセンサーと Trend Analysis でのトピックの重要度決定の比較を行う．詳細度 1 でのトピックの重要度決定では，どちらも"national"が高い重要度を示していた．従って詳細度 2 でのトピックの重要度決定は"national"に含まれるトピック同士の比較を行った．"national" はその下に "incident", "calamity", "policy", "trial", "etc"の 5 つのトピックから構成されている．これらは上記と同様，タイトルタグ部分に含まれるキーワードであるのでタイトルタグ部分の解析により，トピックごとの分類を行った．以下の図 8 は詳細度 2 によるトピックセンサーと Trend Analysis の結果である．

詳細度 1 での結果と比べると，各トピックの重要度の移り変わりが激しくなっているのがわかる．さらに 03/12/25 でのトピックセンサーと Trend Analysis の結果を比べると，両者で trial の重要度

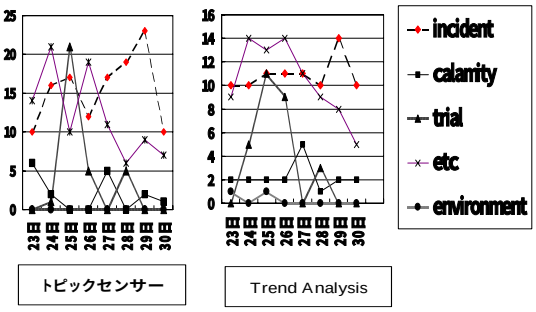

図 8.　トピック"national"での重要度の比較

が突然高くなっていることがわかるが，トピックセンサーの方がその trial の変化を明確に検出できていることがわかる．

- **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．ここでは，詳細度 1 と 2 によるトピックでの比較で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を行った．以下の表 1 は 03/12/25 でのトピックセンサーでのトピックの重要度決定の結果を表したものであり，表.2 は 03/12/25 での 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．

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-浅利慶太-四季-婚姻届 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 社会-事件-千葉 国際-イラク-バグダッド 経済-市況・経済統計-ダウ 経済-市況・経済統計-NY 経済-市況・経済統計-東証	2
74件	1

表 1.　詳細度 3 でのトピックの比較

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ぼ一定であり，また重要度自体も低いことがわかる．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 位のトピックは高校生が永久磁石で鉄球を浮遊させたトピックで，これも電磁気学の定理を覆す大発見として多く報じられていたトピックである．さらに 3 位，4 位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．

４．3. トピックセンサーの評価のまとめ

トピックセンサーを Trend Analysis と比較した場合，詳細度 1 と 2 でのトピックの重要度決定の結果に大きな差異は見られなかった．しかし詳細度 3 によるトピックの重要度決定の結果では両者の間に大きな差異が見られた．

詳細度 1 と 2 では "national", "business", "incident", "trial"など意味が漠然とした普通名詞がトピックとなっており，分類されるトピックの種類自体も少数であるため，1 つのトピックに分類される記事が多くなる．このため，Trend Analysis でも日ごとに重要度の移り変わりが顕著に見られたのだと考えられる．しかし，トピックを表すキーワードが漠然としすぎて，詳細度 1 と 2 でのトピックごとの重要度の結果を利用者が見てもわからないという欠点があるといえる．これに対して，詳細度 3 ではトピックを構成するキーワードに意味がはっきりとした固有名詞が用いられており，利用者にとってもわかりやすく，分類されるトピックの種類自体も多くなっている．さらに本稿では 1 日ごとのトピックの重要度決定を行ったため Trend Analysis では, 分類されたトピック間で重要度の差異はほとんど見られなかったと考えられる．これに対し，トピックセンサーは 1 つのトピックの中に 1 つの詳細記事しか分類されていない場合でも，その記事へのアクセス回数が多ければ重要度が高くなるので，表.1 のような重要トピックの検出ができたと考えられる．

このように，"national"などの大きな概念でトピックが分類されている場合にはトピックセンサーと Trend Analysis では大きな違いは見られないが，その結果は利用者にとってはわかりづらいものである．一方トピックを詳細に定義して分類されるようになったり，1 日ごとのような短い期間でトピックの重要度決定を行う際にはトピックセンサーは重要なトピックを検出することができ，利用者にとっては現在話題になっているトピックを把握することができる．このようにトピックセンサーは利用者への推薦に適用する場合には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，履歴ウェブとトピックの寿命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開発した．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，短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．しかし扱ったデータは 8 日間，特に詳細なトピックでの検証は 1 日分しか実行していないため，さらに期間を長くして実験を行い Trend Analysis に対するトピックセンサーの優位性を検証する必要がある．また, 実験に用いたアサヒ・コムと，利用を検討した YOMIURI ON-LINE,京都新聞へのアクセス数が予想外に少なかったため，信頼性の点で十分な結果は得られていないといえる．従ってこれら以外でよりアクセス数が多いニュースサイトを用いることも検討すべきであると思われる．

また，本稿ではトピックの重要度決定の新たな概念としてトピックの寿命を考案した．しかし，扱ったデータが 8 日間という短い期間だったため過去のデータを使えるような定期的なトピックは存在せず, トピックの寿命を考慮した重要度決定の検証はできなかった．さらにトピックの重要度決定の結果に地域性の特色も反映させることができると期待したが, 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．これらの点からも，今後はより長期間に渡った実験，さらには複数のニュースサイトを用いた実験を行う予定である．

さらに実験評価としてトピックセンサーを用いた利用者への推薦の有効性を検証したが，トピックセンサーをウェブキャッシュに適用することにより，ネットワークのトラフィックを軽減させるというウェブキャッシュの効率化を目指す．

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,” Proceedings of the Broadcast News Transcription and Understanding Workshop1998

[2] K. Rajaraman and Ah-Hwee Tan:”Topic Detection,Tracking and Trend Analysis Using Self-organizing Neural Networks,”PAKDD2001

[3] J. T. Robinson and M. V. Devarakonda: “Data Cache Management Using Frequency-based Replacement,” Performance Evaluation Review, 18(1):134-142, May 1990

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

[5] http://www.yomiuri.co.jp

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

[7] 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

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

[9] C. Cortes and D. Pregibon: “Giga-Mining,” Knowledge Discovery and Data Mining, pp174-178, 1998

[10] Y. Kambayashi and K. Cheng: “Capacity Bound-free Web Warehouse,” CIDR2003

[11] 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