Vol.2014-ICS-175 No.15
2014/3/14

個人情報と一般的重要性に基づく時事情報提供システムの構築

南光[†1]　芋野美紗子[†2]　土屋誠司[†3]　渡部広一[†4]

近年，情報技術の発展により時事情報が爆発的に増加しているため，興味のある情報を取得することが困難になっている．そのため，Web 上に存在する時事情報を取得し，ユーザに合った情報を提供するシステムが求められている．本研究では，個人情報と一般的重要性を用いて柔軟に時事情報を提供する手法を提案する．一般的重要性は，時事情報自体の重要性と性別・年代の嗜好を考慮して判断する．

Developing the Topical Information Offering System Based on Personal Information and General Importance

AKIRA MINAMI[†1]　MISAKO IMONO[†2]
SEIJI TSUCHIYA[†3]　HIROKAZU WATABE[†4]

One of the solutions of means to improve the efficiency of information gathering is that computer voluntarily offers useful current information. Therefore, in the present study, it proposes the system that collects news articles from Web and presents useful topical information for user. The proposal system flexibly offers topical information by associating the word based on personal information and general importance. General importance is investigated in consideration of the importance of current-events information, and the taste of sex and an age.

1. はじめに

情報社会の発展に伴うインターネットの普及により，人間は容易にニュース記事などの時事情報を得ることが可能となった．しかし時事情報は無数に存在し，短時間で大量に更新される．また興味を惹かれる時事情報は，ユーザの個人情報・嗜好によって大きく異なると考えられる．よって，ユーザが自身の求める時事情報を即座に入手することは非常に困難である．人間が情報収集の効率化を図る手段として，コンピュータから自発的に有益な時事情報を提供してもらうことが考えられる．そのため，本研究では Web を用いてニュース記事を収集し，ユーザにとって最も有益であると考えられる記事を提示するシステムを提案する．

本研究における有益な時事情報とは，ユーザの嗜好に合った時事情報と一般的重要性があると考えられる時事情報である．人間は自分の嗜好に合った情報ばかりを集めているわけではなく，台風に関する時事情報など一般的に重要であると考えられる情報も集め，日常生活に活かして行動している．よって，これら二種類の時事情報を提供することでユーザ自身の求める時事情報を網羅できると考えられる．一般的重要性がある記事とは，重要と考えられる記事でかつユーザと同年代の人物が重要視していると考えられるものである．

†1†2 同志社大学大学院　理工学研究科
Graduate School of Science and Engineering, Doshisha University
†3†4 同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

2. 時事情報提供システムの概要

本研究では個人情報と一般的重要性に基づく時事情報提供システムの構築を目的とする．人間がニュース記事を閲覧する際には，その見出し・タイトルを見て興味を持つかを判断し，興味を持ったものに対してのみ詳細を見るということが多いと考えられる．よって本研究で提供する時事情報は，新聞社の Web サイトに存在しているニュース記事の見出し・タイトルを表す文とする．システムは個人情報と一般的重要性の両方を考慮することで時事情報の獲得・点数付けを行い，その点数順にランキング形式の出力を行う．一般的重要性は頻出度から判断した時事情報自体の重要性と，ユーザと同じ性別・年代の人達の嗜好を利用することで判断する．システムの概要を図 1 に示す．

個人情報を考慮した処理では，時事情報と個人情報との関連性を定量化する．時事情報の本文に含まれる語を利用し，時事情報それぞれについての概念を作成する．個人情報については年齢・職業・出身地などの基本的な項目や，嗜好情報といった様々な情報を用いる．これらの情報と時事情報との関連性を調べることで，時事情報に対して点数付けを行う．

一般的重要性を考慮した処理では，ユーザの性別・年代に応じた嗜好や時事情報の頻出度を利用する．旬感ランキング[1]と呼ばれる，性別・年代別に興味のあるキーワードをまとめたランキングを用いて，ユーザの性別・年代の嗜好を探る．そうすることで，ユーザと同年代の人物がいったいどのような情報を重要と考えているかを判断する．ま

た，時事情報の見出しに存在する語それぞれの頻出度から時事情報自体の重要性を調べる．この二つの観点を考慮して，時事情報に対して点数付けを行う．システムの出力は，二つの観点によりつけられた点数が高い順に時事情報を並べたものとなる．

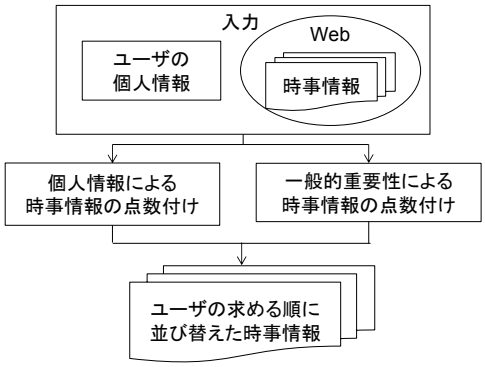

図 1　本研究における時事情報提供システムの概要

3. 使用技術

3.1 概念ベース

概念ベースとは，複数の国語辞書や新聞等から機械的に構築した語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念 *A* に付与される属性には，その重要性を表す重みが付与されている（式 1)．概念ベースには，87242 語の概念が収録されており，1 つの概念あたり平均 38 個の属性が付与されている．本研究では概念ベースに登録されていない概念を未定義語と定義する．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (1)$$

各概念に付与されている属性は，概念ベースに概念として登録されている語であるため，各属性を一つの概念表記としてみなした場合，さらにそれを表す属性を導くことができる（表 1)．このように，概念は概念ベースにより n 次の属性連鎖集合として定義する．また，n 次の属性集合を n 次属性と呼ぶ．

表 1　概念ベースの構成

語	属性
雪	(雪, 0.61), (白い, 0.30), (下る, 0.27), (結晶, 0.25), (雪肌, 0.19)…
白い	(雪, 0.16), (白地, 0.14), (色, 0.14), (白髪, 0.12), (白, 0.12)…
下る	(低い, 0.23), (雪, 0.21), (雨, 0.20), (下る, 0.18), (降参, 0.17)…
…	…

3.2 関連度計算方式

関連度計算方式[2]とは，概念ベースに登録されている 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の間の実数値で表され，概念間の関連が強いほど大きな数値となる．例えば概念「本」に対して「書物」,「雑誌」,「運動」の関連の強さを表 2 のように数値化できれば，コンピュータは「本」と関連がより強いのは 3 つの内,「書物」であるということを判断できる．

関連度計算方式には概念の表記的な特徴を利用する表記関連度計算方式と，お互いの概念が持つ属性の一致度と重みを利用する意味関連度計算方式の 2 つが主としてある．

表 2　関連度計算の具体例

基準概念	対象概念	関連度
本	書物	0. 868
	雑誌	0. 224
	運動	0. 007

ここで述べる関連度計算方式の定義は意味関連度計算方式のものである．以下，関連度計算方式を使うために必要な一致度，およびそれを計算に含めた関連度計算方式について述べる．

3.2.1 一致度

概念 *A*,*B* の属性を a_i，b_j，対応する重みを u_i，v_j とし，それぞれ属性が *L* 個，*M* 個あるとする($L \leqq M$).

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L)\} \quad (2)$$
$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M)\} \quad (3)$$

このとき，概念 *A* と概念 *B* の一致度 *DoM(A,B)*を以下のように定義する．

$$DoM(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases} \quad (4)$$

$a_i=b_j$ は属性同士が一致した場合を示している．すなわち，一致した属性の重みのうち，小さい方の重みの和が一致度となる．このとき各概念の重みの総和は 1 になるように正規化する．よって，一致度は 0.0~1.0 の値をとる．

3.2.2 関連度

関連度 *DoA* は，対象となる二つの概念において，一次属性の組み合わせについて一致度を求め，これを基に概念を構成する属性集合としての一致量を計算することで算出される．

具体的には，一致する属性同士($a_i=b_j$)について，優先的に対応を決定する．他の属性については，全ての一次属性の組み合わせにおいて一致度を算出し，一致度の和が最大となるように組み合わせを決定する．一致度を考慮することにより，属性同士の一致だけではなく，一致度合いの近い属性を有効に対応づけることが可能となる．

また，概念 *A*,*B* 間の一致する属性($a_i=b_j$)については，以下の処理により別扱いとする．$a_i=b_j$ なる属性があった場合，それらの属性の重みを参照し，$u_i>v_j$ となる場合は，a_i の重み u_i を u_i-v_j とし，属性 b_j を概念 *B* から除外する．逆の場合は，同様に b_j の重み v_j を v_j-u_i とし，属性 b_j を概念 *B* から除外する．一致する属性が T 組あった場合，概念 *A*, *B* はそれぞれ *A'*, *B'*として以下のように定義し直され，これらの属性間には一致する属性は存在しなくなる．

$$A'=\{(a'_1,u'_1),(a'_2,u'_2),\cdots,(a'_{L-T},u'_{L-T})\} \quad (5)$$
$$B'=\{(b'_1,v'_1),(b'_2,v'_2),\cdots,(b'_{M-T},v'_{M-T})\} \quad (6)$$

一致した属性の関連度を *DoA_com(A,B)*とし，以下の式で定

義する．

$$DoA\_com(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases} \quad (7)$$

次に，一致する属性を除外した *A'*，*B'*の関連度を *DoA_def(A',B')*とする．*DoA_def(A,B)*を算出するために，属性数の少ない方の概念 *A'*の並びを固定し，属性間の属性一致度の和が最大になるように概念 *B'*の属性を並べ替える．この時，対応にあふれた属性は無視する．概念 *A'*の属性 a'_i と概念 *B'*の属性 b'_x が対応したとすると，概念 *B'*は以下のように並び換えられる．

$$B'=\{(b'_x,v'_x),(b'_{x+1},v'_{x+1}),\cdots,(b'_{x+L-T},v'_{x+L-T})\} \quad (8)$$

この結果，一致する属性を除去した属性間の関連度 *DoA_def(A',B')*を以下の式によって定義する．

$$DoA\_def(A',B')=\sum_{s=1}^{x+L-T}DoM(a'_s,b'_s)\times\frac{\min(u'_s,v'_s)}{\max(u'_s,v'_s)}\times\frac{u'_s+v'_s}{2}$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases},\max(\alpha,\beta)=\begin{cases}\alpha(\alpha\ge\beta)\\ \beta(\alpha<\beta)\end{cases} \quad (9)$$

このように，一致する属性間の関連度 *DoA_com(A,B)*と，それら以外の属性間の概念関連度 *DoA_def(A',B')*をそれぞれ算出し，合計を概念 *A*，*B* の関連度 *DoA(A,B)*とする．

$$DoA(A,B)=DoA\_com(A,B)+DoA\_def(A',B') \quad (10)$$

関連度も，一致度と同様 0.0~1.0 の値をとる．1.0 に近いほど，関連の度合いが強いことを示す．

3.3 TF・IDF

TF・IDF 法[3]とは，語の頻度と網羅性に基づいた重み付け手法である．TF はある文書 d に出現する索引語 t（文書の内容を表す要素）の頻度 *tf(t,d)*を表す尺度である．IDF はある索引語が全文書中のどれくらいの文書に出現するという特定性を表す尺度である．なお，*N* を検索対象となる文書集合中の全文書数，*df(t)*を索引語 *t* が出現する文書数とする．また，文書 *d* における単語の総数を *W*，索引語 *t* の出現回数を *n* とする．このとき IDF は式 11 で，TF は式 12 で定義される．

$$idf(t)=\log_2\frac{N}{df(t)}+1 \quad (11)$$

$$tf(t,d)=\frac{n}{W} \quad (12)$$

3.4 未定義語の属性獲得手法

未定義語の属性獲得手法[4]とは，未定義語 X（概念ベースに定義されていない概念）の意味的特徴を表す単語（属性）とその重要性を表す重みの組を Web を用いて自動的に構成する手法である．まず，ロボット型検索エンジン[5]を用いて未定義語の検索を行う．そして，獲得した検索結果ページから形態素解析を行い，自立語を概念ベースに存在する語に限定して抽出する．その後，獲得した検索結果ページ内での自立語の出現頻度と Web-IDF を用いて，TF・Web-IDF 重み付けを行う．Web-IDF とは Web 上の文書を利用した IDF であり，式 11 の *N* を Google が保有している日本語のページ数，*df(t)*を索引語 *t* の Google で検索を行った際のヒット件数とすることで求めている．本研究では，未定義語の属性獲得手法を，オートフィードバック（*Auto Feedback : AF*）と呼ぶ．具体例を表 3 に示す．ここでは入力として「同志社」，「ミッキーマウス」，「スマホ」を設定するとそれらの語に関係する属性と重みが出力されている．

表 3　**オートフィードバック出力例**

入力語	オートフィードバックの出力
同志社	(研究, 117.4), (大学, 106.1), (学生, 95.5), (キャンパス, 94.6)…
ミッキーマウス	(キャラクター, 99.1), (マーチ, 83.5), (魔法, 68.1), (おもちゃ, 61.5)…
スマホ	(スマート, 431.5), (フォン, 360.2), (通話, 75.4), (機種, 66.5)…

4. 提案システム

提案システムは時事情報と個人情報との関連性や，時事情報の一般的重要性を考慮することでユーザが興味を惹かれると考えられる時事情報を選別する．システムの流れを図 2 に示す．

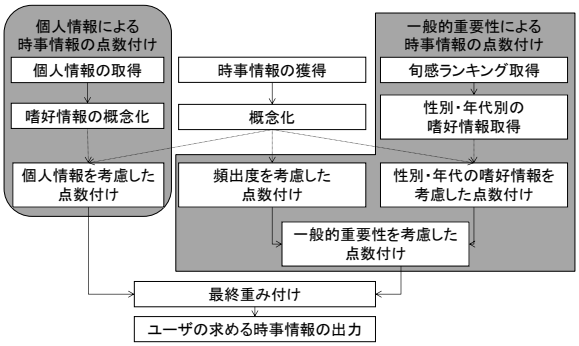

図 2　提案システムの流れ

4.1 時事情報の獲得と概念化

本システムは最初に最新の時事情報の見出しとその時事情報の記事本文全体を新聞社の Web サイトから獲得する．ニュース記事は短期間で更新されるため，時事情報獲得は 1 時，7 時，13 時，19 時の 1 日 4 回行う．

なお，新聞社の情報源として「asahi.com（朝日新聞）[6]」，「毎日 jp（毎日新聞）[7]」，「YOMIURI ONLINE（読売新聞）[8]」の 3 社のニュースを利用する．1 社だけのニュース記事のみを使用している場合ニュースの傾向やジャンルなどが偏ってしまう恐れがあるため，3 社の新聞社の Web サイトから提供されるニュースを用いることにより，情報の信頼性を保証している．

時事情報を取得した後に，見出しを概念，本文に存在する自立語を属性として時事情報の概念化を行う．概念化を行うことで，時事情報の内容とユーザに関する情報との関連性を直接調べることが可能となる．以下に，概念化の詳しい手法について記す．

まず，本文中から自立語を抜き出す．具体的には，時事

情報の本文に対して形態素解析ソフト「茶筌」[9]を用いて形態素解析を行い，本文中に含まれる自立語を抽出する．「茶筌」で形態素解析を行った場合，文は最小単位での意味を持つ自立語に区切られる．そのため，「条例改正」のように名詞の連続した単語が「条例」と「改正」に分けて抽出される．しかし，これでは時事情報中の語句が持つ本来の意味を失う可能性がある．そこで，名詞が連続して存在する場合には，自立語を接続し一語として抽出する．このようにして取得した語句をこの時事情報の属性とする．また，それぞれの属性に対して過去一月分の時事情報の本文を使用して TF・IDF 値を求め，その値を重みとして利用する．IDF については一記事ごとに数える．図 3 に記事の概念化の例を示す．

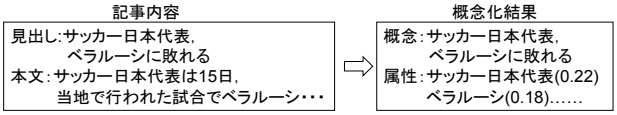

図 3　記事の概念化の例

4.2 個人情報による時事情報の点数付け

4.2.1 個人情報の取得と嗜好情報の概念化

時事情報獲得に適していると思われる情報として，45 個の項目を個人情報として利用する．個人情報の項目は人物情報と嗜好情報に分かれている．人物情報には名前や学校名といった基本的な項目が存在する．嗜好情報は好きなものと嫌いなものに分かれており，食べ物の項目ならば好きな食べ物と嫌いな食べ物の二つを入力してもらう．それぞれの項目に対してユーザ自身に記入してもらい，結果をそのユーザの個人情報として用いる．本研究で用いられる項目を表 4 に示す．

表 4　使用する項目

項目			
人物情報		嗜好情報（好きなもの，嫌いなもの）	
名前	出身地	食べ物	色
学校名	職業	スポーツ	飲み物
取得資格	ペット	昆虫	動物
特技	勤務先	季節	花
持病	国籍	国	教科
現住所	趣味	アーティスト	キャラクター
性格		作家	映画
		本	その他

個人情報を取得した後に，その一部である嗜好情報について概念化を行う．嗜好情報に存在する好きなものと嫌いなものの二種類について，それぞれを分けて概念化する．属性にはそれぞれの項目に格納されている語を使用し，重みは全て 1 とする．この概念化を行うことで，嗜好情報全体と，時事情報との関連度を調べることが可能となる．概念化の例を図 4 に示す．

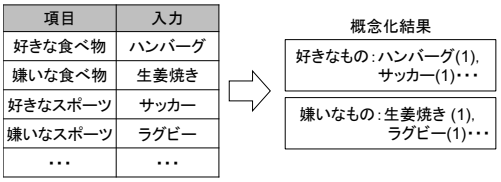

図 4　嗜好情報の概念化の例

4.2.2 個人情報を考慮した点数付け

時事情報と人物情報との関連性，時事情報と嗜好情報との関連性をそれぞれ調べ，合計した値を個人情報による点数とする．人物情報と嗜好情報で関連性を求める処理が違うため，それぞれの処理について以下に記す．

自分の出身地について載っている記事が存在する場合は，自分の嗜好に関係が無い内容でも興味を示しやすいと考えられる．このようにユーザは自身に関連する語が時事情報内に存在する場合，その時事情報に対して興味をもつと考えられる．そこで，人物情報については表記一致を用いることで時事情報内に同じ語が存在するかを判断する．具体的な手法としてユーザの人物情報に存在する語が時事情報内にいくつ存在するのかを調べ，存在した語数の割合を人物情報による点数とする．割合を取っているのは，結果となる値を 0.0 から 1.0 に収めることで，後に使用する関連度の値とのバランスをとるためである．

嗜好情報については，好きなものと嫌いなものそれぞれと時事情報について関連度を調べる．自分の好きなものについて書かれている場合，その時事情報に興味を示すと考えられる．嫌いなものについて書かれている場合は，逆に興味を示さないと考えられる．そこで，好きなものとの関連度の値から，嫌いなものとの関連度の値を引いた値を嗜好情報による点数とする．こうすることで，好きなものに書かれた時事情報は提供されやすく，嫌いなものに書かれた時事情報は提供されにくくなる．点数付けの例を以下の図 5 に示す．

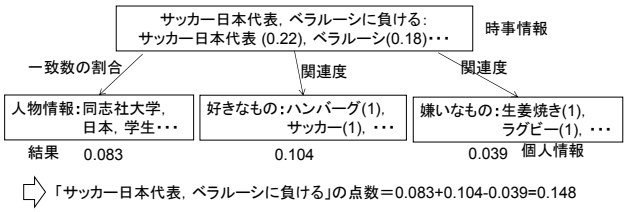

図 5　個人情報を考慮した点数付けの例

同志社大学，日本，学生といったような語が人物情報に存在するため，時事情報内にそれらの語が存在するかを調べてその割合である 0.083 を取得している．嗜好情報については両方とも関連度を取っており，好きなものについては 0.104，嫌いなものについては 0.039 という値をとった．よって，嗜好情報による点数は 0.104 から 0.039 を引いた値である 0.065 となる．そして人物情報による点数を足し合わせた値である 0.148 という値が，個人情報による時事

情報への点数となる．

4.3 一般的重要性による時事情報の点数付け

4.3.1 頻出度を考慮した点数付け

頻繁に報道されているニュースほど重要性が高いと考えられる．記事の見出しに存在する名詞が一日の時事情報の中にどれだけ記述されているかを調べることで，頻繁に報道されているかを判断できる．そこで，記事の見出しに存在する語について一日に記事にされた回数を調べることで記事の重要性を判断する．

まず，一日分の記事の見出しを集めたリストを作成する．ある一つの見出しに存在する名詞全てを取得し，名詞それぞれについてリスト内に何回出力されているかを調べる．見出しに存在する名詞の個数を i 個，それぞれの名詞のリスト内での重複回数を a_i だとすると，記事の頻出度を調べる式は以下のようになる．

$$\text{記事の頻出度} = \frac{1}{9i}\sum_{h=1}^{i} a_h \tag{13}$$

見出しのみを利用しているのは，見出しはその記事の内容を端的に表した名詞のみが存在するため，記事本文に比べ雑音となる語が少ないと考えられるからである．総和を名詞の個数に 9 をかけた値で割るのは，そのまま総和を点数としてしまうと見出し中に存在する名詞が多い記事ほど値が高くなりやすいからである．9 という値は，様々な値で実験を行うことで最適化した値である．

4.3.2 旬感ランキングとキーワードの抽出

旬感ランキングとは，BIGLOBE サイトが提供する検索エンジンによって検索されたキーワードを集計し，性別・年代別で 10 代から 50 代までの急上昇ワード上位 20 位までをランキング形式にまとめたものである．

旬感ランキングに存在するキーワードそれぞれがその性別・年代にとって興味のある語だとみなし，データとして取得する．過去一週間分の旬感ランキングのキーワードを各性別・年代の嗜好情報取得に使用する．

4.3.3 性別・年代別の嗜好情報抽出

旬感ランキングを使用して興味のあるジャンルを集めたジャンル概念と興味のあるキーワードを集めたキーワード概念を作成し，嗜好情報として取得する．

まず，ジャンル概念について述べる．先ほど取得したキーワードそれぞれがどのような意味合いを持つ語であるかを調べることで，嗜好を知ることができる．そのため，それぞれのキーワードに対して *AF* を行うことで属性を取得する．属性は *AF* 対象となっている語の意味合いを示す．同じ属性が多数存在するようであれば，その意味合いに関する語が何度もランクインしているということであり，その属性が示すようなジャンルに関して興味，関心があると考えられる．属性の重複回数を調べ，その回数が多い順に属性を並べる．上位 20 位までの属性を，その性別・年代の興味のあるジャンルとして取得する．その後上位 20 位までに入ったジャンルを属性として使用することでジャンル概念を作成する．重みは 1 位なら 20，2 位なら 19 といったように順位に沿ってつけられる．

次に，キーワード概念について述べる．旬感ランキングは検索急上昇ランキングであるので，短い期間中に同じキーワードが存在することはほとんどない．一度ランキングに載ると，もう一度載るには以前よりも多くの検索回数が必要となるからである．よって，もし複数回同じキーワードが存在する場合はそのキーワードに対しての興味度合いが高くなり続けていると考えられる．そこで，重複して出現しているキーワードをその性別・年代の興味のあるキーワードとして取得する．重みにはそのキーワード自身の重複回数を使用する．キーワードとジャンルで概念を分けているのは，ジャンルの重み 20 と，キーワードの重み 20 では全く意味が異なるからである．例を図 6 に示す．

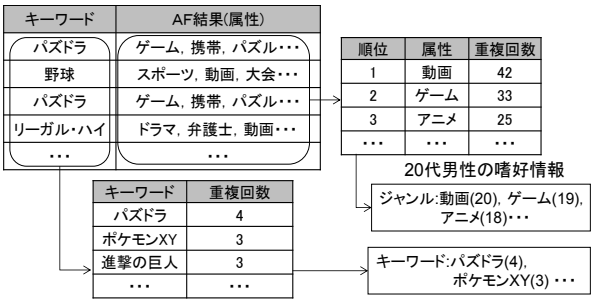

図 6　20 代男性の嗜好情報取得の例

この例では，20 代男性におけるキーワードを使用して嗜好情報の取得を行っている．まずパズドラといったようなキーワードから属性であるゲームや携帯などといった語を取得し，重複回数を調べている．その重複回数の多い属性をジャンル概念の属性に使用している．また，キーワード自身の重複回数も調べ，多いキーワードを属性としたキーワード概念を作成している．

4.3.4 性別・年代の嗜好情報を考慮した点数付け

時事情報と 4.3.3 項で取得した性別・年代の嗜好情報との関連度を調べることで点数付けを行う．その際，ユーザと同じ性別・年代の嗜好情報を使用する．興味のあるジャンル，キーワードともに概念化しているので，直接関連度を計算することができる．ジャンルとの関連度とキーワードとの関連度を足し合わせたものをその時事情報の点数とする．

4.3.5 一般的重要性を考慮した点数付け

4.3.1 項で取得した頻出度による点数と 4.3.4 項で取得した性別・年代の嗜好情報による点数の両方を考慮することで，一般的重要性を調べることが可能となる．これらの点数のどちらかが低ければ，ユーザにとって一般的に重要な時事情報だとは言えない．よって，両方の点数を掛け合わせた値をその時事情報の一般的重要性とする．こうすることで，どちらの点数も高い時事情報が提供されやすくなる．

4.4 最終重み付けと出力

ユーザ自身の嗜好と一般的重要性の両方を考慮することで，時事情報の順位付けを行う．

本研究の目的は冒頭で述べたとおり，ユーザの嗜好に合った時事情報や一般的重要性のある時事情報の提供である．よってどちらかが低くても片方が高ければ，それは提供されるべき時事情報であると考えられる．足し合わせた結果を最終重みとするべきであるが，個人情報による点数は加減算のみで求められるのに対し，一般的重要性による点数は掛け算を使用しているため値が低くなってしまうという問題が存在する．そこで，一般的重要性の点数にさらに 1.5 を掛けたものと個人情報による点数と足し合わせることで最終重みを決定する．これにより，バランスの良い時事情報の提供が可能となる．1.5 という数値は，様々な値で実験を行うことで最適化した値である．

5. 評価実験

個人情報と一般的重要性に基づく時事情報提供システムの出力について，評価実験を行った．

被験者はあらかじめ，実験を行う日の全ての時事情報の見出しと本文を見て，それぞれの時事情報が本人にとって興味を惹かれるものであるかの判断を行っている．被験者が興味を惹かれると判断した時事情報を，正解となる情報とみなす．

実験には 2013 年 12 月 2 日，3 日，4 日の 3 日間に収集した時事情報と，11 月 25 日から 12 月 3 日までの旬感ランキングを用いた．5 人の被験者より個人情報を収集し，正解となる情報の判断も行ってもらった．被験者 5 人と実験 3 日間，合計で 15 種類の出力結果を収集し，評価を行った．20 位までに正解となる時事情報がいくつ存在するか，その割合を評価指標として使用する．本システムの出力は時事情報を点数順に並び替えたものであるが，ユーザは上位に存在する時事情報しか目を通さないと考えられる．そのため，上位にどれだけ正解となる時事情報が存在するかが重要となる．20 という数値は，一日に取得できる時事情報の約 10%の数値である．また，今回は比較対象として，ユーザの個人情報だけを考慮して提供された結果と一般的重要性のみを考慮して提供された結果を用いる．

5.1 評価結果

本研究と比較対象それぞれの結果について，全ての評価結果の平均を以下の表 5 に示す．

表 5　全体の平均精度の比較

	本研究	個人情報のみ	一般的重要性のみ
平均精度	46.7%	41.3%	41.6%

5.2 考察

比較対象と比べ，精度が 5%ほど上昇している．上位により有益な時事情報を提供できた結果である．

内容の考察に移る．比較対象よりも精度がよくなっていた被験者 C の 12 月 2 日の結果のうち，正解となる時事情報を表 6 に，個人情報の一部を表 7 に示す．

表 6　被験者 C の結果における正解となる時事情報

順位	時事情報
3	日本郵便、定形外や速達を値上げ…消費増税で
4	京都・鴨川ステージ計画、周辺の商店主ら反発
7	中国と不測の事態恐れ…米３航空は飛行計画提出
11	中国「防空圏」撤回へ連携…米副大統領が来日へ
12	英海軍参謀長、日本の立場「支持」…中国防空圏
16	ローソン;景品マグカップを回収
17	防空識別圏;首相「撤回求めて、米国とも連携」
18	訪印の両陛下、５３年前に植樹の菩提樹をご覧に
20	毅然と・冷静に…首相、防空圏で日米連携強調

表 7　被験者 C の個人情報の一部

項目	個人情報	項目	個人情報
名前	被験者C	国籍	日本
出身地	奈良	趣味	読書
好きな国	日本	嫌いなスポーツ	陸上競技
好きな教科	歴史	嫌いな教科	数学・体育

被験者 C の個人情報は国籍が日本であり，また好きな国も日本と示されているため，日本に関して書かれている時事情報に対しての点数が高くなった．特に表 6 における 3 位の記事である日本郵便に関しては，嗜好情報のみを考慮した場合は 1 位となっていた．しかし，日本以外の記事に関しては挙がりにくくなってしまっていた．12 月 2 日は中国の防空圏について情報がよく報道されており，一般的重要性の観点から見た場合はこの話題に関する記事が上位に挙がっていた．被験者 C も防空圏に関する時事情報について興味があるとしている．しかし，記事はどれも中国と日米における防空圏についての論争が書かれていたため，個人情報からの観点のみでは，日本のみについて記述された記事よりも点数が低くなっていた．よって個人情報のみを使用した場合は上位に挙がってこなかったが，一般的重要性と個人情報の両方を利用することで，0.146 という値で防空圏に関する記事を上位に挙げることができた．

一方で，精度が低くなっていた被験者 A の 12 月 4 日の結果のうち，正解と判断された時事情報を表 8 に，個人情報の一部を表 9 に示す．

表 8　被験者 A の結果における正解となる時事情報

順位	時事情報
6	対戦したことないチームと…ザック、抽選に出発
7	日本、北中米勢と対戦せず…サッカーW杯
9	サッカー・ＪＦＬ、来季は１４チームに
14	日本の学習到達度、全分野で上昇…脱ゆとり成果
18	奈良女大管理職、日常的に罵声「お前はだめだ」
19	リーガルハイ;松平健が再びゲスト出演 古美門・堺と最終対決！

表 9　被験者 A の個人情報の一部

項目	個人情報	項目	個人情報
名前	被験者A	国籍	日本
出身地	奈良	趣味	ゲーム
好きなスポーツ	サッカー	嫌いなスポーツ	野球
好きな教科	数学	嫌いな教科	国語
好きな動物	犬，鳥	嫌いな動物	スカンク

被験者Aは好きなスポーツをサッカーだと個人情報で示している．結果として，ほとんどの出力でサッカーに関する記事が上位に挙がっていた．一方で，10，11，12，16位に為替に関する記事が存在した．これは，一般的重要性に関して何度も報道されたような記事が存在せず，結果として株式や為替といったような定期的に情報が提供される語に関しての記事の頻出度が高くなってしまったのが原因である．被験者Aは為替等に関する時事情報については特に興味を示さなかった．しかし，個人情報では好きな教科として数学が挙げられているため，経済的な時事情報である為替との関連度は低いものではなかった．結果として，個人情報と一般的重要性の両方を考慮することで，為替等の時事情報が0.081という値で上位に挙がってきてしまった．このように，特に重要と考えられる時事情報が存在しない場合は定期的に提供される時事情報が上位に挙がってしまうという問題点が存在する．

6. 今後の展望

6.1 情報源の改善

5.2節では定期的に出力される経済に関する時事情報が上位に挙がるといった問題点があった．その対策の一つとして，情報源の改善が挙げられる．

本研究では時事情報の情報源として読売，朝日，毎日の三社のWebページを利用した．この三社は社会的な時事情報が豊富に存在するため，政治や経済といったジャンルの時事情報の一般的重要性が高くなりやすい．一方でエンタメといった一部のジャンルの時事情報については数が少なく，一般的重要性は高くなりにくい．そのため，政治的に重要とされる時事情報が存在しない場合は，定期的に出力される経済の時事情報が上位に挙がる可能性が高くなっているのである．

解決策として，エンタメなどの社会的な時事情報以外のものをメインに扱っているWebページを情報源として追加することが挙げられる．エンタメに関する記事の一般的重要性を上げることで，定期的に出力される時事情報が上位に挙がる可能性を下げることが可能となる．また，個人情報には好きな本や映画，アーティストといったエンタメに通じる項目が多く存在する．それらを活用することで，評価をさらに上げることができると考えられる．

6.2 閲覧履歴の利用

現在の個人情報の項目だけでは，すべての時事情報に対して興味のあるなしの判定を行うことは不可能である．また，個人情報は事前入力であり，ユーザに新たな興味が出た場合にその語句を得ることはできない．ユーザの嗜好がどのような時事情報に存在するのかを調べ，嗜好を適宜変更していく仕組みが必要だと考えられる．

時事情報が提供された後，ユーザは時事情報の本文を読むと考えられる．ユーザが閲覧した時事情報については閲覧したという記録を残し，その時事情報を活用して新たな嗜好を探っていく．最初は閲覧の記録がないため情報が少ないだろうが，それを続けていくうちに数多くの時事情報の閲覧の記録が残るはずである．それらの時事情報の本文に存在する語を調べることで，ユーザにとって重要な語句を調べられるのではないだろうか．その語をユーザの嗜好として保存することで，項目からでは拾えない新しい嗜好を探ることができると考えられる．

7. おわりに

本研究では，Webから時事情報を獲得し，そこからユーザが求める時事情報を選出する手法を提案した．具体的には，個人情報からユーザの嗜好を探ったと同時に，ユーザと同性別・年代の人たちの嗜好や頻出度から時事情報の一般的重要性を調べた．そしてユーザの嗜好と一般的重要性の両方を考慮することにより，よりユーザに適した時事情報を出力する手法を提案した．

提案手法を用いることで，46.7%という精度で，個人の趣味・嗜好に沿った時事情報の選別，提供を行うことができるようになった．一般的重要性を考慮することで，ユーザの嗜好だけでは提供できなかった時事情報に関しても，提供できるようになった．更なるシステムの改良により，ユーザの情報収集における負担を軽減することが可能となると考えられる．

謝辞 本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

1) “BIGLOBEサーチ旬感ランキング”, http://search.biglobe.ne.jp/ranking/
2) 渡部広一，河岡司，“常識判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001.
3) 徳永健伸，“言語処理と計算5情報検索と言語処理”，東京大学出版会，1999.
4) 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，第18回人工知能学会全国大会論文集，2D1-01，2003.
5) “Google”, http://www.google.co.jp/
6) “asahi.com：朝日新聞社の速報ニュースサイト”, http://www.asahi.com/
7) “毎日ｊｐ - 毎日新聞のニュース・情報サイト”, http://www.mainichi.jp/
8) “ニュース　速報　YOMIURI ONLINE（読売新聞）”, http://www.yomiuri.co.jp/
9) 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明，“日本語形態素解析システム『茶筌』version1.0 使用説明書”,NAIST TechnicalReport, NAIST-IS-TR97007 ，1997.