Vol.2013-ICS-170 No.1
2013/3/10

嗜好に基づく時事情報推薦システムの構築

山本達也[†1]　芋野美紗子[†2]　土屋誠司[†3]　渡部広一[†4]

本システムでは，テレビの視聴履歴およびソーシャルネットワークシステムである Twitter での発言履歴を収集し，個人の嗜好を分析する．その上で，Web から獲得した時事情報からの話題となる語と嗜好情報との関連性を計算することで個人の嗜好に基づいた時事情報を推薦するシステムを提案する．

Developing the Topical Information Offering System Based on Personal Taste Information

TATSUYA YAMAMOTO[†1]　MISAKO IMONO[†2]
SEIJI TSUCHIYA[†3]　HIROKAZU WATABE[†4]

The proposal system flexibly offers topical information by associating the word based on personal taste information. The system uses taste information from record of viewing TV and record of tweeting on “Twitter”.

1. はじめに

近年，人間同士のコミュニケーションにおいて時事情報というものは欠かせないものとなっている．また，インターネットの発展・普及により他のメディア(TV・新聞など)と比べ，インターネットは容易に天気やニュース記事などの時事情報を手に入れられるようになった．しかし，Web 上の時事情報は他のメディアと比べ，更新頻度が高くその情報量は日々増加しており，興味がある時事情報だけを取得することが困難になってきている．そこで，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．

既存システムとして，Web を用いてニュース記事を収集し，ユーザに最も有益であると考えられる時事情報を推薦する「テレビ視聴履歴を基にした時事情報推薦システム」[1]が存在する．しかしこのシステムではテレビを視聴していない人への時事情報の推薦が行えずユーザが既存システムを使用するにはテレビの視聴履歴を取る期間が必要であるという問題があった．

よって本研究では，テレビの視聴履歴に加えてソーシャルネットワークサービスである Twitter[2]での発言から情報を取得し嗜好情報の幅を広げるとともに，視聴履歴を取得せずとも Twitter からの発言履歴があれば即座にシステムを使用することができるように改善を加え精度向上を図る．

嗜好情報として，テレビは多数の番組が同時に放送されており，ユーザは自分の見たいものを常に視聴しているという考えから，視聴履歴をシステムに学習させることで個人の嗜好情報を抽出できると考えられる．また Twitter でのつぶやき(発言)には自分の嗜好に共感できる物事・事象についての発言が含まれていると考えられるので，発言を収集することで個人それぞれの嗜好の傾向が取得できる．

2. 時事情報提供システムの概要

本論文では，テレビの視聴履歴および Twitter での発言を基に個人の嗜好を抽出し，時事情報との関連性を計算することで個人それぞれに適した時事情報を提供するシステムを提案する．本システムの概要を図 1 に示す．

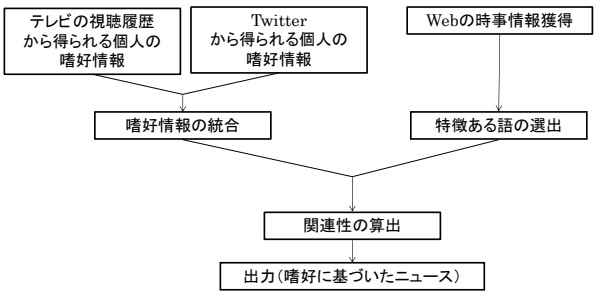

図 1　本システム概要図

個人の嗜好を解析するために，本システムではテレビの視聴履歴および Twitter の発言履歴を収集する．テレビ視聴履歴は，ユーザが視聴した番組に関係する語をテレビ王国[3]の Web サイトから収集する．また，Twitter 発言履歴を解析し個人の嗜好情報を取得する．この２つの嗜好情報を統合することによって個人の嗜好情報として扱っていく．

また，Web から時事情報を獲得していくのだが，人間が時事情報を閲覧する際には，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，本研究で扱う時事情報は，新聞社および株式会社オリコンの 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．

Web ニュースの見出し記事を獲得しその中から特徴となる単語（以下，話題語とする）を見つけ出す．その後，嗜

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

好情報と話題語との関連性を定量化していく．最後に，話題語が表記されている時事情報に対して，嗜好情報との関連性を定量化した値を重要度として付与した上で，値が高い順に時事情報を出力していく．

3. 使用技術

3.1 概念ベース

概念ベース[4]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 3.1)．概念ベースには，約 9 万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．なお，本論文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．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (3.1)$$

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[5]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．

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	0.86
	雑誌	0.23
	運動	0.007

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．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($L \leqq M$).

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L)\} \quad (3.2)$$
$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M)\} \quad (3.3)$$

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．

$$DoM(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases} \quad (3.4)$$

ただし，$a_i=b_j$ 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．

$$A'=\{(a'_1,u'_1),(a'_2,u'_2),\cdots,(a'_{L-T},u'_{L-T})\} \quad (3.5)$$
$$B'=\{(b'_1,v'_1),(b'_2,v'_2),\cdots,(b'_{M-T},v'_{M-T})\} \quad (3.6)$$

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．

$$DoA\_com(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases} \quad (3.7)$$

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．

$$B' = \{(b'_x, v'_x), (b'_{x+1}, v'_{x+1}), \cdots, (b'_{x+L-T}, v'_{x+L-T})\} \quad (3.8)$$

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．

$$DoA\_def(A', B') = \sum_{s=1}^{x+L-T} DoM(a'_s, b'_s) \times \frac{\min(u'_s, v'_s)}{\max(u'_s, v'_s)} \times \frac{u'_s + v'_s}{2}$$
$$\min(\alpha, \beta) = \begin{cases} \alpha(\alpha \le \beta) \\ \beta(\alpha > \beta) \end{cases}, \max(\alpha, \beta) = \begin{cases} \alpha(\alpha \ge \beta) \\ \beta(\alpha < \beta) \end{cases} \quad (3.9)$$

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．

$$DoA(A, B) = DoA\_com(A, B) + DoA\_def(A', B') \quad (3.10)$$

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 Web-IDF

IDF[6] 法とは，一般的な文書（新聞や書籍など）を用いて索引語の特定性を考慮する手法である．IDF の中でも特に，Web-IDF は Web 上にある文書のみを用いて索引語の特定性を考慮する手法である．Web-IDF では式 3.11 のように算出される．

$$idf(t) = \log_2 \frac{N}{df(t)} + 1 \quad (3.11)$$

なお, N は 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 で検索を行ったときのヒット件数としている．ここでは，Google が保有している日本語のページ数を，日本語の文書として最も使われている「は」で検索を行ったヒット件数約 5,580,000,000 件（2014 年 1 月 4 日現在)）としている．この理由として，Google が全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていないためである．

3.4 未定義語の属性獲得手法

未定義語の属性獲得手法[7]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[8]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．本論文では，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. 個人の嗜好に基づいた時事情報提供処理

時事情報と嗜好情報との関連性を求め，ユーザが興味を惹かれると考えられる時事情報を出力するシステムの流れを図 2 に沿って示していく．

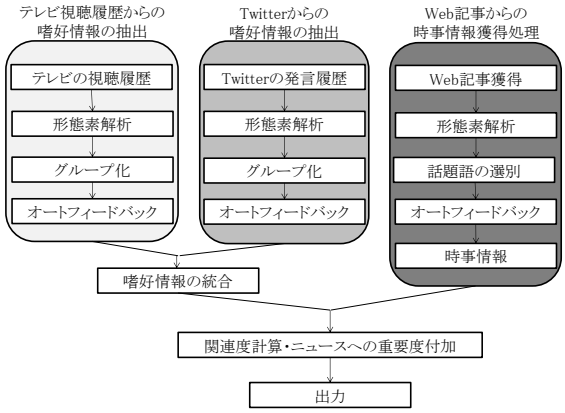

図 2 **本システムの流れ**

4.1 Web 記事からの時事情報獲得処理

時事情報を提供するために，本システムでは最新の時事情報を新聞社および株式会社オリコンの 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）[9]」，「毎日 jp（毎日新聞）[10]」，「YOMIURI ONLINE（読売新聞）[11]」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に，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．また，新聞社だけでは「芸能」，「音楽」などのサブカルチャ系の情報が乏しいため，株式会社オリコンの Web サイト「Oricon Style[12]」を情報源に加えている．

Web サイトから取得する時事情報の中で，その時事情報を特徴づける単語である話題語を取得する方法を説明する．Web から獲得してきた 1 日分の時事情報に対して形態素解析ソフト「茶筌」[13]を用いて形態素解析を行い，時事情報の文中に含まれる自立語を話題語として抽出する．「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，名詞の連続や，「男性の遺体」の

ような「名詞」と「名詞」の間に格助詞「の」を挟んだ連続した語が存在する場合に，自立語を接続し話題語として抽出する．このとき，「ページ」や「リンク」など，Web 上では頻繁に出現するがニュースとは関係のない語を取り除く必要がある．そこで，Web-IDF を用いて，閾値である 3.0 未満の Web-IDF 値を持つ話題語を削除する．閾値は既存研究[11]での実験によって出力が特徴ある語だけになるように最適化された値である．Web-IDF が 3.0 以上の話題語に対しオートフィードバックを利用し属性と概念の組のセットを時事情報として保存する．

4.2 個人の嗜好情報の抽出

この節では，本システムにおける個人の嗜好情報の抽出方法について説明する．テレビの視聴履歴および Twitter の発言履歴から自立語を取得し，グループに分け，*AF* を利用し属性と概念の組の集合を個人の嗜好情報として収集する．

4.2.1 テレビ視聴履歴からの嗜好情報の抽出

個人の嗜好情報を取得するために, 本研究では 20 代の男女 5 名の被験者からテレビの視聴データを以下のグループに分けて収集した．なお，視聴データとは番組のタイトルおよび Web の電子テレビ表から得られる番組の紹介文である．

・グループ◎：集中して視聴していた番組（録画も含む）
・グループ○：視聴したかったが見られなかった番組
・グループ△：なんとなく視聴していた番組

また，本研究では視聴データとして得られた番組タイトル・番組紹介文から形態素解析によって自立語を取得し，オートフィードバック(*AF*)法を用いてその自立語に対する属性と重みを獲得する．その自立語の属性と重みを被験者の嗜好情報として使用する．

4.2.2 Twitter からの嗜好情報の抽出

本研究では，Twitter を日頃から利用している被験者から，被験者自身が発言しているツイート（呟き）を取得する．ツイート取得には，アプリケーションである「Tween[14]」と「TweenPopup[15]」を利用した．取得したツイートから形態素解析により自立語を獲得し，得られた自立語に対してオートフィードバック(*AF*)を用いてその属性と重みを獲得する．その自立語の属性と重みを被験者の嗜好情報として使用する．なお，Twitter における発言には良い印象のある発言・悪い印象のある発言・そのどちらでもない発言の 3 通りの発言があると推測される．ゆえに，上記の発言種類ごとに発言を収集し嗜好情報として収集する．発言分別には各発言の形容詞による表記一致により行う．「良い」「きれい」「おいしい」などの表記があるものは良い印象，「うざい」「めんど（い）」などの表記があるものは悪い印象，その他をどちらでもないものとして収集した．表記一致に使用した形容詞をリスト 1，リスト 2 に示す．

リスト 1　Twitter における良い印象と分別される形容詞

ありがたい　いい　よい　きれい　かわいい　かっこいい　うれしい　うまい　おいしい　たのしい　すばらしい　やさしい　めずらしい　まちどおしい　ほしい　なつかしい　つよい　えらい　うつくしい　うらやましい　このましい

リスト 2　Twitter における悪い印象と分別される形容詞

うざい　めんどうくさい　わるい　きたない　くどい　くやしい　くらい　けだるい　きみわるい　ひどい　つまらない　まずい　みにくい　ぼろい　ばからしい　にくい　なさけない　ださい　とろい　ずるい　うっとうしい　いやらしい　グロい　ケバい　きらい　さむい　あつい

4.2.3 嗜好情報の統合

本研究では，視聴履歴から得られる嗜好情報と Twitter から得られる嗜好情報を統合し，被験者の嗜好情報として保存する．具体的には，Twitter から得られる良い印象にグループ化された *AF* 群をテレビ視聴履歴のグループ◎と足し合わせ，またどちらでもない印象に分別された Twitter から得られる *AF* 群をテレビ視聴履歴のグループ○に足し合わせていく. 悪い印象のものは嗜好情報から外す. なお, Twitter での発言において，どちらでもない印象のものを使用する理由は，形容詞が含まれていなかったとしても個人に関係する自立語が含まれている可能性が高いためである．

4.3 話題語への重み付け

個人の嗜好に基づいた時事情報を提供するために，話題語と統合された嗜好情報との関連性を求め，話題語毎の重みとして値を付与する．

あらかじめ取得した話題語とテレビ視聴履歴及び Twitter でのつぶやきの自立語の AF から得た属性と重みを利用し，関連度を計算することで話題語毎に関連度を求める．これらの関連度を嗜好情報の総数で割った値を嗜好区分に応じた重みに掛け合わせることでその話題語の重みとする．なお，区分に応じた重みは，実験により個人の嗜好情報を最も推薦することのできた値の◎(Twitter:良い印象 TV:集中して視聴していた)は 2.0，○(Twitter:どちらでもない印象 TV:見たかったが見られなかった)は 1.0, △(Twitter:該当なし TV:なんとなく見ていた)は 0.7 とした．話題語への重み付けの具体例を図 3 に示す.

図 3 では，テレビの視聴履歴からの嗜好情報として「報道・日本・国語」，Twitter からの発言履歴からの嗜好情報として「テニス・お菓子」があり，時事情報からの話題語として「大リーグ」といったものがあると想定している．最初に，嗜好情報「報道・日本・国土・テニス・お菓子」と話題語「大リーグ」との関連度を計算をする．その際にそれぞれの嗜好情報に付随している嗜好区分(◎:2.0，△:0.8)によって補正がかけられる．次に，でてきた関連度を全て足し合わせ，総嗜好数(図 3 では 5 個)で割った値

0.1036 が大リーグの重要度となる.

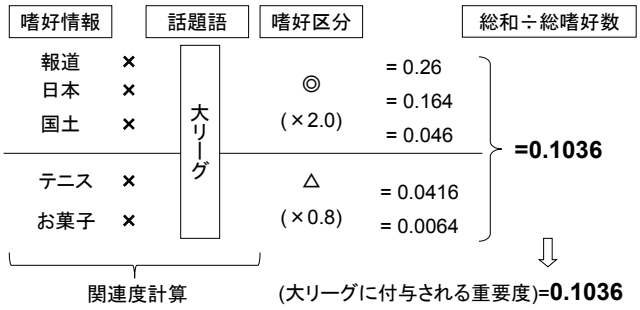

図 3 話題語への重み付け

4.4 時事情報への重要度付与

4.3 節で個人の嗜好情報によって重み付けを行った話題語を用いて，それぞれの時事情報に重要度を付与する．時事情報中に話題語が存在するかどうかを表記一致により調べ，時事情報中に存在している話題語の重要度をその時事情報の重要度とする．最後に，出力として時事情報への重要度が高い順に時事情報をソートし，上位から順に出力する．

5. 評価

個人の嗜好情報に基づいた時事情報提供システムの評価は，被検者を募りテレビ視聴履歴および Twitter での発言履歴を収集し、興味がある時事情報がシステムの出力上位にどの程度あるのかについて評価を行った．

5.1 評価方法

評価には 2012 年 12 月 12 日から 25 日の 13 日間に収集したテレビの視聴履歴および 25 日における被験者の最新の Twitter からのつぶやき 200 件, 2012 年 12 月 25 日におけるニュースの見出し情報を用いた．5 人の被験者より個人情報を収集し，また，正解となる時事情報の判断も行った．なお, 評価には 2012 年 12 月 25 日の時事情報 462 個を使用した．

被験者はあらかじめ 2012 年 12 月 25 日の時事情報を見て，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，正解となる時事情報とみなす．正解となる情報を参照しシステムの出力による順位を調べることで評価を行う．評価指標は，上位 30 位における興味がある情報の割合とする．

5.2 評価

個人の嗜好に基づいた時事情報獲得における被験者毎の評価を表 4 に示す．表 4 より TV・Twitter からの正解率とはシステムから出力された重要度の高い順に上位 30 個を調べたときの被験者の興味のある時事情報の割合という意味である．また，TV からの嗜好のみ・Twitter からの嗜好のみの項目については,評価実験を行った期間の嗜好情報から該当の嗜好以外の嗜好情報を取り除いた状態で評価を取った場合の評価である．最後に,ランダムで取得した場合の平均正解数とは，全時事情報中に被験者が興味のあった時事情報がどの程度の割合で含まれているかを調べ，その割合に応じて時事情報を 30 件取得した場合に正解となる時事情報がどの程度含まれている割合を示す値である．

表 4 評価実験結果

	被検者A	被検者B	被検者C	被検者D	被検者E	平均
TV･Twitterの嗜好からの正解率	56.7%	26.7%	33.3%	66.7%	50.0%	46.7%
TVからの嗜好のみでの正解率	33.3%	20.0%	30.0%	50.0%	33.3%	33.3%
Twitterからの嗜好のみでの正解率	26.7%	13.3%	30.0%	43.3%	30.0%	28.7%
462件の全時事情報中の正解数	98個	63個	76個	152個	85個	94.8個
ランダムで30件 取得した場合の平均正解率	21.2%	13.6%	16.5%	32.9%	18.4%	20.5%

5.3 考察

被験者毎における出力において，ランダムで取得した場合の平均正解率と TV・Twitter からの正解率の差が最も高かった被験者 A において，TV および Twitter からの嗜好情報に基づいて出力された時事情報上位 30 位中で, 興味があると判断した時事情報を表 5 に示す.

表 5 上位 30 件までの被験者 A の興味がある時事情報

順位	時事情報
6	紅白歌合戦、大トリはＳＭＡＰが３年連続
7	ＮＨＫ紅白大トリ、３年連続でＳＭＡＰ
8	ルーベンス傑作、進む修復　来春、日本初公開向け
10	銀座で最古…松坂屋銀座店が来年６月で営業終了
14	女子ゴルフ宮里藍は９位変わらず　世界ランキング
15	大竹しのぶ、『24時間チャリティラジオ』完走
16	平清盛、視聴率最低…「薄汚れた画面」酷評も
17	平清盛最終回9．5%　期間平均は史上最低12．0%
18	仕事帰りに労働相談できます　新橋で２６日夜、無料
20	東日本大震災ブイ式海底津波計、三陸で運用開始
21	東京駅の投影ショー、２５日以降も中止　観客殺到で
22	北京–広州…世界最長の高速鉄道が全線開通へ
23	PRICELESSキムタクの月9　最終回視聴率18．7%
24	しずちゃん、ガックリ…不戦勝に「申し訳ない」
26	外食産業、客数減もファミレスなどで売上増
27	ＮＨＫ大河「平清盛」、平均視聴率過去最低に
28	Ｂ７８７の操縦席窓にひび、岡山空港に引き返す

この被験者 A の TV 履歴を見てみると,「ミュージックステーション」などの音楽番組や「ダウンタウン DX」などのお笑い・バラエティが多く見られた．この結果から表 5 の順位 6, 7 位における「SMAP」や順位 24 位における「しずちゃん」などの表記がある時事情報との関連度が高くなり重要度が他の時事情報より高くなったものと推測される．また，被験者 A の Twitter 発言履歴からの自立語の一部をリスト 3 に表記する．被験者 A においては海外旅行へ行った事が Twitter のつぶやき 200 件に含まれていたため外国の国名や地方の名前が多く含まれていた．その事から順位 2 位「国連分担金日本は大幅減…１３～１５年加盟国中最大」,3 位「２０１５年に国連防災世界会議　日本開催は３回目」などの国際記事が上位に来たと推測される．

また今回，視聴履歴・Twitter からの嗜好情報を組み合わせたことで表 5 の順位 16, 17, 27 位のような京都にゆかりのある「平清盛」とテレビと関連性の強い語の「視聴率」といった独自の時事情報を推薦することができた．これは

表 7 における TV からの嗜好のみでの推薦および Twitter からの嗜好のみでの推薦ではシステム出力の上位 30 位に入ってこなかった新しい時事情報であった.

リスト 3　被験者 A の Twitter からの自立語

イタリア ロンドンの空港 本場ナポリ ピザ カプリ島 青の洞窟 ローマ イタリア 旅行 最終日 Mac 東京 ソフバン iPhone5 リーク 情報 NFC ナポリ 安心感 カプリ島 ミラノドリア ナポリ 京都駅 風邪 健康体 来週末 資格試験 四条 ジャクソン 不吉な予感 フォーム ジャンプ コンビニ ダッシュ レッスン スクール キレ スピード スピンサーブ 京都就活生 テニス

一方で，被験者毎における出力からランダムで取得した場合の平均正解率と TV・Twitter からの正解率の差が最も低かった被験者 C において，Twitter からの嗜好情報に基づいて出力された時事情報上位 30 位までの被検者が興味あると判断した時事情報のみを表 6 に示す.

表 6 上位 30 件までの被験者 C の興味がある時事情報

順位	時事情報
3	紅白歌合戦、大トリはＳＭＡＰが３年連続
4	ＮＨＫ紅白大トリ、３年連続でＳＭＡＰ
7	銀座で最古...松坂屋銀座店が来年６月で営業終了
12	実在選手で夢のチーム...「実況パワフルプロ野球２０１２　決定版」
18	マリオの菓子「チョコエッグ（ＮｅｗスーパーマリオブラザーズＷｉ）」
22	中山雅史さん木梨JAPANでなでしこサッカー対決　新春特番に参戦
27	フィギュア真央、羽生ら余裕の舞　エキシビション
28	東京駅の投影ショー、２５日以降も中止　観客殺到で

この被験者 C の TV 履歴を見てみると,「日清食品 The MANZAI2012」や「ネプリーグ」などのお笑い・バラエティが多く見られた．この結果から順位 9 位「しずちゃん、ガックリ…不戦勝に「申し訳ない」」におけるお笑い芸人「しずちゃん」の表記がある時事情報との関連度が高くなり重要度が他の時事情報より高くなったものと推測される.

一方で，被験者 C の Twitter 発言履歴からの自立語の一部をリスト 4 に表記する．被験者 C の主なつぶやきは「TV ゲーム」に関することが多数見られたため表 6 における順位 12，18 位における「実況パワフルプロ野球」「マリオ」などゲームに関する時事情報が上位にきたことが伺える.

リスト 4　被験者 C の Twitter からの自立語

スクエニ アカウント キャラ スクエニ タイトル 商標登録 ドメイン ブレイブリーデフォルト 白 回避 火力 カリスフェザー タウマス クリティカル サンホラ リンホラ　フレンド アビセア 経験値稼ぎ 卒業 研究 ゲーム　ドロス　メタル　お金　お金稼ぎ　コサージュ　ファミマ　初音　一番クジ　クッション　カレンダー　ゼミ　長時間　ヘイスト　ff11　ミーブル

しかし，被検者 C の興味がなかった時事情報における順位 6 位「天元第２回スポーツアコードワールドマインドゲームズが…」，21 位「ザッケルゥ～ニィ原解散→ピンでＲ１参戦」，22 位「中山雅史さん木梨 JAPAN でなでしこサッカー対決　新春特番に参戦」，25 位「訃報山本隆造さん５６歳＝日本野球機構審判技術委員」などにおける時事情報を見ると，本来の TV ゲームとは関係が薄いと思われるスポーツなどの試合(ゲーム)に関するものが上位にあがってしまい評価が下がる結果となった.

6. おわりに

本論文では，Web から時事情報を獲得し，時事情報に対するユーザの興味を満たす情報を選出する手法を提案した．具体的には，時事情報中の単語とテレビの視聴履歴および Twitter の発言履歴から得られた嗜好情報との関連性を求め，時事情報中の重要な語句に着目した上で，話題となる語に重みを付与する．これらの処理によって時事情報に重要度を付与し，より有益な時事情報を出力する手法を提案した.

その結果，提案手法を用いることで，無作為に時事情報を選ぶより平均で 2.3 倍の個人の趣味・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.

また，今回の評価実験において順位 1 位だがどちらの被検者にも興味がなかった時事情報「お面強盗、店員逃げレジ開かず…たばこ１箱奪う」のように嗜好情報に含まれていないような時事情報が上位に来てしまっている結果となった．これは「店員」といった話題語が比較的どの嗜好情報とも関連度が高いため重要度が高くなってしまったからだと推測される．ゆえに，今後はこういった語句を除去するような手法が求められると推測される.

謝辞
本研究の進行，論文の執筆にあたり，ご指導頂きました本学の渡部広一教授, 並びに土屋誠司准教授に心から感謝致します. また，様々なアドバイスを頂いた，芋野美紗子氏をはじめ知識情報処理研究室の皆様に，厚く御礼申し上げます.

参考文献

1) 山本達也，芋野美沙子，土屋誠司，渡部広一，　“テレビの視聴履歴に基づく時事情報提供システム”，第 10 回情報科学技術フォーラム 2011
2) “Twitter”，https://twitter.com/
3) “テレビ王国”，http://tv.so-net.ne.jp/
4) 奥村紀之，北川晋也，渡部広一，河岡司，“概念ベースの分析と精錬”, 同志社大学理工学研究報告, Vol.46, No.3, pp.133-141, 2005.
5) 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.
6) 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.
7) 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”, 人工知能学会全国大会, 2D1-01, 2003.
8) “Google”，http://www.google.co.jp/
9) “asahi.com : 朝日新聞社の速報ニュースサイト”, http://www.asahi.com/
10) “毎日ｊｐ：毎日新聞のニュース・情報サイト”, http://www.mainichi.jp/
11) “ニュース YOMIURI ONLINE : 読売新聞ニュースサイト”, http://www.yomiuri.co.jp/
12) “Oricon Style:株式会社オリコンのニュースサイト”, http://www.oricon.co.jp/news/
13) 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.
14) “Tween”, http://sourceforge.jp/projects/tween/
15) “TweetPopup for Tween”, http://www.asahi-net.or.jp/~tz2s-nsmr/soft/tweetpop/tweetpop.htm