E-002

テレビの視聴履歴を基にした時事情報提供システムの構築

Development of Current Information Offering System based on the History of Viewing TV

山本　達也†　芋野　美沙子†　土屋　誠司‡　渡部　広一‡
Tatsuya Yamamoto　Misako Imono　Seiji Tsuchiya　Hirokazu Watabe

1　はじめに

時事情報を取得することは日常生活を過ごす上で欠かせないものである．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，ユーザに最も有益であると考えられる時事情報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，テレビ視聴履歴に基づいた個人の嗜好情報に対するユーザの興味を満たす時事情報とする．

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．

2　時事情報提供システムの概要

本論文では，テレビの視聴履歴を基に個人の嗜好を抽出し，時事情報を提供するシステムの構築について述べる．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．システムの概要を図 1 に示す．テレビ視聴履歴を基にした時事情報獲得処理では，視聴履歴からの嗜好情報と時事情報との関連性を定量化する.時事情報の中から話題となる単語(以下，話題語とする）を抽出し，話題語ごとに嗜好情報との関連性を求める．関連性を計算した上で，話題語が表記されているニュースに重要度を付与していく．

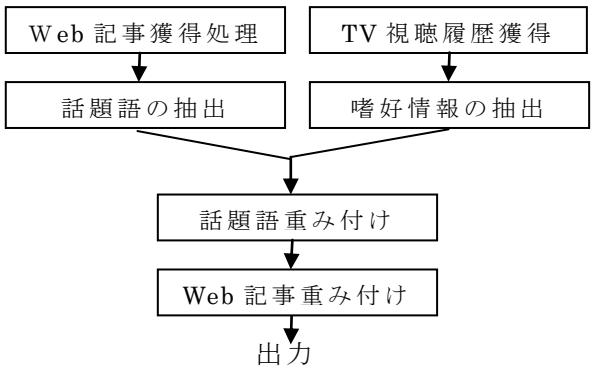

図 1　時事情報獲得システムの概要

†同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University
‡同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

3　使用技術

3.1　概念ベースと関連度計算方式

概念ベース[1]とは，語（概念）の特徴を表す語（属性）を大量に集めたものであり，属性には重みが定義されている．本研究では，複数の国語辞書や新聞などから抽出した概念や属性を加えた約 12 万の概念からなる概念ベースを使用する．なお，本稿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．

関連度計算方式[2]とは，概念と概念の関連の強さを定量的に評価するものである．各概念を 2 次属性まで展開し，重みを考慮した属性集合の一致度合いを計算する．

3.2　未定義語の属性獲得手法

概念ベースにおける未定義語の属性獲得手法[3]とは，未定義語 X の意味的特徴を表す属性（単語）とその重要性を表す重みの組を，Web を用いて獲得する手法である．

本稿では，この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．

3.3　TF・IDF

TF・IDF 法[4]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．このとき，IDF は式 1 で定義される．

$$idf(t)=\log_e \frac{N}{df(t)}+1 \qquad (1)$$

3.4　Web-IDF

3.3 節で説明した IDF 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．一方，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 1 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 2,010,000,000 件（2011 年 1 月 30 日現在））としている．これは，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．

4　嗜好情報を考慮した時事情報獲得処理

時事情報と個人の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選別する．

4.1　話題語の抽出

最新の時事情報を，新聞社の Web サイトから獲得する．Web から獲得してきた 1 日分のニュース記事に対し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．次に，Web-IDF を調べ，閾値である 3.0 未満の Web-IDF 値を持つ話題語を削除する．閾値は過去の研究[5]での実験によって最適化された値である．

4.2　個人嗜好情報の抽出

個人の嗜好情報を取得するために，本研究ではテレビの視聴履歴を以下のように時間系列毎に収集する．

- 集中して視聴していた番組（録画番組も含む）
- 何となく視聴していた番組
- 視聴したかったが見れなかった番組

また、本研究では，Web による電子テレビ番組表を使い，番組タイトル・番組紹介文から得られる自立語のオートフィードバックを利用する．

4.3　話題語への重み付け

話題語とテレビの視聴履歴から得られる嗜好情報との関連性を求め，話題語毎に重みとして値を付与する．話題語への重み付けの具体例を図 2 に示す．

視聴履歴から得た嗜好情報	話題語	関連度	視聴区分
大阪	大リーグ	0.025	◎
野球		0.258	◎
サッカー		0.184	◎

大リーグに付与される重み
（関連度の平均）×（視聴区分に応じた重み）
(0.025+0.258+0.184)/3*2.0=0.31

図 2　話題語への重み付け

まず，話題語，テレビの視聴履歴データの自立語に対して，AF を用いて属性と重みを獲得する．次に，話題語と視聴履歴から得た嗜好情報との関連度を行う．これらの計算結果を嗜好情報の総数で割った値を視聴区分に応じた重みを掛け合わせることでその話題語の重みとする．

なお，視聴区分に応じた重みは，実験試行より集中して視聴した：◎は 2.0，視聴したいができなかった：○は 1.0，なんとなく視聴した：△は 0.7 とした．

4.4　時事情報への重要度付与

時事情報中に表記一致によって話題語が存在するか調べ,話題語の重みの平均値を重要度とする．重要度の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

5.1　評価方法

テレビ視聴履歴に基づく時事情報提供システムの出力について，評価を行う．

実験には 2011 年 1 月 23 日から 30 日の 7 日間に収集したテレビの視聴履歴と 30 日におけるニュースの見出し情報を用いた．5 名の被験者より視聴履歴を収集し，嗜好にあった時事情報であるかどうか判断して頂いた．

被験者はあらかじめ，実験を行う日に取得できたすべての時事情報合計 200 個を見て，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みなす．時事情報に対して重要度を求め，重要度の高い順に並べ変える．そして，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．評価指標には，平均精度を使用する．出力結果に対する平均精度 *AP* は式 2 のように定義される．

$$AP = \frac{1}{S}\sum_{i=1}^{n}\frac{z_i}{i}\left(1+\sum_{k=1}^{i-1} z_k\right) \qquad (2)$$

5.2　結果及び考察

個人情報を考慮した時事情報獲得における被験者毎の *AP* 評価を表 1 に示す．

表 1 被験者別の AP 値

	A	B	C	D	E	平均
AP値	0.637	0.503	0.541	0.583	0.601	0.573

AP 値が最大である被験者 A は，視聴履歴において「サッカー」に関するものを多く視聴しているため，出力結果にサッカーに関するものが多く推薦され評価が高かった．一方で，AP 値が最低である被験者 B は，視聴履歴において「バラエティ」に関するものを多く視聴していたため，出力結果に芸能人に関するものが多く推薦された．しかし、被験者 B にとって興味のない芸能人に関することまで推薦されてしまい評価が低くなった．

6　おわりに

本論文では，Web から時事情報を獲得し，ユーザの興味を満たす時事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，有益な時事情報を出力する手法を提案した．

謝辞

本研究の一部は，科学研究費補助金(若手研究(B)21700241)の補助を受けて行った．

参考文献

[1] 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.

[2] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.

[3] 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.

[4] 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.

[5] 藤田晴樹，渡部広一，河岡司，“コンピュータ日常会話のための Web からの時事情報獲得技術”，情報処理学会研究報告，2007-ICS-147(22)，pp.145-150，2007.