D-007

個人情報を重視した時事情報提供手法
Current Information Offering Method Focused on Personal Information

津田 健太郎† Kentaro Tsuda　後藤 和人† Kazuto Goto　土屋 誠司‡ Seiji Tsuchiya　渡部 広一‡ Hirokazu Watabe

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になった．しかし時事情報は短時間で大量に更新され，インターネット上に無数に存在するため，ユーザが自身の求めている有益な時事情報を即座に入手することは困難である．これに対し，ユーザ自身に関する個人情報，ユーザの好き嫌いを示す嗜好情報，そして一般的重要性を利用して，コンピュータがユーザにとって有益と考えられる時事情報を提供するシステム[1]が存在する．一般的重要性は，時事情報自体の重要性の高さと，ユーザと同じ性別・年代の嗜好から考慮される．しかしこのシステムには個人情報を十分に生かせていない問題点がある．本研究ではこの問題点を改善し，ユーザへより有益な時事情報を提供するシステムの実現を目指した．

2. 関連技術

2.1 概念ベース

概念ベース[2]とは，複数の国語辞書や新聞などから機械的に構築した語(概念)とその意味・特徴を表す語(属性)，属性の重みの集合からなる知識ベースである．概念ベースには，約 9 万語の概念が収録されている．なお，本稿では概念ベースに登録されていない語を未定義語と呼ぶ．

2.2 関連度計算方式

関連度計算方式[3]とは，概念 A と概念 B の関係の深さを定量的に表す方法である．それぞれの概念が持っている属性と重みによって関連度計算は行われ，意味の近さは関連度という数値で表すことができる．関連度は 0～1 の連続的な実数で表され，関連度の値が高いものが意味の近い語となる．

2.3 オートフィードバック(AF)

オートフィードバック[4]は概念ベースに定義されていない未定義語の属性とその重要度をあらわす重みの組を，Web を用いて獲得する手法である．

2.4 TF・IDF

TF・IDF[5]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 はある文書中 d に出現する語 t(文書の内容を構成する要素)の頻度を表す尺度であり，式(1)で定義される．ただし，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．IDF はある語が全文書中のどれくらいの文書に出現するか(特定性)を表す尺度であり，式(2)で定義される．なお，N は検索対象となる文書集合中の全文書数，df(t)は語 t が出現する文書数である．

$$tf(t,d) = \frac{n}{W} \quad (1)$$

$$IDF(t) = \log \frac{N}{df(t)} + 1 \quad (2)$$

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

3. システムの概要

本システムでは，Web 上から取得した時事情報とユーザが予め登録したユーザ情報を入力とし，ユーザ情報と一般的重要性の 2 つの観点において付与された点数の順に時事情報を並べ替えて出力する．システムの流れを図 1 に示す．

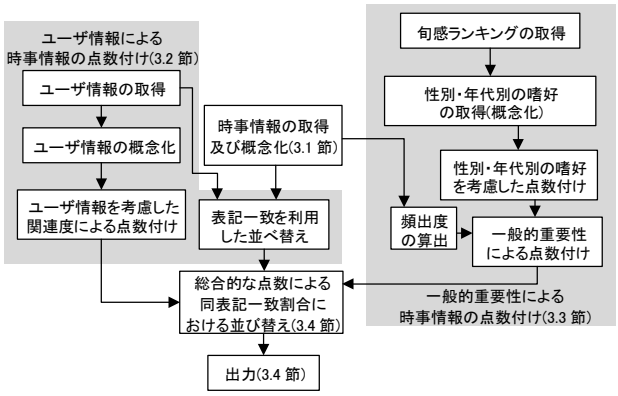

図 1　システムの流れ

3.1 Web からの時事情報の取得と概念化

「朝日新聞デジタル[6]」「YOMIURI ONLINE[7]」「毎日新聞[8]」の 3 社のニュースを利用して時事情報の収集を行う．

まず，各ニュースサイトから記事の見出しと本文を取得し，その後見出しを概念，本文に存在する自立語を属性として概念化を行う．属性の重みには TF・IDF 値を用いる．図 2 に概念化の例を示す．

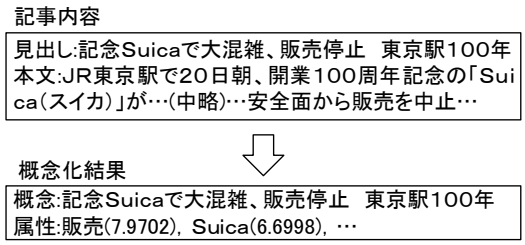

図 2　時事情報の概念化の例

3.2 ユーザ情報による時事情報の点数付け

予めユーザにユーザ情報を登録してもらい，どの時事情報に対してユーザが興味を持つのかの判断を行う．ユーザ情報の内容は大きく分けて個人情報と嗜好情報に分かれており，各項目に語を入力する．嗜好情報には好きなもの・嫌いなものをそれぞれ入力する．項目の一覧を表 1 に示す．

Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 1　ユーザ情報の項目の一覧

ユーザ情報				
個人情報		嗜好情報(好きなもの，嫌いなもの)		
名前	勤務先	食べ物	色	スポーツ
学校名	出身地	飲み物	昆虫	動物
現住所	国籍	季節	花	国
取得資格	免許	アーティスト		キャラクター
職業	趣味	教科		作家
特技	ペット	映画		本
今気になる話題		その他		

3.2.1 表記一致による時事情報の並び替え

表記一致により，個人情報に登録された語のうち，いくつ記事内に出現するかの割合を求めた後，その割合の順に時事情報を並べ替える．この時点では，同順位，すなわち同じ割合となっている時事情報が複数存在することになるため，3.2.2 項の関連度を用いたユーザ情報の点数付け及び 3.3 節の一般的重要性による点数付けによって同順位内の並び替えを行う．

3.2.2 関連度によるユーザ情報を考慮した点数付け

個人情報及び嗜好情報について，登録内容の語を属性としてそれぞれ概念化する．嗜好情報については好きなもの・嫌いなものについてそれぞれ概念化を行う．これらと概念化した時事情報との関連度により時事情報への点数付けを行う．個人情報と好きなものの関連度の合計から嫌いなものの関連度の合計を引いた値を個人情報の点数とする．

3.3 一般的重要性による時事情報の点数付け

時事情報の頻出度合，及び性別・年代別の嗜好を考慮することで，一般的重要性による時事情報の点数付けを行う．次に示す頻出度による点数，性別・年代別の嗜好による点数を掛け合わせた値を一般的重要性による点数とする．

3.3.1 頻出度による点数付け

頻繁に報道される時事情報ほど，一般的重要性が高いと考えられる．その日の記事の全見出しに出現する名詞を取得し，各名詞の出現頻度によって点数を付ける．

3.3.2 性別・年代別の嗜好による点数付け

旬感ランキング[9]を利用して性別・年代別の嗜好情報を取得し，時事情報との関連性を調べることで時事情報の点数付けを行う．旬感ランキングとは，BIGLOBE が提供する検索エンジンによって検索された語を集計し，男女別で 10 代~50 代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．本研究では，過去一週間分の急上昇ワードをその性別・年代の嗜好情報を示すキーワードとして取得する．そして，キーワードから嗜好情報を取得する．例を図 3 に示す．

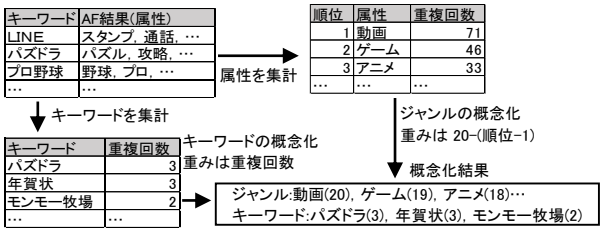

図 3　性別・年代別の嗜好情報取得の例

まず取得したキーワードを集計し，出現回数の多いキーワードを属性に持つキーワード概念を作成する．また，取得したキーワード全てに AF を行い，得られた属性を集計する．出現回数の多かった属性上位 20 語を新たな概念の属性に用い，その性別・年代が興味を持っているジャンルの概念を作成する．時事情報とキーワード，時事情報とジャンルの関連度をそれぞれ求め両方の合計を点数とする．

3.4 結果出力

3.2.1 項において時事情報を表記一致の割合の順に並び替えた後，同じ表記一致の割合の記事群の中で関連度による個人情報の点数と一般的重要性による点数の合計の順に並べ替える．その後，順位の高い順に時事情報(記事の見出し)を出力する．

4. 精度評価

本システムの出力について評価を行った．評価実験には 3 日分の時事情報(1 日当たり約 150 件)と，各日 1 週間分の旬感ランキングを使用した．被験者は予め全ての記事の見出しを見て，ど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．被験者が興味ありと記入した時事情報がシステムの出力の上位(興味ありと記入した時事情報の数と同じ順位まで)にどれだけ存在するかの割合で評価を行った．本研究・既存システムについて，全ての評価日・被験者における結果の平均を表 2 に示す．

表 2　全体の平均精度の比較

	本研究	既存システム
平均精度	26.9%	22.3%

5. 考察

3.2 節に示した個人情報を重視した時事情報の並べ替えによって，平均精度が大きく上がった被験者とほとんど上がらなかった被験者が存在した．これは，ユーザや時事情報によって，ユーザが時事情報の入手においてどの要素を重視するかが変わってくるためと考えられる．したがって，ユーザによって点数付けの配分を変えることや時事情報や急上昇ワードの情報源の改善が必要になると考えられる．

6. まとめ

本研究では個人情報に重点を置いた時事情報の並べ替え処理の改善を行うことによって，ユーザに合った時事情報を提供する方法を提案した．結果として，既存手法と比較して約 4.6%精度が向上した．

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] 南光, 芋野美紗子, 土屋誠司, 渡部広一，“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”，第 175 回知能システム研究発表会，2014.
[2] 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.
[3] 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.
[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”，人工知能学会全国大会，2D1-01，2003.
[5] 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.
[6] 朝日新聞デジタル，http://www.asahi.com/，2015/6/1 参照
[7] YOMIURI ONLINE，http://www.yomiuri.co.jp/，2015/6/1 参照
[8] 毎日新聞，http://mainichi.jp/，2015/6/1 参照
[9] BIGLOBE サーチ旬感ランキング，http://search.biglobe.ne.jp/ranking/，2015/6/1 参照

Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.