E-010

話題語評価に基づく Web ニュースからの時事情報獲得技術
The Acquisition Method of Current Information from Web News based on Evaluation of Topic Word

河合　智弘†　　渡部　広一†　　河岡　司†
Tomohiro Kawai　Hirokazu watabe　Tsukasa kawaoka

1. はじめに

人間とロボットがコミュニケーションをとるための手段として，ロボットから有益な時事情報を提供することが考えられる．そのため，本研究では話題となっている時事情報を蓄えた知識ベース（以下，時事情報知識ベースとする）の構築と更新を行う手法を提案する．

本研究における時事情報は，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やタイトルなどの短文とする．提案手法では，時事情報の話題性を定量化するために，時事情報の中から話題となっている単語（以下，話題語とする）を抽出し，話題語同士の関連性，話題語の特定性を考慮した上で，出現頻度が高いものほど重要となるよう重みを付けた．重み付きの話題語を用いて時事情報の話題としての重要度を決定し，時事情報知識ベースを構築している．また，時間の経過と共に重要度を減衰させている．

2. 時事情報獲得技術の概要

時事情報知識ベースの構築および精練を行うシステムを提案する．本システムは，獲得処理と精練処理の 2 つの処理に分けられる．本システムの概要を図 1 に示す．

獲得処理では，Web 上の時事情報の話題性を定量化し，現在話題であると考えられる時事情報を獲得する．

精練処理では，現在では話題ではないと判断できる情報を削除する．

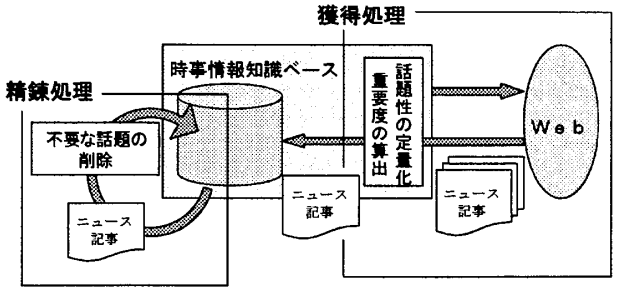

図 1　システムの概要

3. 関連技術

3.1 概念ベース

概念ベース[1]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．概念ベースには，約 12 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概念を未定義語と呼ぶ．

概念は，ある語Aを属性a_iと重み$w_i(>0)$の対の集合として式 3.1 によって定義する．

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

ここで，属性a_iを概念A の一次属性と呼ぶ．また，属性a_iも概念ベースに登録されている 1 つの概念である．従って，a_iからも同様に属性を導くことができる．a_iの属性a_{ij}を概念Aの二次属性と呼ぶ．

3.2 関連度計算方式

関連度計算方式[1]とは，概念ベースに定義された語と語の関連の強さを，同義性，類似性のみに関わらず定量化する手法である．

関連度は，0 以上 1 以下の連続的な実数で表され，概念同士の関連が大きいほど関連度は高くなる．関連度は，それぞれの概念を二次属性まで展開し，その重みを利用した計算によって最適な一次属性の組み合わせを求め，それらが一致する属性の重みを評価することで算出する．

3.3 TF・IDF

TF・IDF法[2]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 はある文書に出現する索引語の頻度を表す尺度である．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か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 3.2 で定義される

$$idf(t)=\log\frac{N}{df(t)}+1 \tag{3.2}$$

3.4 未定義語の属性獲得手法

未定義語の属性獲得手法[3]は，未定義語の属性とその重要性を表す重みの組を概念ベースに基づきWeb を用いて自動的に構成する手法である．

4. 時事情報の獲得処理

獲得処理の流れは，次の通りである．

1. Web からの時事情報獲得
2. 話題語と重みの抽出
3. 話題語を用いた時事情報への重要度付与
4. 時事情報知識ベースへの格納

4.1 Web からの時事情報獲得

新聞社のWebサイト[4, 5, 6]から，1 日分のニュース記事を獲得する．

4.2 話題語と重みの抽出

時事情報の中から話題語を抽出し，重みを決定する．以下に，話題語の抽出と重み付けの方法について述べる．

4.2.1 話題語の抽出

時事情報の中から自立語を話題語として抽出し，1 日分のニュース全体における話題語の出現頻度を初期重み FW とする．

4.2.2 IDF 値を用いた重み付け

話題語に付与した初期重み FW に，その話題語の IDF 値を掛け合わせた重み IW を求める．IDF 値を求めるために利用する文書は，過去 1 年分すべての時事情報とする．

† 同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

4.2.3 グループ化

表記の違いや同義語に対し，出現頻度による重みの差をなくすために，話題語同士の関連を考慮する．

抽出した全ての話題語に対し，関連度計算方式を用いてグループ化を行う．そして，グループに基づいて話題語の重み *TW* を求める．*TW* は以下の計算式で求められる．

$$TW = IW + averageGroupIW$$

（*averageGroupIW*：
グループ内の IDF 値を掛けた重みの平均）

4.3 時事情報への重要度付与

重み付けを行った話題語を用いて，時事情報へ重要度を設定する．時事情報中に存在する話題語の重みを用いて重要度を決定する．存在する話題語の重みの和を重要度とする手法を加算手法，積を重要度とする手法を乗算手法とする．

4.4 時事情報知識ベースへの格納

重要度を付与した時事情報を時事情報知識ベースに格納する．

5. 時事情報の精練処理

過去数日分の話題語を獲得し，時事情報知識ベース内の時事情報が現在でも話題であるかを判断する．

5.1 精錬処理手法

精錬処理を行う時事情報が，精錬処理を行う当日から n 日前の時事情報であるとすると，2n 日分の時事情報の話題語に対して，獲得処理と同様の重み付けを行う．

5.2 時事情報への重要度付与

時事情報への重要度付けは 2n 日分の時事情報の話題語を用いて，獲得処理と同様，加算手法，乗算手法によって行う．

また，精練処理のみ，これらとは別の重み付けの手法を提案する．これは，話題語の一日の出現頻度に着目したものである．精練処理では，数日分の話題語の重みの合計を最終的な重みとしている．このとき，話題語が出現した記事の日数で割ることで，話題語の重みを求める．この後，加算手法，または乗算手法を用いて，時事情報に重要度付けをする．これを平均重み利用加算手法，平均重み利用乗算手法とする．

5.3 時事情報知識ベースへの格納

重要度が付け直された時事情報を再度時事情報知識ベースへ格納する．

6. 評価・考察

提案システムを用いて，時事情報知識ベースに格納した時事情報の話題性の評価を行う．評価は，重み付けを行った上位 50 件の時事情報について行う．獲得処理を行い獲得したある新聞社の時事情報 A に対して，同様であると考えられる時事情報が他の新聞社においても掲載されているかを評価基準として用いる．

6.1 評価基準

10 月 5 日に獲得した時事情報について，10 月 3～7 日の時事情報から抽出した話題語を用いて，精練処理を行った．なお，本稿では IDF 値およびグループ化による重み付け手法に加えて，加算手法を用いて重要度付与を行った手法を基準手法とする．

6.2 比較実験

基準手法のうち，IDF 値を用いた重み付けを行わなかった場合，グループ化処理を行わなかった場合について，それぞれの評価を表 1 に示す．また，4.3 節，5.2 節で挙げた 4 種類の重要度付与の手法についての評価を表 2 に示す．

表 1　話題語重み付けによる手法の比較

	2 社以上に存在	3 社に存在
基準手法	56%	30%
IDF 値無し	56%	26%
グループ化無し	54%	30%

表 2　重要度付与における手法の比較

	2 社以上に存在	3 社に存在
基準手法	56%	30%
積手法	62%	32%
平均重み利用加算手法	54%	30%
平均重み利用乗算手法	58%	32%

6.3 考察

表 1 の結果より，IDF 値を用いた重み付けとグループ処理を行った基準手法が，処理を省いたものに比べ，最もよい結果となった．

また，表 2 の結果では，話題語の重みの積によって重要度付けを行う，乗算手法の結果が今回の実験の中で最もよい結果となった．乗算手法において，時事情報の重要度は，その時事情報が保有している話題語の数に大きく依存する．ここから，話題語を多く保持している時事情報は，より多くの情報を保持しているため，重要となることが考えられる．

7. おわりに

本稿では，Web から時事情報を獲得し，時事情報知識ベースを構築・更新する手法として，話題語の重み付けによる獲得処理と精練処理を提案した．提案手法を用いることで，膨大な時事情報の中から，話題となる情報を 62%の精度で時事情報知識ベースに格納することができることを実験により示した．

参考文献

[1] 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）
[2] 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）
[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”，第 18 回人工知能学会全国大会論文集，2D1-01（2003）
[4] “ asahi.com（朝日新聞）” ：http://www.asahi.com/
[5] “ 毎日 jp（毎日新聞）” ：http://www.mainichi.jp/
[6] “ YOMIURI ONLINE（読売新聞）” ：http://www.yomiuri.co.jp/