

URL の類似性に着目した WWW 空間からの関連語自動収集手法

A Collection Method of Related Words Automatically from WWW by the Similarity of URL

獅々堀 正幹¹ 山本 一晴²
小泉 大地³ 北 研二⁴

Masami SHISHIBORI Issei YAMAMOTO
Daichi KOIZUMI Kenji KITA

本論文では、数個のシーズとなる単語（基底単語）を準備し、その基底単語群の関連語を WWW 空間から効率的に自動収集する手法を提案する。WWW 空間から関連語を収集するにあたり、単語が存在するページの URL に着目する。URL には各ページ間の階層関係を含んでいるため、基底単語群が存在するページの URL 集合と関連語が存在する URL 集合間に類似性があると予測される。そこで本手法では、基底単語群を既存の WWW 検索システムに入力することで得られる検索結果の URL に対して、URL のパス毎に出現頻度を重み付けすることにより、基底単語群の URL 集合と類似した URL 集合を有する単語を関連語として収集する。

In this paper, we propose the method to collect related words from WWW space by using related basis words, which have the same meaning. We paid attention the URL of the page where each word appears, because there is the relationship between the set of URL of basis words and related words if basis words have the same meaning. This method gives the frequency weight to each pass of URL and collects related words from Web sites where have high relation to basis words.

1. はじめに

関連語の自動収集に関する研究は、自然言語処理システムにおける言語知識辞書の構築[1], [2], また、情報検索システムにおける検索質問拡張など、様々な分野で有効活用されている。特に近年、インターネット技術の発達に伴い、WWW 空間から関連語を自動収集する研究が活発に取り組みられている[3]-[5]。本研究では、数個のシーズとなる単語（基底単語）を準備し、基底単語群に意味的に関連した単語をWWW

空間から自動収集することを目的とする。ここで、関連語とは基底単語と共起関係にある単語（関連候補語）の中でも特に意味的なつながりがある単語を表す。

従来のWWW空間からの関連語収集手法は、Webページ内の出現単語の頻度情報を利用するものがほとんどであった。特に、基底単語とその関連候補語との意味的な関連性を双方の単語と共に出現する共起単語の頻度情報に基づいて相互情報量やJaccard係数により計算する手法[3], [4]が主であった。しかし、これらの手法では関連候補語と共起する単語群を得るために、WWW検索エンジンを用いて関連候補語が存在するページを検索し、それら大量のページにアクセスしなければならず、関連語収集に莫大な時間コストを必要としていた。例えば、100個の関連候補語を対象にして各単語につき検索結果上位100件のページを取得すると、10,000件ものページをダウンロードすることになる。

この問題に対して、我々はWWW上の各ページにURLが付随していることに着目し、URLを手がかりにして、WWW空間から関連語を自動収集する手法を提案する。URLには各ページ間の階層関係を含んでいるため、基底単語群が存在するページのURL集合と関連語が存在するURL集合間に類似性があると予測される。また、URLに関する情報は、検索結果のサマリページから取得可能であるため、個々のページをダウンロードする必要がなく、高速な関連語収集が可能になる。

本手法では、既存のWWW検索システムに対する各関連候補語の検索結果から得られるURL集合と、基底単語群から得られたURL集合との間でパス毎の一致性が高ければ、その関連候補語を関連語として採用する。例えば、基底単語“本塁打”と“ホームラン”から得られたURL集合には、同一のサイトやホスト名に類似性をもつサイトが多数出現する。そして、関連候補語“松井秀喜”の検索結果から得られるURL集合が基底単語群のURL集合と高い類似性をもっていれば、この関連候補語は関連語であると判断する。

2. 従来の関連語収集手法

従来の代表的な関連語収集技術を大別すると、単語の共起情報を基に求めた相互情報量により関連語を収集する方法[1], [6], 及び検索結果内に出現する単語の類似性により関連語を収集する方法[3]に分類できる。

まず、相互情報量を用いる手法では、出現頻度の極端に低い固有名詞との共起頻度が悪影響を及ぼすため、WWW空間における関連語収集手法としては不適切である。次に、出現単語の類似性による関連語収集手法について説明する。これは、2つの単語 x と y をそれぞれWWW検索システムを用いて検索し、検索結果から得られる共通単語に対して、頻度ベクトル間の類似度を式(2)で得られるJaccard係数で評価する。

$$\alpha(CF(x), CF(y)) = \frac{\sum_{i=1}^n CS_i \cdot CW_i}{\sum_{i=1}^n CS_i^2 + \sum_{i=1}^n CW_i^2 - \sum_{i=1}^n CS_i \cdot CW_i} \quad (2)$$

ここで、単語 x と y の共通単語に対する頻度ベクトルを

$$CF(x) = (CS_1, CS_2, \dots, CS_i, \dots, CS_n)$$

$$CF(y) = (CW_1, CW_2, \dots, CW_i, \dots, CW_n)$$

とする。

Webの内容解析による関連語の収集手法は、関連候補語数の影響でWWW空間へのアクセス数が多くなり、収集に膨大な時間を費してしまう。また、基底単語の検索結果における出現単語と関連候補語の検索結果における出力単語の共起単

¹ 正会員 徳島大学大学院ソシオテクノサイエンス研究部

bori@is.tokushima-u.ac.jp

² 非会員 徳島大学大学院工学研究科博士前期課程

issei@is.tokushima-u.ac.jp

³ 非会員 徳島大学大学院工学研究科博士後期課程

koizumi@is.tokushima-u.ac.jp

⁴ 非会員 徳島大学高度情報化基盤センター

kita@is.tokushima-u.ac.jp

語に着目して関連度を計算するため，“今日”や“私”など一般的に用いられる語句も関連度に影響を及ぼしてしまうという問題点がある。

3. URLの類似性に基づく関連語収集手法

3.1 本手法の概要

図1に提案する関連語自動収集手法の概要を示し、収集手順を説明する。なお、手順2で示すURLデータベースの構築方法[6]、及び手順5で示す関連度の計算方法については3.2, 3.3で詳しく述べる。

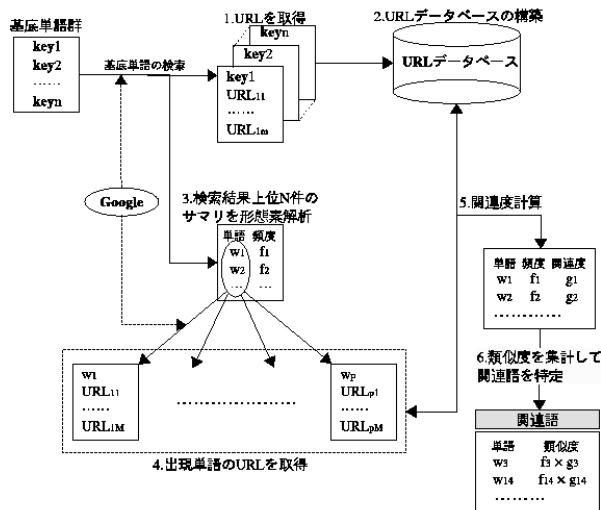


図1 本収集手法の概要図

Fig.1 Outline of the Proposed Method.

[関連語収集アルゴリズム]

手順1：基底単語が存在するページを検索

あらかじめ人手で登録した基底単語群 key_i ($1 \leq i \leq n$) を WWW 検索システムに入力し、各基底単語毎に上位 m 件の検索結果 URL_{ij} ($1 \leq j \leq m$) を得る。

手順2：URLデータベースの構築

手順1で得た検索結果 URL_{ij} からパス毎の頻度を計算し、各頻度に対して WWW 空間全体の大域的頻度を用いて正規化を行い、URLデータベースを構築する。

手順3：ページ内容の解析（単語出現頻度の計算）

手順1で得た検索結果のサマリページを形態素解析し、名詞系単語 w_k ($1 \leq k \leq p$) の出現頻度 $Freq(w_k)$ を集計する。ここで、出現頻度上位の単語を関連候補語とする。

手順4：出現単語のURLを取得

各関連候補語 w_k を WWW 検索システムに入力し、単語毎に上位 M 件の検索結果 URL_{kl} ($1 \leq l \leq M$) を得る。

手順5：関連度の計算

URLデータベースと URL_{kl} とのマッチングを行い、各関連候補語 w_k の関連度を求める。

手順6：関連語の特定

手順5で求めた関連度と出現頻度 $Freq(w_k)$ の積から類似度を求め、類似度上位の関連候補語を関連語と特定する。

以上に示す通り本手法では、上記アルゴリズムの手順2において、基底単語群が出現するURL集合を特定し、手順4において関連候補語が出現するURL集合を求め、手順5において双方のURL集合の類似性を計算している。

3.2 URLデータベースの構築方法

3.1の手順1で得られた URL_{ij} に対し、WWW空間中のURL出現頻度で正規化する。これより、基底単語群と関連性の低いWebサイトの検出を抑え、関連性の高いと思われるWebサイトを特定することができる。以下にURLデータベースの構築手順を示す。

[URLデータベース構築アルゴリズム]

手順1：部分URL毎の出現頻度の計算

URL_{ij} 内に出現する部分URLの出現頻度を求める。部分URLは、“/”を区切りとして分割したものである。例として、“http://www.tokushima-u.ac.jp/G-life/main.htm”のURLに対して部分URLを求めると“www.tokushima-u.ac.jp”、“www.tokushima-u.ac.jp/G-life”の2つの部分URLが作成される。これらの部分URLの各パスの共通部分の頻度を出現頻度とする。

手順2：部分URLの大域的頻度の取得

各部分URLをWWW検索システムのURL検索機能に入力し、検索結果内の「検索件数」を部分URLがWWW空間中に存在する大域的出現頻度とする。

手順3：部分URLの出現頻度の正規化

手順1の出現頻度を以下のように大域的出現頻度で正規化し、その値を関連度とする。

関連度 = 部分URLの出現頻度 / 部分URLの大域的出現頻度

図2に上記の手順に従い、部分URLの出現頻度の正規化を行う例を示す。この例では、3つのURLから作成される部分URLが登録されている。部分URLは

(a) www.tokushima-u.ac.jp

(b) www.tokushima-u.ac.jp/G-life

の2つであり、(a)のデータベース内での出現頻度は3、(b)は2である。次に、各部分URLをWWW検索システムの入力して検索を行うと(a)は8570件、(b)は78件の検索結果を得る。最後に正規化を行うと、(a)は0.00035、(b)は0.0256となる。この関連度は、基底単語群が出現しやすいWebサイトとの関連性を示している。

http://www.tokushima-u.ac.jp/sitemap.htm		
手順1:	3	URLデータベース内の出現頻度
手順2:	8570	WWW空間中での出現頻度
手順3:	0.00035	部分URLと基底単語群との関連度
http://www.tokushima-u.ac.jp/G-life/main.htm		
	3	2
	8570	78
	0.00035	0.0256
http://www.tokushima-u.ac.jp/G-life/New_INFO.htm		
	3	2
	8570	78
	0.00035	0.0256

図2 部分URL出現頻度の正規化の例

Fig.2 Normalization of Partial URL Frequency.

3.3 部分URLマッチングによる関連度計算

3.1の手順5では、構築したURLデータベースと関連候補語の検索結果から取得したURL集合間で、部分URL毎にマッチングを行い、マッチングに成功した部分URLの関連度の総和を求める。例として、図3のURLデータベースに用いて、関連候補語のURL集合が以下の(1)~(3)であった場合の処理内容を示す。

(1) www.tokushima-u.ac.jp

(2) www.tokushima-u.ac.jp/G-life

(3) www.tokushima-u.ac.jp/a2/G-life

まず、(1)の URL は URL データベースと照合すると、“www.tokushima-u.ac.jp”まで一致しているため、関連度は 0.00035 となる。次に、(2)は “www.tokushima-u.ac.jp/G-life”まで一致しているため 0.0256 となる。(3)については、“www.tokushima-u.ac.jp”まで一致しているため、(1)と同じ 0.00035 となる。このように、部分 URL のマッチングを行い、この関連候補語の URL 集合における関連度は、(1)~(3)の URL の関連度を総和とする。なお、本手法では、URL データベース内において部分 URL とのマッチングを効率的に行うため、共通接尾辞を併合できるトライ構造[7]によって URL データベースを構築している。

4. 評価

4.1 実験条件

本手法の有効性を確かめるため、7 種類の分野(野球, 車, 有害, 競馬, アイドル, サッカー, 相撲)の基底単語群について単語数を変化させて (1 個, 3 個, 5 個) 収集した関連語を評価した。まず、各基底単語群を Google サーチエンジン[9]に入力して得たサマリを形態素解析[10]し、名詞系単語のうち出現頻度上位 500 件の単語を関連候補語とした。そして、提案手法と従来手法を関連候補語に適用した際の上位 100 件の関連語に対する平均適合率[11]を評価した。なお、従来手法としては Jaccard 係数を用いる手法[3]を採用し、正解データは人手で判定した。

4.2 従来手法との収集精度の比較

図 3~図 5 に提案手法と従来手法との比較実験結果を示す。各グラフ内で、“jaccard_summary”が検索結果のサマリに対して従来手法を適用した結果、“jaccard_html”がサマリからリンクする各 HTML ページに対して従来手法を適用した結果、“url_base”が提案手法を示す。

まず、検索結果のサマリに対して従来手法を適用した場合の精度が極端に悪くなっている。これは、基底単語と関連候補語のそれぞれのサマリに出現する共通単語が少なく、Jaccard 係数では的確に関連性を評価できないことが要因となっている。次に、HTML ページに対して従来手法を適用した場合と提案手法を比較すると、ほとんどの分野において提案手法の平均適合率が従来手法よりも上回っており、提案手法は収集精度に関して有効であるといえる。

しかし、提案手法における分野毎の結果を比較すると、精度に大きな違いが生じている。特に、アイドル分野とサッカー分野で平均適合率が低下している。これは、アイドル分野におけるアイドルの名字・名前が的確に切り出せていないことや、“画像”や“壁紙”など曖昧な意味を有する単語が頻繁に出現しノイズになっていることが原因と考えられる。また、サッカー分野では形態素解析の辞書に登録されているサッカー用語、国名や外国人選手の氏名が少なく、形態素解析の失敗が精度を低下させる原因となっていた。

次に、基底単語の個数を増加させると、平均的に収集精度が上昇した。これは、基底単語の個数を増やすことで、サマリに出現する適合単語の種類や頻度も増加したために収集精度が上昇したと考えられる。また、URL データベースを構築する際の URL データが増加し、より高精度のデータベースが構築できたと考えられる。

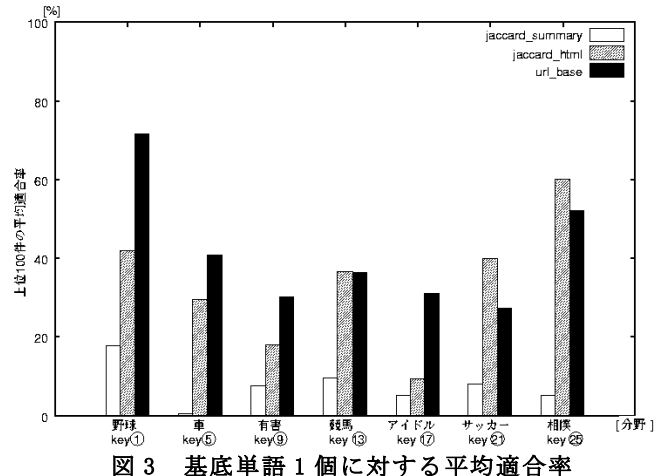


Fig. 3 Graph of the Average Precision on One Basis Word.

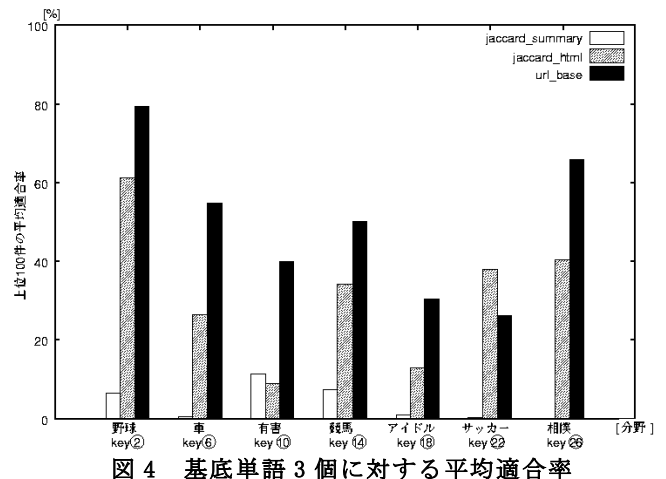


Fig. 4 Graph of the Average Precision on Three Basis Words.

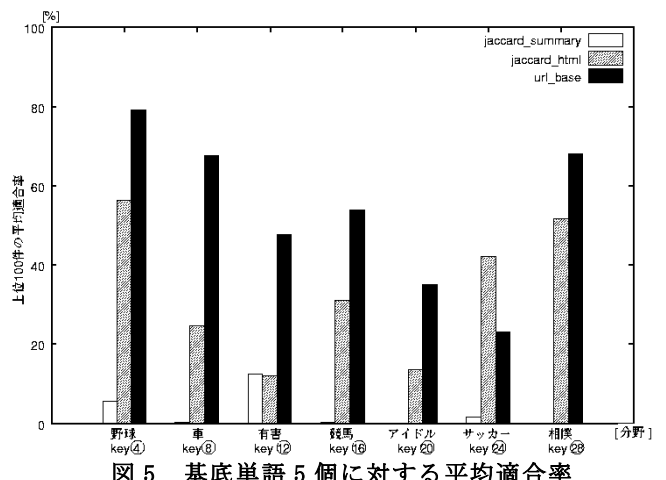


Fig. 5 Graph of the Average Precision on Five Basis Words.

4.3 従来手法との収集時間の比較

今回の実験では、関連語収集の際に両手法が WWW 空間にアクセスした回数により比較を行った。表 1 に、両手法の基底単語数の違いによる平均 WWW 空間アクセス数を示す。

表 1 に示すように、サマリに対する従来手法が最もアクセス数を少ない。しかし、4.2 で示したように、サマリに対する従来手法は収集精度が極端に悪くなってしまふ。そこで、HTML に対する従来手法とサマリに対する提案手法を比較すると、提案手法の方が少ないアクセス数で関連語を収集できている。すなわち、検索結果のサマリに対して提案手法を適用した場合が短時間、かつ、精度よく関連語収集を行うことができる。

これは、従来手法において関連候補語を検索する手順と提案手法の手順 4 における WWW アクセス数の差が要因となっている。HTML に対する従来手法では、関連候補語の検索結果に含まれるすべてのページにアクセスしなければならない。一方、提案手法では、関連候補語の検索結果として URL 一覧が掲載されたサマリページのみを取得するだけで関連語が特定できる。例えば、関連候補語の検索結果 100 件の URL を対象とした場合、従来手法では 100 回のアクセスが必要であるが、検索結果 1 ページ内に 100 件の URL が表示されると仮定すると、提案手法では 1 回のアクセスで手順を進めることができる。ただし、提案手法では、URL データベースを構築する際、各部分 URL の大域的頻度を得るためにパス毎の URL 検索を行う必要がある。そのため、提案手法では、サマリに対する従来手法よりも URL 検索のためのアクセス数が増加する。

表 1 関連語収集に必要な平均 WWW アクセス数
Table 1 Average Number of WWW Accesses Required for Collection of Related Words.

基底単語数	従来手法 (summary)	従来手法 (html)	提案手法 (summary)
1 個	501	50,100	733
3 個	503	50,300	1,095
5 個	505	50,500	1,455

5. まとめと今後の課題

本論文では、特定の分野に関連する基底単語群を用いて、WWW 空間から関連語を自動収集する手法を提案した。本手法では、WWW 空間の各ページに URL が付随することに着目し、基底単語群が存在するページの URL 集合と類似した URL 集合を有する単語を関連語として収集する。本手法を実装する上で、各単語が存在する URL 集合は、WWW 検索エンジンの検索結果のサマリページのみから取得できるため、高速に関連語を収集することも可能になる。評価実験では、提案手法を用いることにより、従来手法よりも関連語収集の精度と速度が向上することを示した。

今後は、形態素解析辞書に登録されていない新語や未知語についても高精度に収集するアルゴリズムを組み入れて評価を行う予定である。更に、収集した関連語を用いたアプリケーションを開発し、本手法の有効性を高めたい。

[謝辞]

本研究の一部は、科学研究費補助金基盤研究(B)(17300036)、科学研究費補助金基盤研究(C)(17500644)を受けて行われた。

[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] 池野篤司, 濱口佳孝, 山本英子, 井佐原均: Web 文書集合からの専門用語獲得, 情報処理学会論文誌, Vol.47, No.6 (2006).
- [3] 小原恭介, 山田剛一, 絹川博之, 中川裕志: ウェブを利用した関連語収集, 第 3 回情報科学技術フォーラム (FIT2004), E-033, pp.183-184 (2004).
- [4] 岡田信哉, 村上淳哉, 渡部広一, 河岡司: Web を用いた新概念の自動学習, 第 3 回情報科学技術フォーラム (FIT2004), F-001, pp.195-198 (2004).
- [5] 大塚真吾, 豊田正史, 喜連川優: 大域ウェブアクセスログを用いた関連語の発見方法に関する一考察, 情報処理学会: データベース, Vol.46, No.SIG 8, pp.82-92 (2005).
- [6] 北研二, 中村哲, 永田昌明: 音声言語処理-コーパスに基づくアプローチ-, 森北出版 (1996).
- [7] 小泉大地, 獅々堀正幹, 中川嘉之, 柘植覚, 北研二: WWW 画像検索システムにおける有害画像フィルタリング手法, 情報処理学会: データベース, Vol.47, No.SIG 8, pp.147-156 (2006).
- [8] 山本一徳, 獅々堀正幹, 柘植覚, 北研二: パトリシアライの一次元配列構造への圧縮方法, 言語処理学会第 11 回年次大会, P4-4, pp.688-690 (2005).
- [9] Google, <http://www.google.co.jp/>.
- [10] Mecab, <http://mecab.sourceforge.jp/>.
- [11] 北研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム, 共立出版 (2002).

獅々堀 正幹 Masami SHISHIBORI

徳島大学大学院ソシオテクノサイエンス研究部准教授. 1993 徳島大学大学院工学研究科博士前期課程修了. 博士 (工学). マルチメディア情報検索, 自然言語処理の研究に従事. 著書「情報検索アルゴリズム」共立出版, 情報処理学会第45回全国大会奨励賞受賞. 電子情報通信学会, 情報処理学会, 言語処理学会, 日本データベース学会正会員.

山本 一晴 Issei YAMAMOTO

2007 徳島大学大学院工学研究科博士前期課程知能情報工学専攻修了. 2005 徳島大学工学部知能情報工学科卒業. 自然言語処理の研究に従事. 現在, 大坂ソフトハウス勤務.

小泉 大地 Daichi KOIZUMI

2007 徳島大学大学院工学研究科博士後期課程情報システム工学専攻修了. 2002 徳島大学工学部知能情報工学科卒業. マルチメディア情報検索, 自然言語処理の研究に従事. 現在, ㈱ジャストシステムに勤務. 情報処理学会正会員.

北 研二 Kenji KITA

徳島大学高度情報化基盤センター教授. 1986 早稲田大学理工学科学科卒業. 博士 (工学). 自然言語処理, マルチメディア情報検索等の研究に従事. 平成 6 年日本音響学会技術開発賞受賞. 著書「確率的言語モデル」東京大学出版会, 「情報検索アルゴリズム」共立出版など. 電子情報通信学会, 言語処理学会正会員.