URL マッチングによる WWW 空間からの関連語収集手法

中出訓規[1]　獅々堀正幹[2]　北研二[3]
K.Nakade　M.Shishibori　K.Kita
(徳島大学大学院　先端技術科学教育部)[1]
(徳島大学大学院　ソシオテクノサイエンス研究部)[2]
(徳島大学　高度情報化基盤センター)[3]

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や検索結果の分類など [3] 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まれている [2, 4]．本研究では，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．

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 C-value[4]，Jaccard 係数により計算する手法 [2] 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 検索システム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．

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 集合間に共通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共通性が高ければ，その関連候補語を関連語として採用する．

2. 従来の関連語収集技術

従来の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や C-value により関連語を収集する方法，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．また, 単語の類似性による手法では Jaccard 係数が有名であるが, 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求めて関連語を収集する．しかし, 各 Web ページの内容を解析して関連語を収集するため, 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，基底単語による検索結果と関連候補語による検索結果の双方に共起する単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．

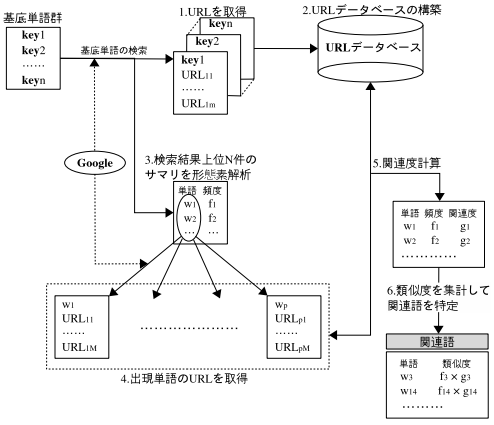

図 1: 本収集手法の概要

3. URL の共通性に基づく関連語収集手法

本研究では，WWW 上に存在する文書には URL が付随していることに着目し，URL の共通性を計ることで関連語の収集精度を向上し，かつ，収集時間を短縮する手法を提案する．本手法では，基底単語の検索結果から URL データベースを構築し，関連語候補語の検索結果から得られる URL 集合と URL データベースとのマッチングを行い，共通性の高い URL を有する関連語候補語を関連語として特定する．

3.1 本手法の概要

提案する関連語収集手法の概要を図 1 に示し，収集手順を説明する．なお，収集手順 2 で示す URL データベースの構築方法，及び収集手順 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．

[関連語収集アルゴリズム]

収集手順 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 $\mathrm{key}_i (1 \leq i \leq n)$ を WWW 検索システムに入力し，基底単語毎に上位 m 件の検索結果 $\mathrm{URL}_{ij} (1 \leq j \leq m)$ を得る．

収集手順 2: URL データベースを構築

収集手順 1 で得た検索結果 URL_{ij} からパス毎の頻度を計算し，求めた各頻度に対して 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．

収集手順 3: ページ内容の解析（出現頻度の計算）

収集手順 1 で得たサマリページを形態素解析し，名詞系単語 $\mathrm{w}_k (1 \leq k \leq p)$ の出現頻度 $\mathrm{Freq}(w_k)$ 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．

収集手順 4: 関連候補語の URL 集合を取得

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 $\mathrm{URL}_{kl} (1 \leq l \leq M)$ を得る．

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

収集手順 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．

収集手順 6: 関連語の特定

収集手順 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，類似度上位の単語を関連語とする．

【手順終了】

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

3.2 URL データベースの構築方法

3.1 の収集手順 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．これにより，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]

構築手順 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．

構築手順 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．

構築手順 3: 部分 URL の出現頻度の正規化

構築手順 1 の出現頻度を以下の式により大域的出現頻度で正規化し，その値を関連度とする．

$$\text{関連度} = \frac{\text{部分 } URL \text{ の出現頻度}}{\text{部分 } URL \text{ の大域的出現頻度}}$$

【手順終了】

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った例を示す．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

1. www.tokushima-u.ac.jp
2. www.tokushima-u.ac.jp/G-life

の 2 つであり，(1) のデータベース内での出現頻度は 3，(2) は 2 である．次に，各部分 URL を WWW 検索システムの URL 検索機能に入力して検索を行うと，(1) は 8570 件，(2) は 78 件の検索結果を得る．最後に正規化を行うと，(1) は 0.00035，(2) 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．なお，同様の方法で完全 URL の関連度を算出すると 1 になり，部分 URL の関連度との差が大きくなる．本手法では，URL 間の部分一致度を重視するため，完全 URL の関連度は URL データベース内に格納していない．

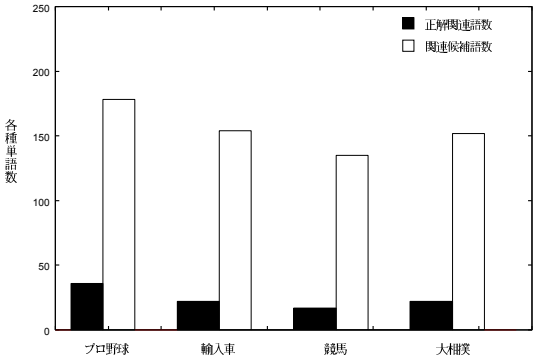

図 3: 基底単語毎の正解関連語数と関連候補語数

3.3 部分 URL マッチングによる関連度計算

3.1 の収集手順 5 に示すように，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 毎に照合を行い，照合に成功した部分 URL の関連度の総和を求める．例として，図 2 の URL データベースを用いて，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/a2/G-life
3. www.tokushima-u.ac.jp/G-life/main.html

まず，(1) の URL は URL データベースと照合すると，“www.tokushima-u.ac.jp” まで一致しているので関連度は 0.00035 となる．次に (2) については，部分 URL で一致しているのが “www.tokushima-u.ac.jp” までなので，(1) と同じ 0.00035 となる．最後の (3) については，URL データベース内の 2 番目の URL と完全一致しているが，部分 URL としては “www.tokushima-u.ac.jp/G-life” までが一致しているので 0.0256 となる．このように，URL データベース内に完全一致する URL が存在しても，最長で一致する部分 URL に対する関連度を取得する．最終的に，この関連候補語の関連度は (1)〜(3) の URL 関連度の総和とする．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，表 1 に示すような 4 種類の分野（野球，車，競馬，相撲）の基底単語について収集し，関連語を評価した．各基底単語に対して正解とすべき関連語の指標を表 1 内に示す．これらの基底単語を Google サーチエンジン [5] に入力して得たサマリを形態素解析 [6] し，出現頻度が 2 以上の名詞系単語を関連候補語とした．関連候補語内の正解とすべき関連語は，表 1 の指標に従って人手で判定した．各基底単語から収集した関連候補語数と正解関連語数を図 3 に示す．各関連候補語に各手法を適用した結果の上位 100 件の関連語を対象にして，上位から 5 件毎の結果に対して再現率と適合率 [7] を評価した．なお，従来手法としては，関連候補語を検索結果サマリ内の出現頻度を用いる手法,tf-idf 法 [7],Jaccard 係数手法 [2] をベースラインとして提案手法と比較した.

4.2 出現頻度・tf-idf との収集精度比較

検索結果のサマリに対して提案手法が有効かどうか検証するため, 基底単語（4 種類）で収集し精度比較実験を行った.

表 1: 実験で用いた基底単語群

分野	基底単語	正解とすべき関連語の指標
野球	プロ野球	野球用語，チーム名，選手名等，但しアマチュア野球に関する用語は除く
車	輸入車	メーカー名，車種名，販売会社，サイト名等，但し国産車に関する用語は除く
競馬	競馬	競馬用語，競馬場名，馬名，重賞名，競馬専門サイト名等
相撲	大相撲	相撲用語，親方名，力士名等，但しアマチュア相撲に関する用語は除く

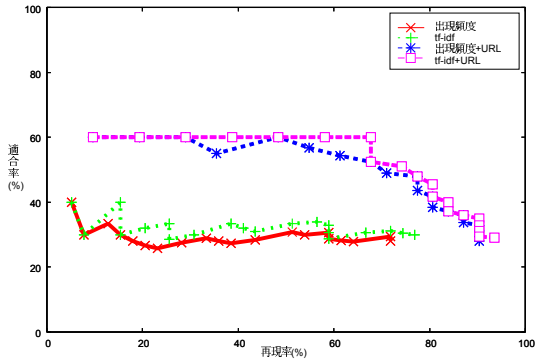

図 4: 基底単語 “プロ野球” に対する再現率・適合率曲線

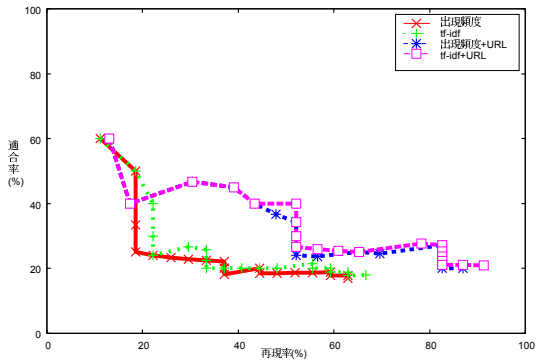

図 5: 基底単語 “輸入車” に対する再現率・適合率曲線

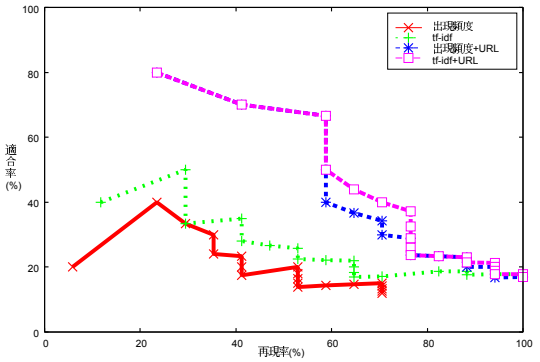

図 6: 基底単語 “競馬” に対する再現率・適合率曲線

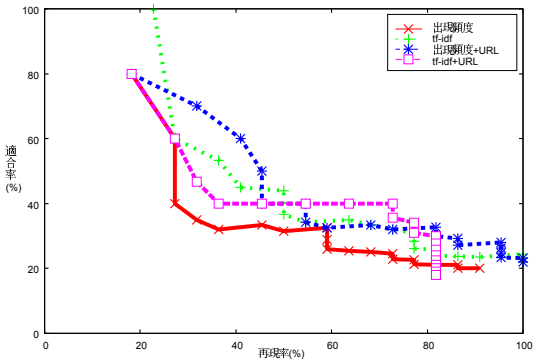

図 7: 基底単語 “大相撲” に対する再現率・適合率曲線

結果を図 4〜図 7 に示す．これらは，各手法による結果の上位 5 件毎の再現率と適合率をプロットした再現率・適合率曲線である．各グラフ内で，“出現頻度” が関連候補語を出現頻度順にソートした結果，“tf-idf” が関連候補語の出現頻度に対して tf-idf 法を適用した結果，“出現頻度+URL” が出現頻度に提案手法を加えた結果，“tf-idf+URL” は出現頻度に tf-idf 法を適用したものに提案手法を加えた結果を示す．また収集精度を比較するため，実験結果に対する F 値 [7] を表 2 に示す．表内の値は，上位 100 位までの F 値を 5 件毎に求め，それらの平均値を示している．大きい値を取れば取る程収集精度が良い事を示しており，どの分野においても提案手法を加える事により精度向上が見受けられ，提案手法は検索結果のサマリに対する関連語収集において有効である結果が得られた．しかし，基底単語 “大相撲” では，他と比べて F 値があまり向上していない．これは，“tf-idf” のみによる収集精度が良い事から提案手法を加えずとも既にある程度収集されていたため精度があまり伸びなかったと考えられる．

4.3 jaccard 係数との収集精度比較

次に，基底単語に対して（3 種類）提案手法と従来法である Jacard 係数との収集精度に関する比較実験結果を図 8〜図 10 に示す．これらのグラフは，各手法による結果の上位 5 件毎の再現率と適合率をプロットした再現率・適合率曲線である．

表 2: サマリにおける各手法の収集精度（F 値（%））

基底単語	出現頻度	tf-idf	出現頻度+URL	tf-idf+URL
プロ野球	27.70	30.97	48.35	49.50
輸入車	22.89	27.68	35.56	36.40
競馬	28.46	33.23	39.81	41.06
大相撲	35.33	40.74	42.11	39.73
平均	28.60	33.16	41.46	41.68

表 3: Jaccard 係数と提案手法の収集精度（F 値（%））

基底単語	サマリ	HTML	tf-idf+URL
プロ野球	44.30	44.18	49.54
輸入車	34.07	51.50	36.40
競馬	24.52	43.55	41.06
平均	34.30	46.41	42.33

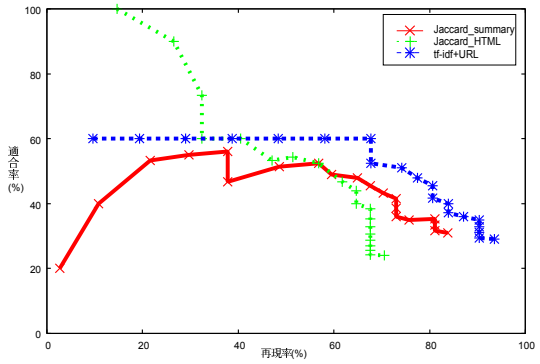

図 8: 基底単語 “プロ野球” に対する再現率・適合率曲線

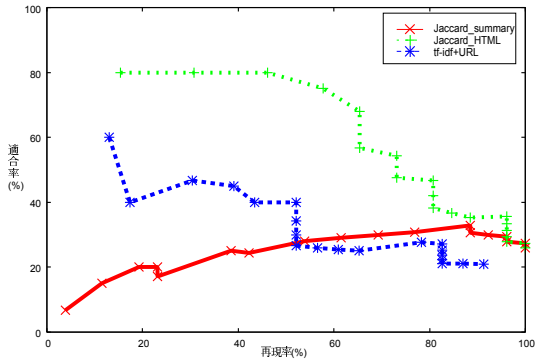

図 9: 基底単語 “輸入車” に対する再現率・適合率曲線

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，“tf-idf+URL” が検索結果のサマリに対して提案手法を適用した結果を示す．また収集精度を比較するため，実験結果に対する F 値を表 3 に示す．

まず，どの基底単語に対してもベースラインとして示した “出現頻度” の結果が最も低くなっており，単語の出現頻度のみでは，正しい関連語の収集が行えないことが伺える．“Jaccard_summary” については，“競馬”，“輸入車” に対する収集精度が極端に悪くなっており，サマリだけを用いた Jaccard 係数では的確に関連性を評価できないといえる．一方，提案手法では，サマリ内に共通単語が少なくなるような基底単語に対しても頑健に関連語を収集できており，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，基底単語毎に提案手法の結果を比較すると，精度に大きな違いが生じている．まず，“プロ野球” に対する精度は，再現率が優れている反面，上位での適合率が劣っている．これは，プロ野球の選手名（姓だけ）のように同姓の有名人が存在するため，意味的な曖昧性を有し関連度が低下してしまう点や，基底単語の検索結果サマリに意図していない分野の URL が混入したため，ノイズ単語の関連度が上昇してしまう点が理由として挙げられる．例えば，“プロ野球” の検索結果サマリ内には，スポーツ（野球）分野の URL だけでなく，ゲームソフト（野球）分野のノイズ URL も大量に混入されてしまい，その結果，“ファミスタ”，“ナムコスターズ” といったノイズ単語が上位にランキングされていた．

4.4 各手法との収集時間の比較

提案手法と従来手法との関連語収集時間を評価するため，関連語収集の際に両手法が WWW 空間にアクセスした回数を

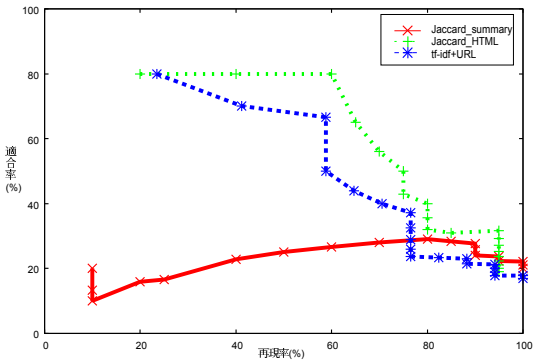

図 10: 基底単語 “競馬” に対する再現率・適合率曲線

表 4: 関連語収集に必要な平均 WWW アクセス数

基底単語数	従来手法 (summary)	従来手法 (html)	提案手法 (summary)
1 単語	166.5	16,551	465.5

比較する．表 4 に，両手法の平均 WWW 空間アクセス数を示す．表 4 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が少ない．また,HTML に対する従来手法と提案手法を比較すると，提案手法の方が従来手法に比べて 2.3%～2.8%の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．ただし，提案手法では，各部分 URL の大域的頻度を得るためにパス毎の URL 検索を行うためアクセス数が増加する．

5. まとめ

本論文では，特定の分野に関連する基底単語を用いて，WWW 空間から各ページに URL が付随すること着目し，関連語を自動収集する手法を提案した．評価実験を行い，提案手法が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，収集した関連語を用いて，WWW 検索システムの検索結果をフィルタリングするアプリケーションの開発に取り組む計画である．

6. 謝辞

本研究の一部は，科学研究費補助金基板研究 (B)(17300036)，科学研究費補助金基板研究 (C)(17500644) を受けて行われた．

参考文献

[1] 佐藤理史, 佐々木靖弘ウェブを利用した関連語の自動収集, 情報処理学自然言語処理研究会資料, NL-153-8, pp.57-64 (2003).

[2] 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, FIT2004(第 3 回情報科学技術フォーラム), E-033, pp.183–184 (2004).

[3] 安川美智子, 横尾英俊: クエリログから獲得した関連語のクラスタリングに基づく Web 検索, 電子情報通信学会論文誌, Vol.J90-D, No.2, pp.269-280 (2007).

[4] 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, pp.1717-1727 (2006).

[5] Google,http://www.google.co.jp/.

[6] Mecab,http://mecab.sourceforge.jp/.

[7] 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).