2005－NL－170（21）
2005／11／22

出現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

山本 一晴[†]　獅々堀 正幹[†]　柘植 覚[†]　北 研二[‡]

[†]徳島大学 工学部 知能情報工学科
[‡]徳島大学 高度情報化基盤センター

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，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 空間を対象とした，関連語を自動収集する研究が活発に取り組まれている．本稿では，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．本手法では，基底単語群が一連の同じ意味を有する場合，基底単語群を既存の検索システムに入力して得られる検索結果の URL 集合と，関連語を入力として得られる URL 集合との間に類似性があることに着目した．そして，パス毎の URL 出現頻度に重みづけを行うことにより，URL 集合内の類似性を計算し，基底単語群の URL 集合と類似した URL 集合を有する単語を特定することで関連語を収集する．実際に，3〜5 個の基底単語から成る 4 種類の基底単語群に対して実験を行った結果，従来法に比べて平均 17%程度の精度向上が認められ，収集速度に関しても有効性を示すことができた．

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

Issei Yamamoto[†]　Masami Shishibori[†]　Satoru Tsuge[†]　Kenji Kita[‡]

[†]Department of Information Science & Intelligent Systems, Faculty of Engineering
Tokushima University
[‡]Center for Advanced Information Technology, Tokushima University

The method to gather related keywords automatically is used in the construction of the dictionary on natural language processing systems and query expansion on information retrievals. In recent years, the gatheing methods from WWW space have been studied activity. In this paper, we propose the gatheing method from WWW space by using related basis words. On this method, we paid attention that the URL of retrieval result has commonness, if basis words has the same meaning. Then, weight is given at the time of each passing URL, and Web site is specified that it has a high related basis words. And, related words are collected there. Actually, it experimented to four kinds of basis words that consisted of the 3〜5 words. As a result, accuracy of about 17% on the average has improved compared with the conventional method. And,it is effective for the speed when related words are collected.

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，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 空間を対象とした，関連語を自動収集する研究が活発に取り組まれている．本研究では，数個のシーズとなる単語 (基底キーワード) を準備し，基底キーワード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を利用するものが殆どであった [2][3]．しかし，我々は，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 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．

本手法では，まず，基底キーワード群を既存の WWW 検索システムに入力する．ここで，基底キーワード群が一連の同じ意味を有すると仮定すると，検索結果内には，基底キーワード群に関連性の高いページが多く含まれると考えられる．従来は，この関連性の度合いをページ内に共起する単語 (関連候補語) の頻度情報を用いて評価していた．しかし，Web ページは一般の文書と異なり，各ページに URL が付随している．そこで本手法では，URL 集合間の類似度により関連性の度合いを評価する．つまり，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，基底キーワード群から得られた URL 集合との間で類似性が高ければ，その関連候補語を関連語として採用する．

例えば，基底キーワード“ 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．ここに，関連候補語“ 松井秀喜 ”の検索結果から得られる URL 集合が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できる．

また，本手法を用いれば URL をマッチングをするだけで，単語の関連性を短時間に取得することができ，効率的な関連語収集を実現できる．

2　従来の関連語収集技術

従来の代表的な関連語収集技術として，単語の共起情報を基に相互情報量を求め，この値により関連語を収集する方法 [1]，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することができる．

以下に，各手法の概要を説明する．

2.1　相互情報量による関連語収集

単語の共起情報に基づく相互情報量による関連語収集手法について説明する．これは，単語 x と y が同時に観測される確率 $P(x,y)$ と x,y が独立に観測される確率 $P(x),P(y)$ から式 (1) で単語の関連性を評価する．

$$I(x;y) = log\frac{P(x,y)}{P(x)P(y)} \tag{1}$$

これは，WWW 空間をキーワードの収集対象とした場合，出現頻度の極端に低い固有名詞などの単語がノイズとなる問題が生じるため，WWW 空間における関連語収集手法として不適切である．

2.2　出現単語の類似性よる関連語収集

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頻度ベクトル間の類似度を Jaccard 係数 (2) で評価する．

$$\sigma(CF(x),CF(y)) = \frac{\sum_{i=1}^{n} cs_i \cdot cw_i}{\sum_{i=1}^{n} cs_i^2 + \sum_{i=1}^{n} cw_i^2 - \sum_{i=1}^{n} cs_i \cdot cw_i} \tag{2}$$

図 1 に出現単語の類似性による関連語収集のシステムの概要を示し，手順を説明する．

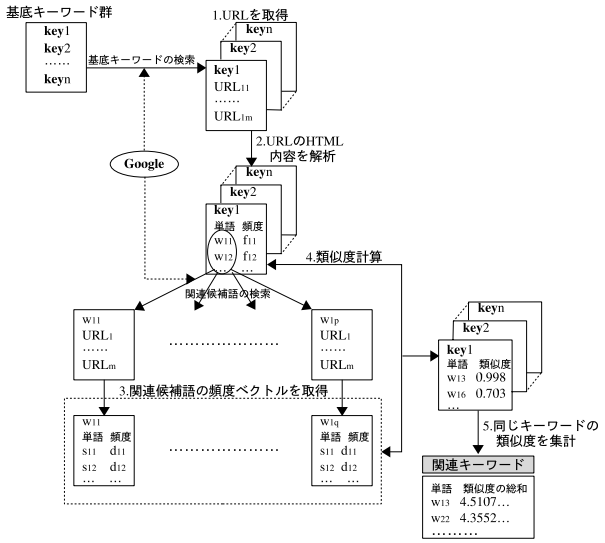

図 1: 出現単語の類似性による関連語収集システム

手順 1: 基底キーワードが存在するページを検索

あらかじめ人手で登録した各基底キーワード $\mathrm{key}_i(1 \leq i \leq n)$ を WWW 検索システムに入力し，各キーワード毎に上位 m 件の検索結果 $\mathrm{URL}_{ij}(1 \leq j \leq m)$ を得る．

手順 2: ページ内容の解析 (基底キーワードの頻度ベクトルを取得)

手順 1 の検索結果 URL_{ij} に対応する HTML を形態素解析し，各出現単語 $\mathrm{w}_{ik}(1 \leq k \leq p)$ を関連候補語とする．また，総出現頻度ベクトル $CF(key_i) = (cw_{i1}, cw_{i2}, \cdots, cw_{ik}, \cdots)$ を集計する．ただし，cw_{ik} は，キーワード w_{ik} の出現頻度とする．

手順 3: 関連候補語の頻度ベクトルを取得

手順 2 で取得した関連候補語 $\mathrm{w}_{ik}(1 \leq k \leq p)$ について，同様に手順 1，手順 2 を行い，各関連候補語の検索結果 $\mathrm{s}_l(1 \leq l \leq q)$ に対しても総出現頻度ベクトル $CF(w_{kl}) = (cs_{k1}, cs_{k2}, \cdots, cs_{kl}, \cdots)$ を集計する．

手順 4: 類似度計算

基底キーワードと関連候補語の頻度ベクトルを用いて，式 (2) により類似度を求める．

手順 5: 関連キーワードの特定

各基底キーワードから求めた関連語において，同一の単語が存在すれば，それぞれの類似度で和をとる．w_{ik} の類似度 $\sigma(CF(key), CF(w))$ でソートし，上位の単語を関連キーワードとする．

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．そこで，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 の共通性を用いることで関連語の収集時間を短縮する．その手法として，本稿では基底キーワードの検索結果 URL から構築する URL データベースを用いて，検索範囲を限定して関連語を収集する．さらに，URL による関連度の判定により，関連語の収集効率を上昇させている．

3 URL の類似性に基づく関連キーワード収集手法

3.1 本収集手法の概要

本稿では，URL の類似性を用いた関連語収集を提案する．図 2 に提案する関連キーワード自動収集手法のシステムの概要を示し，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 [4]，および手順 5 で示す関連度の計算方法については 3.2 で詳しく述べる．

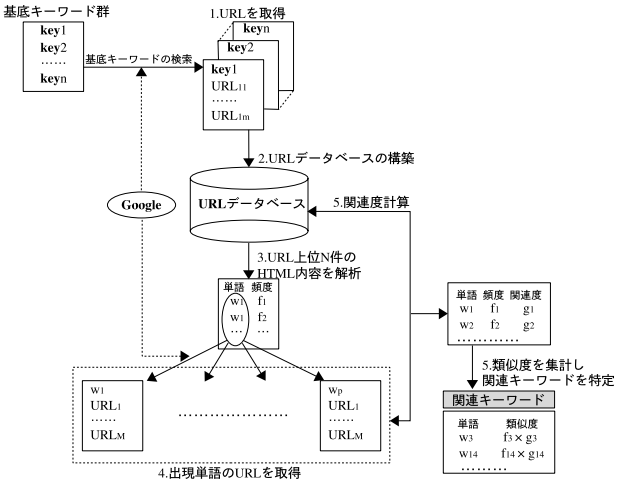

図 2: 本収集手法のシステム

手順 1: 基底キーワードが存在するページを検索

あらかじめ人手で登録した各基底キーワード $\mathrm{key}_i(1 \leq i \leq n)$ を WWW 検索システム

に入力し，各キーワード毎に上位 m 件の検索結果 $\mathrm{URL}_{ij}(1 \leq j \leq m)$ を得る．

手順 2: URL データベースを構築

手順 1 で得た検索結果 URL_{ij} から部分 URL を抽出し，WWW 空間全体の URL 出現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 (出現単語頻度の計算)

URL データベース中の正規化された URL 出現頻度の上位 N 件の URL に対応する HTML を形態素解析し，出現単語 $\mathrm{w}_k(i \leq k \leq p)$ と出現頻度 $\mathrm{Freq}(w_k)$ を集計する．

手順 4: 出現単語の URL を取得

出現単語 w_k を WWW 検索システムに入力し，出現単語毎に上位 M 件の検索結果 $\mathrm{URL}_{kl}(1 \leq l \leq q)$ を取得する．

手順 5: 関連度の計算

URL データベースと URL_{kl} のマッチングを行い，出現単語 w_k の関連度を求める．

手順 6: 関連キーワードの特定

関連度と $\mathrm{Freq}(w_k)$ の積から類似度を求め，ソートして関連キーワードを特定する．

上記アルゴリズムの手順 2 において，基底キーワード群が出現する URL 集合を特定している．また，手順 4 において出現単語 (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．出現頻度の正規化で，基底キーワードと URL における関連性の強弱を判別することができる．これにより，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．以下に正規化の手順を示す．

手順 1: 部分 URL 毎の出現頻度の計算

URL データベース内に出現する部分 URL 毎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 ”と“ 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 の出現頻度を式 (3) 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．

$$\text{関連度} = \frac{\text{部分 } URL \text{ のデータベース内での出現頻度}}{\text{部分 } URL \text{ の } WWW \text{ 空間中での大域的出現頻度}} \quad (3)$$

図 3 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．図 3 の URL データベースには 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 と (b)www.tokushima-u.ac.jp/G-life の 2 つであり，部分 URL(a) のデータベース内での出現頻度は 3，(b) は 2 である．つぎに各部分 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部分 URL(a) は 8570 件，(b) は 78 件の検索結果を得る．最後に，式 (3) により正規化した出現頻度を求める．部分 URL(a) は 0.00035，(b) は 0.0256 となる．

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と基底キーワード群との関連度
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 3: 出現頻度の正規化の例

この関連度は，基底キーワードが出現しやすい Web サイトとの関連性を示している．3.1 の手順 5 では，関連度により構築した URL データベースと出現単語の URL 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．さらに，手順 6 で出現単語の関

連度と出現頻度で積をとった類似度により関連語を特定している．

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 [5] によって URL データベースを構築している．

4 評価実験

4.1 実験条件

本手法の有効性を確かめるために，今回は“ 野球 ”に関連するキーワードを収集して評価を行った．あらかじめ人手で登録した基底キーワード数を 3 件，5 件と変化させ，既存の WWW 検索システムに入力して得られた検索結果の URL をデータベースに登録した．また，URL データベースの上位 100 件の URL に対応したページから関連候補語を収集した．関連候補語は，出現頻度が 5 以上の単語である．以下に実験で使用した基底キーワードを示す．

1. key①={ 本塁打，打率，打点 }
2. key②={ 本塁打，ホームラン，打者 }
3. key③={ 野球，ベースボール，ホームラン，打者，野手 }
4. key④={ 本塁打，打率，打点，打者，三冠王 }

また，収集結果上位 100 件までの単語に対して，基底キーワードとの関連性を人手で判定し，関連キーワードの収集精度を式 (4)(5) に示す再現率と適合率 [6] を用いて評価した．

$$\text{再現率} = \frac{\text{システムが出力した単語に含まれる正確関連語数}}{\text{収集結果上位 100 件に含まれる全正確関連語数}} \quad (4)$$

$$\text{適合率} = \frac{\text{システムが出力した中単語に含まれる正確関連語数}}{\text{システムが出力した単語数}} \quad (5)$$

4.2 WWW 検索システムの比較

現在，様々な WWW 検索システムが開発されているが，本手法を適用するのに最適な検索システムは，入力する基底キーワード群によって異なってくることが予想される．この点に関して，基底キーワードが与えられた時点で，そのキーワードに最適な検索システムをユーザに提示することが重要となる．そこで，各基底キーワードと検索システムとの組合せを変化させた際の収集精度を検索結果内の単語の異なり数といった尺度で評価した．

今回は，2 種類の検索システムとして，Google [7] の Web Search と Image Search を用いて比較実験を行った．以下に，検索結果内の異なり数の計算方法について説明する．

手順 1: 基底キーワードが存在するページを検索
あらかじめ人手で登録した各基底キーワード $\mathrm{key}_i(1 \le i \le n)$ を WWW 検索システムに入力し，各キーワード毎に上位 m 件の検索結果 $\mathrm{URL}_{ij}(1 \le j \le m)$ を得る．

手順 2: ページ内容の解析 (出現単語頻度の計算)
m 件の URL に対応する HTML を形態素解析し，出現単語 $\mathrm{w}_k(i \le k \le o)$ を集計する．

手順 3: 異なり数の計算
URL_{ij} で取得した単語に他の基底キーワードが含まれている数と m 件中他の基底キーワードが出現したファイル数を取得する．これを，式 (6) に代入し，異なり数を求める．

$$\text{異なり数} = \frac{key \text{ の出現数}}{key \text{ が出現したファイル数}} \quad (6)$$

異なり数は，ある基底キーワード群において，1 つのキーワードの検索結果に他のキーワードがどれだけ出現しているかという割合である．したがって，異なり数の値が大きければ，そのキーワードは他のキーワードを含んでいることが多く，関連性が大きいキーワードであると判断することができる．すなわち，基底キーワード群における異なり数の平均値は，一連の同じ意味を持つキーワード集合である評価となる．

表 1 の上段に異なり数の平均，下段に収集した関連語上位 100 件に対する平均適合率を示し，図 4〜7 にキーワード毎の再現率，適合率グラフを示す．

表 1: 異なり数と平均適合率との関係

		Web	Image
key①	異なり数	0.9603	0.7899
	平均適合率 (%)	88.63	73.15
key②	異なり数	0.6034	0.6164
	平均適合率 (%)	60.43	74.05
key③	異なり数	0.9059	0.9518
	平均適合率 (%)	59.69	71.55
key④	異なり数	1.3905	1.0461
	平均適合率 (%)	92.60	76.95

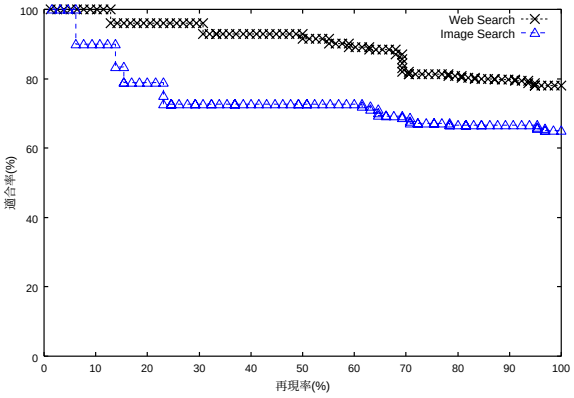

図 4: key①の再現率/適合率グラフ

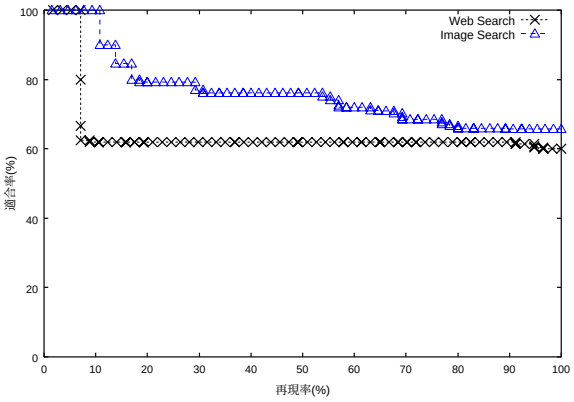

図 5: key②の再現率/適合率グラフ

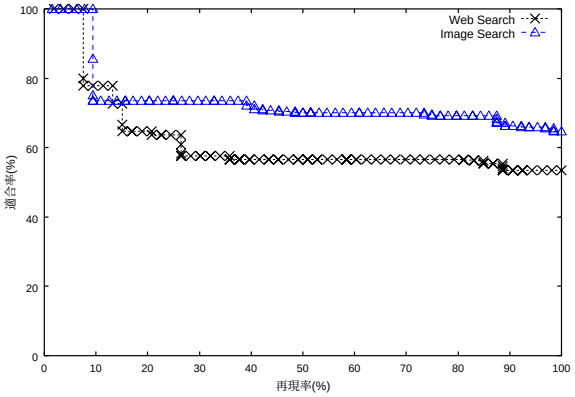

図 6: key③の再現率/適合率グラフ

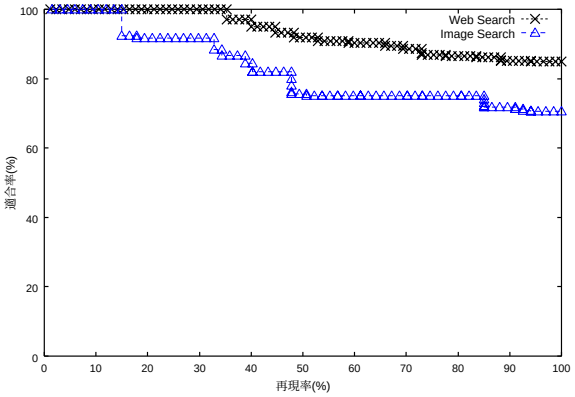

図 7: key④の再現率/適合率グラフ

異なり数と適合率の平均値を照合すると，異なり数の値が大きい検索方法を用いる方が高精度に関連語収集できることがわかる．つまり，異なり数と適合率の相関関係により，様々な WWW 検索システムの中から本手法の適用に最適な検索システムを特定することができる．

4.3 従来手法との比較

従来手法として“ ウェブを利用した関連用語収集 ”[2] を用い，本手法で収集した関連語に対する精度比較，及び，収集速度について理論的に検証した．

4.3.1 関連語収集精度の比較

表 2 に収集した関連語の上位 100 件に対する平均適合率を示す．また，図 8〜10 に異なり数によ

り選択した WWW 検索システムを用いて収集した関連語に対しての再現率/適合率グラフを示す.

表 2: 関連語収集の平均適合率

基底キーワード群	本手法	従来法
key①	88.63	70.03
key②	74.05	62.19
key③	71.55	60.76
key④	92.60	66.27
平均	81.71	64.81

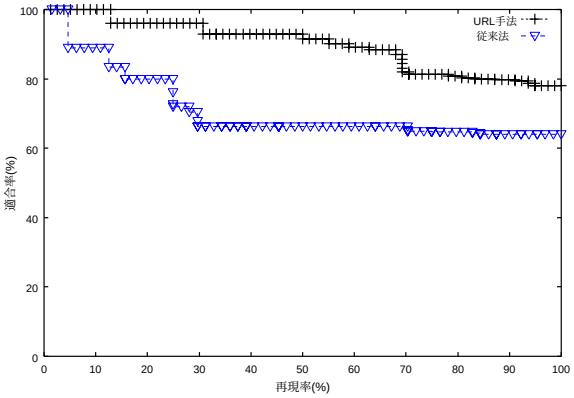

図 8: key①再現率/適合率グラフ

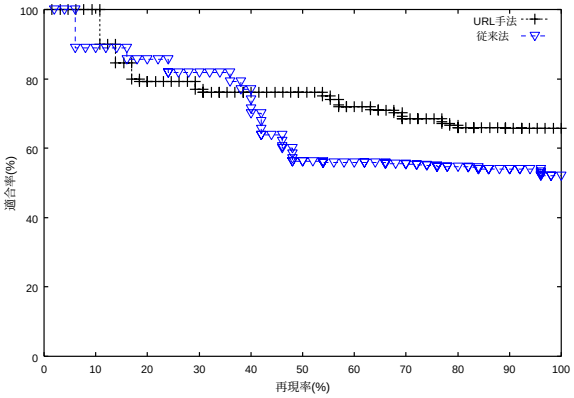

図 9: key②再現率/適合率グラフ

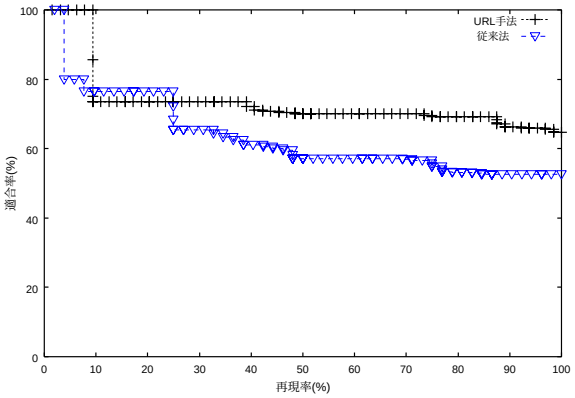

図 10: key③再現率/適合率グラフ

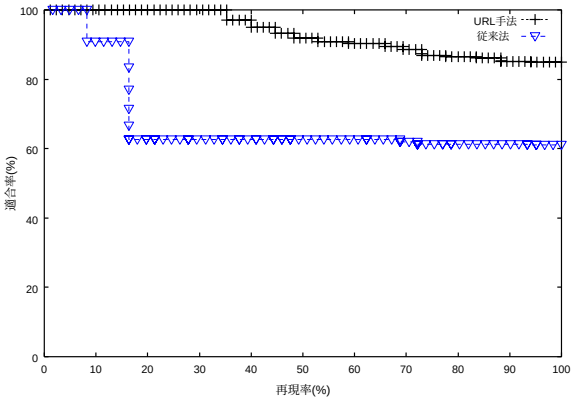

図 11: key④再現率/適合率グラフ

表より，本手法と従来法の平均適合率は提案手法が勝っており，本手法は収集精度において優れていることがわかる．また，図中より，key①と key④が高水準の適合率で推移しており，関連語の収集精度が高いことがわかる．これは，基底キーワード数 3 件と 5 件の中で最も異なり数の平均値が大きい基底キーワード群であり，意味的関連性の大きい基底キーワード群であることを示している．一方，異なり数の平均値が小さい基底キーワード群の中には，他のキーワードと比べて小さい異なり数の値の単語が存在する．このような単語は意味的多義性をもつ単語であるため，意味の異なる要素がノイズとして混入し，関連語の収集精度を低下させる要因となる．すなわち，異なり数で個々の基底キーワードの意味的関連性を評価すれば，基底キーワード群の精錬化をすることができ，さらに関連語収集精度を向上させることも可能であると考えられる．

4.3.2 関連語収集時間の比較

あらゆる検索や情報収集において，検索・収集時間の短縮は必要不可欠な要素である．今回の評価実験では，関連語収集時の両手法が WWW 空間にアクセスした回数により比較を行った．表 3，4 に，両手法の基底キーワード数の違いによる WWW 空間アクセス数の変化を示す．

表 3: WWW アクセス数：Web Search

基底キーワード数	本手法	従来法
key①	6,473	178,725
key②	6,131	228,181
key③	11,659	361,646
key④	5,816	354,820
平均	7,519.8	280,843

表 4: WWW アクセス数：Image Search

基底キーワード数	本手法	従来法
key①	12,835	343,245
key②	10,620	291,585
key③	9,667	675,255
key④	11,140	538,860
平均	11,065.5	462,236.3

表に示すように，従来法より提案手法が少ない WWW アクセス数で関連語収集をしていることがわかる．すなわち，提案手法の方が短時間で関連語収集を行うことができる．これは，従来法の 2.2 手順 3 と本手法の 3.1 手順 4 における WWW アクセス数の差が要因となっている．従来法は，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する．一方，本手法では，関連候補語の検索結果の URL を取得するだけである．すなわち，関連候補語の検索結果 100 件の URL を対象とした場合，従来法では 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，本手法では 1 回のアクセスで手順を進めることができる．ただし，本手法では，URL データベースを構築する際，各 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，本手法ではこの処理に対するアクセス数が増加する．

5 まとめ

本研究では，特定の分野に関連するキーワードを用いて，WWW 空間内における URL の出現頻度に着目し，関連語を自動収集する手法を提案した．評価実験では，本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，さまざまな分野における評価実験，関連語収集の精度や速度向上を検討し，収集した関連キーワードを用いたシステムやアプリケーションの開発を通じ，本手法の有効性を更に高めたい．

謝辞

本研究の一部は，科研費基盤研究 (B) 17300036，科研費基盤研究 (C) 17500644 を受けて行われた．

参考文献

[1] 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.

[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 .

[3] 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,FIT2004(第 3 回情報科学技術フォーラム), F-001, pp.195–198 .

[4] 中川嘉之, 獅々堀正幹, 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, 言語処理学会第 11 回年次大会, 2005.

[5] 山本一徳, 獅々堀正幹, 柘植覚, 北研二: パトリシアト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, 2005.

[6] 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.

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