出現URLの類似性に着目したWWW空間からの関連キーワード自動収集手法

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

竹安 真紀夫[†] 獅々堀 正幹[†] 中川 嘉之[†] 柘植 覚[†] 北 研二[‡]
Makio Takeyasu Masami Shishibori Yoshiyuki Nakagawa Satoru Tsuge Kenji Kita

1. はじめに

本稿では，数十個程度のメタなキーワードを入力として与え，入力キーワード群に関連した大量のキーワードをWWW空間から効率的に自動収集する手法について述べる．従来の関連キーワードの収集手法としては，相互情報量を尺度にした単語間の共起性に基づく手法が主流であった．しかし，WWW空間を収集対象とした場合，出現確率が極端に低い単語を共起単語として選択し，ノイズを生じることが問題となっていた．

本研究では，個々のキーワードの共起ペアを収集するのではなく，一連のキーワード群と意味的に関連したキーワードを収集することを目的とする．入力キーワード群が同じ意味をもつと仮定すると，それらが出現するページは限定され，限定されたページ集合と同じようなサイト内で共起する単語を関連キーワードとして収集できると考えられる．

そこで本手法では，まず入力キーワード群が出現するページのURL集合を特定する．そして，URL集合内で共起する単語の中から，URL集合とより類似したサイト内に出現する単語を関連キーワードとして収集する．本手法により再帰的に増幅したキーワード集合を用いると，そのキーワードに関連したコンテンツを含むページのみをWWW空間から自動収集することも可能になる．

2. 関連キーワード収集技術

2.1 WWW空間からの関連キーワード収集技術

我々は，WWW空間における有害情報フィルタリングシステムを開発している[1]．このシステムは，有害情報を含むURLに類似するURLを有害サイトとしてフィルタリングするシステムである．既存のフィルタリング手法は，大きく(1)URLチェック方式，(2)キーワードチェック方式の2つに分類される[2]．

(1) URLチェック方式…有害情報を含むURLをデータベース化し，完全にマッチするURLに対してフィルタリングを行う．

(2) キーワードチェック方式…有害情報を表すキーワードをあらかじめ登録し，登録キーワードが頻繁に出現するページに対してフィルタリングを行う．

(1),(2) いずれの方式も高精度なフィルタリングを行うには，改善の余地がある．特に(2)キーワードチェック方式によるフィルタリング手法は，事前にキーワードを登録しておく必要があるが，膨大な情報が存在するWWW空間を対象に，人手によってキーワードを収集するには限界がある．その問題を解決するため，有害サイトに関連するキーワードをWWW空間内でのキーワード出現頻度を考慮し，自動収集する技術が有効である．この技術は，最初に基底となる数十個のキーワード(基底キーワード)を用意し，基底キーワードに意味的に関連のあるキーワードを収集する技術である．

キーワードの自動収集技術は有害情報フィルタリングに限らず，特定分野に関するページのみを収集することも可能にする．また，WWW収集システムについてもその分野の充実したキーワードが多ければ多いほど収集精度は向上すると考えられる．このようにWWW空間からの関連キーワード収集技術は言語知識を自動獲得するものであり，数多くのアプリケーションに有効である．

2.2 単語の共起関係を用いた関連キーワード収集手法

従来の関連キーワード収集手法は，主に相互情報量を尺度にした単語間の共起関係を用いた手法であった．相互情報量は単語の共起や関連を表す尺度として用いられ，式(1)で表すことができる[3]．相互情報量$I(x;y)$は2つの単語xとyとが同時に観測される確率$P(x,y)$をx,yが独立に観測される確率$P(x),P(y)$と比較する．

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

図1に単語間の共起関係を用いた関連キーワード収集システムを示し，手順を説明する．なお，通常，単語が出現するページを収集するにはクローラーを用いるが，時間コストがかかるため本稿では既存のWWW検索システムの検索結果を利用する．

†徳島大学工学部
‡徳島大学高度情報化基盤センター

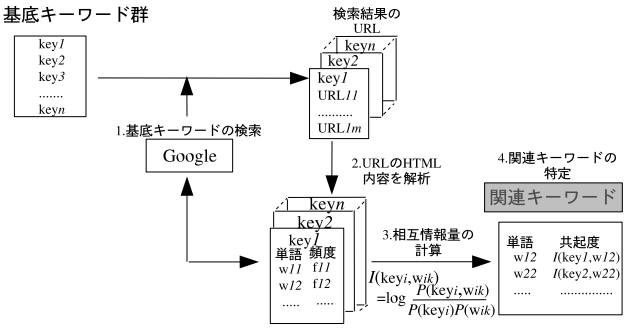

図 1: 単語の共起関係を用いた関連キーワード収集システム

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 検索エンジン (Google) 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．また，検索結果内に含まれる「検索件数」の値を key_i の WWW 空間内での出現頻度 $\mathrm{Freq}(\mathrm{key}_i)$ とする．

手順 2: ページ内容の解析 (出現単語頻度の計算)

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 $\mathrm{w}_{ik}(1 \leq k \leq o)$ と key_i との共起頻度 $\mathrm{Freq}(\mathrm{key}_i, \mathrm{w}_{ik}\mathrm{xs})$ を集計する．

手順 3: 相互情報量の計算

key_i と w_{ik} との相互情報量を計算する．ただし，w_{ik} を WWW 検索エンジンで検索し，その「検索件数」を $\mathrm{Freq}(\mathrm{w}_{ik})$ とする．$\mathrm{Freq}(\mathrm{key}_i, \mathrm{w}_{ik})$，$\mathrm{Freq}(\mathrm{key}_i)$，$\mathrm{Freq}(\mathrm{w}_{ik})$ を用いて式 (1) より相互情報量を計算する．

手順 4: 関連キーワードの特定

各 w_{ik} について手順 3 で求めた相互情報量をソートし，上位の単語を関連キーワードとする．

単語の共起関係を用いた手法は，WWW 空間を収集対象にした場合，人名に代表される固有名詞などの極端に出現確率の低い単語と共起する単語がノイズとなることが問題になる．そこで，本稿では関連したキーワードは同じような URL をもつサイト内で共起するのではないかと考え，入力キーワードと共起する単語の中から，入力キーワード群が出現する URL 集合とより類似したサイト内に出現する単語を関連キーワードとして収集する手法を提案する．

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

図 2 に本稿で提案する関連キーワード自動収集手法を示し，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，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．

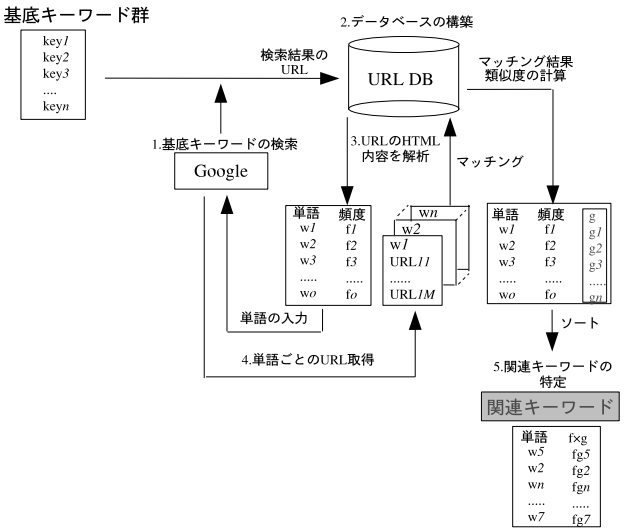

図 2: 本提案手法の概要

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 検索エンジン (Google) 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 を作成し，WWW 空間中の 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 件の部分 URL に対応する HTML を形態素解析し，出現単語 $\mathrm{w}_k(i \leq k \leq o)$ 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 検索エンジン (Google) に入力し，出現単語毎に上位 M 件の検索結果 $\mathrm{URL}_{kl}(1 \leq l \leq M)$ を得る．

手順 5: 関連キーワードの特定

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．関連度と $\mathrm{Freq}(w_k)$ の積を求めソートし，上位の単語を関連キーワードとする．

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 (関連キーワード候補) 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} に対して，WWW 空間中の URL 出現頻度を用いて正規化を行う．出現頻度の正規化を行うことで，基底キーワードとの関連が弱い Web サイトの検出を抑えることができる．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 データベース内の部分 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 検索エンジンの URL 検索機能に入力し，検索結果内の「検索件数」を部分 URL の WWW 空間中での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 (2) 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．

$$\text{関連度} = \frac{\text{部分 } URL \text{ のデータベース内での出現頻度}}{\text{部分 } URL \text{ の } WWW \text{ 空間中での大域的出現頻度}} \quad (2)$$

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 (1)www.tokushima-u.ac.jp と (2)www.tokushima-u.ac.jp/G-life の 2 つであり，部分 URL(1) のデータベース内での出現頻度は 3，(2) は 2 である．つぎに各部分 URL を WWW 検索エンジンの URL 検索機能に入力して検索を行うと部分 URL(1) は 8570 件，(2) は 78 件の検索結果を得る．最後に，式 (2) により正規化した出現頻度を求める．部分 URL(1) は 0.00035，(2) は 0.0256 となる．

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，この関連度付き URL データベースと出現単語の URL とのマッチングを部分 URL 毎に行い，マッチングに成功した URL の関連度の総和を求める．

```
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

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，今回は有害情報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 N を 10，30，50 と変化させ，基底キーワード群を Google Image Search 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，HTML を形態素解析する URL は，正規化した URL データベースの上位 100 件を使用した．表 1 に基底キーワード件数毎の URL 数，形態素解析で得られた出現単語数を示す．

表 1: 実験データ

基底キーワード数	URL 数	出現単語数
10	2300	6118
30	8181	6213
50	15933	5066

4.2 関連キーワードの適合率

基底キーワード数 N を 10，30，50 件と変化させたときの関連キーワードの適合率を式 (3) より求めた．出現単語と基底キーワードとの関連があるかないかの判断は人手により行った．

$$\text{適合率} = \frac{\text{システムが出力した中で関連がある出現単語数}}{\text{システムが出力した出現単語数}} \quad (3)$$

今回の実験では，基底キーワード数の違いによる収集精度の変化を評価した．まず，基底キーワード数の何倍の出現単語を収集すると，どの程度の精度変化があるか調べるため，システムの出力単語数 (式 (3) の分母の

値) を $N \times k$ とし，k の値を 1〜10 まで変化させながら適合率を求めた．実験結果を図 4 に示す．

図中より，基底キーワードが多くなれば，適合率が低下していることがわかる．基底キーワードの増加に伴い，ノイズが増えているといえる．一方，基底キーワード数が 10 件の場合は，適合率 0.8 前後で推移しており，$k = 10$ 倍までの関連語を収集するうえでは，ノイズの影響をあまり受けないことがわかる．これは、$N = 10$ の場合には基底キーワードが出現するサイトの URL が限定されるのに対して，$N = 50$ になると，出現頻度が広がり，URL の類似性を判定しにくくなることが原因であると考えられる．特に，意味的多義性をもつ単語が基底キーワードに存在すると，意味の異なるページの URL がノイズとして URL データベースに混入され，精度低下の原因となる．

次に，図 4 の結果では，$N = 10$ と $N = 50$ とでは対象となる関連キーワード候補数が大きく異なるため単純に比較することはできないので，$k = 1$ のときほぼ同じ適合率であった $N = 10$ と $N = 30$ の適合率をシステムの出力単語数を変化させて比較した． 図 5 より，上位 50 件までの適合率は，ともに約 0.8 あり，基底キーワード数による違いは見られないが，上位 50 件を超えると，$N = 30$ は多くのノイズを含むようになり，適合率が低下している．一方，$N = 10$ は上位 100 件までは適合率 0.8 を維持している．しかし，$N = 10$ の場合も上位 100 件を超えると適合率が徐々に低下している．これは，少数の基底キーワードのみでは収集単語数を増加するとノイズ混入が大きくなることを示しており，適時，収集した単語のレイティングを行ったり，再帰的にシステムを稼動することで精度向上が期待できる．そのために，今後，$N = 10$ で収集した上位 50 件の関連キーワードを基底キーワードとして再帰的にキーワードを収集した結果と，図 4 で示した $N = 50$ の収集結果との精度比較を行う必要がある．

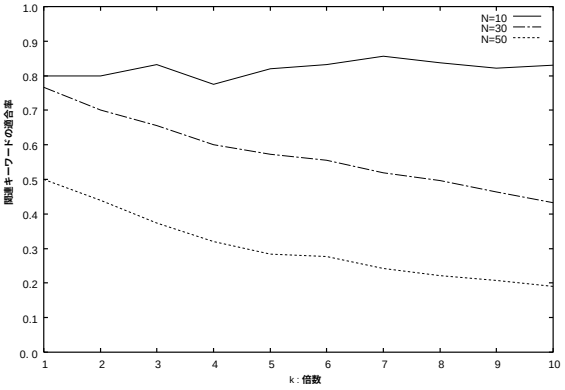

図 4: 基底キーワード毎の適合率の変化

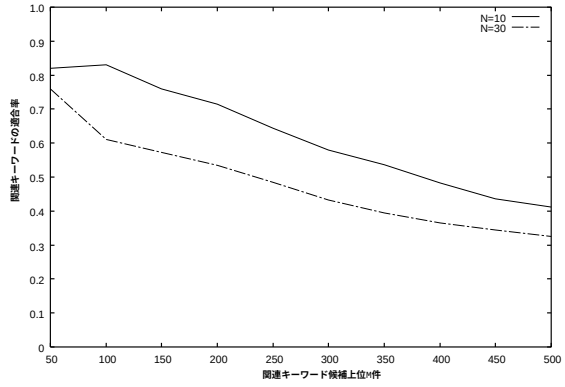

図 5: 基底キーワード 10 件と 30 件の適合率の変化

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 WWW 空間内でのキーワード出現頻度を考慮し，自動収集する手法を提案した．評価実験では，基底キーワード数による収集精度を示した．今後は，さまざまな分野における評価実験，収集した関連キーワードを用いたフィルタリング実験を行い本手法の有効性を更に検討したい．

参考文献

[1] 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.

[2] 井ノ上直己, 帆足啓一郎, 橋本和夫: 文書自動分類手法を用いた有害情報フィルタリングソフトの開発, 信学論 D-II, J84-D-II, No.6, pp.1158–1166, 2001.

[3] 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.