D-010

Webマイニングのためのトピック主導型クローラの試作
A Topic-focused Crawling Method for Web Mining

金子 昌弘† 安藤 一秋†
Masahiro Kaneko Kazuaki Ando

1 はじめに

近年，インターネットの普及に伴い，Web 上には様々なトピックに渡る膨大な情報が存在するようになった．そこで，Web 上のデータを巨大な情報源とみなし，マイニングすることで有用な知識を発見する技術（Web マイニング[1]）が注目を浴びている．

Web マイニングを行うためには，クローラを使って Web 上の情報を収集する必要がある．しかし，ロボットやクローラなどで無作為に収集した場合，目的のトピック以外の情報も含まれるため，マイニングのための情報源としてはノイズを含む可能性が高い．

以上の問題点を解決するため，我々は Web マイニングのための情報源の構築を目的としたトピック主導型クローラを提案した[2]．提案手法では，トピックランクによるランキングと Web ページの評価，シソーラスを利用した単語の類似度評価，正例・負例による学習型の単語スコアを実装した．シソーラスの利用により，トピックを限定しながら幅広い収集が可能となること，収集トピックに関連する単語を学習する単語スコアにより，収集精度が向上することを確認した．しかし，シソーラス上に存在しない単語（未登録語）は類似度計算に利用していなかった．

そこで本稿では，複数の特徴語と未登録語との相互情報量を計算することで，未登録語の類似度を計算し，トピックとの関連性を評価する手法を検討する．シソーラスには一般的な語だけしか登録されておらず，トピックに関する専門用語や固有名詞などは，ほとんどが未登録語となる．未登録語とトピックとの関連性を評価することで，収集精度の向上を目指す．

2 関連研究

石川らの研究[3]では，データベースを利用して Web 上から特定地域の飲食店情報を収集するクローラを提案している．しかし，石川らの研究では実データの登録されたデータベースの構築をしなければならない問題点がある．異なるトピックを収集する場合，再度データベースを構築する必要がある．また，冨山らの研究[4]では，One man& his dog システムと呼ぶトピック判定とリンク選出戦略機能を連携させるシステムを提案している．冨山らの研究は，ページ全体の単語を利用しており，ノイズとなる単語が含まれる可能性がある．また，ターゲット文書を判定するのはキーワードによるマッチングである．ため，キーワードを含まない Web ページは収集されない．

本研究の目的は，Web マイニングのための情報源の構築であるため，幅広くかつノイズの少ない情報源を構築する必要がある．そこで，シソーラスを用いた単語の類似度計算と，単語スコアによる学習機能により，これらの問題点を解決する．

† 香川大学大学院工学研究科

3 トピック主導型クローラの概要

文献[2]で提案したクローラの概要を説明する．本研究では，Web ページと収集トピックとの関連性を示す指標として，トピックランクを定義する．トピックランクは，URL のランキングのためのランキング値，クローリングのためにユーザが設定したキーワードとタイトル・本文との類似度の総和である．トピックランクが閾値以上であれば，そのコンテンツはトピックと関連していると判定する．ランキング値は，リンク元ページのトピックランク，アンカー文字列の類似度，後述する単語スコアの総和であるランキング値を基に URL リストがランキングされるので，クローラはリストの上位から順にアクセスする．トピックランク及ランキング値における類似度計算には，シソーラスを利用する．

クローラが収集開始する初期ページは重要である．本研究では Yahoo!検索 API[5]を利用し，キーワードで検索した上位 10 件を初期ページとして利用する．

以下，シソーラスによる類似度計算と単語スコアの計算法について説明する．

3.1 シソーラスによる類似度計算

キーワードとの完全一致による判定の場合，キーワードの含まれないページは収集されない．そこで，シソーラスを用いることで，この問題点を解決する．以下に手順を示す．

1. 形態素解析を行い形態素単位に区切る．
2. 日本語語彙体系[6]により各単語の意味属性を求める．
3. キーワードとの類似度 r を求める．

ある二つの意味属性 a，b の類似度 r は，意味属性 a の深さを d_a，意味属性 b の深さを d_b とし，a，b 共通の上位概念の深さを d とすると以下の式で定義する．

$$r = \frac{2d}{d_a + d_b}$$

3.2 単語スコアの計算

リンク周辺の文字列はそのリンク先の内容を表していると仮定する．そこで，ランキング値で利用されるアンカー文字を除いたリンクの周辺文字列の各単語にスコア付けを行い，収集精度の向上を図る．以下に手順を示す．

1．リンク周辺の文字列を形態素単位に区切る．
2．リンク先のトピックランクが閾値以上であれば各単語を正例，閾値以下であれば負例としてカウントする．
3．単語スコア $score_i$ を計算する．

単語 i に対するスコア $score_i$ は以下の式で計算する．

$$score_i = \frac{x_i - y_i}{x_i + y_i}$$

但し，x_i は単語 i の正例カウント数，y_i は単語 i の負例カウント数である．

4 未登録語の処理

3 で説明した手法において，未登録語は類似度の計算に利用されない．しかし，未登録語の数は多く，トピックに関係する専門用語や固有名詞は，Web ページの関連性を評価する上で重要な情報である．そこで，未登録語とトピックの関連性を評価することで収集精度の向上を図る．

本稿では，検索エンジンの hit 数を基に，未登録語とトピックの関連性を評価し，類似度計算の結果に代用する方法を検討する．また，ユーザが設定したキーワードだけでなく，キーワードと関連の強い複数の語を「特徴語」として設定し，トピックに関連するページを精度高く収集することを目指す．

未登録語を用いて，シソーラスによる類似度計算の代用値を計算する手順を以下に示す．

a. 未登録語を無視しながら，1000 件程度のページを収集する．
b. a の結果から任意数の特徴語を選出する．
c. hit 数を利用して，シソーラスによる類似度の代用値を求める．

シソーラス類似度の代用値として，共起計算に用いられる相互情報量，Jaccard 係数，T 検定の 3 つを予備実験により検証した結果，相互情報量が適切であると判断した．

単語 w の出現確率を $p(w)$ とすると，相互情報量 I は以下で定義できる．

$$I = \log\left(\frac{p(w \cap w')}{p(w)p(w')}\right)$$

検索エンジンの総ページ数を N，単語 w の hit 数を $n(w)$ とすれば，

$$p(w) = \frac{n(w)}{N}$$

となる．実際に I を利用する際は類似度同様に 0～1 となるように正規化を行う．なお，本研究では検索エンジンとして Yahoo!検索 API を利用する．

5 簡易実験

未登録語の評価に用いる特徴語の決定法およびそれに基づく未登録語評価の有効性を確認するために簡易実験を行った．キーワードに“ワイン”，特徴語の選出数を 5，選出方法を，A．類似度上位，B．単語スコア上位，C．類似度×単語スコア上位，D．キーワードのみ（一単語），の 4 通りとした．

約 1000 個の未登録語を評価するまでクローリングを行い，それぞれの結果を比較した．5 つの単語から得られた評価値の上位 3 つの平均を，その未知語の評価値とする．また，日本語ということを考慮し，$N=6\times10^9$ とした．実験で用いた単語を表 1 に示す．

A，B の上位 5 件を表 2，C，D の上位 5 件を表 3 に示す．いずれの結果も，ワインの種類名，産地，葡萄の品種名が上位となった．A についてはノイズが多く見られ，また値も全体的に低かった．D でもノイズが見られたが，これは

表 1 実験に用いた単語

A	B	C	D
ビンテージ	赤	ワイン	**ワイン**
ワイン	価格	ラム	
アルコール	グラス	葡萄	
ジン	葡萄	ベリー	
冷酒	ボルドー	焼酎	

単語 1 つで相互情報量を利用したため，検索エンジンの hit 数のばらつきの影響があったためといえる．B，C についてはノイズの量も少なく，良好な結果といえる．また,キーワードを“レシピ”に設定し，同様の実験を行ったが同じ傾向が見られた．

以上より，複数の特徴語との相互情報量による未登録語の評価は有効であるといえる．

表 2 (A)類似度，(B)単語スコアの結果の上位

A		B	
デキャンタ	0.555	ビオデナミワイン	0.700
トロピカルフィズ	0.525	レゼルブピノノワール	0.657
カベルネ	0.493	マルベック	0.634
シャルドネ	0.488	ウィリアムフェーブル	0.633
シュワァーツ	0.476	ビオデナミ	0.628

表 3 (C)類似度×単語スコア，(D)キーワードの結果の上位

C		D	
チェンバロンゴ	0.735	コパーウウィール	1
ブドバイゼ	0.692	フォレスタワイン	1
ヴィゴレッロ	0.690	赤ワイン	1
マルベック	0.680	知りあい	1
テンプラニーリョ	0.673	ワインセラー	1

6 おわりに

本稿では，Web マイニングのための情報源の構築を目的に，トピック主導型クローラを提案した．特に，未登録語の評価について行った簡易実験では，複数の特徴語との相互情報量による未登録語の有用性を確認できた．

今後の課題として，未登録語の利用によるクローリング精度への影響調査，検索回数の低減手法の考案などを検討している．

参考文献

[1] George Chang , Marcus J. Healey , James A. M. McHugh , Jason T. L. Wang “Mining the World Wide Web: An Information Search Approach”，共立出版
[2] 金子，安藤，”トピック主導型クローラの試作” 2007 電気関係学会四国支部連合大会 15-4
[3] 石川，張，黒川，北川，“地域ウェブ情報源の収集のためのクローリング手法の提案”，DEWS2005 4B-i5，2005.
[4] 冨山，伊東，廣川，“自己学習型トピッククローラーの開発と評価”，DEWS2006 3B-i11，2006.
[5] Yahoo!デベロッパーネットワーク http://developer.yahoo.co.jp/
[6] 日本語語彙体系 CD-ROM 版，株式会社岩波書店 1999.