WebをコーパスとしたSetExpansionの改善

伊藤　淳†　筧　捷彦‡

早稲田大学理工学術院 基幹理工学研究科 情報理工学専攻†　早稲田大学理工学術院‡

1　背景

近年，検索エンジンは Web ページのリストを提示するだけではなくなってきた．例えば，Google[1]では，住所がクエリとして入力されると，Google Maps[2]が検索結果の上位に表示される．

このようなサービスが，検索エンジンとどのように関連づけられているかは明らかにされていない．しかし，検索クエリが何を意味しているのか，どんなカテゴリに含まれるのかといった情報をあらかじめ辞書として保有しておき，検索があったときに辞書参照を行うことで実現しているのではないかと推測できる．この辞書作成には大変な手間がかかるため，人の手間をかけずに，精度よく，カテゴリに含まれるキーワードを抽出できる手法が重要となる．そのような手法の中で，我々は，新語が現れやすく，入手が容易な，Web ページをコーパスとする手法に着目し，抽出精度の向上を試みた．

2　関連研究

Cafarella ら[1,2]の KnowItNow, Ghahramani ら[3]の Bayesian Sets, Richard ら[4,5]の Set Expander for Any Language (SEAL) が関連研究としてあげられる．また，Google が行っている実験的なサービスである，Google Sets[6]も関連研究としてあげられる．

2.1 SEALの概要

我々の手法は，SEAL に基づいているので，SEAL の概要を簡単に述べる．SEAL は次の 6 ステップでキーワード群の抽出を行う．

1. あるカテゴリからシードとしてキーワードをいくつか選択する
2. シードをクエリとして Google 検索を行い，トップ N ドキュメントを取得する
3. シードの左右に出現する文字列のうち，すべてのシード由来で，かつ最長となるものの組をパターンとして抽出する
4. パターンをもとにキーワードの抽出を行う
5. 3, 4 をすべてのドキュメントで行う
6. 4 で得られたキーワードを GraphWalk アルゴリズムによってランキングする

3　Synonome アルゴリズム

SEAL では，パターンを抽出するためにドキュメントの始めからシードの左までと，シードの右からドキュメントの終わりまでの文字列でトライ木を生成する．このトライ木のうち，実際にパターンとして抽出される部分はごく一部である．表 1 で示される我々の予備実験によると，0.03〜0.05%しか利用されていないという結果が出ている．これはメモリ効率の面で無駄がある．

また，GraphWalk アルゴリズムは抽出数によらず一定の試行回数でランキングが終了するという利点がある一方で，実行のたびにランキング結果が変わるという問題点がある．さらに，ランキング結果を解析した結果，多くの場合，被リンク数によるランキングでも同様の結果が得られるということが分かってきた．

そこで，我々は，パターン抽出とランキング手法に改善を加えた，Synonome アルゴリズムを考案した．

3.1 パターン抽出における改良

トライ木を生成する代わりに，次のようなステップでパターン抽出を行った．

1. シードの右の文字でグループ分けする
2. 取得した文字がすべてのシード由来となっているグループのみを残す
3. ひとつ右の文字を取得する．もし，すべてのシード由来でなければ，それまでの文字列をパターン右文字列として確定し，共通でなかったものをグループから除外する
4. 2, 3 を繰り返す
5. パターン右文字列とシードの組み合わせが出現する場所において，シードの左の文字

Improvement of Set Expansion using the Web
† Jun Ito, Department of Computer Science and Engineering, Fundamental Science of Engineering, Waseda University
‡ Katsuhiko Kakehi, Faculty of Science and Engineering, Waseda University

[1] http://www.google.co.jp
[2] http://maps.google.co.jp

でグループ分けする
6. パターン右文字列抽出のときと同様の操作を行う．パターン左右文字列が揃ったら，その組をパターンとして抽出する

3.2 ランキングにおける改良

キーワードの被リンク数と，パターンが持つリンク数に着目し，我々は次のようなステップでランキングを行った．

1. あるキーワードを抽出したパターンが持つリンク数の平均値が小さい順に並べる
2. あるキーワードを抽出したパターンが多い順にならべる
3. トップ R を結果として提示する

なお，このランキングは各ドキュメントについて行い，さらに統合段階でも行う．統合段階においては，パターンは文字列ではなく，どのドキュメントから返されたランキングであるかになる．我々は R の値として100を用いた．

4 評価実験

4.1 実験設定

我々は，アメリカの州名，歴代大統領名，都道府県名，歴代首相名の4カテゴリにおいて，SEAL，Boo!Wa![5]，Google Sets[6]との比較実験を行った．比較にはMean Average Precision (MAP) という指標を用いた．MAPはAverage Precision (AP) の平均値である．APは次のような式で定義される．

$$AP(L) = \frac{\sum_{r=1}^{|L|} Prec(r) * NewEntity(r)}{\#TrueEntities}$$

L はランキング結果の集合であり，$Prec(r)$ はランキング r 位の適合率である．$NewEntity(r)$ は r 以降にそのキーワードを含まず，かつ正解データだった場合に1を，それ以外は0を返す関数である．正解データは論文で用いられているものに最新のキーワードを反映したものを使用した．実験では，カテゴリからランダムに3シードを選択してYahoo!ウェブ検索API[3]のクエリとし，検索結果の最大200URLをコーパスとして用いた．これを5回実行してMAPを計算した．また，L の大きさは100とし，GraphWalkは論文と同様の条件で行った．なお，実験は2010年1月5日に行った．

4.2 結果と考察

表2に実験結果を示す．これを見ると分かるとおり，すべてのカテゴリにおいてSEALよりSynonomeの精度が高い．しかしながら，SEALのWebアプリケーション実装であるBoo!Wa!には，都道府県名カテゴリ以外で及ばなかった．

この理由として，使用している検索エンジンの違いがあげられる．論文ではGoogle検索を用いていると記されているが，我々はYahoo!検索を用いた．現在利用できるGoogle AJAX Search API[4]では最大64件までしかURLを取得できないためである．また，Boo!Wa!では論文に記されていないフィルタリングが行われているとも考えられる．HTMLタグを含むような，不適切な結果が返されないことから予想できる．

表1. トライ木における文字列長

	左文字列	右文字列	パターン左文字列	パターン右文字列
最大長	416737	1286025	126	126
平均長	37363.5	42671.4	9.7	19.5

表2. 実験結果

	SEAL	Synonome	Boo!Wa!	Google Sets
アメリカの州名	0.9225	0.9484	1.0000	1.0000
歴代大統領名	0.6977	0.9556	0.9634	0.9982
都道府県名	0.9972	0.9993	0.9962	1.0000
歴代首相名	0.9499	0.9517	0.9959	0.1923

5 まとめ

本研究では，SEALにおけるパターン抽出とランキング手法を改良したSynonomeアルゴリズムを提案した．我々の実験の結果，SEALよりもすべてのカテゴリで抽出精度向上を達成できた．しかし，SEALのWebアプリケーション実装であるBoo!Wa!には一部のカテゴリ以外で及ばなかった．今後は，抽出されたキーワードのフィルタリングや，実行速度の向上を目指す．

参考文献

[1] M. J. Cafarella, D. Downey, S. Soderland, and O.Etzioni, "KnowItNow: Fast, Scalable Information Extraction from the Web", in EMNLP, 2005.
[2] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised Named-Entity Extraction from the Web: An Experimental Study", Artificial Intelligence, vol. 165, pp. 91-134, 2005.
[3] Z. Ghahramani and K. A. Heller, "Bayesian Sets", in Advances in Neural Information Processing Systems, 2005.
[4] Richard C. Wang and William W. Cohen: "Language-Independent Set Expansion of Named Entities using the Web", in Proceedings of IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA. 2007.
[5] Boo!Wa!, http://boowa.com/, 2010.
[6] Google Sets, http://labs.google.com/sets, 2010.

[3] http://developer.yahoo.co.jp/webapi/search/

[4] http://code.google.com/intl/ja/apis/ajaxsearch/