LF-007

検索エンジンを用いた特異なウェブページの分類
Classifying Unique Web Pages Using a Search Engine

広瀬 雅之 Masayuki Hirose
鈴木 英之進 Einoshin Suzuki

横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース
Department of Electrical and Computer Engineering, Yokohama National University

1. はじめに

インターネットが急速に普及した現在，利用者はウェブページやウェブ検索エンジンを用いて，多様かつ大量の情報を容易に入手することができる．今後10年の間には，科学と経済に関するさらに多くの情報が，ウェブで広く無償で公開され，利用できるようになるという予測もある [Lawrence 99]．それ以外の分野についてもより多くの情報が利用できるようになることは想像に難くない．それら増大するウェブデータに対し，計算機を用いて，利用者に有用な情報を取り出す技術が必要となる．

ウェブデータに対して，データマイニング技術を用いて情報を発見，抽出することをウェブマイニングと呼ぶ [Kosala 02]．ウェブマイニングは対象とする情報，使用するリソースに応じて3種類に分類される．ウェブページの文書コンテンツから情報を取り出すウェブ内容マイニング (web content mining)，ページ間のリンク関係から情報を取り出すウェブ構造マイニング (web structure mining)，そして検索エンジンのログなど Web 利用ログから情報を取り出すウェブ利用マイニング (web usage mining) である [Kosala 02]．

データマイニングのアプローチは主としては，データ中の特異データを取り除くことによって，必要な情報を得る．しかし，特異なものにこそ，価値ある情報が埋もれている場合がある [Knorr 97]．たとえばクレジットカードの履歴中，急に支払いの傾向が変化する場合には盗難などの犯罪が予測される．このように問題によっては，特異な情報を発見することで重要な情報が得られると期待できる．

ウェブマイニングにおいても日々増加していくウェブページ中，内容が特異なウェブページを発見することで重要な情報が得られることが期待される．ありきたりと思われる項目についても，希少な事例が発見できれば，それを元に，ビジネスなどで利益を得ることが期待できる．他にも，ユニークな技術や知識をいち早く分類，発見することは，技術，商業的な観点から見た場合などに価値があることが期待される．

そこで，本論文はあるトピックについて書かれているウェブページが同トピックを持つウェブページ中，特異かどうかのクラス分類を目標とする．ここでの特異というのは，他の一般的なウェブページと比較して、テキストの内容が珍しい、特異であるということである．その際，特異性を判断するための手段として一般に利用されている検索エンジンを用いる．

横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース 鈴木研究室，〒240-8501 横浜市保土ヶ谷区常盤台 79-5，Tel: 045-339-4135，E-mail: mayusaki@slab.dnj.ynu.ac.jp

2. 単純手法

本論文は，ウェブページが特異かの判断を“あるトピック単語 w_{key} について書かれているウェブページ p は，w_{key} をトピックに持つウェブページの中で特異かどうか”という問題として考える．したがって，入力 p, w_{key} に対して出力 c を得る分類問題とする．c は特異か否かを表す2値のクラスである．

この問題に対する単純な手法には，単語を手がかりにして内容を判断する方法がある．特定のキーワード w_{key} をトピックとして持つページを収集する．収集したページのテキスト部に着目し，これを単語の集合ととらえて，その分布をもとに分類する．以下，詳細を述べる．

2.1 文書のモデル化

分類に用いるテキスト情報をモデル化するのに，ページ集合 $P = \{d_1, d_2, ..., d_D\}$ 全体から，各ページを特徴づけるのに有用と思われる単語 (内容語) w_i(総数 M) のリスト W を導出する．ウェブページテキストの分類を行う場合，日本語ならば形態素解析を行い，品詞情報をたよりに単語に分割する．英語の場合，3人称単数形や複数形などで語尾変化した語を同一視するための語末処理を行う．ここから，ストップワードと呼ばれる，一般的な語を取り除く．ストップワードは事前に一般性の高い，文書の内容を表さないと思われる語を人間がリストアップしておくもので，ここでは，SMART システム [Salton 83] に採用されているものを使用する．その結果，次式で表される W を得る．

$$W =< w_1, w_2, ..., w_M > \quad (i \neq j \text{ なら } w_i \neq w_j,)$$

この W を元に，各ページ d_j をモデル化する．ある内容語 w_i について各 d_j 中の出現回数を導出することで，リスト N_j を得る．

$$N_j =< n_{j,1}, n_{j,2}, ..., n_{j,M} >$$

この操作を集合 P 全体に施し，全てのページについて N_j を得る．

2.2 索引語の導出

内容語の内，特にページ内容を特徴づけている l 語を索引語として抽出する．索引語は $TFIDF$ 値の大きい物とする．$TFIDF$ による索引語導出は，[Billsus 99, Dumais 98] 等のウェブ内容マイニングでも利用されている．$TFIDF$ 値とは，内容語のページ中の出現頻度 TF とページ集合全体における，その語を含むページ数の割

合である文書頻度の逆数 IDF を掛け合わせて導出される値である．

以上の手順を式で表す．ページ d_j における内容語 w_i の頻度 $TF(d_j, w_i)$ は次式で表される．Num_{d_j} はページ d_j 中に出現する総内容語数で，$Num_{d_j} = \sum_{i=1}^{M} n_{j,i}$ とする．

$$TF(d_j, w_i) = \frac{n_{j,i}}{Num_{d_j}}$$

次にページ集合中 w_i が出現するページ数の割合 $DF(w_i)$ と，その逆数である $IDF(w_i)$ は次式のように表される．D はページ集合の全ページ総数，D_i は w_i を含むページの総数である．

$$DF(w_i) = \frac{D_i}{D}$$

$$IDF(w_i) = \frac{1}{DF(w_i)} = \frac{D}{D_i}$$

この $TF(d_j, w_i)$ と $IDF(w_i)$ の積より，各内容語の重みを導出する．このとき，各語の値の差分を調節するため，対数で正規化する．

$$TFIDF(d_j, w_i) = TF(d_j, w_i)\{1 + \log(IDF(w_i))\}$$

ページ中の全ての内容語について $TFIDF$ 値を導出する．各語の値の大きさでの上位 l 語を特に索引語集合 $I_j = \{w_{j,1}, w_{j,2}, \cdots, w_{j,l}\}$ とする．この I_j を特異ページ分類に使用する．

特異分類には索引語の DF 値と $TFIDF$ 値を利用する．各内容語の DF は，文書頻度であり，その語が出現するページ数が少ないほど小さい値となる．したがって，DF の小さい内容語は文書集合においては特異な単語であるといえ，値の小ささが特異さの尺度と考える．索引語集合に特異な単語を多く含み，また，各語 DF の値が小さい場合，特異なページであるとする．具体的には，閾値 b を設定し，索引語集合から次式で導出される値が b より小さい場合に，ページ d_j は特異であるとする．

$$\sum_{k=1}^{l} \left(DF(w_{j,k}) \frac{TFIDF(w_{j,k})}{\sum_{g=1}^{l} TFIDF(w_{j,g})} \right) < b$$

各索引語の度数をその語の重みと見なし，各語の特異さの平均値を導出し各ページの特異さを表すことにする．

2.3 問題点

この手法の問題点としては，結果が収集したページの内容に著しく依存する事があげられる．収集したページに，本来トピックとの共起性の低い語が含まれるページがたまたま多く含まれる場合やその逆の場合などが主に問題となる．このとき，共起性をページ集合だけを見て導出するため，精度が悪くなる．これは，限定された数のページ集合をもとに $TFIDF$ を導出するため，起こりうる問題である．ウェブ全体における分布を反映させるためにはページ集合の拡張，検定交差なども考えられるが，そのために大量のページ収集を行わねばならないため，効率が悪くなってしまう．

その他に，あらかじめ指定した語でだけ，一般性の強い語を除いているため，ウェブに対応した語が指定されていない場合，結果に誤差が生じる可能性がある．たとえば “home,” “link” などは通常のテキスト分類ではストップワードにはならない語であるが，ウェブ上では出現頻度がきわめて高い語である．

このような問題に対処することで，分類精度を向上させることを考えた．その手法を次節で提案する．

3. 提案手法

前章の問題に対処する手法として，ウェブページを広く網羅している検索エンジンを使用して，より大きいウェブデータに基づく分布の導出を提案する．そして，文書頻度 DF に代わり，ウェブ上のページの分布に基づく文書頻度 GF とその導出法，それを利用した特異ページの分類手法を提案する．

3.1 検索エンジンを用いた文書頻度導出

ウェブから情報を得る方法のひとつに，検索エンジンがある．検索エンジンは検索語を入力すると，検索語を含むページのリストを出力する．つまり，検索語に基づいた文書数と，文書リストを得ることができる．検索エンジンは最新に近い，大容量のページ集合を保持しているため，収集の手間をかけずに実用的なウェブ上の単語分布を再現できると考えられる．

検索エンジンの中で今回は広く一般に利用されている Google [Google] を利用する．Google は，検索語を入力すると，その語を含むページの件数と，URL の一覧を出力する．Google で，ページのトピックであるキーワード w_{key} を検索語として入力した場合に出力される件数はすなわち，ウェブ全体にページ集合を拡張した場合の，ページ集合中のページ総数 D にほぼ相当すると期待される．キーワード w_{key} と，分類対象のページに含まれる各内容語 w_i をそれぞれ AND 検索して得られるページ件数がウェブ全体から得られるページ集合中の w_{key} と w_i を含むページ総数 D_i に相当する．

Θ を 1 語以上の単語の集合と定義し，$Google(\Theta)$ を Google で Θ に含まれる全ての単語を AND 検索した際の出力件数と定義する．キーワード w_{key} に関するページ d_j に出現する各内容語 w_i の Google 頻度 $GF(w_i, w_{key})$ を次式のように定める．

$$GF(w_i, w_{key}) = \frac{Google(\{w_{key}, w_i\})}{\min(Google(\{w_{key}\}), Google(\{w_i\}))}$$

この Google 頻度 GF を，文書頻度 DF の代わりに使用する．分母に min を取るのは，内容語 w_i がキーワード w_{key} に用いられる専門用語などで，共起性が高い場合を考慮するためである．GF の逆数 IGF を IDF の代わりに使用する．

$$IGF(w_i, w_{key}) = \frac{\min(Google(\{w_{key}\}), Google(\{w_i\}))}{Google(\{w_{key}, w_i\})}$$

この利点は，

1. ウェブからの最新に近い文書情報に基づいた，統計情報を使用している，
2. 広く一般的な収集がなされているため，文書情報に偏りがない，
3. DF 導出に必要な文書集合 P の取得の手間を省く．

ことが挙げられる．

3.2 Google を用いた索引語の最適化

ウェブ上では，検索エンジンを用いることで単語を含む文書数を得ることができるというのは上述の通りである．ある単語 w_i に対して Google での検索件数がきわめて大きく，ある閾値を超えた場合，その単語はストップワードと考える．ストップワードはその単語が広く一般に使用されているため，文書を特徴づけることがほとんど無い．言い替えれば，その語を含む文書が非常に多いという事である．この方法を使用することで，随時，ウェブ上における一般性だけを考慮したストップワードの発見と除去ができる．

本研究では，あらかじめ定めた閾値 Num_{border} に対して，各内容語 w_i の Google の検索件数 $Google(\{w_i\})$ が，

$$Google(\{w_i\}) \geq Num_{border}$$

を満たす時，内容語 w_i はストップワードであると判断した．

3.3 提案手法

特異ページの判断は以下の手順で行う．キーワード w_{key} で検索し見付かったページ d_a を，日本語の場合形態素解析を行い，また英語の場合は語末処理を行って内容語集合 W と各語の度数リスト d_j を得る．W の要素から，閾値 Num_{border} を、日本語は 10,000,000、英語は 15,000,000 として判別されたストップワードを除去して得られた各 w_i について，まずは $TFIGF(w_i, d_a)$ を導出する．IGF については，各キーワードによる差を低減するために対数で正規化した値を用いる．

$$\begin{aligned} &TFIGF(w_i, d_a) \\ &= TF(w_i, d_a)\{1 + \log(IGF(w_{key}, w_i))\} \end{aligned}$$

この値の大きさから上位 l 語を，$I_a = \{w_{a,1}, w_{a,2}, \cdots, w_{a,l}\}$ として導出し，ページの索引語集合とする．索引語集合について，GF によって各索引語の特異さを表し，集合の特異さの値を次式で導出し，あらかじめ定める閾値 b と比較する．GF はその語を含むページの割合であり，値が小さいほど，w_{key} について書かれているページに使われにくい語であると考えられるためである．

$$\sum_{k=1}^{l} \left(GF(w_{a,Tk}, w_{key}) \frac{TFIGF(w_{a,Tk}, d_a)}{\sum_{g=1}^{l} TFIGF(w_{a,Tg}, d_a)} \right) < b$$

上式を満たす場合，D_a は特異であるとする．

4. 実験

4.1 実験

「車椅子+仕様」「ハムレット」というキーワードで収集した日本語ページ集合2種類と、「yokozuna」「castle」「toyota」というキーワードで収集した英語のページ集合3種類と計5種類のデータについて実験を行った．まず，特異か否かのクラス分けを手作業で行っておく．$l = 10, b = 0.1$ として単純手法と提案手法とそれぞれ特異ページの分類を行った．結果について，事前のクラスとの再現率，適合率で比較した．実験に使用したページ集合は表 1，2 の通りである．なお、"+" は AND 検索を行ったことを意味する．

表 1: 実験データ：日本語ページ，特異数は内数

キーワード	車椅子+仕様	ハムレット
ページ数	86	94
特異数	10	19

表 2: 実験データ：英語ページ，特異数は内数

キーワード	yokozuna	castle	toyota
ページ数	86	94	70
特異数	28	28	20

4.2 結果と考察

5種類のデータについての再現率および適合率を表3，4に示す．英語のページについては，再現率で最大17.9%，適合率で最大7.9%向上している．このことは，ウェブ上の情報によって，特異なウェブページの分類において，提案手法は単純手法と比較してある程度有効であることを示している．単純手法より良い結果が得られるのは，限られたページ集合だけによらない，単語の特異判定を行ったためであったと考えられる．例えば，単純手法では "yokozuna" "software" の組み合わせを共起しやすい組み合わせと見なしていたが，実際はそうではない．

英語に比べて日本語の結果は単純手法，提案手法とも適合率が低い．これは主に，日本語ページについて，単語を組み合わせた複合語がはじめて特異になる場合に対処できていないためであった．すなわち複数の語の共起性によって特異となるページに対して，単純手法，提案手法とも形態素解析器を使用して単語を分けているため，対応できていないことがあったためと考える．とくに顕著な例として，「自動意思伝達装置」を装備した車椅子のページがあり，事前のクラス分けでは，この語をもって特異とした．形態素解析では，「自動」，「意思」，「伝達」，「装置」とそれぞれ分かれてしまい，その共起性を考慮してはじめて特異と考えられる項目だったため，特異とは分類されなかった．他に，「運転補助装置」「助手席」などの語にも同様の問題が見られた．

英文では単語が区切られており，複合語の問題がほとんど起きなかった．単語の特異さとページの特異さが直接関与しているため，各単語の検索エンジンの出力を元

表 3: 実験結果：日本語

キーワード	「車椅子+仕様」		「ハムレット」	
ページ数	86		94	
特異数	10		19	
手法	提案	単純	提案	単純
抽出数	10/61	8/20	17/44	15/38
再現率	100%	80.0%	89.4%	78.9%
適合率	16.4%	40.0%	38.6%	39.4%

表 4: 実験結果：英語

キーワード	「yokozuna」		「castle」	
ページ数	86		94	
特異数	28		28	
手法	提案	単純	提案	単純
抽出数	26/52	21/44	24/52	21/55
再現率	92.9%	75.0%	85.7%	75.0%
適合率	50.0%	47.8%	46.1%	38.2%
	「toyota」			
ページ数	70			
特異数	20			
手法	提案		単純	
抽出数	19/38		17/32	
再現率	95.0%		85.0%	
適合率	50.0%		53.1%	

に特異判断を行った提案手法が効力を発揮した結果となった．しかしながら，“toyota” のページについては適合率で負けている．これは，ストップワードの除去によって低頻度ながらも希少な語が索引語となる場合が多く見られたためである．その分，抽出に成功しているページも増加しているため，再現率を向上させた結果でもある．残りの “yokozuna”,“castle” のページについては，再現率で最大 17.9%，適合率で最大 7.9 したがって，単語の分かれ目の判別が容易なウェブページの特異分類には有効な手法であると考えられる．

5. おわりに

ウェブマイニングにおいても特異な情報を発見することは有意義である．そのためのアプローチとして，ウェブページをある程度収集し，その傾向に基づいた分類によって特異なページの分類を行う方法が考えられるが，各内容語のウェブ上の共起性などを十分に考慮しておらず，その精度に問題があった．

本論文はウェブ内容マイニングのアプローチの一つとして，特異なウェブページの分類を行った．その際に，上述の問題への対処として，より大きいウェブデータを使用する手法として，検索エンジンを用いた方法を提案した．加えて，ウェブデータを用いたストップワードリストを生成する方法を提案した．

5種類のウェブページ集合を用いて単純手法と提案手法で実験を行った．その結果，単語単体で内容を判断できる英語ページに関しては再現率，適合率の結果が提案手法が勝るか同程度であることがわかった．

参考文献

[Billsus 99] D. Billsus and M. Pazzani. A Hybrid User Model for News Story Classification. In *Proceedings of the Seventh International Conference on User Modeling,* pp.99-108, 1999

[Dumais 98] S. Dumais, J. Platt, D. Heckerman and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the 1998 ACM Seventh International conference on Information and Knowledge Management (CIKM),* pp.148-155, 1998

[Eirinaiki 03] M. Eirinaiki, M. Vazirgiannis and I. Varlamis. SEWeP: Using Site Semantics and Taxonomy to Enhance the Web Personalization Process. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.99-108, 2003

[Ester 02] M. Ester, H. Kriegel and M. Schubert. Web Site Mining: A New Way to Spot Competitors, Customers and Suppliers in the World Wide Web. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.249-258, 2002

[Google] Google. http://www.google.com/

[北 02] 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

[Knorr 97] E. M. Knorr. On Digital Money and Card Technologies. *Technical Report 97-02,* University of British Columbia, Canada, 1997

[Kosala 02] R. Kosala and H. Blockeel. Web Mining Research: A Survey. In *ACM SIGKDD Exploration, Issue 2* , pp.1-15, 2000

[Lawrence 99] S. Lawrence and L. Giles. Accessibility of Information on the Web. In *Nature,* Vol.400, pp.107-109, 1999

[Salton 83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval,* McGraw-Hill, New York, 1983

[Yi 03] L. Yi, B. Liu and X. Li. Eliminating Noisy Information in Web Pages. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.296-305, 2003