リンクの共起関係を用いたURL関連度計算手法の一検討

3X−04

杉崎 正之　牧野 俊朗　稲垣 博人
日本電信電話株式会社　NTTサイバーソリューション研究所

1 はじめに

インターネットなどに代表されるコンピュータネットワークの普及により、大量のテキスト情報が不特定多数に対して公開されている。その代表格がHTMLファイルであり、分散して存在する多量のHTMLファイルの中から欲しい情報を容易に取り出せるようにするため、検索や分類、ナビゲーション等、多くの方式が研究されている。これらを効果的に行なうには、単語と文書、あるいは文書と文書の関連付けが重要である。本稿では、ハイパーリンクを用いることにより、Webページを関連付ける手法について検討する。

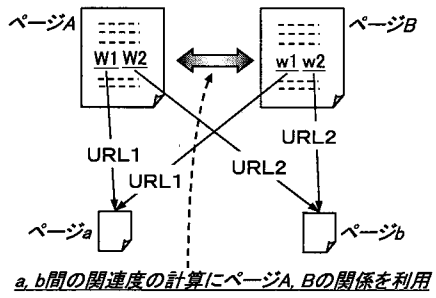

図 1: ハイパーリンクの共起の考え方

2 リンクの共起に基づく関連付け

2.1 概要

基本的な考え方は、単語間の関連度を求めるのに文内での共起関係を利用するのと同様な方法である。すなわち、同一のWebページに複数のハイパーリンクが存在するとき、それらのリンク先のWebページどうしは互いに関連しているとみなす(図1の「ページa, b」が該当)。

WWWの特性として、自分の興味に合致し、かつ何度もアクセスするページをリンク集として作成公開しているユーザが多く存在する。その場合、同じ分野へのリンクはまとまっていることが多く、さらにそれらが恒常的な情報の場合は多くの「リンク集」Webページで内容が書き換えられることがない。また、恒常的な情報でなくても企業間の提携情報などがWebページで公開される場合に、関連URLとして各企業へのリンクが同一ページ内に記述されたりする。

このような特性を生かし、同一ページ内での共起情報を利用することで、関連のあるURLが抽出できることを確認した[1]。さらに、Webページ内において、出現位置情報を利用して共起を調べる範囲を設定することにより、検索サイトのように複数のWebページに大量に存在する可能性のあるリンク先との関連度が大きくなってしまう問題(トピックドリフト問題)もある程度解消できることも分かった。

しかし、上記の手法の問題点として、もともと関連のない2つのURLを同一ページ内に距離的に近く記述したページを大量に作成しておくことで、意図的に関連度を高くすることが可能となってしまう(図1でいえば「ページA」「ページB」に当たる)。また、意図的でないにしても「問い合わせ先」や「ヘルプページ」、「トップページ」へのハイパーリンクが記述された大量のWebページを所有するサイトから単純に共起関係を抽出すると、偏った関連度が計算されてしまう。

そこで、それらを避けるために2つのURL間の関連度を定義しているWebページどうしの情報の利用を検討する。

2.2 改良手法

まず、2つのURL(u, vとする)に対して次のような関数を定義する。

$$f^s(u,v) = 1 \quad if\ u,v\text{が同一ページ上の}\ s\text{バイト以内}$$
$$= 0 \quad \text{それ以外}$$

従来は、関数$f^s(u,v)$の単純な総和とした。今回は、a, bが共起しているWebページ、すなわち$f^\infty(a,b) = 1$となるWebページの集合を$S(a,b)$として、集合$S(a,b)$の

A method of calculating the degree of relation between URLs using co-citation analysis of hyperlinks
Masayuki SUGIZAKI, Toshiro MAKINO, and Hirohito INAGAKI
NTT Cyber Solutions Laboratories, NTT Corporation

処理したWebページ数　　:1,035,753
抽出されたリンク数　　:1,016,684
関連度が確認されたURL　　:　235,148
関連度が確認されたURLの組:4,393,662

図 2: 抽出結果

要素数を $|S|$ とし、ユニークなドメイン数を $u\_dom(S)$ として、2つのURL$(a, b$ とする$)$の関連度 $rel(a, b)$ を、

$$rel(a,b) = \sum f^s(a,b) * \frac{u\_dom(S)}{|S|} \quad (1)$$

とする。上記の式は、互いに異なる多くのページからWebページから参照されている場合に値を大きくすることで、上記の問題点を解消しようとしている。

3　実験と考察

実際にWebページを収集し実験を行なった。2001年11月に収集した日本語のWebページのうちの約100万ページを利用した。同一ドメイン内の相対リンクやCGIへのリンクは削除することにし、抽出時の s の値は800とした。抽出結果を図2に示す。

関連度の高いURLの抽出結果の違いについて検討する。図3をみると、従来手法では「Yahoo! Japan (http://www.yahoo. co. jp)」と企業間で関連があるサイト (2, 3) や自社内の一サービスのURL(7) との関連度が高かった。一方、改良手法ではいわゆる「検索サイト」と呼ばれるURLが上位を占めた。また、図4をみると、従来手法では「goo (http:// www.goo.ne.jp)」を運営している会社のサイト (1) や自社内のサービスのURL(3, 4, 6) との関連度が高かった。一方、改良手法では図3の時と同様なURLが上位を占めた。

今回抽出元となったWebページ内には図3や図4の従来手法に出現したサイト内から取得したページも含まれており、運営会社へのリンクや同社が行なっている別サイトでのサービスとの共起が多くなるという問題が実際に生じている。が、改良手法を用いることで、多くの人が関連があると考えているWebページを抽出できていることが確認できた。

また、図5は「日本書籍出版協会 (http://www.books.or.jp) 」に対する関連URLを関連度の高い順に並べたものであるが、従来手法と改良手法での出力に大きな違いは生じなかった。これは、関数 $f^s(a, b) = 1$ となるWebページの集合においてそれぞれのページが別のドメイン上に存在しており、この場合は今回の改良による悪い影響が出なかった。

従来手法	改良手法
1. http://www.goo.ne.jp/	1. http://www.goo.ne.jp/
2. http://www.zdnet.co.jp/	2. http://www.infoseek.co.jp/
3. http://www.softbank.co.jp/	3. http://www.lycos.co.jp/
4. http://www.infoseek.co.jp/	4. http://www.excite.co.jp/
5. http://www.lycos.co.jp/	5. http://www.yahoo.com/
6. http://www.excite.co.jp/	6. http://www.google.com/
7. http://chat.yahoo.co.jp/	7. http://navi.ocn.ne.jp/

図 3: 「http://www.yahoo.co.jp/」と関連があるURL

従来手法	改良手法
1. http://www.nttx.co.jp/	1. http://www.yahoo.co.jp/
2. http://www.yahoo.co.jp/	2. http://www.infoseek.co.jp/
3. http://community.goo.ne.jp/	3. http://www.lycos.co.jp/
4. http://shop.goo.ne.jp/	4. http://www.excite.co.jp/
5. http://www.infoseek.co.jp/	5. http://www.nikkei.co.jp/
6. http://channel.goo.ne.jp/	6. http://japan.infoseek.com/
7. http://www.nikkei.co.jp/	7. http://www.google.com/

図 4: 「http://www.goo.ne.jp/」と関連があるURL

4　今後の課題

今回は関連度の計算に「ユニークなドメイン数」を用いたが、同一ドメインで別の個人のWebページから張られているリンク情報がうまく利用されておらず、今後「ドメイン単位」ではなく「サイト単位」での関連度の計算を検討したい。また、sの値によるURLのカバー率への影響や、HITS[2] などの再帰的な手法との比較検討を行なっていきたい。

参考文献

[1] 大久保, 杉崎, 田中: リンクの共起関係を用いたWebページ分類方式の検討情処第59回全大 (3), pp.81-82, 1999.9

[2] J.Kleinberg: Authoritative sources in a hyperlinked environment, Proc. of 9th ACM-SIAM Symposium in Discrete Algorithms, pp.668-677,1998

従来手法	改良手法
1. http://www.maruzen.co.jp/	1. http://www.maruzen.co.jp/
2. http://bookweb.kinokuniya.co.jp/	2. http://bookweb.kinokuniya.co.jp/
3. http://webcat.nacsis.ac.jp/	3. http://www.trc.co.jp/trc-japa
4. http://www.trc.co.jp/trc-japa	4. http://webcat.nacsis.ac.jp/
5. http://www.ndl.go.jp/	5. http://www.ndl.go.jp/
6. http://www.amazon.com/	6. http://www.amazon.com/
7. http://www.asahi.com/	7. http://www.asahi.com/

図 5: 「http://www.books.or.jp/」と関連があるURL