リンクの共起関係を用いたWebページ分類方式の検討

3P－2

大久保 雅且　杉崎 正之　田中 一男

NTTサイバーソリューション研究所

1. はじめに

WWW情報の激増[1]により、ほしい情報の発見や入手がますます困難になってきている。この問題に対処するために、検索や分類、ナビゲーション等、多くの方式が提案されている。これらを効果的に行うためには、単語と文書、あるいは文書と文書の関連付け方式が重要である。

関連付けは、従来、各文書に含まれる単語の種類や頻度に基づいて計算されてきた[2]。しかし、同じ概念を異なる表記で表す同義語や、逆に同じ表記でありながら異なる意味を表す多義語などにより、単語の統計的な処理による関連付けの精度は必ずしも高いわけではない。本稿では、WWWに特有のハイパーリンクを用いることにより、Webページを関連付けて分類する手法について検討する。

2. リンクの共起に基づく関連付け

2.1 基本的な考え方

自然言語処理では、単語間の関連度を求めるために共起関係を利用することがよくある。すなわち、同一の文やフレーズなど、局所的に同時に頻出する単語どうしは関連が強いとみなす。この考え方をハイパーリンクに適用する。すなわち、同一のWebページに複数のハイパーリンクがあるとき、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1参照）。

WWWでは、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いる。また、同じ趣味を持ったユーザ（Webページ）どうしで、お互いにリンクをはること（相互リンク）もよくある。このため、同じ分野へのページへのリンクは、1ページにまとまっていることが多いので、同一のページからはられているリンク先のページどうしは、強い関連があると予想される。

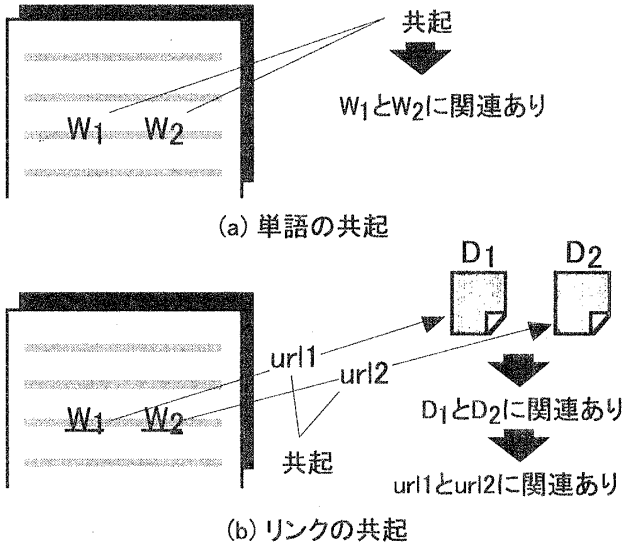

図1　ハイパーリンクの共起の考え方

また、ハイパーリンクによって関連付けられたWebページは、そのページの作成者が「リンクすべき」と判断した結果としての参照先であるため、より高い質のページであることが多い。特に、多くのユーザから参照されているページは、その分野における定番ページともいえる。

これらのことから、リンクの共起に基づく関連付けは、単語の表記に依存しないだけでなく、同じ分野の良質なWebページが関連付けされると期待できる。

2.2 アルゴリズム

上記で明らかなように、2つのハイパーリンクuとvの関連度は、同一のページに出現する頻度や、ページ内での位置関係に依存する。そこで、uとvのページ上での位置やレイアウトに応じて値を決定する関数を、f(u,v)とし、すべてのページについて集計したΣf(u,v)を、uとvの関連度とする。

3. 実験

以下に示す2種類のf に関して、日本語のWebページ約10万ページを対象として関連度の計算を行った。

(a) f (u, v) = 1 if u,vが同一ページ上にある
= 0 otherwise

Web page clustering based on cooccurrence of hyper links

Masaaki OHKUBO, Masayuki SUGIZAKI, and Kazuo TANAKA

{ohkubo, sugizaki, tanaka}@aether.hil.ntt.co.jp

NTT Cyber Solutions Laboratories

表1 抽出されたURLの数と関連度>1のURLの組の数

	(a)	(b) s=200
関連があるURLの数	243, 436 (98. 1%)	243, 084 (97. 9%)
関連度>1のURLの組の数	83, 408, 137	56, 851, 767

(b) f (u, v) = 1 if u,vが同一ページ上のsバイト以内
= 0 otherwise

表1に実験結果を示す。(b)はs = 200とした。抽出されたURL数は248,194であった。「関連があるURLの数」とは、抽出されたURL(248,194)の中で、他の少なくとも1つのURLと関連があったURLの数である。

4. 考察

表1より、(a)(b)いずれの場合でも、抽出されたURLの約98%が、他のURLと関連付けられたことがわかる。このうち、他のURLと関連を持つ数の多いURLを図2に示す。これらは、多くのWebページに現れるURLであることを意味しており、検索やダウンロードなどのサービスを提供しているWebページが多い。また、お互いに共起することも多く、ポータルサイトなどを同一ページから参照しているユーザが多いことがわかる(図3参照)。

次に、関連URLの違いについて検討する。例えば、日本書籍出版協会(http://www.books.or.jp)に対する関連URLを、関連度の高い順に図4に示す。(a)では、検索(a,1)(a,2)やダウンロードサービス(a,4)のサービスを提供してるWebページや、新聞社(a,3)(a,5)など、多くのWebページに出現するURLとの関連度が高くなっていた。すなわち、「同一ページ内に共起する」ことが多いURLどうしでも、実際に内容的な関連があるとは限らないことが分かる。一方、(b)は図書総合目録データベースの検索サイト(b,1)や書籍検索サイト(b,2)、仮想書店のサイト(b,3)(b,4)や国立国会図書館(b,5)などのように、書籍に関するURLとの関連度が高い。さらに、表1を見ると、(a)(b)で関連のあるリンクの数はあまり差がないが、共起を調べる範囲を制限することで組数は70%にまで減少しており、内容的な関連の低いURLの組が削除されていると考えられる。以上のことから、文書における単語の共起と同様に、共起を調べる範囲を設定することによって、より効率的かつ効果的にWebページの関連度を定義できることが検証できた。

(a) 1. http://kids.yahoo.co.jp
2. http://www.yahoo.co.jp
3. http://www.goo.ne.jp
4. http://www.forest.impress.co.jp
5. http://www.vector.co.jp

(b) 1. http://www.yahoo.co.jp
2. http://kids.yahoo.co.jp
3. http://www.vector.co.jp
4. http://www.infoseek.co.jp
5. http://www.goo.ne.jp

図2 他のURLとの共起が多いURLのランキング

(a) 1. http://www.infoseek.co.jp/Titles
2. http://www.goo.ne.jp/default.asp
3. http://www.yahoo.co.jp
4. http://www.goo.ne.jp/ie4.0/msResult.asp
5. http://www.vector.co.jp

(b) 1. http://www.yahoo.co.jp
2. http://www.goo.ne.jp/default.asp
3. http://japan.infoseek.com
4. http://navi.ntt.co.jp
5. http://www.infoseek.co.jp

図3 「http://www.goo.ne.jp」と関連があるURL

(a) 1. http://www.goo.ne.jp
2. http://www.yahoo.co.jp
3. http://www.asahi.com
4. http://www.forest.impress.co.jp
5. http://www.mainichi.co.jp

(b) 1. http://webcat.nacsis.ac.jp
2. http://www.trc.co.jp/trc-japa
3. http://bookweb.kinokuniya.co.jp
4. http://www.maruzen.co.jp/index-j.html
5. http://www.ndl.go.jp

図4 「http://www.books.or.jp」と関連があるURL

Webページの自動分類は、URL間の関連度が定義できればクラスター分析手法などを用いることにより可能である。今回の実験で、リンクの共起情報を用いることにより、精度のよいWebページ分類を行える見通しを得た。

5. おわりに

同一Webページ内に存在するパイパーリンクの共起情報を用いて、URL間の関連度を定義する方法について述べた。今後は、レイアウトなどの影響を考慮してより精度の高い関数f の決め方を検討していきたい。

参考文献

[1] S.Lawrence and C.L.Giles, "Searching the World Wide Web", Science, Vol.280, No.5360, p.98, 1998

[2] G. Salton, "Automaic Text Processing", Addison Wesley, Reading, Mass, 1989.