社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－94 (2)
2009－ＤＤ－70 (2)
2009／3／25

Webリンク活用のためのアンカーテキストの自動分類の研究

大塚 博紀† 吉岡 真治‡

‡北海道大学大学院情報科学研究科

Web 文書の特徴はリ ンクによりお互いの文書の関係が示されている点にあり、この情報はPageRank に代表される Web 空間の解析手法に用いられている。しかし、これらの解析では、サイト内、サイト外といった分類は考慮されているものの、ページの中でのアンカーテキストの役割には、注目していない。本研究では、このアンカーテキストの役割の違いに注目し、適切に分類することで、リンク構造解析や Web ページ上でのユーザーの行動解析に役立てる事を目標としている。本稿では、人手による役割分析の結果 に基づいたアンカーテキストの自動分類システムを提案する。また、その有効性を検証するために、実際に人手 の分類とシステムの出力を比較した結果について報告する。

Research on the Automatic Classification of Anchor Texts for utilizing Web links

Hiroki Ohtsuka† Masaharu Yoshioka†

†Graduate School of Information Science and Technology, Hokkaido University

A Link Structure among Web documents represents the relationship among these documents. Web structure analysis methods such as PageRank use this information. However, most of these methods pay little attention to the type of link (e.g., most of the methods use simple classification such as link to the same site or not). In this research, we proposed a method to classification criteria of anchor texts based on the role analysis of the text in Web documents for better link structure analysis and the users' behavior analysis on the Web page. In this paper, we propose an automatic classification method based on the result of manually classification experiment. In order to evaluate this method, we compare the result of automatic classification results with manually classified ones.

1.はじめに

本研究では、アンカーテキストを役割に応じて分類し、リンク構造解析や Web ページ上でのユーザーの行動解析に役立てる事を目的としている[2]。我々はこれまでの研究で、実際にどのような役割のアンカーテキストが存在するのか、またどのような分類定義が妥当であるかを調査し、分類基準を定義した。しかし、人手による分類実験を行ったところ、分類基準の修正が必要なことが分かった[情報処理学会]。本稿では、新たに修正した分類基準による分類実験の結果と、その結果に基づいて構築したアンカーテキストの自動分類システムについて報告する。

2.Web 情報活用のためのアンカーテキストの分類と利用の研究

本研究では、アンカーテキストの性質の違いに注目して、サイト内、サイト外からのリンクの区別だけではなく、さらなる分類の詳細化を行い、以下の 8 つの分類を提案している。[1]

1. リンク先の内容を表すテキスト (内容)：「Yahoo!JAPAN」などの、リンク先の名前を示すテキスト
2. ページの機能を表すテキスト (機能)：リンク元のページの内容を前提としたとき、アンカーテキストからリンク先の内容を把握できるテキスト
3. リンク先との関係を表すテキスト (関係)：「発行者 Web サイト」などのリンク先とリンク元のページの関係を示すテキスト
4. トップページを指示するテキスト (トップ)：「HOME」、「ホーム」、「 TOP」などのトップページを示すテキスト

[1]以下の本文では、()内の表記で参照する。

5. ナビゲーションを指示するテキスト　(指示)：「戻る」、「次へ」、「こちら」などのリンク先のページと関係なく用いられるテキスト
6. インデックスを表すテキスト (インデックス)：「1」、「2」、「3」、「あ行」、「○」などの、幾つかの関係するページをまとめるためのテキスト
7. URL(URL)：URL をそのまま利用しているテキスト
8. その他 (その他)：アダルト サイトなどが「18 歳未満」を Yahoo にリンクするような、リンク先のページとまったく関係ないテキスト

この分類の有効性を検証するために、 NTCIR の Web テストコレクション nw100g[3]を利用し、アンカーテキストの抽出と、利用頻度の高いアンカーテキストについて分析を行った。その結果、これらのアンカーテキストが上記の 8 つの分類により分類可能である事を確認した。ただし、一つのアンカーテキストが複数の分類に属する場合がある(例：「Yahoo!JAPAN のホームページ」　→1,4)。

3. アンカーテキストの分類実験

3.1. 分類基準の修正と人手による分類実験

我々は、既に、前節で述べた分類基準の一貫性を調査するために、実際のWebページのデータであるnw100gから抽出したアンカーテキストを用いた分類実験を行っている[4]。しかし、前回の実験では、一部の分類基準に曖昧性が存在したために、一部の分類が一貫して行われなかった。特に、問題となったのが、分類1,2,3,6であり、これらを考慮して、以下のような形で分類基準の修正を行った。

■分類の重複について
分類4,5のみ、分類1,2,3,6,7と重複してよいが、それ以外の分類は重複しない。
■分類基準の詳細化
- 分類1(内容)
  リンク元のページの内容に関わらず、リンク先の内容が判断できるアンカーテキスト。
- 分類2(機能)
  リンク元のページ(サイト) の内容を前提としてそのアンカーテキストに注目すると、リンク先の内容(意味) が分かるアンカーテキスト。
- 分類3(関係)
  リンク元のサイトのコンテンツの中身に関係なく、リンク元との関係のみを把握できるアンカーテキスト。
- 分類6(インデックス)
  リンク先のページとして意味を成さないものであり、あるインスタンスの集合体を表すものである。リンク先に多くのリンクが存在するようなページへのアンカーテキスト

また、前回の実験では、被験者が二人であったために、分類基準の説明が不足したことに起因する個人的な勘違いが起きた場合と、本当に判断に迷うような分類の差が分かりにくいという問題があった。そこで、今回は、被験者を3名とすることで、各個人ごとの分類結果の傾向を比較することにより、その影響についても分析を行った。

今回の分析対象としたものは、前回の実験と同様に、nw100gから抽出したアンカーテキスト17000件で、前回の実験の考察を受け、複数のサイトからテキストを集めるように注意した。この分類実験における分類結果の一致度を表1に示す。

表1：人手による分類作業結果

分類	3名が一致	2名が一致	1名のみ
1	2072	3225	3238
2	1259	4255	4682
3	53	39	86
4	464	66	0
5	1189	112	130
6	1817	1297	3014
7	78	0	532
8	7	74	1242

まず、最初に被験者の違いによる影響について考える。今回のデータで、3名が一致したものが6939件、2名が一致したものが9068件であった。特に、2名が一致という場合を分析すると、特定の1名のみの判断がずれるということがほとんどであり、その1名の判断を除いた2名の一致した件数は11770件(69.24%) であった。

次に、分類項目ごとの一致度について考える。分類3(関係)、分類4(トップ)、分類5(指示)、分類7(URL)については、三者とも一致する確率が非常に高く、一貫性が見られた。また、一貫性が高いと考えられる3名が一致したデータを

分析したところ、分類2から6に関しては、これらの分類において頻度の大きいキーワードが存在することが確認できた。これらのキーワードの多くは、その分類に特徴的に存在するキーワードであり、アンカーテキストの分類に役立つと考えられる（表2）。

表2：分類2から6の高頻度語

分類2(機能)	**地図**	65
	ヘルプ	62
	Help	60
	お問い合わせ	58
	what's new	48
分類 3(関係)	pdf	19
	English	14
	Japanese	7
	テキスト版	7
	text-only version	6
分類 4 (トップ)	Home	129
	ホーム	127
	トップ	105
	TOP	103
分類 5(指示)	**前**	307
	次	280
	こちら	144
	next	114
	prev	101
分類 6 (インデックス)	2	46
	3	45
	1	41
	[あ]	25
	Page001	17

一方で一貫性が低いアンカーテキストを分析すると、分類の際にその単語の意味やリンク先のテキストを理解するための背景知識が必要であるものがあった。このような分類に関しては、判断する際に個人の知識に依存するため、一貫性を確保するのが困難な場合がある。

3.2. 各分類の特徴分析

前節で行った分類実験結果のうち、比較的判断のゆれが少ないと考えられる三者の判断一致したアンカーテキストに対して、各分類の特徴を次の 3 つの観点から分析した。

1. 分類毎に特徴的なキーワード

表 2 に示したように、分類 2 から 6 には、各々の分類に特徴的なキーワードが存在するので、このキーワードリストを利用した分類を行う。ただし、分類 3(関係) 、分類 4(トップ)、分類 5(指示)については、大半のアンカーテキストが、このキーワードリストに含まれる単語を含む。一方、分類 2(機能)、分類 6(インデックス)については、アンカーテキスト中にこれらのキーワードリストに含まれる単語を持たないものも存在する。

2. アンカーテキストとリンク先文書の比較

分類 1(内容),分類 2(機能)については、多くの場合アンカーテキストがリンク先のタイトルや文書中に含まれることが確認された。ただし、アンカーテキストで書かれている表記とは異なる類義語などが用いられているケースも散見された。

3. リンク先、リンク元の URL の特徴

I. サイト内リンク・サイト外リンク

アンカーテキストが記述されているページの URL とアンカーテキストのリンク先の URL を比較することにより、サイト内リンクとサイト外リンクを判別した。各分類におけるリンクの違いを表 3 に示す。

表3：分類ごとのリンク先の違い

分類	サイト内	サイト外
1	545	1527
2	1106	153
3	53	0
4	448	16
5	1158	31
6	1810	7
7	5	73
8	0	7

II. リンクのディレクトリ構造の類似性

分類 3(関係)においては、 単に、リンク先の URL がサイト内であるだけでなく、 リンク先の文書の URL のディレクトリ部分が、リンク元の文書の URL のディレクトリ部分に一致していた。これは、異なるバージョンのテキストをまとめて管理しているからであると考えられる。

III. アンカーテキストとリンク先 URL

分類 7(URL)は、すべてのアンカーテキストがリンク先 URL と一致する。

4. アンカーテキストの自動分類実験

4.1, 自動分類規則の作成

前節で述べた各分類の特徴を考慮して、アンカーテキストの自動分類を行うための規則を作成した。各々の分類規則では、前節の分析を踏まえ、次の 3 つの基準を組み合わせて、判定を行うこととした。

●キーワードリストとの比較

分類ごとに高頻度語からキーワードリストを作成する。アンカーテキストに対して、形態素解析を行った結果、得られた単語リストにキーワードリストの語を含む場合には、各々の分類と判定する。

●リンク先文書との比較

アンカーテキストとリンク先の文書を比較する。この場合も、キーワードリストの場合と同様に形態素解析を行った結果で比較をする。

●URL 情報の利用

分類 7 では、リンク先の URL とアンカーテキストを直接比較する。分類 3 では、リンク先の URL とリンク元の URL のディレクトリ部分を比較し、同一かどうかを判定する。

表 4.分類毎の特徴の利用

分類	キーワードリストとの比較	リンク先文書との比較	URL
1		○	
2	○	○	
3	○		○
4	○		
5	○		
6	○		
7			○
8			

複数の特徴を利用する分類 2, 3 については、キーワードリストとの比較を先に行った上で、リンク先の文書との比較もしくは URL 情報の利用を行う。

次に、分類手順について述べる。分類 4(トップ), 分類 5(指示)は、他の分類と重複する可能性があり、判断基準は他の分類には関係なく、独立している。よって、まず分類 4(トップ), 分類 5(指示)について分類の判断を行う。

次に、排他的分類である分類 1 から 3、6 から 8 について考える。次に、キーワードリストとの比較が可能である分類 2(機能),3(関係),6(インデックス)ならびに、URL との比較により判断が可能な分類 7(URL)について分類を行う。さらに、残ったアンカーテキストに対し、分類 1 かどうかを判断する。最後に、残ったテキストのうち、分類 4,5 と判断されていないテキストを分類 8(その他)と判断することとした(図 1)。

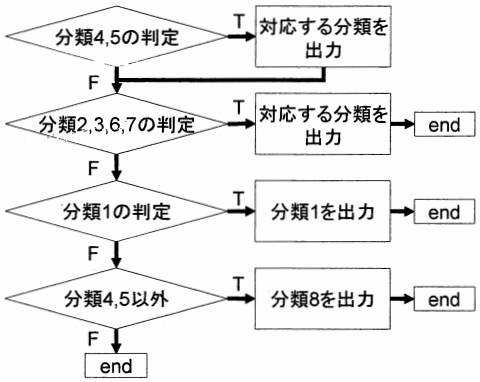

図 1：自動分類の流れ

4.2. 自動分類実験

上記ルールの妥当性を調査するために、排他的に分類を行うことができるアンカーテキストに対して、自動分類実験を行った。

実験データは事前実験により得られた人手による分類において、三者とも一致したアンカーテキスト(6939 件)を用いた。

本自動分類で用いる分類ごとのキーワードリストであるが、分類 4(トップ)に関しては、「トップ」、「top」、「HOME」、「ホーム」の 4 種類の語を必ず含むので、キーワードはこれらの 4 つの語に設定した。

また、分類 2,3,5,6 に対するキーワードリストについては、次の手順で作成した。

1. 分類ごとのアンカーテキストの収集

4 節で述べた人手による分類実験の分類情報を用いて、アンカーテキストを収集する。

2. 形態素解析による高頻度語の抽出

集めたアンカーテキストに対し、形態素解析を行い、高頻度語を各分類のキーワードリストとして抽出する。全体の件数が少ない分類 3 以外では、出現頻度が 10 件以上のものを利用する。分類 3 については、表 2 に示す 5 種類のテキストしか存在しなかったため、これをキーワードリストとした。

上記のようにして作成したキーワードリストを用いて、自動分類を行った結果を表 5 に示す。

表 5.自動分類結果

分類	正解	出力	一致	再現率	精度
1	2072	410	215	10.38	52.4
2	1259	326	290	23.03	89
3	53	53	53	100	100
4	464	464	464	100	100
5	1189	1551	1071	90.08	69.1
6	1817	910	760	41.83	83.5
7	78	78	78	100	100
8	7	3051	7	100	0.23

4.3. 実験結果の分析

分類 3(関係), 分類 4(トップ), 分類 7(URL)は再現率、精度は共に 100%となった。一方で、分類 1(内容),2(機能),6(インデックス)の再現率がかなり低く、結果として、分類 8(その他)に分類されるデータが多くなってしまった。

このうち、リンク先文書との比較を行っている分類 1(内容)、分類 2(機能)については、キーワードの表記ぶれや類義語の問題であり、これらへの対応が必要であると考えている。

また、キーワードリストを用いている分類 2(機能)、分類 5(指示)、分類 6(インデックス)については、キーワードリストの網羅性の問題がある。現時点では対応する分類のアンカーテキスト中で高頻度の単語からキーワードリストを作成しているが、少ない回数しか出ていないが、各々の分類に対応するキーワードが存在するため、網羅性に欠ける。ただし、キーワードリストの網羅性を上げるために、単純に分類に属するアンカーテキストに存在するキーワードを利用すると、再現率の向上が期待できる半面、精度の低下が心配される。

次に、精度に関しては、分類 1(内容),5(指示)において問題がある。

また、分類 5(指示)の精度が低い理由としては、収集したキーワードリストに「 >」や「<」といった記号が含まれており、アンカーテキストがそれらの記号と他の記号の組み合わせである場合、インデックスに属する分類と混同していること(例えば、<1>、<2>...)が考えられる。

次に、分類 1 の精度が低い問題についてであるが、これは、分類 1 の問題というよりも、分類 2,6 のキーワードリストの網羅性の低さに伴う再現率の低さに起因する問題であり、先に指摘した分類 2, 6 の再現率が向上することによって改善されることが期待される。

4.4. システムの改良指針について

今回の分類実験では、3.2 節で説明したアンカーテキストの情報を利用して、どの程度自動分類が行えるのかを確認することが目的であった。現時点では、 件数が比較的多い、分類 1(内容),2(機能),6(インデックス)で再現率が低いため、全体としての性能が思わしくない状況である。

今後のシステムの改良方針は、以下の通りである。

- 異表記、類義語への対応

類義語辞典・英語からカタカナへの翻字などを行うことにより、異表記、類義語への対応能力を向上させる。

- リンク元テキスト情報の活用

現在のシステムでは、リンク元のテキストの情報をうまく活用していない。例えば、同じ、テキスト中に、分類 6(インデックス)と判定されるリンクが多いのであれば、そのリンクの近くに並んでいるリンクは同じ分類である可能性が高いといった情報を利用する。

- 分類規則による段階的判別から、確率的な判別へ

現在のキーワードリストとの比較を行う分類では、最初に作成するリストに入らなかったアンカーテキストは対応できず、結果として、再現率の低下につながっている。上記のリンク元テキスト情報の活用結果なども踏まえ、確率的に判別する方法についても検討する必要がある。

また、今回の実験では、抽出するキーワードの設定は、手動で分類を行ったデータ(6939 件)全てから、ある程度の出現頻度が見られるものを利用しており、分類規則作成のために用いたデータと同じデータを使って評価を行っている。キーワードリストの網羅性などを検証するためにも、分類規則作成に利用していないデータに対する評価も行う必要があると考えている。

5. まとめ

本研究では、アンカーテキストの役割に応じた自動分類システムの提案を行っている。本稿では、まず、これまでに提案していた分類規則の修正を行い、新しい規則を用いることによって、より一貫性の高い分類が行えることを確認した。また、分類実験の結果として得られた分類データから分類ごとの特徴を分析し、それらの特徴

を利用した自動分類システムの提案を行った。しかし、現在の自動分類システムでは、用いている情報が不十分であるため、分類性能があまり高くないという結果となった。今後は、現在のシステムの問題点を考慮しながら、システムの改良を行っていきたいと考えている。

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[2] 吉岡真治, Web 情報活用のためのアンカーテキストの分類と利用情報処理学会情報学基礎研究会, 2006-FI-84, pp. 27-33, 2006.

[3] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama. An evaluation of the web retrieval task at the third ntcir workshop. SIGIR Forum, Vol. 38, No. 1, pp. 39–45, 2004.

[4] 大塚博紀,吉岡真治, ハイパーリンク活用のためのアンカーテキストの役割分析と分類, 情報処理学会第 70 回全国大会論文集, pp. 5-209-5-210, 2008.