

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の研究

大塚 博紀[†] 吉岡 真治[‡]

[†]北海道大学工学部電子工学科 [‡]北海道大学大学院情報科学研究科

1 はじめに

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る。この Web ページのリンクを利用する研究として、リンク構造に注目した PageRank[1]やアンカーテキストを利用した Web ディレクトリの構築支援[2]などがある。しかし、実際にアンカーテキストに書かれている内容を体系的に分析した研究はほとんどない。我々はアンカーテキストの内容を分類し、目的に応じた利用をすることによって、既存のシステムの性能の向上が可能であると考え、分類の基準を提案している[3]。本稿では、実際に多くのアンカーテキストについて手動で分類を行う事により分類基準の一貫性について検証を行う。

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

我々は、アンカーテキストの内容を分析する事によって、以下の 8 種類に分類することを提案している[3]。

1. リンク先の内容を表すテキスト：「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：サイト全体におけるリンク先のページの機能的役割を示すテキスト
3. リンク先との関係を表すテキスト：「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：「HOME」、「ホーム」、「TOP」などのトップページを示すテキスト
5. ナビゲーションを指示するテキスト：「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：「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 をそのまま利用しているテキスト

8. その他：アダルトサイトなどが「18 才未満」を Yahoo にリンクするような、リンク先のページと全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[4]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ります。(例：「Yahoo!JAPAN のホームページ」→1,4)より詳細な分類定義が必要である事も判明した。

3.アンカーテキストの分類実験

[3]では主に高頻度のアンカーテキストに限定して分析を行ったが、アンカーテキストの網羅的な分析のためには、低頻度のものを含め、多くのアンカーテキストについて分析する必要がある。また、先の実験において、一つのアンカーテキストが一つ以上の分類に属する場合があったように、現時点の分類基準は、実際の分類を行うための基準として、一貫性がかけている可能性が考えられる。この基準の一貫性を分析するために、同一のデータに対して二人の作業者が作業を行い、その一致度についても分析を行った。また、分類の際には、一つのアンカーテキストが、複数の種類に関係すると判断される場合には、その全てを分類の情報として与えることとした。

4.実験結果

本実験では、[3]の分析と同じnw100gを利用し、ファイルの先頭から15120件のアンカーテキストを抽出して手作業による分類対象とした。二人の作業による分類の一致度を調べると、完全に一致したものは11202件(73.98%)、部分一致は307件(2.03%)という結果であった。部分一致とは、候補が複数存在した場合、少なくとも一つの分類が一致した場合である。

また、各分類について作業員A,Bが付与したアンカーテキストの数とその一致度を表1に示す。

Research on Role Analysis and Classification of Anchor Text for Utilizing Hyperlinks

[†]Hiroki Ohtsuka, Faculty of Engineering, Hokkaido University

[‡] Masaharu Yoshioka, Graduate School of Information Science and Technology, Hokkaido University

表1.アンカーテキストの分類結果と一致度

分類	A	B	完全一致	部分一致
1	2247	2943	1581	102
2	2026	3127	1519	28
3	2	1	0	0
4	1362	1277	1126	77
5	7355	6334	5929	100
6	1731	1027	496	0
7	551	552	551	0
8	2	4	0	0

次に、各分類に対し、分類の一致・不一致に関するテキストの具体例について分析を行った。

「1.リンク先の内容を表すテキスト」、「2.ページの機能を表すテキスト」については、「～のホームページ」や「～に返信」、「～表示」など、の特定のパターンに当てはまるものについては一致度が高かった。また、「4.ホームページを指示するテキスト」、「5.ナビゲーションを指示するテキスト」、「7.URL」については、他の分類に比べ、全体的に一致度は高かった。

一致度が低いテキストについて分析すると、「1.リンク先の内容を表すテキスト」、「2.ページの機能を表すテキスト」、「6.インデックスを表すテキスト」の間での分類間違いが多く見られた。「1.リンク先の内容を表すテキスト」と、「6.インデックスを表すテキスト」については、アンカーテキストが固有名詞である場合、分類を混同している場合が多々あり、不一致が生じた。「2.ページの機能を表すテキスト」「5.ナビゲーションを指示するテキスト」についても同様に、「詳しく見る」や、「もっと見る」といったテキストなどの場合、どちらに属するか、で分類が分かれる結果となった。「2.ページの機能を表すテキスト」「4.トップページを指示するテキスト」についても同様に間違いやすい分類として、「表紙」といったアンカーテキストが存在した。「3.リンク先との関係を示すテキスト」「8.その他」については、該当するアンカーテキストがほとんど存在しなかったため、今回の実験では考察が行えなかった。

また、今回扱ったアンカーテキストは、nw100gより抽出したアンカーテキストのうち、先頭から順に分類を行ったため、サイトのバリエーションが少なく、特定のテンプレートから生成されたと考えられるものが多くみられた。

5.分類基準の詳細化

前節の実験により、幾つかの分類については一

致度の高い分類が行えたが、まだ分類基準が曖昧なものが存在する。この曖昧性を解消するためには、分類基準の詳細化が必要である。

「2.ページの機能を表すテキスト」、「インデックスを表すテキスト」については、定義をより明確にすることで、他の分類と区別をし易く修正する。

「～に戻る」といったテキストの場合、「5.ナビゲーションを指示するテキスト」には該当しないように分類を行ったが、そのことにより、「5.ナビゲーションを指示するテキスト」と「2.ページの機能を表すテキスト」の分類のどちらに属するかでずれが生じた。従って「～に戻る」といったテキストも、「5.ナビゲーションを指示するテキスト」に分類することで、ずれを解消できると思われる。「7.URL」については、個人によるずれは生じない。リンク先の URL とアンカーテキストを比較することで分類できるので、予め除外してよいと思われる。

6.まとめ

本稿では、実際の Web ページから得られたアンカーテキストを手作業で分類することにより、基準の一貫性の検証を行った。約 75%のアンカーテキストは一貫して分類されたが、一貫性に欠ける分類が散見された。また、分類誤りの実例を分析し、基準の詳細化を行った。

今後は新しい基準の一貫性を検証するために、再度の分類実験を行う予定である。また、その実験の際には、サイトごとの偏りの少ないデータに対して分析を行う予定である。

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] 鈴木祐介, 松原茂樹, 吉川正俊. アンカーテキストを用いた web ディレクトリの構築. 情報処理学会自然言語処理研究会, 2005-NL-168, pp.75-80, 2005.
- [3] 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.
- [4] 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.