2P-4

Web 文書におけるアンカーテキストの役割分析と自動分類の研究

Research on Role Analysis and Automatic Classification of Anchor Texts in Web Documents

大塚博紀[*1]　吉岡真治[*1]
Hiroki Ohtsuka　Masaharu Yoshioka

[*1]北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology,Hokkaido University

Anchor text of a web document is a text that characterises relationship between the page and linked one. This type of information is useful for web text indexing and web structure analysis. However, most of the researches pay little attention to the type of anchor texts. In this research, we have already proposed to classify role of anchor texts from the view point of hyperlink utilization. In this paper, we report the results of the manual classification experiment of the anchor texts information that are extracted from the web test collection NTCIR nw100g. We also discuss the further direction for the automatic classification.

1. はじめに

Web 文書の特徴は，複数のページがハイパーリンクと，そのリンクの内容を示すアンカーテキストにより関係付けられている点にある．この Web 文書の特徴を生かした情報システムの研究としては，リンク構造に注目した PageRank[Brin 98] の研究やアンカーテキストを利用した検索インデックスの作成 [McBryan 94] などがある．しかし，これらの研究では，アンカーテキストが持つ役割について，あまり分析が行われておらず，サイト内リンクとサイト外リンクといった区別はあるものの，全てのアンカーテキストやリンクをほぼ同等に取り扱っている．これに対し，本研究では，実際にアンカーテキストの持つ役割を考慮した分類を行い，目的に応じて使い分けをすることにより，既存のシステムの性能の向上が可能であると考え，分類の基準の提案 [吉岡 06] を行うと共に，人手による分類実験 [大塚 08] を行ってきた．本稿では，[大塚 08] の分類実験の結果問題となった分類作業基準について見直しを行うと共に，再度，分類実験を行うことで，その有効性を検証する．最後に，今後の自動分類に向けての方針について述べる．

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

2.1 アンカーテキストの役割分類

アンカーテキストには，組織・商品名と official site のようなページの内容を表すようなテキストが用いられたり,「戻る」や「次」といったようなページ内でのナビゲーションを目的にするようなテキストが存在する．我々は，Web のテストコレクション NTCIR の nw100g[Eguchi 04] 中のアンカーテキストの情報を参考にすることにより，以下の 8 種類の分類を提案している [吉岡 06].

1. リンク先の内容を表すテキスト：「Yahoo!JAPAN」などの，リンク先の内容を示すテキスト
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 歳未満」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

2.2 アンカーテキストの分類実験

この 8 つの役割分類に基づいたアンカーテキストの自動分類を行うためには，提案している基準が一貫していて，安定的な分類が可能であることが望まれる．この分類の安定性・一貫性を分析するために，二人の作業者により，nw100g から抽出したアンカーテキストに対し分類実験を行い，その一致度を調べる実験を行った [大塚 08].

具体的には，nw100g の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分類を行った．アンカーテキストには，2.1 節で提案した役割の内，複数の役割に属するものが存在したため，被験者は，一つのアンカーテキストに対して，最大 2 つまでの役割の情報を付加することとした．

二人の作業者による分類の一致度を調べると，完全一致したものは 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者 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.

2.3 分類実験の考察

表 1 からも分かるように，特定の役割ては，非常に高い一致度を実現している一方で，一致度が低い分類基準も存在する．

連絡先: 大塚博紀，北海道大学大学院情報科学研究科，〒 060-0814 札幌市北区北 14 条西 9 丁目，電話 011-706-7575, ohtsuka@kb.ist.hokudai.ac.jp

表 1: 分類実験の結果

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

これは，現時点での役割分類の基準が曖昧であり，作業者が人手で分析をしたとしても，一貫した分類が行えないということを意味している．

役割の分類基準について分析するために，一致度の高いテキスト，低いテキストの傾向を分析することにより，分類基準についての考察を行った．

まず，一致度が高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」については,「〜のホームページ」や「〜に返信」,「〜表示」など，分類ごとにある程度の特定のパターンに当てはまるものが存在した．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，全体的に一致度が高く，特定のキーワードの存在や，決まったフォーマットなどの情報が確認された．

次に，一致度が低いテキストについて分析すると，特定の分類間での分類の不一致が多く見られた．具体的には,「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」,「3. リンク先との関係を表すテキスト」,「6. インデックスを表すテキスト」の間での分類の不一致が多く存在した．これは，これらの役割の分類基準が曖昧であることが原因であると考えられ，分類基準の詳細化が必要であることを示している．

また，今回扱ったアンカーテキストが，nw100g より抽出したアンカーテキストのうち，上位のもの (nw100g の ID) 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同一のページに対する同一のアンカーテキストが多数存在し，一致度の分析などを行う際に，あまり適切でない結果となってしまった．

3. 分類基準の詳細化と再実験

前節の実験ならびにその考察を踏まえた分類基準の詳細化を行った．また，その新しい基準による再度の分類実験を行ない，新しい基準の妥当性の検証を行った．

3.1 分類基準の詳細化

前節での考察を踏まえ，次の 4 つの役割については分類基準の見直しを行った．また，これらの 4 つの役割については，分類を排他的なものとした．

1. リンク先の内容を表すテキスト
   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．

2. ページの機能を表すテキスト
   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．

3. リンク先との関係を表すテキスト
   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．

6. インデックスを表すテキスト
   リンク先のページが，テキストの内容に直接関係する情報ではなく，関連するページや電話番号といったインスタンスの集合体を表す場合のテキストである．

3.2 アンカーテキスト分類の再実験

前節で行った実験では，ファイルの先頭からからアンカーテキストを抽出したために，特定のサイトからのアンカーテキストが多く存在し，特定のアンカーテキストとリンクのペアの分析結果が全体の結果に大きく影響を与えるような場合が見受けられた．この問題を解決するために，この再実験では，nw100g よりランダムに抽出した 17000 件のアンカーテキストを分類の対象とした．

また，前回の実験で，被験者ごとの解釈の違いの存在が明らかになった．様々な解釈の可能性を考慮した上で，判定の一貫性を分析するために，被験者の数を 3 名に増やした．また，3 名とすることにより，分類の不一致が起こった際に，どちらの被験者の情報が一般的であるかを判断する材料が得られ，典型的な分類事例，やや判断にまよう事例，判断が困難な事例などを考えることが可能となる．

各分類について作業者 C,D,E が付与したアンカーテキストの数を表 2 に示す．また，作業者 C，D の完全一致は 69.24%(11770 件)，部分一致は 4.20%(714 件)，作業者 C,E の完全一致は 53.97%(9144 件)，部分一致は 6.09%(1036 件)，作業者 D,E の完全一致は 49.32%(8384 件)，部分一致は 6.29%(1069 件) という結果になった．

表 2: 作業者ごとの分類件数

分類	C	D	E
1	4173	4209	9564
2	6922	8459	1576
3	233	63	335
4	1378	1242	1202
5	1945	2320	2502
6	3510	3519	2496
7	306	82	578
8	209	15	308

3.3 実験の考察

表 2 や，被験者ごとの一致度の分析結果から，被験者 C,D では一致度が高い一方で，被験者 E に関しては，一致度が低いという状況が発生した．これは，現時点での分類基準のマニュアルの記述が不十分であり，その理解の状況に応じては，異なる分類になってしまう危険性があることを示している．

また，前回の実験と同様に，一致度の高いテキスト，低いテキストの傾向を分析することにより，分類基準についての考察を行った．

まず，一致度が高いテキストについては，前回の実験と同様の傾向が確認されたが，分類対象のアンカーテキストの抽出方法を変更したことにより，多くのバリエーションのデータを獲得することができた．また,「6. インデックスを表すテキスト」については，定義を明確にしたこともあり，Yahoo!のディレクトリへのリンクなどのアンカーテキストを一致度が高く分類することができた．

次に，一致度が低いテキストについて具体例をあげながら分析を行う．

1. 前回と同様に，「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」における不一致事例についての不一致が多く見られた．この不一致は，ページの内容の専門性が高いときに多く見られ，被験者がリンク先のページの内容を理解しないとうまく分類できない場合があることを示している．

2. 「〜のページへ」といった，ページの内容を示すテキスト＋誘導をしていると考えられる助詞「へ」の組合わせの場合に，「5. ナビゲーションを指示するテキスト」に分類するかどうかで不一致が存在した．

3. 「3. リンク先との関係を表すテキスト」のアンカーテキストについては，特定のアンカーテキストの解釈の違い (例えば，「フレームあり」，「フレームなし」を「2. ページの機能を表すテキスト」と分類するか，「3. リンク先との関係を表すテキスト」と分類するか) が多く見られた．

上記の問題の内，2,3 番目の問題は，不一致の事例は多いものの，そのパターンに特徴があるため，具体的な不一致事例について，事例集のようなものを作成することにより，大部分の不一致が解消できると考えられる．

1 番目の問題については，被験者の持っている知識や読解力に左右されることから，完全にこれらの役割の間での不一致をなくすことは困難であると考えられる．ただ，これらの役割においても，一致度の高いアンカーテキストの典型的な分類事例は存在している．これらの典型的な分類事例が持つ特徴について整理をしていくことにより，判断が困難な曖昧な領域を狭めていくことが求められる．

3.4 アンカーテキストの自動分類に向けて

これまでの 2 回の実験を通して，各々の役割分類に対応する典型的なアンカーテキストの情報を数多く収集することが出来た．特に，「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，全体的に一致度が高く，特定のキーワードの存在や，決まったフォーマットなどの情報が確認されていることから，主にアンカーテキストの情報を主体とした分析を行うとともに，アンカーテキストの自動分類が可能になると考えられる．

これに対し，「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」などについては，ページ先の内容との対応関係が重要になるので，アンカーテキストだけではなく，リンク先のテキストの情報も利用した分類を行う必要がある．

今後は，上記の考察を踏まえて，アンカーテキストに関する特徴情報を付加する方法の検討を行うと共に，役割ごとに分類に役立つ基準が異なる可能性があることも考慮して，機械学習システムを選択し，一致度の高いアンカーテキストの情報を用いた分類システムの構築を行いたいと考えている．

4. まとめ

本論文では，Web 文書の特徴であるアンカーテキストの付加されたリンク構造を活用するためのアンカーテキストの役割分類の考え方を紹介すると共に，人手による分類実験の結果とその分析結果について述べた．現時点において，分類基準に多少の曖昧性は残るものの，役割分類に対する多くの典型的な分類事例を得ることが出来た．

今後は，事例集の作成などを行うことによって，役割分類の一貫性の向上をはかると共に，これまでに行った分類実験の結果得られた典型的な分類事例を元にした自動分類システムの構築を行いたいと考えている．

参考文献

[Brin 98] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117 (1998)

[Eguchi 04] Eguchi, K., Oyama, K., Ishida, E., Kando, N., and Kuriyama, K.: An Evaluation of the Web Retrieval Task at the Third NTCIR Workshop, *SIGIR Forum*, Vol. 38, No. 1, pp. 39–45 (2004)

[McBryan 94] McBryan, O. A.: GENVL and WWWW: Tools for taming the Web, in *Proceedings of the First International Conference on the World Wide Web* (1994)

[大塚 08] 大塚博紀, 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, 情報処理学会第 70 回全国大会講演論文集, pp. 5–209–5–210 (2008)

[吉岡 06] 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)