参照元HTMLテキストからのWebサイト紹介文抽出

3V−1

原田 昌紀　風間 一洋　佐藤 進也

日本電信電話株式会社　NTT未来ねっと研究所

1　はじめに

今日のWebサーチエンジンの多くは，検索されたWebページの要約文として，テキストの先頭部分を利用している．しかし，こうした要約文では，利用者がWebページの概要を把握し，閲覧すべきかを判断することは難しい．また，重要文抽出法などの自動要約手法[1]は，テキスト量の少ない一般的なWebページには適していない．

一方，Amitayらは，アンカー（HREF属性を持つA要素）の近傍に，参照先Webページの概要を紹介する文章が多く見られることに着目し，そうした紹介文を要約文として利用する方法を提案している[2]．この方法は人間が記述した自然で読みやすい文章が得られる点で優れているが，要約文の得られるWebページが十分に多くなければ実用的でない．そこで本稿では，Amitayらの方法の精度を改善し，実際に大量のWebページに適用した結果を述べる．

2　紹介文候補の抽出方法

本節では文献[2]で提案された紹介文候補の抽出方法の一実現方法を説明する．

一般に紹介文は図1に示すように，アンカーの直後に紹介文が記述され，それらの組のみで独立した段落となることが多い．こうしたWebページの記述の慣習を利用することで，複雑な言語処理を用いずに紹介文の候補を抽出できる．

しかし，Webページは必ずしも論理的にマークアップされているとは限らない．そこで，空行，罫線，表の枠など，レイアウトの記述に用いられるHTMLの要素によって，段落の境界を判定する．

抽出手順は次の通りである．まず，HTMLの要素を(A)前後で空行を出力する要素，(B)直前のみ空行を出力する要素，(C)現在の行を改行する要素，(D)改行と関連しない要素の4種類に分類し，HTMLテキストを空行で区切られたテキストに変換する．そして，図2に示すような，直前が空行でありアンカーを含む行と，それに続く空行ではない行の並び（0行以上）からなるテキストを紹介文候補として抽出する．具体的には(A)はP, HR, TABLE, UL, OL, DL, H1 ～ H6, (B)はDT, LI, TR, (C)はBRのみ，(D)はそれ以外とする．ただし，A要素の中の改行は無視し，一つのアンカーが複数行にまたがらないようにする．

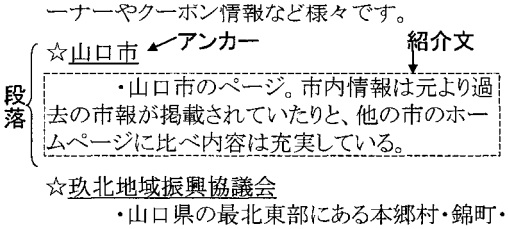

図1: 一般的な紹介文の記述例

3　紹介文の選別方法

抽出された紹介文候補には要約文に適した紹介文テキストを持たないものが含まれている．また，一つのWebページを参照する紹介文候補が複数ある場合もある．そのため，Amitayらのシステムでは動詞の用法や人称代名詞の有無など，多数の表層的な特徴を評価することで紹介文テキストを選別している．

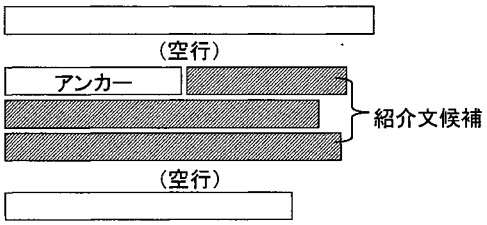

図2: 紹介文候補として抽出するパターン

Extracting descriptive hyperlinks for web site summaries

Masanori Harada, Kazuhiro Kazama, Shin-ya Sato
NTT Network Innovation Laboratories, NTT Corporation

しかし，紹介文テキストの表層的な特徴だけでは，要約文に適した客観的な文章を選別することは難しい．そこで我々は，アンカーテキストおよび参照元Webページも評価に加えることで，より高い精度で紹介文候補を選別する方法を提案する．すなわち，紹介文候補の要約文としての適切さを表わすスコアSを次のように定義する．

$$S = S_t \times S_a \times S_r$$

以下，本稿では計算方法の概略のみを述べる．

S_tは紹介文テキストのスコアであり，長さが50字から150字で，句読点がそれぞれ2～4個程度含まれている場合に高い値とする．また，「ホームページ」「サイト」等の語句があれば加点し，句読点以外の記号があれば減点する．

S_aはアンカーテキストのスコアであり，長さが5字から20字で，多くのアンカーテキストに共通した一般的な表記の場合に高い値とする．URL文字列や，インライン画像を含むアンカーは減点する．

S_rは参照元Webページのスコアであり，抽出される紹介文候補の数が一定数以上の場合に高くする．これは少数のWebページのみに言及するWebページでは，参照元の文脈に依存した主観的な紹介文が多いのに対して，リンク集などでは客観的な紹介文が期待されるためである．

図3に山口市のWebサイトのトップページを参照する紹介文候補を評価した例を挙げる．

要約文に適した紹介文候補（S = 144）

アンカーテキスト: “やまぐち”
紹介文テキスト: “山口市の遊び場、福祉、イベント情報等が掲載されており、市民にはお勧めです。もちろん史跡案内もありますので観光客にもお役立ちサイトです。”
http://www.interwoman.co.jp/chihou/chihou_chugoku.htm

S_t=180 紹介文68文字，句点2個，読点3個，「サイト」
S_a=0.8 アンカーテキスト4字
S_r=1.0 抽出された紹介文候補数: 23個

要約文に適さない紹介文候補（S = 19.2）

アンカーテキスト: “山口市”
紹介文テキスト: “山頭火の庵跡　風来居　文学・記念碑、記念館等　種田山頭火句牌”
http://www.joho-yamaguchi.or.jp/bteprise/san.htm

S_t= 64 紹介文27文字，句点0個，読点1個
S_a=0.6 アンカーテキスト3字
S_r=0.5 抽出された紹介文候補数: 5個

図3: `http://www.urban.ne.jp/home/cityyama/` を参照する紹介文候補の例

4　紹介文の抽出実験

WWWロボットで収集した約1,345万URLのHTMLテキストに対して，第2節で述べた方法を適用したところ，約1,864万個の紹介文候補が抽出された．このうち，参照元Webページと参照先Webページが同一ホスト上にあるものは，記述が客観的でない可能性が高いため除外すると，518万個の候補が得られた．

ここで，いくつかの紹介文候補を目視で選別し，スコアのしきい値Tを，T以上のスコアを持つ紹介文候補の9割以上が，要約文として利用可能な紹介文テキストを持つように決めた．その結果，T以上のスコアを持つ紹介文候補は約175万個となり，約80万URLのWebページに対して一つ以上の要約文が得られた．

最終的に要約文が得られたWebページが，収集したWebページの1割にも満たないことから，本手法は他の自動要約手法に対して補助的に利用する必要があるといえる．

しかし，Webサーバのトップページを無作為に1,000URL選んで調べたところ，621URLに対して要約文が得られていることがわかった．今日のWebサーチエンジンは，トップページなどの被参照数の大きいWebページを検索結果の上位にランキングするため，本手法で得られる要約文が利用者に提示される確率は高くなると予想される．

また，紹介文が抽出された参照元Webページを検索結果と共に提示することで，サーベイ的な検索を支援することも考えられる．

5　まとめ

参照元WebページのHTMLテキストから，参照先Webサイトを客観的に説明する紹介文を抽出し，要約文として用いる方法を述べた．また，その実用性について，予備的な評価をおこなった．

今後は，本手法をWebサーチエンジン上に実装し，要約文の有効性を評価する予定である．

参考文献

[1] 奥村 学, 難波 英嗣: “テキスト自動要約に関する研究動向”, 自然言語処理, Vol.6, No.6, pp.1-26, 1999.

[2] Amitay E., Paris C.: “Automatically Summarising Web Sites - Is There A Way Around It?”, ACM 9th International Conference on Information and Knowledge Management (CIKM 2000), 2000.