ウェブ文書からの見出し構造抽出精度の向上と検索結果表示への応用

Accuracy improvement of headline structure extraction in Web pages and application to a search engine results page

テーマ：インターネット技術とその応用　　教養学部　情報科学科
指導教員：松本 章代　　1057231　菅原 宇規

1. 研究背景および目的

本研究では，Web ページをブラウザ上に表示した際の見た目の階層構造に着目し，見出しを階層構造ごと抽出することを目的とする．

Google や Yahoo!などの検索エンジンの検索結果には，検索キーワードの前後の文章が表示される．しかし，それよりページ全体の概要を表示した方が，欲しい情報が記載されているか判断できる場合があると考える．見出しの階層構造を抽出することができればそれが可能である．また，複数の検索キーワードで検索する場合，その検索キーワードの階層関係が分かれば，検索者の欲しい情報が記載されているか判定できる．他にも，見出しとその見出しの支配範囲の文章を正しく対応づけて抽出することができれば，その情報を二次利用することに役立つ．

本年度の研究の到達目標は，第一に階層関係の判別精度を向上させることである．先行研究 [2] により，Web ページの見出しを抽出するプログラムと，先行研究 [3] により，隣接する 2 つの見出し間の階層関係を判定するプログラムが作成されている．この 2 つを統合することによって，最終的に見出し階層構造抽出システムとなる．現時点で階層関係を判定するプログラムの判別精度がシステムとして運用するには不十分である．そのため，階層関係の判定が正しく行えていない原因を判明させるための失敗事例分析を行い，それに基づきシステムを改善して階層関係の判別精度を向上させる．本研究における階層関係の定義を図 1 に示す．

第二に見出し構造システムによって Web ページの要約を作成し，それを検索結果として提示することにより検索効率の向上を図ることである．先行研究 [4] により，Web ページの概要をツリー構造を用いて表示する手法について検討が行われた．この手法と見出し構造システムを統合することによって，検索結果における Web ページの概要をツリー構造で表示する．このことにより，Web ページを開かなくともその内容をわかるようにし，検索の効率を向上させる．

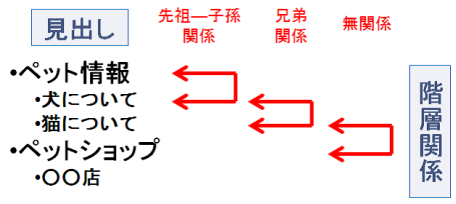

図 1.　階層関係の定義

2. 階層関係判定の精度改善

2.1 利用する正解データ

本研究では，タグの階層構造ではなく，Web ページの「見た目」の階層構造に着目している．「見た目」は主観であり，人によって異なる可能性があるため，「正解データ」の作成が重要である．そこで先行研究 [2] では，以下の手順で「正解データ」を作成した．

(1) Web ページ 200 件を用意し，各 3 部プリントアウトする．
(2) 1 件につき 3 人の作業者を割り当て，(1) の用紙に見出しとその支配範囲を記入してもらう．
(3) 3 人の意見が一致した見出しと支配範囲を「正解」とみなし，HTML ファイルに見出しとその支配範囲のタグ付けを行う．

本研究でもこの「正解データ」を利用する．

2.2 判別式の作成

本研究では，作成した判断材料を用いて階層関係が先祖－子孫関係なのか兄弟関係なのか無関係なのかを自動判定するための判別式の精度を向上させることを目指す．

先行研究では判別式の作成に数量化理論 2 類を用いていたが，本研究では主に C4.5[6] を用いる．その理由は，C4.5 を用いて作成された決定木は，人間が見て理解し易いため，失敗事例分析に利用するのに適しているためである．

数量化理論 2 類や C4.5 で階層関係を判定するための判別式を作成するためにまず，装飾情報の差（判定材料）を数値データとして出力し独立変数とする．「正解データ」の見出しの装飾情報を抽出し従属変数とする．そのデータを半数ずつ訓練用データ，テスト用データの 2 つに分ける．訓練用データを用いて数量化理論 2 類，もしくは C4.5 で階層関係を判定するための判別式を生成した後，テスト用データを用いて判別式を適用し精度を確認する．

2.3 失敗事例分析および改善

2.3.1 文字サイズの判定に関する改善

先行研究 [3] の時点では，比較する 2 つの見出し間の文字サイズ指定方法が同じでないと，装飾情報として使われていなかったことが判明した．たとえば，font-size:medium という指定方法の見出しと fontsize:10px という指定方法の見出しは比較できず，装飾情報の差を出すことができなかった．そこでそのような場合でも比較できるようにするため，絶対サイズのキーワード指定と%，px で指定された文字サイズの関係を調査し，単位を統一した．

また，文字サイズの装飾情報はページ全体の基準となる文字サイズと現在の見出しの文字サイズを比較し

ていた．しかし階層関係を判定するにあたって，その 2 つの比較よりも前後の見出し間の文字サイズの差を比較した方が有効であると考えため，そのように修正をした．修正前の C4.5 で測定した判別結果を表 1，修正後の判別結果を表 2 に示す．

表 1.　修正前の判別結果

実際の群＼判別された群	先祖-子孫関係	兄弟関係	無関係	合計
先祖-子孫関係	103	264	10	377
%	(27.3)	(70.0)	(2.7)	(100.0)
兄弟関係	18	5113	22	5153
%	(0.4)	(99.2)	(0.4)	(100.0)
無関係	3	198	66	267
%	(1.1)	(74.2)	(24.7)	(100.0)

表 2.　文字サイズ修正後の判別結果

実際の群＼判別された群	先祖-子孫関係	兄弟関係	無関係	合計
先祖-子孫関係	110	257	10	377
%	(29.2)	(68.2)	(2.6)	(100.0)
兄弟関係	21	5112	21	5154
%	(0.4)	(99.2)	(0.4)	(100.0)
無関係	3	192	72	267
%	(1.1)	(71.9)	(27.0)	(100.0)

2. 3. 2　テキストインデントの判定に関する改善

文字サイズの場合と同様に，テキストインデントも階層関係を判定するにあたって，前後の見出し間のテキストインデントを比較する方が有効だと考えたため，そのように修正を行った．

先行研究 [3] の時点で，テキストインデントは text-indent:-10px のように負の値であれば 1 という数値データ，text-indent:10px のように正の値であれば 2 という数値データが与えられていた．この場合，たとえば，比較する 2 つの見出しのテキストインデントが 10px と 15px のとき，数値データは 2 つとも 2 であるため装飾情報の差は 0 となる．しかし，実際は前の見出しの方が後の見出しより左側にあるので，テキストインデントには差がある．

このような状況の場合でも正しく装飾情報の差を与えられるように修正をした．

またテキストインデントを決めるプロパティとして，margin，padding を判定材料に追加した．

修正後の判別結果を表 3 に示す．

表 3.　インデント修正後の判別結果

実際の群＼判別された群	先祖-子孫関係	兄弟関係	無関係	合計
先祖-子孫関係	105	267	5	377
%	(27.9)	(70.8)	(1.3)	(100.0)
兄弟関係	16	5111	16	5143
%	(0.3)	(99.4)	(0.3)	(100.0)
無関係	4	203	60	267
%	(1.5)	(76.0)	(22.5)	(100.0)

3.　Web ページの概要をツリー構造を用いて表示するプログラム

3. 1　システム概要

見出し階層構造抽出システムと，Web ページの概要をツリー構造を用いて表示する手法を統合する．このプログラムは Web ページの要約を作成し，それを検索結果として提示することにより検索効率の向上を目的とする．このプログラムの階層関係判定では，無関係の場合は再帰的に階層関係を判定していく．

3. 2　システムの流れ

(1) Web ページの HTML と CSS から Web ページのテキストを見出しか否か判定する．

(2) 見出しと判定された 2 つのテキストの階層関係を判定する．

(3) 2 つの見出しが先祖－子孫関係，兄弟関係と判定された場合はツリーを作成していく．

(4) もし無関係と判定された場合は今判定している見出しと前の見出しの先祖の見出しとで再度階層関係を判定する ((2) に戻る)．

(5) すべての見出しの階層関係を判定したらツリーを表示する．

3. 3　結果

図 2 の左が実際の Web ページで，右がプログラムを適用してツリー表示した結果である．なお，ツリーのルートは Web ページのタイトルとする．

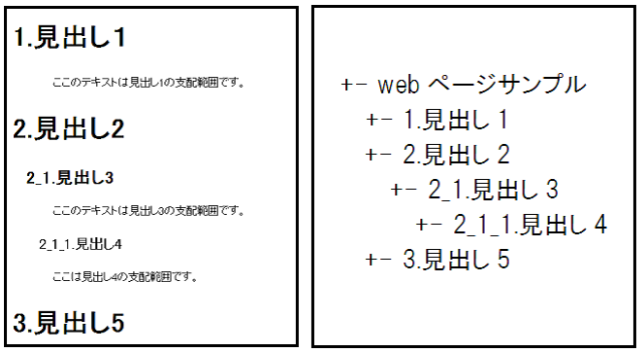

図 2.　実際の Web ページとツリー表示した結果

4.　まとめ

本年度の研究で階層関係判定プログラムの文字サイズとテキストインデントの判定に関する改善を行った．また，Web ページの概要をツリー構造を用いて表示するプログラムを作成した．

今後は引き続き階層関係判定プログラムの判別精度の向上させていくことによってツリー表示をより正確にしていく必要がある．

参考文献

[1] 西口直樹，松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：見出しの階層関係を利用した WWW 検索精度の改善，信学技報，Vol.105，No.595，NLC2005-114，pp.1-6(2006)．

[2] 斎藤 貴大：ウェブページにおける見出し構造の抽出と分析，東北学院大学卒業論文 (2012)．

[3] 松田 駿：ウェブページにおける見出し間の階層関係の判定，東北学院大学卒業論文 (2013)．

[4] 千葉 悠真：Web ページの概要をツリー構造を用いて表示する手法の検討，東北学院大学卒業論文 (2013)．

[5] 池田彰吾，松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：繰り返し構造を考慮した Web ページの見出しの階層構造の解析，情報処理学会研究報告，Vol.2008，No.34，pp.31-38(2008)．

[6] J.R.Quinlan：“C4.5 Programs for Machine Learning”，Morgan Kaufmann(1993)．