4N－8

テキストマイニングシステムにおける文書データとその付加情報の可視化手法

坂入　隆
日本アイ・ビー・エム（株）東京基礎研究所

1. はじめに

計算機上に貯えられた膨大な数のテキストデータから有用な知見を発見するテキストマイニングシステムに関する研究が活発に行われている [1]。消費者からの問い合わせに電話で回答するコールセンターの記録やWeb上のテキストデータなど様々な種類のデータをテキストマイニングの対象とすることができる。

テキストマイニングシステムにおいては、テキストデータを解析することが重要なのはもちろんであるが、解析した結果を利用者にとって理解しやすいように可視化することも同様に重要な意味を持ってくる。

本稿では、文書間の階層構造とそれらの文書の付加情報の重ね合わせという可視化手法と、この手法のサイトマップへの適用について述べる。

2. 文書間の階層構造及び付加情報

この章では、文書間の階層構造とそれらの付加情報について述べる。

テキストマイニングシステムで扱う膨大な数の文書からなるデータは、明示的な階層構造を持っていることがある。例えば、ファイルシステムのディレクトリの構造や、ハイパーテキストのリンクの構造である。また、明示的な階層構造を持っていない場合でも、自然言語処理などの技術を用いて解析することによって、階層構造を持たせることができる。例えば、コールセンターの記録では、問い合わせ元の住所や問い合わせのあった製品の種類などによって階層構造を持たせることが可能である。

また、文書データを解析することによって、キーワードなどの付加情報を得ることができる。この情報も、膨大な数の文書を効率良く利用するために有効である。

3. 重ね合わせによる可視化手法

階層構造の可視化のために、Cone Trees [2] などの手法が提案されている。また、キーワードは、索引という形で可視化することができる。

しかし、既存のシステムでは、階層構造の表示と付加情報の表示が独立した機能として提供されている。そのため、利用者にはこれらの関連がわかりにくい。例えば、新聞社のWebサイトで、スポーツ選手の名前がスポーツのセクションに現れるのは珍しいことではないが、経済のセクションに現れるのは珍しいことである。このようなことを既存のシステムで利用者が発見するのは難しい。

著者は、文書間の階層構造の表示と付加情報の表示を重ね合わせる、という可視化の手法によってこの問題を解決した。

また、階層構造の表示を通して付加情報の表示を変更し、逆に付加情報の表示を通して階層構造の表示を変更するということによって効率的な操作を可能にした。

4. サイトマップへの適用

この章では、前の章で述べた可視化手法の適用例としてWebのサイトマップ [3] について説明する。このシステムは、著者の属するグループで開発した Site Outlining [4] というシステムを修正することによって実装されている。元のシステムでは、サイトマップを表示する機能もキーワードを表示する機能も持っていたが、これらの機能に関連がなかった。そこで、重ね合わせによる可視化手法による改良によって、これらの機能に関連を持たせた。

図1にこのシステムの画面の例を示す。右下のフレームにサイトマップを、左下のフレームにキーワードの一覧を、上のフレームにボタンを表示している。サイトマップは、階層をインデントによって

A Technique for Visualizing Document Data and Those Additional Information in Text Mining Systems
SAKAIRI Takashi
sakairi@jp.ibm.com
IBM Research, Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato, Kanagawa 242-8502, Japan

示されたWebページを表わす行からなる。それぞれのWebページは、0個以上のキーワードを含んでいる。マウスのポインタの下にキーワードを含むWebページを表わす行があるときは、その行の下にキーワードを表示する。

図2にサイトマップ上でWebページを表わす行を選択したときの画面の例を示す。それぞれのキーワードが幾つの選択されたWebページに含まれているかをキーワード一覧に表示し、1つ以上のWebページに含まれているキーワードの色を変えている。

図3にキーワードの一覧上でキーワードを選択したときの画面の例を示す。サイトマップ上のWebページを表わす行に、選択されたキーワードが含まれれば、そのキーワードを表示し、その行の色を変えている。

ここで示した画面例では、非常に単純な方法でサイトマップを表示している。しかし、重ね合わせによる可視化の手法は、ここで示したような表示に限定される訳ではなく、様々な表示法と組み合わせることが可能である。

5. おわりに

本稿では、文書間の階層構造の表示とそれらの付加情報の表示を重ね合わせる可視化の手法について述べた。また、この可視化手法を適用したサイトマップを例として述べた。

テキストマイニングシステムで扱うデータには、明示的な階層構造がない場合でも適切な解析によって階層構造を得る事ができる。また、キーワードなどの付加情報も適切な解析によって得られる。そのため、本稿で述べた手法は、多くのテキストマイニングシステムに適用可能である。

参考文献

1. 那須川, 諸橋, 長野, テキストマイニング - 膨大な文書データの自動分析による知見発見 -, Vol. 40, No. 4, (April 1999) pp. 358 - 364.
2. Robertson, G. G., Mackinlay, J. D., and Card, S. K., Cone Trees: Animated 3D Visualizations of Hierarchical Information, Proceedings of ACM Human Factors in Computing Systems, (April/May 1991) pp. 189 - 194.
3. Sakairi, T., A Site Map for Visualizing Both a Web Site's Structure and Keywords, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, (October 1999).
4. Takeda, K., and Nomiyama, H., Site Outlining, Proceedings of ACM Digital Libraries, (June 1998) pp. 309 - 310.

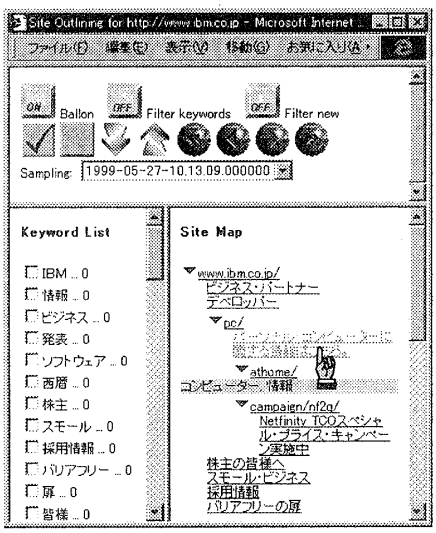

図1. システムの画面の例

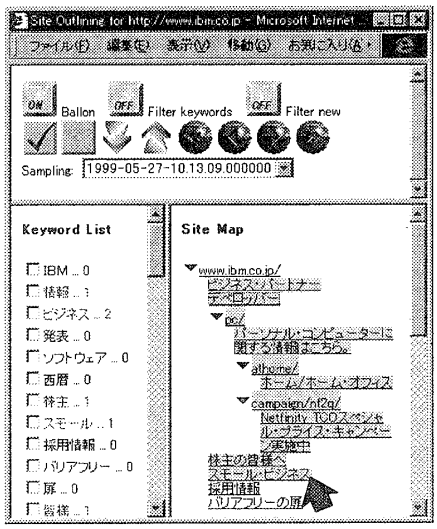

図2. Webページを選択した画面の例

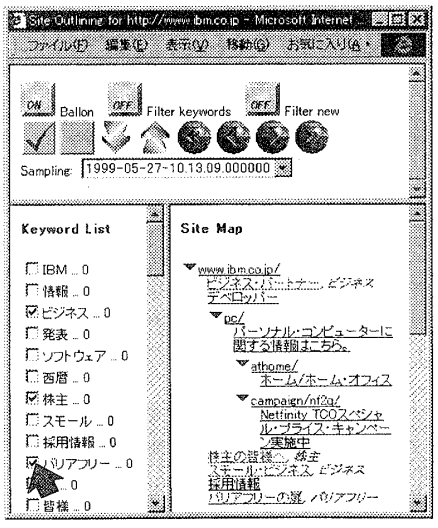

図3. キーワードを選択した画面の例