新聞記事データベースからの話題の抽出 II - 話題の構造の解析*

6E－3

野美山 浩†
日本アイ・ビー・エム株式会社 東京基礎研究所‡

1. はじめに

近年、新聞記事、論文などの巨大なデータベース、あるいは、World Wide Web などを通じたネットワーク上の情報など、多種多様な情報にアクセス可能になってきた。このような種々雑多で膨大な情報源に対しては、まず、「何を探すか」以前に「何があるのか」を知ることが検索者にとって検索を進めるための有効な指標となる。従来のキーワード検索システムが対象にしてきたのは、探すものが明らかである場合、それをどう効率良く探すかという問題であり、「何があるか」を探るための効率的な手段を提供していない。

このような問題に対する1つの解法として、我々は、大局的に情報全体を概観するために、時間軸における特異点を求めることによって、それらをユーザにわかり易く提示する手法を提案した[1]。本稿では、前回提案した話題[1]のモデルを拡張し「話題」の間の構造を持つ新しいモデルを提案し、その解析方法を提案する。

2. 話題のモデル

前回提案した話題のモデルは、1つのキーワードが1つの話題に対応するものであった。しかし、通常、1つの話題に対しては、複数のキーワードが対応するため、「話題」を表示した場合に1つの話題に対するキーワードが分散して表示されてしまうために、全体像が把握しづらい場合があるという問題点があった[2]。

実際の話題に対応するようなまとまりを作るために、従来の「1話題＝1キーワード」モデルの拡張を行なった。まず、1つの話題に対して複数キーワードが対応するように拡張した。次に、「話題」は、独立したものでなく、包含関係があるものとした。

*"Topics Extraction from Newspaper Databases II - Analysis of Topic Structures

†Hiroshi Nomiyama
‡IBM Research, Tokyo Research Laboratory

[1]本稿では、提案する話題のモデルによって言及されるものを「話題」と記述し、一般的な話題と区別する。

[2]通常のクラスタリングと異なり、時間軸に沿って表示されるため、順序がまちまちでも比較的容易に識別できる場合はある（図1参照）。

3. 話題の構造の解析

ここでは、定義した話題のモデルに基づいて、「話題」を解析するための手法について述べる。

我々がこの手法を適用しようとしてる対象は、対話的な検索時のユーザへのフィードバックのためである[2]。そのため、動的に、かつ、対話的に解析する必要がある。新聞記事の検索においては、2,000件程度の記事集合に対して、実用上は、遅くとも1分以内で解析できる必要がある。そのため、厳密な解析をするのではなく、できるだけ高速化できるように処理の簡略化に努めた。処理の手順を順に述べる。

(1) 話題性の高いキーワードの抽出 これは、[1]の方法で行なう。結果として、キーワード＋話題性＋期間の組み（この3つ組を話題要素と呼ぶ。）の話題性の高い順のリストが得られる。以下にその例の一部を示す（最初の要素はキーワード "開発" が話題性が26.843138で、その期間が7-10(8月から11月)であることを示している。）。

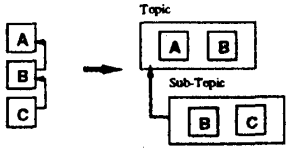
```
開発 26.843138 (7 10)
開発事業 23.539215 (7 10)
```

(2) 話題要素間の包含関係の解析 (1)で得られた順序リストの各話題要素間の包含関係を解析する。ここでは、効率化のために、「話題性の低いキーワードは話題性のより高いキーワードに含まれる」ということを仮定する。この仮定によって、包含関係を検査する回数は話題要素数を n とすると、$\sum(n-1)$ となる。

ある「話題」を決定するために、まず、2つの話題要素の期間に包含関係があるかどうかを調べる。期間の一致がないものは、包含関係がないものとする。次に、2つの話題要素が共起する記事の頻度の割合を調べる。話題要素AとBを調べる場合（AはBより話題性が高いものとする）、$Freq(A\&B)/Freq(A)$ と $Freq(A\&B)/Freq(B)$ がともにある閾値より大きなものを包含関係があるものとし、その包含の度合を $Freq(A\&B)/Freq(B)$ と定義する。

個々の話題要素について、自分より話題性の高い話題要素について、包含関係を調べ、最も包含の度合の高い話題要素に包含されるものとする。

(3)「話題」の構造解析　他の話題要素を含んでいる話題要素は、それらをすべて含む「話題」となる。「話題」が含んでいる話題要素が、さらに他の話題要素を含んでいる場合は、その「話題」の「副話題」として解析される（下図参照）。



(4)「話題」の表示　解析された話題に含まれるキーワードを時系列に沿って表示されるバーの上に列挙する。また、他の「話題」を包含している「話題」で他の「話題」に包含されていないものは、その関連する「話題」を連続して表示し、線でそのまとまりを示す。

4. 実験

1994 年の日経新聞の記事 1 年分に対して、1 カ月毎の頻度について本手法を適用した。キーワード「宇宙」で絞られた 598 件の記事に対して、その分類が「H1: トピック」であるキーワードを用いた、「1 話題＝1 キーワード」の話題抽出の結果を図 1に、拡張モデルに基づく抽出の結果を図 2に示す。

5. おわりに

具体的な有用性の検証は行なっていないが、かなり簡略化した解析方法にもかかわらず、感覚的には、直観的でより理解し易くなったと思われる。さらに、1 つにまとめることで、同じ空間により多くの情報を表示することができるようになった。

今回は、前回と同じような 2 次元的な表示方法を採用したが、「話題」の間の構造をより分かり易く表示するためには、3 次元的な表示が有効であると思われる。また、キーワードの羅列では、その話題の中身が分かりづらい場合がある。これらに適切な名前を付けることも「分かり易さ」を向上させると期待できる。

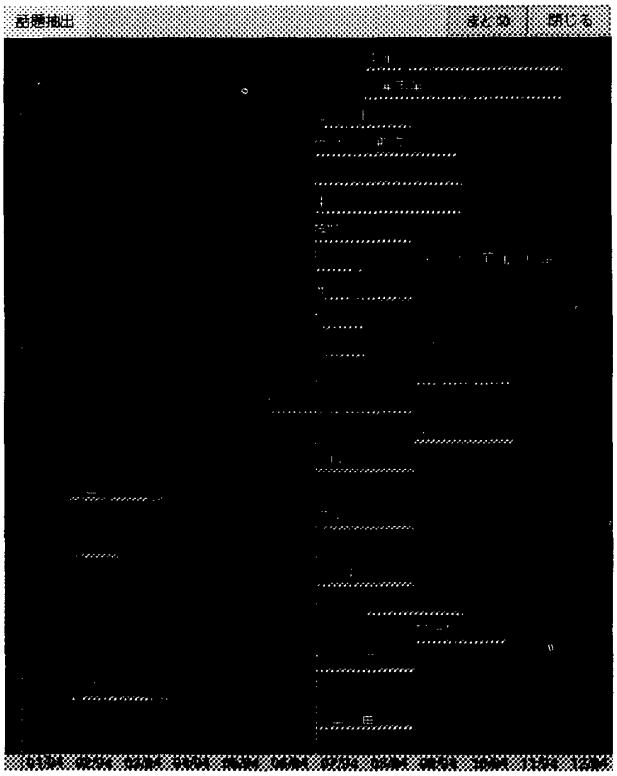


図 1: 「1 話題＝1 キーワード」の話題抽出

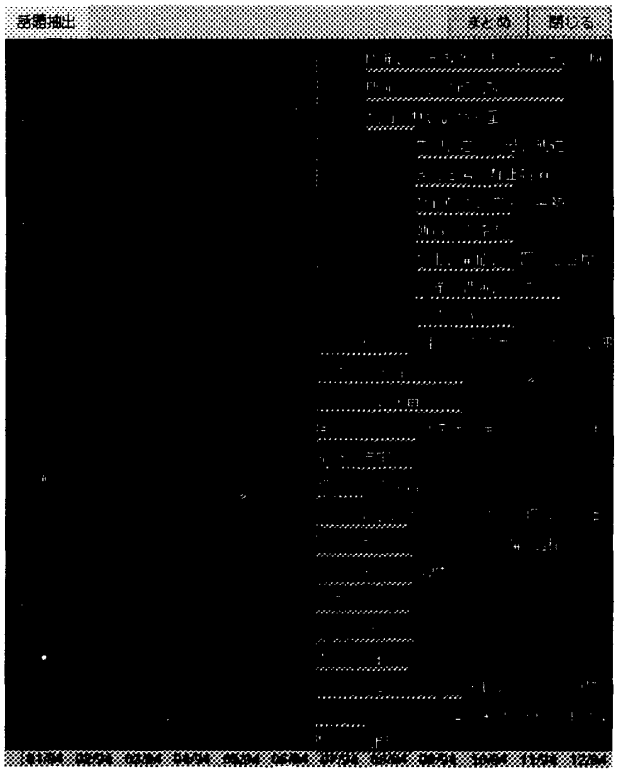


図 2: 拡張モデルに基づく話題抽出

参考文献

[1] 野美山, "新聞記事データベースからの話題の抽出," 情報処理学会第 50 回全国大会,Vol. 4, pp. 45-46, 1994.

[2] Morohashi, M. et al., "Information Outlining - Filling the Gap between Visualization and Navigation in Digital Libraries," Proc. of International Symposium on Digital Libraries,1995.