新聞記事データベースからの話題の抽出*

4F－4

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、新聞記事データベース、あるいは、論文データベースなどの巨大なデータベースが整備され、それらは、非常に有用な情報の宝庫となっている。また、近年のネットワークの発達により、World Wide Web などを通じてネットワーク上の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。

本稿では、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、恒常的な事象を排除し、異質な事象を抽出し、それらをわかり易く提示する手法を提案する。本稿では、特にこの手法を新聞記事システムにおいて適用した例を示す。新聞は、日常の社会のできごとを伝えるものであるが、そこでは、時間を次元とした特異点は一般に「話題」と呼ばれる。

2　構成

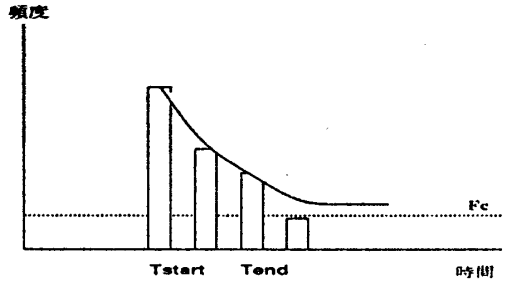

図 1:「話題」のモデル

2.1　アルゴリズム

本手法は、キーワードによって絞り込まれた文書集合に対し、それらに付与された全キーワード集合の中から、(1)「話題」となるキーワードを抽出し、(2) それらに対し話題性の計算を行ない、(3) それらの結果を話題性の高い順に表示する。

本システムの、基本となる検索機構は、キーワード検索であり、キーワードでの絞り込みは、キーワードインデックスを用いて行なわれる。絞られた文書集合に付与されたキーワードの取り出しは、データベースを直接検索するか、あるいは、文書 ID とそれに付与されたキーワードに対してインデックスを作成し、それを用いる。また、文書 ID に対する時間を求めるには、インデックスを用いる。

話題は、あるキーワードによって代表され、ある特定の期間に集中して出現し、その出現頻度が時間に反比例して減少するものと仮定する。本手法では、そのキーワードの出現は図 1に示すようなパターンを持つものとする。図 1は、X 軸が時間、Y 軸がキーワードの出現頻度である。

あるキーワードの出現が話題であるかどうかを判断するためには、本手法が仮定する話題のモデルとの距離が、ある値より小さい場合、それを話題であると判断する。

話題の抽出　まず、最初に話題であるかどうかと、その範囲を決定する。話題の範囲を決定するために、まず、その期間で恒常的に生じている頻度（恒常的出現頻度:Fc）を推定する。これは、その期間の中で、最も少ない頻度と 2 番目に少ない頻度の平均で代表させる（図 1の破線がこの値を表す）。次に、その期間の中で、恒常的頻度に対して最も大きな頻度を持つ時間を求め、それを話題の開始時間 (Tstart) とする。

次に、話題の終了時間 (Tend) を求める。話題の終了時間は、話題の開始時間から、単調に減少している範囲で、かつ、恒常的出現頻度よりも大きな出現を持つ最大の時間である。頻度が増加した場合、あるいは、頻度が恒常的出現頻度より小さくなった場合は、その直前の時間が話題の終了時間となる。

このようにして、抽出された単調減少している区間と、本手法が仮定する話題のモデルとの距離を以下の式で求める。

```
distance2 = 0;
for time = Tstart to Tend;
  if freq(time) > モデル推定量 then do;
   distance2 = distance2 + 「モデル推定量との差」の自乗;
  end;
end;
距離 = square_root(distance2);
```

モデル推定量は以下の式で表される。

$$\text{モデル推定量}(time) = (freq(Tstart) - Fc) * (time - Tstart) + Fc$$

このようにして計算されたモデルとの距離がある閾値より小さいものを話題として識別する。

話題性の計算　識別された話題の重要度を判断するために、話題性という尺度を導入する。話題性は、話題の期間の出現頻度と、その頻度が全体の期間のどのくらいの割り合いを占めるかという観点に基づいて以下の式で計算される。

話題性 = 話題の期間の出現頻度*
(話題の期間の出現頻度/全期間での出現頻度)

*Topics Extraction from Newspaper Databases
†Hiroshi Nomiyama
‡IBM Research, Tokyo Research Laboratory

抽出されたすべての話題について、話題性を計算しそれらを話題性の降順に並べ、横軸を時間とし、話題性の大きな順にその期間をバーで表示する。図2に例を示す。

3 実験

3.1 話題の抽出

日経記事データベース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結果を示す。

キーワード「殺人」を入力し、絞られた1095件の記事に対して、本手法を適用した結果を示す。まず、キーワード「殺人」で絞られた記事集合に付与されているキーワードの中から、その分類が「H1:トピック」であるものを頻度順に上位30語を表示する。最後の欄は、その頻度を表す。

1	容疑者	308	
2	殺人事件	191	
3	県警	187	
4	無職	176	
5	会社	170	
6	捜査本部	162	
7	地裁	152	
8	死亡	143	
9	刺殺	135	
10	公判	129	
11	会社員	118	
12	遺体	100	
13	死体	100	
14	自宅	91	
15	殺人未遂	85	
16	強盗殺人	81	
17	死体遺棄	78	
18	現場	75	
19	殺害事件	73	
20	判決公判	71	
21	名古屋	71	
22	殺人容疑	63	
23	射殺	63	
24	信金OL誘拐殺人	58	
25	職員	57	
26	犯人	56	
27	被害	55	
28	特捜本部	54	
29	住所	53	
30	被害者	52	

これらのキーワードから、本手法によって話題と認定されたものを重要度順に上位30語を示す。それぞれの欄は、順位、キーワード、話題性、Tstart(最初の月、0は、1992年12月、その他は、1993年の月)、Tend(最後の月)を表す。

```
 1 信金OL誘拐殺人 56.017241 8 10
 2 信金OL誘拐殺害事件 41.490196 8 10
 3 容疑者 39.285714 8 10
 4 文民警察官 37.209302 5 6
 5 県警 33.374332 8 10
 6 信金OL 33.230769 8 10
 7 合同捜査 32.818182 8 10
 8 捜査本部 22.969136 8 9
 9 誘拐殺人 22.781250 8 10
10 殺害事件 20.835616 8 10
11 犯人 12.071429 8 9
12 職員 11.859649 8 9
13 被害者 11.076923 8 10
14 刺殺 10.140741 0 2
15 殺人事件 9.235602 8 9
16 住所不定 8.820000 3 6
17 殺傷事件 8.257143 5 7
18 被害 8.018182 8 9
19 自宅 8.010989 8 9
20 判決公判 6.816901 3 6
21 殺人未遂 6.223529 0 1
22 遺体 5.290000 8 9
23 公判 5.240310 3 4
24 住所 4.830189 3 4
25 同僚 4.500000 8 9
26 両親 4.500000 0 2
27 地裁 4.447368 3 3
28 関係者 4.363636 8 10
29 高裁 4.355556 4 6
30 設置 4.170213 7 9
```

「殺人」から容易に想像されるキーワード(「容疑者」、「殺人事件」、「死亡」など)は、順位が下がっている。一方、キーワードの出現頻度では24位だった「信金OL誘拐殺人」が話題性では1位に、また、出現頻度では、39位だった「文民警察官」が話題性では4位となっている。

3.2 結果の視覚的表示

話題抽出の結果を表示したものを図2に示す。

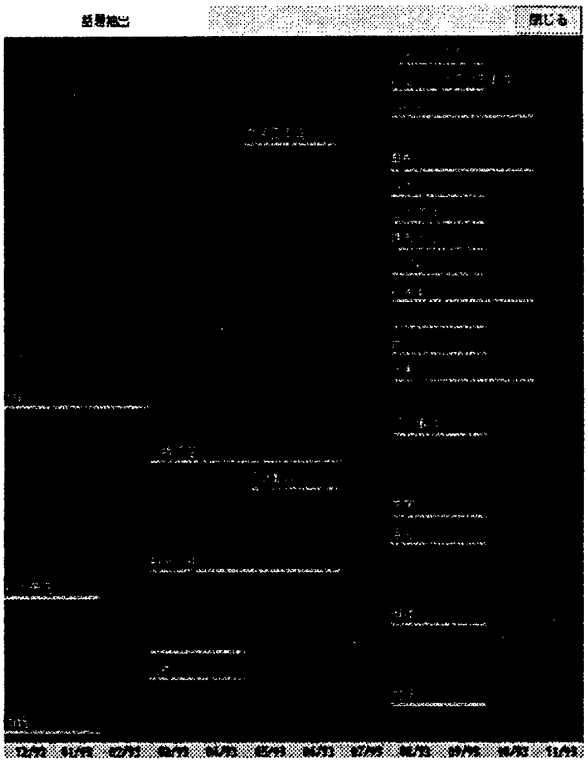

図2:「話題」の表示

4 おわりに

本手法では、対象となる期間全体に渡って「話題」となっているものを抽出することができない。例えば、ある首相の任期期間中に「首相」というキーワードで話題を抽出した場合、おそらく時の首相は、全期間に渡って出現しているため話題として認識されないであろう。この問題は、話題の対象となる期間を広げることによって解消できるが、抽出された話題のほとんどは、時の首相に関連しているので、他の方法で容易に識別できると考える。

また、本稿では、時間的な特異点(「話題」)のみを求めたが、他の次元の特異点(位置座標など)を求めることによっても、有効な情報が得られると予想される。

参考文献

[1] 堤他, "電子図書館 I - IV," 情報処理学会第49回全国大会, Vol. 4, pp.209-216, 1994.