WWW検索サービスにおけるトレンド語抽出

4T－3

河合 英紀　赤峯 享
NEC
ヒューマンメディア研究所

1. はじめに

WWW 検索サービスにおいて、検索に利用された単語は、利用者の情報ニーズを明示的に表わしている。利用者の情報ニーズは非常に多様であるが、とりわけ、事件やイベント、あるいは季節性を反映した情報ニーズを把握することは、新鮮なサービスを提供するために重要である。そこで本稿では、検索ログに出現する単語のうち、事件やイベント、および季節性を反映した情報ニーズを表わす単語をトレンド語と定義し、単語の利用頻度の変化を基にトレンド語を抽出する方法を提案する。さらに、実際の検索ログを用いた評価結果を報告する。

2. トレンド語抽出法

情報ニーズの強さは、単語の利用頻度に比例する。ある話題に関する情報ニーズを測定する方法として、検索ログに出現する単語の利用頻度と、各単語の利用された時間間隔を基に単語の関係を求める研究がある[1]。しかし、単純に利用頻度で単語を順序付けすると、WWW における日常語（「チャット」「ダウンロード」,etc.）など、トレンド性に乏しい単語が上位を占めることになる。

一方、何らかの理由で、ある単語がトレンド語となった場合、短期間に多くの利用者がその単語を検索に使うため、その利用頻度は激しい増加を示すはずである。図 1 に、単語利用頻度の時間変化の例を示す。急激な増加を示す単語を抽出する方法として、一定期間（図1のT0～T4）における平均利用頻度 P0 と、評価対象期間（図 1 の T3～T4）における単語の平均利用頻度 Px の比 Px/P0 を指標とする方法がある。しかし、この方法では T0～T4 の期間にいくつかのピークが含まれる場合、基準となる平均値 P0 が増加してしまうため、その単語をトレンド語として抽出するのは困難である。

そこで、本稿では、T0～T4 の期間において、利用頻度の時間変化が最も安定した期間（図 1 の T1～T2）を基準期間と定め、基準期間での平均利用頻度 Pb と、評価対象期間での平均利用頻度 Px の比 Px/Pb をトレンド語の評価指標として用いることを提案する。これによって、短期間で利用頻度の上下が起きても、基準となる平均値 Pb は変化しないためトレンド語として抽出することができる。

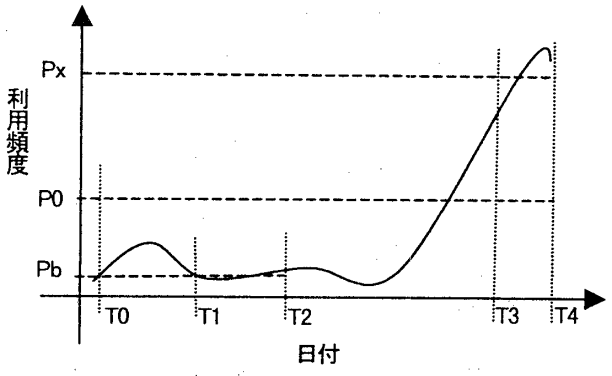

図1 単語利用頻度の時間変化の例

3. 実験方法

本稿で提案したトレンド語抽出法について、実際にBIGLOBE[2]の検索サービス NETPLAZA[3]の検索ログを用い、「ある時点（図 1 の T4）で、最近一週間のトレンド語を、過去 1 ヶ月間の時系列データを基に求める」ことを課題として実験を行った。利用頻度は、IP アドレスとブラウザ名から利用者を区別して計数した。評価期間についての条件を表 1 に示す。基準期間（図 1 の T1～T2）を求めるための安定性の指標として利用頻度の分散を用いた。また、基準期間が短すぎると、「競馬」など 1 週間ごとにピ

A Method of Trendy Words Extraction in WWW Search Engine
Hideki KAWAI, Susumu AKAMINE
Human Media Research Labs., NEC

ークを持つ単語をトレンド語として抽出してしまうため、本稿では基準期間の長さを 2 週間とした。T3～T4 の期間において利用者数が 20 人以上の単語のみを評価対象単語とし、評価対象単語のうち、その単語についての事件、イベント、季節性など、明確な話題が特定できる単語を正解のトレンド語とした。

表 1　評価期間に関する条件

T4	1998/8/21、9/25、10/23、11/27、12/25
T4 - T3	1 週間
T4 - T0	5 週間
T2 - T1	2 週間

以上の評価期間および評価対象単語について、(a)利用者数、(b)Px/P0、(c)Px/Pb をそれぞれ指標として単語を順序付け、再現率-適合率曲線[1]による比較評価を行った。

4. 結果および考察

図 2 に、(a)～(c)のそれぞれの方式における、再現率-適合率曲線を示す。再現率が小さい範囲においては、方式(b)と方式(c)はほぼ同じであるが、再現率が大きくなると、方式(c)が優れていることがわかる。

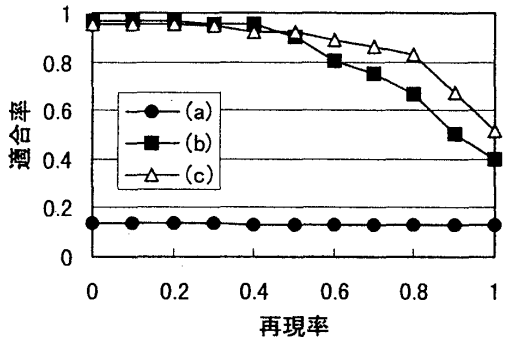

図2 各方式における再現率-適合率曲線

表 2 に、10 月 16 日～10 月 23 日の一週間における評価対象単語 158 語のうち、正解としたトレンド語 16 語、(a)～(c)それぞれの方式において、それが何番目にランク付けされたかを示す。参考までに、それぞれのトレンド語の裏付けとなる話題も記しておく。表 2 より、方式(a)の単純な利用頻度による順序付けでは、トレンド語は下位になってしまうことがわかる。一方、方式(b)と方式(c)では、上位 5 語程度は順位が入れ替わる程度でほぼ同じであるが、台風に関連した単語が、方式(b)では下位に位置づけられているが、方式(c)では上位に順位付けられている点で大きく異なる。これは、9 月 22 日に強い台風 7 号が接近しており、図 1 の T0～T4 の期間で利用頻度の上下が起きたために、方式(b)ではうまくトレンド性を評価できなかったからである。この種の単語の存在が、図 2 の再現率-適合率曲線における差異の原因となっている。

表 2 1998.10.16～10.23 のトレンド語

話題	トレンド語	(a)	(b)	(c)
(*1)情報処理技術者試験 (1998.10.18)	情報処理技術者試験	156(位)	2(位)	3(位)
	情報処理	150	5	4
	情報処理試験	157	3	5
	シスアド	155	7	15
(*2)宅地建物取扱主任者試験 (1998.10.18)	宅建	153	4	2
(*3)英語検定 (1998.10.18)	英検	154	6	9
(*1)(*2)(*3)	解答	158	1	6
プロ野球日本シリーズ (1998.10.18～24)	日本シリーズ	151	11	8
台風10号上陸 (1998.10.17)	台風情報	41	18	1
	台風	31	28	13
	気象庁	136	32	19
	天気	28	92	17
	天気予報	11	87	20
オリジナル着信音のデータ	着信音	147	10	14
	携帯電話	45	21	27
紅葉シーズン	紅葉	129	22	45

5. おわりに

本稿では、事件やイベント、および季節性を反映したトレンド語の抽出について提案し、実際の検索ログを用いて評価実験を行った。その結果、利用頻度が最も安定した期間を基準期間とする方法がトレンド抽出に効果的であることを示した。今回は最近一週間のトレンド語を抽出するために、過去 1 ヶ月間のデータを用いたが、今後、トレンド抽出に適切な評価期間および基準期間の長さについての検討が必要である。

参考文献

[1]大久保, 杉崎, 井上, 田中, "WWW 検索ログに基づく情報ニーズの抽出",情報処理学会論文誌, Vol.39, No.7, 2250 (1998).
[2] http://www.biglobe.ne.jp/
[3] http://netplaza.biglobe.ne.jp/

[1] 本稿では、再現率 = (正解数) / (全評価対象単語中の正解数)、適合率 = (正解数) / (抽出した単語数) とした。