4S-4

時間的要因を考慮した属性獲得手法

波多腰 優斗 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

Web 上の情報を知識源とし，統計量を基に属性を獲得する方法が提案されてきた．しかし，話題語の属性を獲得する場合，逐次増加する記事の影響を受け，一時的に関連が高くなる語を獲得する．本研究では，時間的な影響を考慮した属性の獲得手法を提案する．長期に渡り属性獲得を行い，Web の時間変化が属性獲得に及ぼす影響を調査し，システムを構築する．

2 Web の時間変化により受ける影響

話題語に対して長期に渡り属性獲得を行い，Web を用いた未定義概念に対する属性獲得手法が時間変化の影響をどの程度受けるか調査する．

2.1 話題語の定義

話題 (図 1) とは，社会的認知性と時刻の両軸に対しての大きさを持っており，社会的認知性が増加することによる話題の「広がり」と，その時間的継続である「伸び」があるものとする [1]．話題の「広がり」や「伸び」は，注目度の高い話題ほど大きくなり，Web 上のデータを利用した属性獲得手法に影響を及ぼす．本研究では，このように大事件を伝えるニュースなどの話題の影響を受け，一部の属性との関連が一時的に高まっている概念を話題語と定義する．

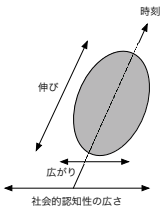

図 1: 話題の定義

2.2 調査方法

話題語 50 語を概念とし，1ヶ月間 (2010 年 11 月 11 日 ~2010 年 12 月 10 日) 属性獲得を行った．1 日あたり 3 時間間隔で 8 回の属性獲得を行っている．獲得した属性の中から 150 個の属性を無作為に抽出し，$tf \cdot idf$ 値，tf 値の時間による変化を調査する．

2.3 獲得した属性の分類

獲得した属性の重みの時間変化の傾向を見たところ，大まかに次の 3 つの特徴を持つ属性に分類された．

- 話題には影響されない属性 (44 個)
- 一時的に話題の影響を受ける属性 (73 個)
- 長期に渡り話題の影響を受ける属性 (33 個)

2.4 統計的な分析

3 つの集団の $tf \cdot idf$ 値の平均と分散，tf 値の平均と分散を求めた (表 1)．表 1 中の値は，集団の中の各属性の値の平均をとったものである．

表 1: 3 つの集団における $tf \cdot idf$ および tf の平均と分散

	$tf \cdot idf$		tf	
	平均	分散	平均	分散
一時的に話題の影響を受ける属性	257.46	44021.18	20.45	116.14
長期にわたり話題の影響を受ける属性	327.29	41837.67	31.09	346.68
話題には影響されない属性	248.62	996.62	18.71	108.76

話題の影響を受ける属性の分散は，話題には影響されない属性の 40 倍以上の値を示した．また，長期にわたり話題の影響を受ける属性の tf の分散は，話題には影響されない属性の約 3.2 倍の値を示した．

3 提案手法

学習頻度，学習期間の 2 つのパラメータを用いて，次のような手順で重み付けを行う．

1. 学習頻度，学習期間を設定する．
2. 学習期間の間は学習頻度に応じて属性獲得を行う．この間は属性の重みを決定せず，値を保持する．
3. 学習期間外は，保持していた値との平均をとり，最終的な重みを決定する．

ここで，学習頻度を f(日) とし，学習期間 T(日) に保持する重み群 W_L は式 (1) のように定義する．

$$W_L = \{W_f, W_{2f}, W_{3f}, \cdots, W_{Nf}\} \text{ ただし, } T \geq Nf \quad (1)$$

式 (1) において，W_f の f 日後の重みが W_{2f} であり，保持している重みの数が N 個である．学習した重み W_L を用いて時刻 t における重み W_t を式 (2) のように定義する．

$$W_t = \frac{1}{N+1}\left(\sum W_L + W\right) \quad (2)$$

ここで，W は時刻 t に属性獲得を試行することによって得られる重みである．

An Acquisition Method of Attributes for Unknown Words Considering Time Factor.

† Yuto HATAKOSHI
† Noriyuki OKUMURA
Nagano National College of Technology, Department Electronics and Computer Science, noriyuki_okumura@ei.nagano-nct.ac.jp (†)

4 評価実験

Web を用いた未定義概念に対する属性獲得手法で話題語の属性を獲得する際に，次のような問題点を解決できない．

- Web の時間変化による属性の重みの変動を抑えたい場合
- 話題の影響を受ける属性を概念を特徴付ける属性として残したい場合とそうでない場合

これらを解決する上で，学習頻度 f，学習期間 T のどのような設定が有効であるか調べる．以下，前者の問題点を解決するための実験 (実験 1) と後者の問題点を解決するための実験 (実験 2) について述べる．

4.1 実験 1

学習頻度 f，学習期間 T にどのような値を設定すると，重みの分散を抑えることができるかを調査する．評価には，一時的に話題の影響を受ける属性と長期にわたり話題の影響を受ける属性を用いる．

表 2 は，各パラメータにおける $tf \cdot idf$ 値の分散を示している．最低でも 2 回の学習を行うと，話題の影響を受ける属性の分散の約 1/17 以下の値となった．話題には影響されない属性の分散は表 2 より 996.62 であるので，表 2 の $f = 6, T = 18$ から下のパラメータは良好である．3 回以上の学習回数があれば，話題には影響されない属性程度まで重みの変動を抑えられることがわかった．

表 2: 提案手法による $tf \cdot idf$ 値の分散

学習頻度 f(日), 学習期間 T(日)	分散
$f = 12, T = 18$	2384.5
$f = 6, T = 12$	1110.01
$f = 6, T = 18$	858.42
$f = 1, T = 9$	154.07
一時的に話題の影響を受ける属性	44021.18
長期にわたり話題の影響を受ける属性	41837.67
話題には影響されない属性	996.62

4.2 実験 2

話題の影響を受ける属性を，概念を特徴付ける属性として残したい場合とそうでない場合の 2 パターンに応じた重み付けが可能なことを示す．評価には，一時的に話題の影響を受ける属性 (50 個を抽出) を用いる．

T の値を変化させ，$tf \cdot idf$ 値の平均がどのように変化するか調べた (図 2)．縦軸が $tf \cdot idf$ 値，横軸が学習期間 T である．また，学習頻度 f の値は，学習回数が 3 回となるように設定した．ここでの $tf \cdot idf$ 値は評価に用いた 50 個の属性の $tf \cdot idf$ 値の平均である．学習期間が長いほど $tf \cdot idf$ 値は小さな値をとり，学習期間が 3 日のときに最大であった．$T = 3$ と設定した場合，一時的に話題の影響を受ける属性の $tf \cdot idf$ 値の平均の約 1.4 倍の値となった．

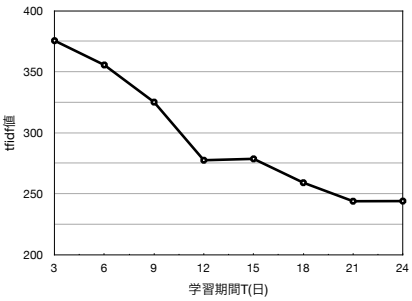

図 2: 学習期間 T と $tf \cdot idf$ 値の関係

5 考察

学習回数が 3 回以上になるようにパラメータを設定することで，話題の影響を受ける属性の重み付けを平均的に安定して行えることを示した (表 2)．しかし，データの分布には大きな違いが見られた．表 3 は表 2 の $f = 6, T = 18$ の場合と，話題には影響されない属性の分散値の分位数である．

表 3: 分散値の分位数

	$f = 6, T = 18$	話題には影響されない属性
最小値	0	0
第 1 四分位数	18.32	394.0
中央値	53.76	651.06
第 3 四分位数	164.50	1313.00
最大値	51310.00	3523.00

$f = 6, T = 18$ の場合，第 3 四分位数までは小さな値を取っているが，最大値は話題には影響されない属性の分散を大きく上回っている．これは，パラメータ設定にが有効に作用した場合と，全く作用しない場合に分かれたことを示している．表 3 の話題には影響されない属性の分散の最大値である 3523 を超えたものは，評価に用いた 106 個の属性の中に 3 個あった．これらの共通点は，$tf \cdot idf$ 値は最大で 2000 を超えていたことである．4 番目に分散が大きい属性の $tf \cdot idf$ 値は最大でも 600 程度であり，$tf \cdot idf$ 値の時間変化による変動が大きすぎたことがパラメータ設定が有効に作用しなかった原因と考えられる．

6 おわりに

本研究では，Web 上の情報を利用した属性獲得手法が，Web の時間変化による影響を受けると仮定し，実際に 1ヶ月間に渡り調査を行い，問題点を挙げた．話題の影響を受ける属性の重み付けの手法として，学習頻度 f，学習期間 T を設定し，複数回の試行から重みを洗練させることを提案した．

参考文献

[1] 「時系列ニュース記事における最新話題語抽出方法」NTT サイバーソリューション研究所: 情報処理学会, 2005.