1P－03

決定木を用いた単語の流行性推定

岡田真*，安藤一秋**，獅々堀正幹*，青江順一*
*徳島大学工学部知能情報工学科
**香川大学工学部信頼性情報システム工学科

1 はじめに

文書データ中の単語には，時間経過に伴って出現頻度が変化するものや，ある特定の期間に出現しやすいものが存在する[1]．このような時系列変動が観察される単語は，各時点での文書作成者の興味を表している場合が多いため，文書の特徴を表す重要語となりやすい．しかしながら，従来の文書処理や文書検索技術などでは，単語の重要度を決定する際に単語の時系列変動を考慮に入れていなかった．

本研究では，過去の文書内での単語の出現頻度変化に基づき，時間軸上での単語の流行性を決定木[2]によって推定する手法を提案する．

2 時系列変動を考慮した単語の流行性

単語の時間軸上での流行性を判定するための指標として，時間経過における安定性という観点から，以下の3つのクラス（安定性クラス）を定義する．

1）新出：時間経過につれて頻度が増加
2）安定：時間経過しても頻度は安定
3）衰退：時間経過につれて頻度が減少

各クラスに属する単語をそれぞれ新出語，安定語，衰退語と呼ぶ．

3 安定性クラス決定に用いられる属性

単語の頻度変化の特徴を定量的にとらえるために，1）回帰直線の傾き，2）回帰直線の切片，3）相関係数，4）2直線間の角度，5）単語の品詞の5つの属性を定義する．なお，ここで定義した各属性の値が，決定木学習の入力データとなる．以下，各属性について説明する．

3.1 回帰直線の傾きと切片

回帰分析は統計的手法の一つであり，二次元直行座標系上の標本値の変化を直線で近似する手法である．その近似した直線を回帰直線と呼ぶ．

回帰直線の傾きより，単語がどのクラスに属するかを判定する．傾きが正ならば新出語，負ならば衰退語，0 に近ければ安定語となる．切片からは，同クラス内での単語の重要性が判定できる．

3.2 相関係数

相関係数は，回帰直線から実際のデータがどの程度ばらついているかを示す値で，回帰直線の信頼性を判断するために用いられる．相関係数の絶対値が1に近いほどばらつきが少ないことを示す．

3.3 2つの回帰直線間の角度

単語の中には，全体としては衰退しているにもかかわらず，全てのデータから求められた回帰直線の傾きが0に近くなるため，安定語と判定されてしまう事例が存在する．この問題を解決するために，最初と次の年の回帰直線と最新のデータまで考慮した回帰直線の2直線間の角度を用いる．

3.4 品詞

本研究では，時間経過の影響が表れやすく，安定性クラスのの特徴が得られやすいと考えられる品詞6種類（普通名詞，固有名詞，組織名，職業・地位，人名（姓），人名（名））に絞って判別をおこなう．

4. 評価

毎日新聞 CD-ROM（1994年から1998年）に登録されている見出しキーワード“プロヤキュウ”を持つ記事から，3.4で定めた品詞を持つ単語を抽出し，人間が修正およびクラス付けをおこない，実験用データとした．表1に実験用データを示す．

それらを学習データとテストデータに分けて決定木学習をおこない，その決定木を用いてテストデータをクラス分けした結果を人間が付けたクラスと比較して評価した．実験結果の評価には，正解率，再現率，F-measure を用いた．

Estimating the Trend of Words by Using the Decision Tree, Makoto Okada, Kazuaki Ando, Masami Shishibori, Jun-ichi Aoe.
* University of Tokushima
** Kagawa University

表1　実験用データ

Table.1 Evaluation data.

		期間	総数	新出語	安定語	衰退語
学習用データ	A	94-97	1069	35	852	182
	B	95-97	1016	46	778	192
	C	96-97	951	46	756	149
テストデータ	X	94-98	796	37	675	84
	Y	95-98	796	37	675	84
	Z	96-98	797	47	661	89

$$\text{正解率（P）} = \frac{\text{システムが抽出できた正解単語数}}{\text{システムが抽出した単語数}}$$

$$\text{再現率（R）} = \frac{\text{システムが抽出できた正解単語数}}{\text{人間が決めた正解単語数}}$$

$$\text{F-measure} = \frac{2 \times P \times R}{P + R}$$

まず，属性の組み合わせを変えて精度を求めた．学習データはAを，テストデータはXを用いた．表2に属性の組み合わせを，図1に属性の組み合わせと分類精度の関係のグラフを示す．

表2　属性の組み合わせ

Table.2 Combinations of attributes.

1	a	傾き，切片，相関係数
	b	傾き，切片，角度
	c	傾き，切片，品詞
2	a	傾き，切片，相関係数，角度
	b	傾き，切片，相関係数，品詞
	c	傾き，切片，角度，品詞
3		傾き，切片，相関係数，角度，品詞

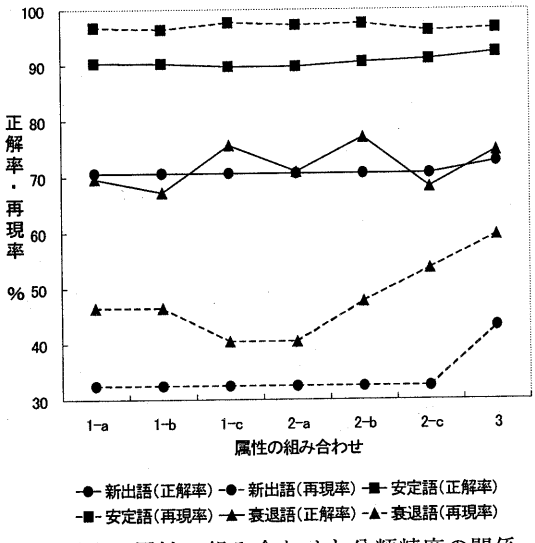

図1 属性の組み合わせと分類精度の関係

Fig.1 Relationship between the various periods of times and the classification precision.

全属性を使用した時の実験結果が最良となることが分かった．また，“品詞”を含む場合は精度が高くなることが分かった．これにより，定義した属性が有効であることが確認できた．

次に，学習期間変化時の実験結果を表3に示す．ハイフン（-）は値がないことを示す．

表3　学習期間と分類精度の関係

Table.3 Relationship between the various periods of times and the classification precision.

		正解率	再現率	F-measure
(C, Z)	新出	73.91	36.96	0.493
	安定	85.29	99.09	0.917
	衰退	−	−	−
(B, Y)	新出	68.97	54.05	0.606
	安定	90.17	95.11	0.926
	衰退	56.36	36.90	0.446
(A, X)	新出	72.73	43.24	0.542
	安定	92.22	96.59	0.944
	衰退	74.63	59.52	0.662

学習用データの期間が長くなるにつれて，全体的に推定精度が向上している．その傾向は衰退語によく現れている．これは，学習期間が長くなるにつれて単語数が増加し，クラスの特徴を明確に把握できるようになるからだと考えられる．学習期間が短い場合に分類精度が低下するのは，相関係数と2直線間の角度が有効に機能しないためと考えられる．

5. 終わりに

単語の流行性の尺度として安定性クラスを設定し，単語の頻度変化を定量的にとらえるために属性を定義した．過去の文書データから抽出した属性値を決定木学習させて，単語の安定性クラスを自動的に推定する手法を提案した．実験により，属性数最大，学習期間最長の時，分類精度が最良となることを確認した．今後の課題としては，今回用いたプロ野球以外の分野の文書についての実験と，実際の文書処理システムに組み込んでの有効性の確認などが考えられる．

参考文献

[1] 久野雅樹：新聞の用字の面による変動と時系列変動，自然言語処理，Vol.7，No.2，pp.45-61（2000）．

[2] Quinlan, J.R.：C4.5: Programs for Machine Learning，Morgan Kaufmann Publishers（1993）．