E-12 ユーザの趣味・嗜好を汲み取ったブックマーク情報の動的分類

Dynamic Classification of Bookmark Information, Considering Interest of a User

柴田 祐助* Yusuke Shibata
村岡 洋一* Yoichi Muraoka

1 はじめに

本稿では、ユーザの趣味・嗜好を汲み取り、動的にブックマーク情報を分類する手法を提案する。

インターネットから入手した情報を効率よく分類・管理する手法が必要となっているが、ユーザの趣味・嗜好に因るところが大きいWebページの閲覧では、ブックマーク情報の分類基準にユーザの主観を取り入れる必要があると考える。

従来のテキスト分類の手法 [1] では、動的に分類階層を生成しているが、ユーザの主観は取り入れておらず、同じテストデータ群に対しては全く同じ分類結果を示す。

本稿での手法は、Webページの主要な構成要素であるテキストデータを解析し、出現する名詞の出現頻度を要素とする特徴ベクトルという値を、ブックマークしたWebページや分類するカテゴリに適用し、これをユーザの意図によって順次修正することでユーザの趣味・嗜好を汲み取る。この手法では、同じテストデータ群に対しても、ユーザの意図によっては違う分類結果を示す。

本システムを搭載したブラウザソフトを実際にユーザに使ってもらい評価を行ったところ、本手法によってユーザの趣味・嗜好を汲み取ったブックマーク情報の動的分類が実現できることがわかった。

2 提案する手法

本稿で構築するシステムの処理の流れと、具体的な手法について述べる。

2.1 処理の流れ

構築するシステムの処理の流れを以下に示す。

1. ユーザがあるWebページにブックマークする。
2. ブックマークしたページの特徴ベクトルと既存のカテゴリの特徴ベクトルを照合する。
3. 初期操作時や、ブックマークしたページと類似するカテゴリが存在しない場合は、ユーザが新規カテゴリを作成し、そのカテゴリにブックマークしたページを登録し、カテゴリ特徴ベクトルを計算する。
4. ブックマークしたWebページの特徴ベクトルと既存のカテゴリの特徴ベクトルが類似する場合は、ユーザに類似するカテゴリに分類してよいか確認をとる。もし、ユーザの意に反する場合は、ユーザは別のカテゴリか、または新規のカテゴリを指定する。
5. カテゴリに含まれるすべてのページからカテゴリ特徴ベクトルを再計算する。
6. 以上の操作を繰り返すことにより、ユーザの趣味・嗜好を汲み取ったブックマーク情報の動的分類が可能となる。

2.2 形態素解析

ユーザがブックマークしたWebページのHTMLソースからテキストデータのみを取得し、形態素解析器にかけ、不要語を除いた名詞を抽出する。ここでの形態素解析には、形態素解析システム「茶筌」[2] を使用した。

2.3 ページ特徴ベクトルの作成

抽出した名詞の出現頻度 tf をTF法 [3] を用いて計算し、ページに出現する全ての名詞についての tf を要素とするベクトルをページ特徴ベクトルとする。

2.4 TF法

TF法とは、文書中の単語の出現頻度を表すものである。文書 d 中に出現する m 種類の単語 $w_1, w_2, ..., w_m$ がそれぞれ $N(d, w_i), (i = 1, 2, ..., m)$ 回出現するとき、文書 d 中の単語 w_i の出現頻度 tf_i を以下のように定義する。

$$tf_i = \frac{N(d, w_i)}{\sum_{w_j \in d} N(d, w_j)}, (j = 1, 2, ..., m)$$

本手法では、Webページの文書の長さに依存しないよう、文書 d 中の単語 w_i の出現数 $N(d, w_i)$ を、文書 d 中に出現する全名詞の出現数の総和で割った値を出現頻度として用いている。

2.5 カテゴリ特徴ベクトルの作成

カテゴリについても同様にTF法を用いて、カテゴリに含まれる全てのページに出現する全ての名詞についての tf を要素とするベクトルをカテゴリ特徴ベクトルとする。

2.6 照合用カテゴリ特徴ベクトルの作成

ユーザがWebページにブックマークした際に、既存のカテゴリ特徴ベクトルの要素のうちブックマークしたページに出現する名詞についての tf を抜き出し、照合用カテゴリ特徴ベクトルを作る。

このとき当然ブックマークしたページには出現するが、カテゴリには出現しない名詞が存在する可能性があるが、その場合については、$tf = 0$ とする。

この操作は次の特徴ベクトルの照合のために、2つのベクトルの次元を揃えるために行っている。

2.7 特徴ベクトルの照合

ブックマークしたWebページの特徴ベクトルと照合用カテゴリ特徴ベクトルを、ベクトル空間モデル [3] を用いて照合する。

ブックマークしたWebページに m 種類の単語が出現し、これらの単語 $w_1, w_2, ..., w_m$ の出現頻度がブックマークした

* 早稲田大学大学院理工学研究科

Web ページで $tf_{p,1}, tf_{p,2}, ..., tf_{p,m}$ であり、照合するカテゴリで $tf_{c,1}, tf_{c,2}, ..., tf_{c,m}$ であった場合、ページ特徴ベクトル $\vec{p}$ と、照合用カテゴリ特徴ベクトル $\vec{c}$ は、それぞれ以下のように表すことができる。

$$\vec{p} = (tf_{p,1}, tf_{p,2}, ..., tf_{p,m}), \vec{c} = (tf_{c,1}, tf_{c,2}, ..., tf_{c,m})$$

このとき、$\vec{p}$ と $\vec{c}$ の偏角 θ は以下のようにして求められる。

$$\cos\theta = \frac{\vec{p}\cdot\vec{c}}{|\vec{p}| \times |\vec{c}|} = \frac{\sum_{i=1}^{m} tf_{p,i} \times tf_{c,i}}{\sqrt{\sum_{i=1}^{m} tf_{p,i}^2} \times \sqrt{\sum_{i=1}^{m} tf_{c,i}^2}}$$

この $\cos\theta$ に重み付けを行った値が閾値以上で最大であるカテゴリを、ブックマークした Web ページが属すべきカテゴリとしてユーザに提示する。

3 評価

本稿では、ユーザの趣味・嗜好を汲み取ったブックマーク情報の動的分類を目指しているので、構築したシステムを搭載したブラウザソフト [4] を実際にユーザに使ってもらい、本システムの評価実験を行う。

3.1 評価判定法

ここで、本稿での評価判定法を定義する。

ユーザが Web ページにブックマークし、そのページの特徴ベクトルに類似する特徴ベクトルを持つカテゴリが存在する場合は、それをマッチするカテゴリとしてユーザに提示する。ブックマークしたページを提示されたカテゴリに分類することがユーザの意図通りであった場合は「Good 判定」となり、ユーザの意図通りでなかった場合は「Bad 判定」となる。

また、マッチするカテゴリが存在しなかった場合、ブックマークしたページを登録するカテゴリの指定をユーザに要求するが、このときユーザが新規カテゴリを指定した場合は、余計なカテゴリ提示を行わなかったとして「Good 判定」となり、既存のカテゴリを指定した場合は、ユーザが分類を希望するカテゴリを提示できなかったとして「Bad 判定」となる。

ただし、図 1 の※で示す「Bad 判定」において、ブックマークした Web ページと、ユーザが登録を指定したカテゴリとに解析上何の関連性も見受けられなかった場合は「除外判定」として特別に扱う。

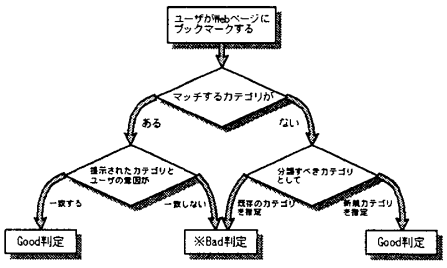

図 1: 評価判定フローチャート

3.2 評価実験結果

インターネットを頻繁に利用しているのべ 16 人のユーザに対し実験を行った。これらのユーザは表 1 の作成したカテゴリの例の通り様々な分野の趣味・嗜好を持っている。

表 1: ユーザが作成したカテゴリの例

サッカー	スロット	JAVA	メタル	落語
猫	日本酒	旅行	Linux	釣り

表 2: 評価実験結果

		全体	最低	最高	平均
Good 判定数		183	5	15	11.4
除外判定数		30	8	0	1.88
除外判定を除外	Bad 判定数	27	2	0	1.69
	Good 判定率	87.1%	71.4%	100%	87.1%
除外判定を含む	Bad 判定数	57	10	0	3.56
	Good 判定率	76.3%	33.3%	100%	76.3%

実験では、これらのユーザに日本語中心のページを 15 ページずつブックマークしてもらい、Good 判定数,Bad 判定数, 除外判定数を計測し、除外判定を Bad 判定として計算しない場合とする場合の Good 判定率を計測した。この評価実験結果は表 2 に示す通りである。表中の最低・最高は Good 判定数が最低数と最高数になるケースについて示してある。

4 考察

表 2 から明らかなように、除外判定を Bad 判定として換算する場合の方が Good 判定率は低いが、除外判定が起こる原因は、ユーザの意図を Web ページに出現する名詞の解析からは汲み取れないことにある。

例えばある人物 A が、友人 B の音楽関連の Web ページと、友人 C のスポーツ関連の Web ページをブックマークし、「友人」というカテゴリに分類したい場合、これら 2 つのページは、人物 A にとって人物 B,C が友人であるということを分類条件として用いているが、このことは形態素解析を行い名詞を抽出して、Web ページやカテゴリの特徴ベクトルを計算している本システムでは分析不可能である。このことが除外判定となる原因であり、分類精度を落としている原因でもある。

以上のことから、本システムはユーザの分類条件が概念的でなく、具体的であるほど高精度の分類が可能であることがわかる。

5 まとめ

本稿では、ユーザの趣味・嗜好を汲み取ったブックマーク情報の動的分類を実現するシステムについて述べた。また、実験において多くのユーザの意図通りにブックマーク情報を分類できることを確認した。

参考文献

[1] 筒井, 福井, 真鍋:分類階層の自動生成機能を備えた文書分類システムの構築, 情報処理学会, 第 62 回全国大会公演論文集 (3),pp.133-134.

[2] 奈良先端大情報科学研究科 松本裕治研究室:形態素解析システム「茶筌」,http://chasen.aist-nara.ac.jp/index.html.ja.

[3] 川前, 青木, 安田:情報理論的モデルを用いた情報検索, 電子情報通信学会技術研究報告,Vol.101,No.192,2001.

[4] ブックマーク情報を自動分類するシステムを搭載したブラウザ,http://yooce.muraoka.info.waseda.ac.jp.