D-5-3

トピック数に基づく Web ページの段階的分類を用いたユーザの興味の推定

Inference of user's preferences using phased classification of Web pages based on number of topics

林　真暢　　福田直樹　　新谷虎松
Masanobu HAYASHI　Naoki Fukuta　Toramatsu SHINTANI
名古屋工業大学 知能情報システム学科
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology

1. はじめに

我々は，ユーザが関心を持った文書の集合をクラスタリングすることによってユーザの興味を推定し，推定された興味に基づく協調フィルタリングによってユーザの情報収集を支援するシステムの開発を行ってきた [2]．文書には複数の話題により構成されるもの（マルチトピック文書）と単一の話題により構成されるもの（単一トピック文書）とがある．一般に，異なる話題では使われるキーワードも異なるため，キーワードの類似性のみに基づくクラスタリングではマルチトピック文書をユーザにとって適切な形に分類できないことがある．この問題は我々のシステムにおいて推薦精度を低下させる原因の 1 つとなっていた．

本稿では，文書に含まれる話題が単一か複数かによってクラスタリング手法を変更することにより，ユーザの興味に沿った文書の分類を実現する手法を提案する．

2. 文書の段階的分類

本研究では，(1) マルチトピック文書を除外し，単一トピック文書のみをクラスタリングする．(2) 形成されたクラスタに基づいてマルチトピック文書を分類する．この 2 段階の分類により，ユーザの興味に沿った分類の実現を目指す．

2.1 マルチトピック文書と単一トピック文書の識別

マルチトピック文書と単一トピック文書とを識別するために，文書の表層的情報を利用する．文献 [1] の手法を参考にして，テキストの n 文目と $n+1$ 文目の間にポイント $point(n, n+1)$ を設け，ポイントごとにパラメータを抽出する．表 2.1 に，パラメータ p_i とその抽出方法を示す．各パラメータには人手で決められた重み w_i が付けられる．パラメータ数を N とするとき，式 (1) で示される値 S の大きい $point$ でトピックが変わると判断される．

$$S = \sum_{i=1}^{N} w_i p_i \qquad (1)$$

2.2 マルチトピック文書の分類

マルチトピック文書は，単一トピック文書により形成されたクラスタに類似する話題がある場合はそのクラスタに所属する．類似する話題とクラスタとの組が複数存在する場合は，該当する全てのクラスタに所属する．類似したクラスタが無いトピックは無視される．トピックとクラスタとの間の類似度は，各々のキーワードベクトル間の類似度で表される．クラスタのキーワードベクトルは，式 (2) で示される重み w_t に基づいて作成される．$idf(t, N)$ は対象文書集合 Y における語 t の idf 値を表す．N_{all} は全単一トピック文書の集合を，$N_{cluster}$ は対象クラスタ内の全文書集合を表す．

$$w_t = \frac{idf_{all}(t, N_{all})}{idf_{cluster}(t, N_{cluster})} \qquad (2)$$

p	分類	抽出方法
1	助詞	「は」の付く文の前
2		「は」の付く文の後
3		「が」の付く文の前
4		「が」の付く文の後
5	接続詞	文頭の「添加」の前
6		文頭の「強調」の前
7		文頭の「説明」の前
8		文頭の「順接」の前
9		文頭の「逆接」の前
10		文頭の「転換」の前
11	前方照応	文頭の「あ」型の前
12		文頭の「こ」型の前
13		文頭の「そ」型の前
14	同一文タイプ	叙述文→叙述文の間
15		判断文→判断文の間
16		断定文→断定文の間
17		その他→その他の間

表 1. 使用するパラメータ

3. Web ページ推薦システムへの適用

従来のシステムでは，ブックマーク及び閲覧履歴から取得した Web ページを区別することなく一度にクラスタリングしていた．しかし，今回は取得した Web ページを単一トピックのページとマルチトピックのページとに分け，前節で提案した段階的クラスタリングを適用するように変更した．

この変更による Web ページ分類の改善例を挙げる．従来の手法では Windows 系のクラスタに分類され Macintosh 系のクラスタには所属しなかった Macintosh ファンによる Windows 批評ページが，Macintosh 系のクラスタにも所属するようになった．Windows や Microsoft という単語が多いために Windows 系クラスタに分類されていた Web ページから，トピック分割により Macintosh に関する記述が抽出されたためと考えられる．

4. まとめ

本稿では文書をクラスタリングにより分類する際に生じる問題を回避するために，マルチトピック文書と単一トピック文書とを区別し，各々に応じてクラスタリング手法を切り替える手法を提案した．

参考文献

[1] 望月, 本田, 奥村, 複数の知識の組み合わせを用いたテキストセグメンテーション, 情報処理学会自然言語処理研究会報告 109 pp.47-54 (1999)

[2] 林, 福田, 新谷, Web ページのクラスタリングに基づくユーザの興味の推定, 第 60 回情報処理学会全国大会講演論文集 (3) pp.3-4 (2000)