2P-4

Web閲覧履歴に表れる単語の連想関係を考慮した情報推薦のためのユーザモデルの生成

福井康俊† 佐藤晴彦† 小山聡† 栗原正仁†
北海道大学大学院情報科学研究科

1 はじめに

近年，Amazon.comやGoogle News等に見られるユーザの嗜好に応じて情報を取捨選択する情報推薦が注目を浴びている．情報推薦をする際に使用されるユーザの嗜好を表現したものをユーザモデルまたはユーザプレファレンスと呼ぶ．ユーザモデルにはユーザが興味持つ事柄に対応する単語を扱うものが多く見られ，通常の情報推薦では，その単語に対する興味の度合を考慮して，ユーザに情報が推薦される．

本研究では，従来の単語に対する興味の度合だけでなく，単語間の連想関係にも着目し，ユーザモデルを生成する．Web閲覧履歴はユーザ固有のものであるから，Web閲覧履歴内のテキストを使用して単語に対する興味の度合や単語間の連想関係の強さを抽出することで，ユーザ毎に異なる嗜好を表現出来る．抽出した単語に対する興味の度合や単語間の連想関係の強さをユーザモデルに反映させることで，より正確にユーザの嗜好を表現することを目的とする．

2 提案手法

提案手法は，ユーザが興味を持つ単語に関するグラフを生成し，ユーザ嗜好をモデル化するというものである．グラフのノードはユーザが興味を持つ単語で，その重みは興味の度合を表現する．エッジは単語の連想関係で，その重みは連想関係の強さを表現する．本研究ではユーザ毎に異なる嗜好を表現するために単語を扱うのでコンテンツベースフィルタリングを対象とした．

2.1 単語重要度

ユーザが興味を持つ単語をWeb閲覧履歴から抽出するためにTF・IDF法を用いる．これは文書中における特徴的な単語に対して，高い重要度を与える手法である．ある文書集合が存在するとき，数多くの文書で出現頻度が高い単語の重要度は下がり，特定の文書中にのみ出現頻度が高い単語の重要度を上がるという性質を持つ．情報検索の分野において，文書の代表語を抽出する際に用いられる手法の一つでもある．

$$tfidf(w_i) = \frac{n_i}{\sum_j n_j} \cdot log\frac{N}{|\{d : w_i \in d\}|} \quad (1)$$

ここで n_i は 文書における単語 w_i の出現回数，N は全文書数，$|\{d : w_i \in d\}|$ は単語 w_i を含む文書数である．

2.2 連想関係

連想はある事物から別の事物の情報を記憶の中から想起することで，当人の過去の経験により，何を連想するかは人によって異なる．つまり，一般的に連想される事物がすべてのユーザに当てはまるものではなく，各ユーザに固有の連想関係がある．

本研究では，文という短いセグメントに出現する各単語間は連想関係が強いと定義する．Web閲覧履歴中のあるWebページ本文を適当な区切りで分割し，文(セグメント)集合 S を作成する．文集合 S の各文 s に対し，以下の式で定義される $SCORE(s)$ を与える．

$$SCORE(s) = \sum_{w_i \in s} tfidf(w_i) \quad (2)$$

$SCORE$ の値が高い上位 N 件の文をその文書の重要文として抽出し，重要文に含まれる単語同士を連想関係がある単語ペアとした．これにより，単語重要度が高く，かつ，共起している単語ペアが連想関係があるとして抽出できる．

2.3 ユーザモデル生成

Webページから抽出した重要文に含まれる単語をノードとした完全グラフを作成する．上位 N 件の重要文からそれぞれ完全グラフを作成し，それらグラフを重ね合わせる．重なった部分のノードとエッジの重みは加算することでノードにはその単語の出現回数を表す値が与えられ，その単語の興味の度合を表現する．また，エッジにはその単語ペアが同じ重要文に同時に出現する回数を表す値が与えられ，閲覧履歴にその重要文集合を含むようなユーザにとっての単語間の連想関係の強さを表現する．これにより1つのWebページからそのページの特徴を表現するグラフを作成する。Web閲覧履歴のすべてでこのグラフを作成し，同様に重ね合わせ，それをユーザモデルとする．

Generation of User Model for Information Recommendation considering Keyword Association in the Web Browsing History
†Yasutoshi FUKUI, Haruhiko Sato, Satoshi Oyama, Masahito Kurihara : Graduate School of Information Science and Technology, Hokkaido University

2.4 連想関係の予測

1つの文には同時に出現しなかったが，ある単語を通じて連想関係がある単語ペアを予測する．本研究では，ユーザモデルがグラフで表現されるため，直接エッジが張られていない単語ペアの連想関係の予測に拡散カーネルを用いた．拡散カーネル K は，グラフの隣接行列 A と，ノード次数の対角行列 D から作られたグラフ・ラプラシアン行列 L で次の様に表現される．

$$\begin{aligned} K &= \frac{1}{Z(\beta)} exp(-\beta L) \\ Z(\beta) &= tr(exp(-\beta L)) \\ exp(-\beta L) &= I - \frac{\beta L}{1!} + \frac{(\beta L)^2}{2!} - \frac{(\beta L)^3}{3!} + \cdots \end{aligned} \tag{3}$$

ここで，βは拡散の度合を調節するパラメータである．拡散カーネルの要素の値は行番号のノードから列番号のノードへの関連の強さを表す．

3 実験

本研究では推薦候補 Web ページ群から Web ページを推薦する実験を行った．推薦候補の Web ページと生成したユーザモデルのグラフとの類似度を計算し，類似度が高い Web ページを推薦するというものである．

3.1 類似度の計算

ユーザに Web ページを推薦するためには，それまでに作成したユーザモデルのグラフと推薦候補の Web ページの特徴を表現したグラフの類似度を計算する必要がある．類似度の計算には，コサイン尺度を拡張したものを使用した．グラフ A，B のすべてのノードを含むノード集合 N を考える．n_i はノード集合に含まれるノードで，ノード集合 N に含まれる2つのノードの重複を含むすべての組合せ (n_i, n_j) を考え、以下の式で計算する．$A(n_i, n_j), B(n_i, n_j)$ はノード n_i とノード n_j に対応するグラフ A, B を表現した行列の要素の値で，対応した要素が存在しなければ0とする．

$$Similarity = \frac{\sum_{i=1}^{|N|} \sum_{j=i}^{|N|} A(n_i, n_j) \cdot B(n_i, n_j)}{\sqrt{tr({}^tA * A)}\sqrt{tr({}^tB * B)}} \tag{4}$$

3.2 使用するユーザモデル

今回の実験では，提案手法によって生成したユーザモデルの評価のために，3つのユーザモデルで Web ページの推薦を行った．1つ目は単語に対する興味の度合のみを表現したユーザモデル (モデル A) で，提案手法によって生成したユーザモデルの対角成分だけを用いて，興味の度合のみを考慮したユーザモデルである．2つ目は提案手法によって単語の連想関係も考慮した生成したユーザモデル (モデル B) である．3つ目はモデル B から拡散カーネルにより，直接エッジが張られていないノード間の連想関係の強さを予測したユーザモデル (モデル C) である．

3.3 結果

モデルの評価には，順位付きの結果の評価に良く用いられる平均適合率を使用する．推薦候補集合を2つ用意して，2回の試行を行った．表1に結果を示す．どちらの推薦候補でも単語の興味の度合のみを考慮したモデル A より，単語の連想関係も考慮したモデル B，C による推薦の方が良い結果を示した．拡散カーネルによって直接張られていないエッジを予測したモデル C は推薦候補1ではモデル B より僅かながら良い結果を出しているが，推薦候補2ではモデル B より低い結果となっている．

	推薦候補 1	推薦候補 2
モデル A	0.734	0.381
モデル B	0.950	0.643
モデル C	0.967	0.480

表 1: 平均適合率

4 おわりに

本研究では Web 閲覧履歴から単語に対する興味の度合と単語の連想関係を考慮したユーザモデル生成手法を提案した．単語に対する興味の度合のみのユーザモデルよりも単語の連想関係も考慮したユーザモデルの方が良い推薦が行えることが判明した．拡散カーネルを用いた連想関係の予測は，良い推薦が行われる場合があることが分かった．今後の課題は，連想関係の予測に関する調査と複数のデータを用いた実験などがあげられる．

参考文献

[1] 神嶌敏弘. 推薦システムのアルゴリズム, 人工知能学会誌, Vol. 22-23, No. 6-2, 2007-2008.

[2] 土方嘉徳. 嗜好抽出と情報推薦技術, 情報処理学会論文誌, Vol. 48, pp. 957-965, 2007.

[3] 渡部勇, 三末, 和男. 単語の連想関係によるテキストマイニング., 情報処理学会研究報告. DD, [デジタル・ドキュメント], Vol. 99, No. 57, pp. 57-64, 1999.