E-006

個人化のための興味の偏りと雑多文書を利用したキーワード抽出手法

Extracting Keywords for Personalization using Deviation of Interest and Ill-chosen Documents

松山 学 † 大囿 忠親 † 伊藤 孝行 † 新谷 虎松 †
Manabu Matsuyama Tadachika Ozono Takayuki Ito Toramatsu Shintani

1. はじめに

近年，ユーザ個人に特化した情報を推薦する，パーソナライゼーションに関する研究が注目されている．従来研究では，(1) ユーザの明示的な嗜好入力，(2) 嗜好の絶対的な評価を必要とするものなどシステムがユーザに入力を求めるものが主である．しかし，ユーザの興味は変化するため，ユーザモデルをそれに応じて変更することはユーザにとって負担となる．また，ユーザが持つ複数の興味をいかに考慮するかが重要である．

本研究室で開発された論文推薦システム [1] では，語の共起グラフにおけるスケールフリー性に基づきユーザモデルを構築することにより，ユーザの興味変化を表現している．本論文では，図 1 のように構築されたネットワークでユーザが持つ複数の興味を表現するため，ユーザの研究活動において収集した論文における語の偏りに基づいたキーワード抽出手法を提案する．本論文では，ユーザモデル構築で用いるキーワード抽出に特化して説明する．複数の興味を捉えたユーザモデルを構築することにより各興味に即した情報推薦が可能となる．

以下，2 章で関連研究と議論し，3 章で本手法の詳細について述べる．4 章で評価を行い，5 章でまとめる．

2. 関連研究

パーソナライゼーションにおけるユーザプロファイル構築に関する研究は数年前から盛んに行われている．PVA(Peorsonal View Agent)[2] では，ユーザの興味を人工生命的な考え方を用いて表現している．複数の興味を考慮すること，そして付加的にユーザプロファイルを更新することを学習によるフィルタリングを用いて解決している．語の重み付けに TF-IDF を用いており，本手法のように重み付けにユーザの閲覧履歴を考慮するものではない．IRM[3] は Web の閲覧履歴より頻出語をユーザに身近な語とし，身近な語との共起の偏りが大きい語をユーザの興味として取り出す手法である．重み付けに閲覧履歴を考慮する点は本手法と類似する．しかし，ユーザの複数の興味を扱うことが出来ない点で本手法とは異なる．この他にもさまざまなシステムが提案されている．しかし，本論文のようにユーザが閲覧，収集した文書群に存在する語の偏りを利用し，ユーザの複数の興味を考慮したキーワードを抽出するものではない．

3. キーワード抽出

本論文では，ユーザの Web 閲覧履歴や研究活動において収集した論文に存在する語の偏りにはユーザの興味と何らかの関連があると仮説を立てている (以後，研究活動を例に説明する)．ここで，ユーザが収集した論文集合をユーザの興味分野と捉えると，ユーザの興味分野には興味分野特有の単語が出現するという仮説 [4] に置き換えることができる．

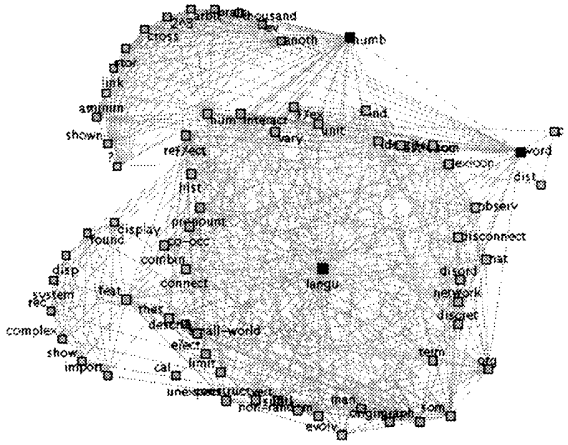

図 1: 語の共起グラフにおけるネットワーク

興味分野中に複数の興味が存在する場合，出現頻度を元に考えると語の偏りは文書数に依存してしまい単一の興味に関するキーワードしか抽出できない恐れがある．そこで，出現頻度ではなく各文書に出現する単語 w_i の出現確率の総和を用いることで興味分野中の語の偏りを平均化させる．頻度平均化は式 (1) により算出される．ここで，N_j は分野 j 中の全文書数，$freq(w_{ik})$ は文書 k 中の単語 w_i の出現回数，n_k は文書 k 中の総単語数を示す．

$$x(w_{ij}) = \sum_{k=1}^{N_j} \frac{freq(w_{ik})}{n_k} \tag{1}$$

$x(w_{ij})$ 値が高い順に並べただけでは，頻出する一般語までもが抽出されてしまうという問題が発生する．また，これだけでは語の偏りを捉えることはできない．そこで，本論文では興味分野とは別に雑多分野を用意する．雑多分野とは，既存に存在する分野から文書をランダムに集めて作った論文集合である．Web 閲覧履歴の場合は，既存サイトに存在するカテゴリからランダムに収集した Web ページ集合となる．興味分野がユーザの興味に偏って収集された論文集合に対して，雑多分野は偏りなく全ての分野が均一に興味あるユーザ，つまり興味に偏りがなく収集された雑多論文集合を表す．

†名古屋工業大学 大学院工学研究科，Graduate School of Engineering Nagoya Institute of Technology

興味分野，雑多分野それぞれの語について，各分野ごとの頻度を標本値とし，帰無仮説として「単語 w_i の出現する確率は興味分野，雑多分野を通じて等しい」と仮定し χ^2 値を求める．ここで，χ^2 値より帰無仮説が採択されれば興味分野に偏りが存在しないと言える．しかし，χ^2 値が十分に大きな値をとるならば帰無仮説が棄却され，「単語 w_{ij} の出現確率はある分野 j にのみ集中して現れる」と言い換えることができる．ここで，ある分野とは雑多分野を偏りのない集合としているため，興味分野となる．つまり，一般的な語はどちらの分野にも出現する可能性が高いため，ここで得られたキーワードは収集論文中の語の偏りを捉えている可能性が高い．χ^2 値は式 (2) により算出される．

$$\chi^2(w_i) = \sum_{j=\{p,p_{not}\}} \frac{(x(w_{ij}) - m_{ij})^2}{m_{ij}} \qquad (2)$$

ただし

$$m_{ij} = \frac{\sum_{j=\{p,p_{not}\}} x(w_{ij})}{N} * N_j$$

ここで，p は興味分野，p_{not} は雑多分野，N は全分野の文書数を示す．本手法では，χ^2 値を検定法として使うのではなく，偏りの程度を示す指標，度合いとして用いている．有意水準の考え方は無視し，式 (1) と χ^2 値による傾きの程度のみを利用する．

4. 評価実験と考察

収集論文中の語の偏りとユーザの興味の関連性を評価するため以下の実験を行った．3 人の被験者にそれぞれ現在研究しているテーマに関連した英語論文を 10 本ずつ合計 30 本収集してもらった．3 つの研究テーマに興味を持っているユーザを想定し 30 本の論文数を変化させキーワードを抽出した．被験者 3 人の研究テーマはそれぞれ被験者 A(オークション，エージェントに関する研究)，被験者 B(テキスト要約に関する研究)，被験者 C(ユーザモデル構築に関する研究) である．なお本実験では雑多分野として AAAI(American Association for Artificial Intelligence) の 1999 年，2002 年の論文合計 511 本を用いた．今回は文書の長さをある程度統一するためアブストラクトの内容のみを対象とする．実験パターンとして，(1) 被験者 A:4 本，被験者 B:6 本，(2) 被験者 B:6 本，被験者 C:3 本，(3) 被験者 A:4，被験者 B:6，被験者 C:3 の 3 つのパターンについて抽出を行った．抽出例として実験パターン (3) において本手法，TFIDF，TF において抽出された上位 10 キーワードと表 1 に示す．また，各手法により抽出された上位 10 キーワードが興味キーワードをどれだけ含んでいるかという割合を表 2 に示す．

表 2 より本手法が TF，TFIDF よりも高い評価が得られた．TF，TFIDF の評価が低かった理由として文書数が少ないものを興味としたときのそれに伴う頻度の減少が考えられる．実験結果よりユーザの収集履歴における語の偏りを利用することでユーザの持つ複数の興味を抽出可能であることが言える．しかし，結果からではどの語がどの興味といった部分までは判別不可能である．本論文では，キーワード抽出に特化して説明したが，図 1 で示したネットワークのノードと本手法で抽出されたキーワードとの関連を調べることで複数の興味に対応し，ユーザの興味変化にも対応したユーザモデルの構築が可能であると考えられる．

表 1: 抽出された上位 10 キーワード

本手法	TF	TFIDF
user	user	topic
interest	system	summary
agent	summary	user
inform	topic	summarization
topic	inform	document
auction	text	text
page	agent	interest
bidder	summarization	page
summarization	document	agent
system	learn	inform

表 2: 各手法の評価

実験パターン	本手法	TF	TFIDF
(1)	0.70	0.40	0.50
(2)	0.60	0.50	0.50
(3)	0.70	0.40	0.60

5. まとめ

本論文では，ユーザの論文収集履歴中に存在する語の偏りを用いたキーワード抽出手法を提案した．実験より語の偏りとユーザの興味との関連性を示した．本手法ではユーザが複数の興味を持つ場合にも万遍なく興味キーワードを取得することができる．今後の課題として，現在提案した手法と語の共起グラフにおけるネットワークとの関連性を調べ複数の興味を表現したユーザモデル構築及び推薦システムへの応用が挙げられる．

参考文献

[1] Satoshi Watanabe, Takayuki Ito, Tadachika Ozono and Toramatsu Shintani, "A Paper Recommendation Mechanism for the Research Support System Papits," DEEC2005, 2005.

[2] Chien Chin Chen, Meng Chang Chen and Yeali Sun, "A Web Document Personalization User Model and System," Workshop on Machine Learning, Information Retrieval and User Modeling, User Modeling, 2001.

[3] 松尾 豊, 福田 隼人, 石塚 満, "ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援," 人工知能学会誌,Vol.18, No. 4E, pp. 203-211, 2003.

[4] 松山学, 新谷虎松, 伊藤孝行, 平岡佑介, 渡邉倫, "論文収集・共有システム MiDoc におけるユーザプロファイル生成のためのキーワード抽出手法," 電気学会論文誌 (部門誌)C, vol.124, no.12, pp.2489-2494, 2004.