収集論文を利用したキーワード抽出に基づくユーザプロファイルの生成について

松山学 † 平岡佑介 † 渡邉倫 ‡ 伊藤孝行 ‡ 新谷虎松 ‡
† 名古屋工業大学 知能情報システム学科 ‡ 名古屋工業大学 大学院工学研究科 情報工学専攻
e-mail: {manabu, hiraoka, watanabe, itota, tora}@ics.nitech.ac.jp

1 はじめに

論文の電子化により，WWW などより広い範囲から研究に関する情報を得ることが可能となった．しかし，自分が必要とする情報を見つけることは困難である．一方で，研究室などの小グループでは，類似した研究を行なうことが多く，必要とする情報も類似している可能性が大きい．つまり，グループにおいて論文共有を行なうことで論文収集活動を効率的に行なうことができる．

本稿では，論文共有においてユーザ同士がどのような興味を持っているかを把握し合うため，収集論文からキーワードを抽出する．その際ユーザの興味を捉えるため，単語の文書頻度を利用する．また，他のユーザにより分かりやすいキーワードを提示するため bigram の統計量を利用し複合語を考慮したユーザプロファイルを生成する．本ユーザプロファイルを利用し論文共有を行うことで，システム利用者の興味を把握し，必要とする論文を効率的に収集することが可能となる．また，システムとして本研究室で開発中の論文収集・共有システム MiDoc を利用する．

本稿では，2 章で文書頻度を用いたキーワード抽出手法，3 章で bigram を利用したユーザプロファイル生成について述べ，4 章で評価を行い，5 章で本稿をまとめる．

2 収集論文における文書頻度を用いたキーワード抽出

本ユーザプロファイル生成手法の流れについて図 1 に示す．本稿では，英語論文に焦点を絞る．まず，前処理として不要語の削除と単語に対して接辞処理を行う．不要語リストには SMART システムで利用されているものを使う．また，接辞処理にはポータ・アルゴリズムを利用する．次に収集論文からユーザプロファイルとして成りうる候補キーワードを選択する．候補

The generation of a user profile based on the keyword extraction using the collection paper
Manabu MATSUYAMA, Yusuke HIRAOKA, Satoshi WATANABE, Takayuki ITO,and Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555, JAPAN.

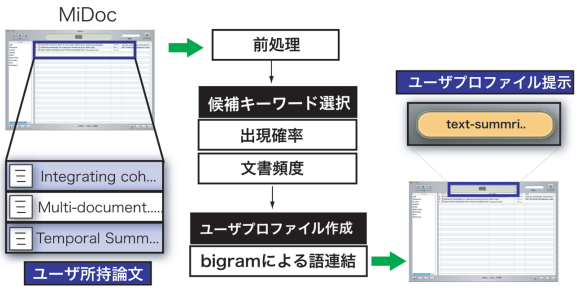

図 1: ユーザプロファイル作成の流れ

キーワード選択に関しては式 (1) を用いて語 w に対して重み付けを行なう．

$$gfp(w) = \frac{1}{N} \cdot \sum_{i=0}^{N} \frac{tf(w)_i}{N} \tag{1}$$

ここで，N は全論文数，$tf(w)_i$ は文書 i における語 w の出現頻度を表す．式 (1) を用いることにより，文書の長短を考慮することが可能となる．重み付けされた語から $gfp(w)$ が高い順に全文書中の語の延べ数 n の 10%の語を候補キーワード群 G とする．

候補キーワード $g(\in G)$ の中には，論文特有の頻出語が含まれることは自明である．ここで，論文特有の頻出語とは，論文の内容語ではなく，論文執筆において必ず含まれるような単語である．例えば，”result”，”document”などがある．論文特有の頻出語を考慮するため，式 (2) ような重み付けを行う．

$$keyword(g) = \begin{cases} gfp(g) & (df(g)/N \leq \alpha) \\ (1 - \frac{df}{N}) \cdot gfp(g) & (df(g)/N > \alpha) \end{cases} \tag{2}$$

式 (2) において α は 0 から 1 の範囲の定数である．また，$df(g)/N$ は語 g の文書全体に出現する確率を表している．つまり，$df(g)/N$ が高い単語はどの文書にも出現している単語であることが分かる．論文特有の単語はどの文書にも出現するため論文特有の一般的な語は $df(g)/N$ も高くなる．しかし，ユーザがある一つの分野に興味をもっており，その分野に関連した論文をたくさん持っている場合，一般的な単語と同様に $df(g)/N$ の値が高くなる．本稿では，bigram を利用した複合語

をキーワードとしているため，分野に関連した単語に対しての考慮は語連結を行う際考慮される．式 (2) によって計算された値を語 g の重みとして候補キーワード群内で順位付けを行う．

3 bigramを用いたユーザプロファイル生成

本稿では，bigram の統計量を用いて複合語をキーワードとして考慮する．bigram を用いたキーワード抽出として文献 [1] などが提案されている．bigram の関連図を図 2 に示す．まず，取り出された語 $g(\in G)$ に対して右連接組み合わせ $bi\_right(g)$ を調べる．ここで得られた連接集合 $R = \{r_1, r_2, ..., r_n\}$ のうち $r_i(\in R)$ の出現回数が 3 回以上のものを取得する[1]．次に右 bigram 頻度 $bf\_right(g, r_i)$ が 0.1 以上のものを候補キーワード群 G 内から捜索し複合語とする．また，右連接組み合わせが上の値を満たしていた場合は，左連接組み合わせに対しても考慮し，左右が連接するかを調べる．ここで，候補キーワード群 G 内に出現回数 3 回以上，$bf\_right(g, r_i) \geq 0.1$ の語が無かった場合は，左連接組み合わせに対して同様の操作を行う．

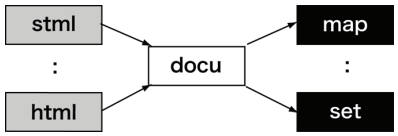

図 2: 左右 bigram の関連図

4 評価

4.1 実験方法と評価方法

本手法のキーワード抽出の精度を評価するために，評価実験を行った．評価実験では，”Agent”，”Auction” に興味があるユーザが執筆した 7 論文と”Bibliography General Summarization Papers”[2]から任意の 8 論文を加えた 15 論文に対してに対して tf と比較した．実験では，各手法でキーワードを 20 個出力し，ユーザが興味を表すキーワードとの precision を見る．ここで，ユーザが興味を表すキーワードとは，あらかじめ著者であるユーザに対して調査したものである．また，実験で抽出される語は単一の語であるため，本実験では，bigram による連結を使わないものとする．次に coverage を調査するため，あらかじめユーザの興味分野を聞いておきその分野に 20 個のキーワードがいくつ含まれるかを調査した．また，ここでは αを 0.5 と設定した．

[1]Apriori 的な手法により，出現回数が 3 回以上の 4-gram までの全てのフレーズを取り出す [2].

[2]要約に関してのリファレンスのみをのせた論文

4.2 実験結果と評価

実験結果を表 1 に示す．precision，coverage ともに tf より高い結果を得た．ここで，注目すべきは coverage である．tf においては”document”，”text”，”system”などの単語が上位に出現することが原因として挙げられる．表 2 より，本手法では”bidder”，”summary”，”auction”などユーザの興味を上位語として万遍なく捉えることができている．また，bigram を利用した結果，順位の最上位に”bidder agent”を抽出することができた．順に，”auction site”，”temporal summaries”といったようにユーザの興味を捉えた結果を得ることができた．以上より，MiDoc において有効なユーザプロファイルが生成されたと思われる．

表 1: 15 論文に対しての precision と coverage

	tf	本手法
precision	0.25	0.35
coverage	0.25	0.55

表 2: 15 論文から抽出されたキーワード (上位 5 個)

順位	$keyword$ 値	頻度	キーワード
1	0.0382	123	bidder
2	0.0240	78	auction
3	0.0192	86	site
4	0.0191	225	summary
5	0.0172	84	negotiation

5 おわりに

本稿では，収集論文から文書頻度を用いてキーワードを抽出し，そこから bigram を用いて複合語を作成しユーザプロファイルを生成する手法を提案した．ユーザが収集した論文を用いるため，それ以外のコーパスを必要としない手軽さが大きな特徴である．本提案手法により，ユーザの興味を捉えたユーザプロファイルを生成できることが確認できた．

参考文献

[1] 湯本紘彰，森辰則，中川裕志：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理研究会報告，2001-NL-145，情報処理学会，(2001).

[2] Fürnkranz, J：A Study Using N-grams Features for Text Categorization，Technical report, Austrian Research Institute for Artificial Intelligence，OEFAITR-98-30，(1998).