1D1-01

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機構の試作とその評価

Implementation of a user profile generation feature
in paper collection and share system MiDoc and the evaluation

松山学[*1] 平岡佑介[*1] 伊藤孝行[*1] 新谷虎松[*1]
Manabu Matsuyama Yusuke Hiraoka Tkayuki Ito Toramatsu Shintani

[*1]名古屋工業大学大学院 工学研究科 情報工学専攻
Graduate School Engineering, Nagoya Institute of Technology

In this paper, we present a new keyword extraction algorithm that can be applied to papers that a user collected for user's research without using a large corpus. We assume that user's interest exists in papers that the user collected. The purpose of this paper is to propose a method for extracting keywords that can express user's interest. Our method can be applied to the paper collection and sharing system, MiDoc. In MiDoc, user profiles are constructed by using the proposed method. We conducted several experiments to show how effectively our method can extract keywords that represent user's interests.

1. まえがき

組織内外から膨大な情報を入手できるようになっている現在，日々蓄積される情報の中から，ナレッジを抽出し，共有・活用することは重要な研究課題となっている．知識共有を支援するシステムの研究・開発は数多く行われているが，そのシステムの多くは，ユーザの知識獲得への負担が大きいことが問題点として挙げられている．特に，ユーザの興味や知識を獲得するためにシステムがユーザに課す負担が大きいことが指摘されている．ユーザの興味や知識を表現した情報は，ユーザプロファイルと呼ばれる．

本論文では，研究室における論文共有に焦点を当てる．研究者は研究活動において論文収集作業を行う．論文収集者は，セッションや論文中のキーワード等を参考に収集作業を行うため，収集された論文には収集者の興味が存在すると考えることができる．つまり，収集論文には収集したユーザの研究に関する興味の偏りが存在すると仮定することができる．本論文では，ユーザが収集した論文を解析することより，収集者の興味を表すキーワードを統計情報を用いた重み付けにより抽出する．ユーザが収集した論文を効果的に利用するため，本研究室で開発中である論文収集・共有システム MiDoc[1] を利用する．MiDoc では，研究室において論文を共有することができる．本手法を MiDoc に組み込むことにより，ユーザの負担は論文を収集し MiDoc において管理するだけであり，ユーザへ負担は極力軽減されると考えられる．

本論文では，2 章で論文収集・共有システム MiDoc の概要，3 章でユーザプロファイル生成機構の詳細について述べ，4 章で評価を記す．

2. 論文収集・共有システム MiDoc

2.1 システム概要

MiDoc とは本研究室で試作する情報共有システムである．システムへの論文の登録や，システムに登録されている論文の閲覧が主な機能となっており，バックグラウンドで動作しているエージェントがダウンロードランキングや論文登録数ランキングを効果的に示し，ユーザに論文の登録や論文のメタ情報の入力，レビューの執筆を啓発することを特徴としている．MiDoc のインターフェースを図 1 に示す．MiDoc のソフトウェアは，iPod^{TM} [*1]のハードディスクにインストールされており，ユーザが iPod^{TM} を計算機に接続することで MiDoc が起動し，システム利用者間で論文情報の共有を行うことが可能となる．

図 1: MiDoc インターフェース

2.2 ユーザプロファイル生成機構

ユーザプロファイル生成機構 (図 2) は MiDoc 内部に Java 言語により実装を行っている．ユーザが MiDoc に論文を登録すると MiDoc 内部で PDF ファイルを自動的にテキストファイルに変換する．変換されたテキストファイルは iPod^{TM} 内部のフォルダに格納される．ユーザプロファイル生成機構では，テキスト化された論文が格納されるたびに，随時新しいユーザプロファイルの自動生成を行う．MiDoc では，論文管理を行う際，研究室内のメンバーがタイトルなどのメタデータを登録

連絡先: 松山学，名古屋工業大学大学院 情報工学専攻 新谷研究室，〒 466-8555 名古屋市 昭和区 御器所町 名古屋工業大学，TEL:(052)735-7968，FAX:(052)735-5477，manabu@ics.nitech.ac.jp

*1 Apple 社の小型軽量の MP3 プレイヤー．外付ハードディスク，および，簡易テキストビューアとしても利用することが可能である．

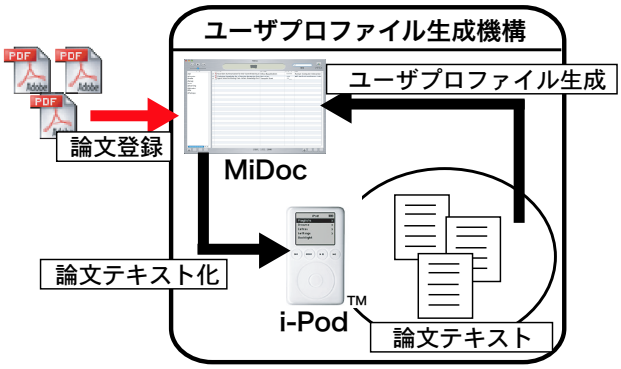

図 2: ユーザプロファイル生成機構

してある場合は自動補完することが可能である．メタデータの自動補完により，論文管理におけるユーザの負担を軽減することができる．論文共有には P2P 型ネットワークを用いており，利用者同士が本ユーザプロファイルやメタデータを参考に双方向で論文をやりとりすることができる．

3. ユーザプロファイル生成手法

本論文では，ユーザが収集した論文中から研究に関するユーザの興味を的確に捉えることを目的としている．そのため，ユーザがいくつかの興味キーワードを持っていた場合，それらのキーワードを高く評価する重み付けを行う必要がある．本手法では，出現頻度と文書頻度を利用した重み付けを行う．以下，図 3 に基づき本ユーザプロファイル生成手法を説明する．

前処理

前処理では，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 [2] 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 [3] を利用する．次に，キーワードを抽出するために利用する以下の値を測定する．

- 出現頻度:$tf(w)$, $gf(w)$
- 文書頻度:$df(w)$
- 右 bigram 頻度と組み合わせ:$bi\_right(w)$, $bf\_right(w, r_i)$
- 左 bigram 頻度と組み合わせ:$bi\_left(w)$, $bf\_left(w, l_j)$

ここで，$tf(w)$ は語 w における単一文書中の出現頻度，$gf(w)$ は語 w における全文章中の出現頻度，$df(w)$ は語 w における文書頻度 (一つの文書中に語 w が一回以上出現する回数)，$bi\_right(w, r_i)(bi\_left(w, l_j))$ は語 w と連続する語群 r_1 ,..., $r_n(l_1, ..., l_m)$ および $bf\_right(w, r_i)(bf\_left(w, l_j))$ は語 w と連続する語 $r_i(l_j)$ との組の頻度を表す．

キーワードに対する重み付け

本手法では，出現確率と文書頻度を組み合わせ重み付けを行う．本手法の特徴として二つ挙げることが出来る．一つは収集論文に含まれるユーザの興味に対して高い重み付けをつけることが出来る点である．もう一つの特徴は，ユーザが幾つかの分野に興味があった場合，万遍なくそれぞれの分野のキーワードに高い重み付けを行える点である．

はじめに，本手法ではユーザプロファイルとしてふさわしい語を効率良く抽出するために，候補キーワード選択処理を行う．本手法では，出現頻度を重要な値と見ている．しかし，文書の長短により語の頻度が変化することを考慮するため語 w に対する出現確率を求める．出現確率 $tfp(w)$ は式 (1) のようになる．

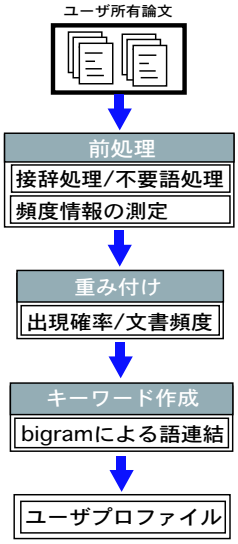

図 3: ユーザプロファイル生成の流れ

$$tfp(w) = \frac{tf(w)}{n_{word}} \tag{1}$$

式 (1) で語 w に関する出現頻度を正規化することによって，語が多い文書と短い文書における語 w の頻度差を無くすことが可能となる．n_{word} は，単一文書中の単語数を示す．次に，本手法ではユーザの興味の偏りを捉えるため全文書においての出現確率を求める．全文書に対する出現確率 $gfp(w)$ は式 (2) のようになる．

$$gfp(w) = \frac{1}{N} \cdot \sum_{i=0}^{N} tfp_i(w) \tag{2}$$

本手法では，式 (2) により重み付けされた語を候補キーワードとする．候補キーワードを選択することにより，ユーザプロファイルとして適した語を絞り込むことができる．候補キーワード選択では，式 (2) が高い語から順に全文書中の語の延べ数 N_{word} の 10%の語を候補キーワード群 G とする．式 (2) において N は全文書数を表す．

候補キーワード $g(\in G)$ は式 (2) により全文書における出現確率を示しているため，収集論文の語の頻度による偏りを捉えることができる．また，頻度情報を利用しているため，候補キーワード群 G 内には，ユーザの興味を表す重要語が含まれると考えられる．しかし，論文中には，論文の内容を表す語でなく，論文執筆において頻出する語も多数存在するため，候補キーワード群 G 内には，論文特有の頻出語も多数存在してしまう．論文特有の頻出語の例として，”result”,”paper”,”document” といった語が挙げられる．ユーザプロファイルとしてこのような語が上位に出現してもユーザの興味を捉えたキーワードとは言えない．そこで，このような論文特有の頻出語を考慮するため，本論文では，式 (3) により候補キーワード群 G 内の語 g に対して再度重み付けを行う．

$$keyword(g) = \begin{cases} gfp(g) & (df(g)/N \leq \alpha) \\ (1 - \frac{df}{N}) \cdot gfp(g) & (df(g)/N > \alpha) \end{cases} \tag{3}$$

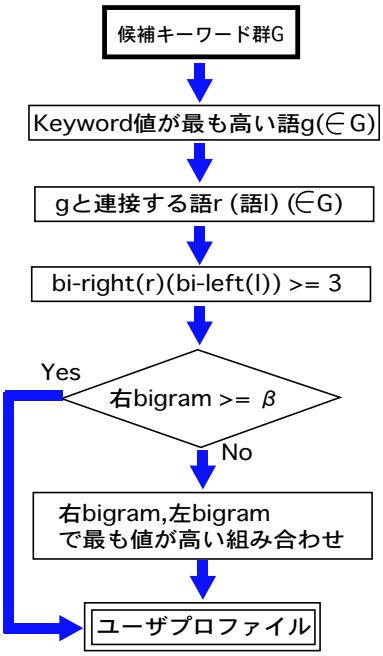

図 4: キーワード作成の流れ

式 (3) において閾値 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．論文特有の語はどの文書にも出現するため $df(g)/N$ の値も高くなる．よって，式 (3) により重みを操作してやることで，論文特有の頻出語が上位にくる問題を解消することができる．また，幾つかの分野に興味がある場合に対しても，式 (3) により，それぞれの分野を捉えた重み付けができる．ここで，ユーザがある一つの分野に興味をもっている場合，論文特有の語と同様に，その分野に関連した語の $df(g)/N$ 値が高くという問題が生じる．しかし，論文特有の語と共に分野に関連した重要な語に関しても重みが操作されるため問題なく適用可能である．

キーワード作成

キーワード作成では，bigram を利用し語連結を行う．語連結では，候補キーワード群 G 内の語に関して行われるため，重みが低くなった重要語に関しても語連結によりキーワードとして考慮することができる．また，複合語を作成するため具体的なキーワードを他のユーザに提示することができる．ユーザプロファイル作成に関しての流れを図 4 に示す．以下，図 4 に基づき流れを説明する．まず，候補キーワード群から式 (3) より得た重みが高いものから順に取り出す．取り出された語 $g(\in G)$ に対して左右の連接語 r, l を候補キーワード群 G より取得する．ここで，語 r, l が $bi-right(r)(bi-left(l)) \geq 3$ かどうかを判断する．ここで，$bi-right(r)(bi-left(l)) \geq 3$ に関しては [Fürnkranz, J][4] を参考にしている．[Fürnkranz, J] では，単一文書中に連続する語が 3 回以上出現する場合，複合語である可能性が高いことを示しいる文献である．本手法では，正確に複合語を取得するため参考にした．右 $bigram$，左 $bigram$ を式 4，5 に示す．

$$左\ bigram = \frac{bfleft(l,g)}{tf(l)} \tag{4}$$

$$右\ bigram = \frac{bfright(g,r)}{tf(g)} \tag{5}$$

次に，右 $bigram \geq \beta$ かを判断し，真が得られれば複合語を作成しユーザプロファイルとする．右 $bigram \geq \beta$ が負であった場合，右 $bigram$，左 $bigram$ で最も値が高いものを複合語としユーザプロファイルとする．

4. 評価実験

4.1 実験

本研究では，収集論文中にはユーザの研究における興味の偏りが存在するという仮説を立てている．本研究における目的は，本手法により収集論文から，興味の偏りを捉えたキーワードを抽出することと，複数の研究に興味がある場合，それぞれの分野を万遍なく捉えたキーワードを抽出することである．そこで，本研究を評価するため，precision と coverage[5] を評価尺度とした．precision は，ユーザに対して予め調査した興味キーワードが抽出されたキーワード中にどれだけ含まれるかを示す値であり，キーワードの精度を表す．coverage は，ユーザに対して興味分野を調査しておき，抽出されたキーワードをユーザに振り分けてもらう．つまり，抽出されたキーワードがユーザの興味分野を幅広く考慮できたかを示す値である．評価実験では，”Auction”及び”Agent”の研究分野に関して研究を行っている被験者が執筆した論文 7 本を収集論文と仮定した場合といくつかの分野に興味があるユーザを想定するため，7 論文に全く別分野である論文を 8 本[*2]加えた場合を考えた．つまり，仮説として，被験者は”Auction”及び”Agent”の研究分野に関して研究を行っているので，”Auction”や”Agent”に関連したキーワードが上位に抽出されるはずである．また，要約関連の論文 8 本を加えてた場合では，”Auction”や”Agent”の分野に加えて”Summary”に関連したキーワードが抽出されるはずである．比較手法としては，tf，$tfidf$[*3]と比較した．各手法でキーワード 20 個を出力し，precision と coverage を測定した．また，閾値は予備実験により $\alpha = 0.8, \beta = 0.1$ とする．

4.2 実験結果と評価

7 本の論文に対して行った実験結果を表 1 に示す．どの手法においても，被験者の研究分野である”Agent”，”Action”に関連した語である”ag(agent)”．”bid(bidding)”や”auct(acution)”を上位に取り出すことができた．本手法では，被験者が定義したキーワード (”biddingbot”) や被験者が執筆した論文において重要な概念を示すキーワード (”negotiate”，”persuasion”) も頻度情報に関係なく上位に取り出すことができた．この結果により，本手法では収集論文がある分野に偏っている場合，研究分野に関連したキーワードだけではなく，論文における重要な語に関しても取り出すことが可能である．次に，7 本の論

表 1: 7 論文に対しての precision と coverage

	tf	$tfidf$	本手法
precision	0.35	0.44	0.45
coverage	0.45	0.60	0.62

*2 要約に関してのリファレンスのみをのせた論文”Bibliography General Summarization Papers”より任意に選択

*3 コーパスは AAAI(American Association for Artificial Intelligence) の 2002 年の論文 50 本とした．また，語 v に対する idf の重み付けは $\log(D/df(w)) + 1$ とした．ただし D は全文書数，$df(w)$ は語 w が出現する文書数とする．

表 2: 15 論文に対しての precision と coverage

	tf	$tfidf$	本手法
precision	0.25	0.34	0.35
coverage	0.25	0.45	0.55

表 3: 15 論文から抽出されたキーワード

順位	重み	頻度	キーワード
1	0.0382	123	bid
2	0.0240	78	auct
3	0.0192	86	sit
4	0.0191	225	summ
5	0.0172	84	negoty
6	0.0158	451	docu
7	0.0139	115	hierarchy
8	0.0127	42	pric
9	0.0117	305	text
10	0.0113	337	top

文に全く別の分野である”要約”を加えた場合の実験結果と本手法により抽出された 10 個のキーワードを表 2 と表 3 に示す．表 2 より，precision と coverage ともに tf，$tfidf$ よりも向上していることがわかる．この理由として，表 3 から tf では”document”，”system”，”text”などの論文特有の頻出語が上位にきてしまい，ユーザの興味分野を上位に捉えることができないことが挙げられる．本手法では，その問題を重み付けにより解消することができたといえる．また，分野に関しては，”auction”，”bid”，”negotiation”，”summary”と被験者の興味分野に加え，要約に関連した分野を均等に取り出せている．この結果から，収集論文に幾つかの興味分野があった場合においても，それぞれの興味分野を均等に捉えたユーザプロファイル構築が可能である．次に，bigram を利用した MiDoc においてのユーザプロファイルの結果を表 4 に示す．表 3 においては”agent”という出現頻度が高く重要な語が上位に取り出すことが出来なかったが，bigram を利用することにより，上位に取り出すことができている．つまり，重み付けに関して発生する誤差を bigram の利用により解消することができることがわかる．本手法により，ユーザの興味を表すキーワード及び，ユーザの研究分野を均等に捉えた結果を得ることができた．実験結果より，収集論文が単一分野に偏っている場合には，出現頻度が低い語に関しても考慮することできている．また，複数の分野に偏っている場合においては，それぞれの分野を万遍なく得る結果となった．

結果を定性的に評価すると，本手法は頻出語を基準とするが，tf による頻出語ですでに十分よいキーワードになっている場合には，本手法の提示する語はおおよそ tf に準ずる場合が多かった．逆に，頻出語が一般的な語である場合やユーザが複数の興味分野の論文を所持している場合には，本手法の提示する語が適切なキーワードとなっているケースが多かった．したがって，収集論文に偏りがある場合においても，ユーザプロファイルとして適切なキーワードを取り出すことができると考えられる．

表 4: 15 論文中から抽出された bigram を利用したキーワード

rank	keyword
1	bidder agent
2	auction site
3	temporal summaries
4	negotiation agent
5	relationship document
6	decid hierarchy
7	reserv price
8	text summarization
9	test topic
10	web site

5. まとめ

本論文では，論文共有・収集システム MiDoc におけるユーザプロファイル生成機構とユーザプロファイル生成手法について述べた．本手法では，出現確率と文書頻度および bigram の統計量を利用したキーワード抽出に基づいている．特徴としては，収集した論文以外に特別なコーパスを用意する必要がないことと，複数の分野に興味があるユーザに対しても，万遍なく興味キーワードを抽出できる点が挙げられる．また，MiDoc を利用することにより，従来ユーザの興味や知識を獲得するためにシステムがユーザに課す負担を軽減することができた．今後，日本語文書においてもキーワード抽出を行えるようにするため改良する必要がある．また，システム側で分野ごとにコーパスを用意しておくことで，ユーザが収集した論文から分野キーワードを獲得することも可能である．キーワード抽出の精度向上に関しては，論文特有の不要語を実験により観測し，システムに利用することで可能である．

参考文献

[1] 平岡佑介，伊藤孝行，新谷虎松：情報携帯可能な論文収集・共有システム MiDoc について，第 66 回情報処理学会全国大会論文集，情報処理学会，(2004).

[2] Salton. G.(ed.) : The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice Hall, (1971).

[3] Porter, M . F：An algorithm for suffix stripping, *Program*, Vol.14, No.3, pp.130-137(1980)
Reprinted *in Readings in Information Retrieval*, Jones,K.S. and Willett,P.(Eds.), Morgan Kaufmann Publishers, pp.313-316, (1997).

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

[5] 松尾豊，石塚満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会誌，Vol.17, No.3, pp.217-223, (2002.5).