類似論文からの関連用語抽出による論文検索支援システムの提案

Academic Paper Searching Support System Based On Related Keyword Extraction From Related Papers

南浦 佑介†　新美礼彦†
†公立はこだて未来大学システム情報科学部

1. はじめに

近年，学術分野の専門分化と，学術に関する情報量の爆発的に増加により，その中から閲覧した論文を見つけだすことが年々困難になってきている．これに対し、論文検索結果の視覚化や、専門用語の自動抽出などの研究が行われており，またCiNii[1]やGoogle Scholarなどの論文検索システムでも様々な工夫が行われている．

論文検索を行う際，調べたい内容が明確である場合や，ある程度の専門知識を有している場合，キーワードや著者情報を適切に用い，目的の論文を見つけるは容易である．しかしながら，専門知識が乏しい特定分野の初学者の場合，思いつくキーワードは少なく，たまたま想起したキーワードから論文を探すため，目的の論文を探すことが困難である．論文検索に関して，2009年度よりCiNiiではウェブAPIコンテスト[2]が行われており，論文検索の方法についての議論の余地は十分にあると考える．

そこで本研究では，特定分野の初学者を対象に，関連用語に着目した論文検索支援を行う手法を提案する．本手法を用いることにより，システムがユーザに関連用語を提示することで専門知識不足によりキーワードを適切に設定できない初学者に対する検索支援に役立つと考える．

2. 関連研究

論文検索支援の類似した研究としては高久[3]らの研究がある。高久らは普段研究に携わらないユーザを対象として，任意のテキストから関連文献を自動的に引き出すことができる検索手法を提案している．入力テキストの頻度と生起確率を掛け合わせた重みを用い，より類似度の高いものから順に論文を提示している．研究対象は類似しているが，関連用語を提示するという点において異なる．

また，論文における関連用語抽出の類似した研究として難波ら[4]の研究がある．難波らは，論文間の引用関係に着目した関連用語の自動抽出を行っている．専門用語の抽出には中川ら[5]の開発したTermExtractというツールを用いている．この手法は「多くの異なる語と連接する名詞から構成される複合語は重要語である」という考えに基づいており，この重要度と引用関係を用いた関連用語の出力を行っている．本研究では同じアルゴリズムを実装したツールとして東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システムtermex[6]を用いる．このツールは重要度に加え，用語の出現頻度もリスト化するので，使い勝手の良さとしてこちらを用いた．本研究との違いとしては，関連する論文を引用関係か

らの抽出ではなく，出現頻度を用いたクラスタリングにより抽出することを提案しており，引用関係にない論文からでも関連用語を抽出することができる。

3. 提案手法

本提案手法のシステム概要，提案アルゴリズムについて述べる．

3-1　提案システム概要

本システム概要を示す．図1は本システムの検索画面である．

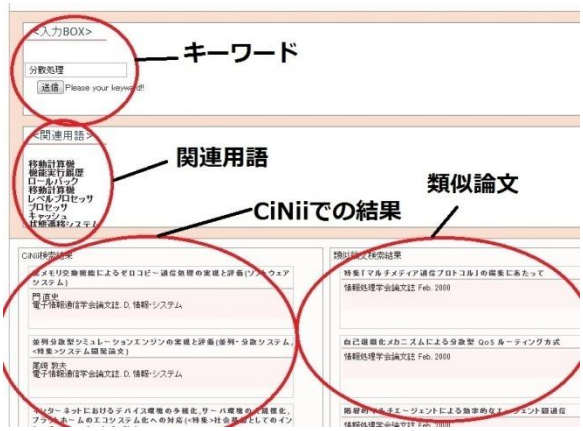

図1　検索画面

ユーザの入力キーワードをクエリとして，関連用語・論文を提示する．関連用語を探しながら，興味のあるものであればすぐに論文を閲覧できるようなインターフェースを提供する．関連用語抽出アルゴリズムと関連論文抽出アルゴリズムについては3-2で述べる．

ユーザがキーワードを入力すると，その関連用語の提示を行う．加えて，ユーザは用語の関連度を任意で指定でき，大まかに検索したいときは関連度を低くし，興味があり他の単語がさらに知りたい場合は，関連度を高くするなど，検索目的に応じて幅広い検索を行うことが可能である．

検索結果の論文提示については，クエリに対するCiNiiでの検索結果と，関連用語に関する類似論文を提示する．ユーザは指定したキーワードを含む論文の検索結果と，キーワードに関連する論文を常に閲覧することができる．

3-2　アルゴリズム

提案システムは，キーワード抽出，クラスタリング，結果出力の３つの要素に分かれる．

事前に，論文からキーワードを抽出する．形態素解析システムChasenと，専門用語抽出ツールtermexを用い，論文テキストデータから用語，頻度，重要度を取得し，それをデータベースに保存する．

タームベクトルを作成し，論文をクラスタリングする．タームベクトルには単語の出現頻度を用いタームベクトルを用いる．作成後，その跡で作成するクラスタリングの精度を上げるため，一つの論文にのみ出現するタームは削除し，タームベクトルを再構築する．ここで作成したタームベクトル用い，本研究では階層的クラスタリングであるward法により論文をクラスタリングする．ユーザが提示される関連用語の関連度を指定できるように，様々な階層でのクラスタリング結果をデータベースに保存する．予備実験において，デンドログラムをもちいてクラスタリング結果を確認し，複数の閾値でのクラスタリング結果を論文データベースに格納した．表1にはデータベースの登録例を示す．

表1　複数閾値でのクラスタリング結果

	大まか ←		→ 細かい	
	結果 A	結果 B	結果 C	結果 D
論文 1	4	10	23	40
論文 2	4	10	26	60
論文 3	6	15	33	73
論文・・・	・・・	・・・	・・・	・・・

表中の大まかとはクラスタリングの際,クラスタ数を絞った場合のクラスタリング結果であり,細かいとはクラスタ数が多いときのクラスタリング結果である.提案システムでは,クラスタリングの粒度を論文の関連度と定義している.提案システムでは,同じクラスタに所属する論文を類似論文と定義し,推薦に用いる.

次に,関連語の提示についてのアルゴリズムを述べる.入力キーワードで先ほどのデータベースに問い合わせ,それを含む論文のクラスタを特定し,termexを用いて算出した重要度の高いものから順に出力する.また,ユーザによる関連度指定では,指定された関連度(クラスタの粒度)により使用するクラスタリング結果を変更することにより,関連語の出力に反映させる.

論文の提示については,先ほど述べたようにクエリに対してCiNiiで検索した結果と類似論文検索による結果の2種類の推薦となる.検索結果ではCiNiiAPI[7]を用いた.

以上が本研究における提案手法である.提案手法を用いたシステムは,関連用語と類似論文の提示により,ユーザの専門知識不足を補い,効率よく論文を検索できると考える.

4. 評価実験

提案手法に対する評価実験について述べる.

提案手法の有用性を証明するため,以下のカバーストーリーを用意し,そのシナリオに準じた状況を想定してもらい,普段論文検索支援システムを使用しないユーザを対象として被験者実験を行う.

「あなたは情報システム専門の研究室に配属され、自分の興味のあるテーマについてレポートを提出することになっています。論文検索システムを用いて,取り上げたいテーマを考え,レポート作成に十分な文献を調査してください.」

以上のストーリーのもと,a)本提案手法とb)CiNii,c)本システムの画面を使用するが,関連用語,類似論文の提示部分をランダム選択にしたものとの比較を行う.被験者には,時間制限の中でテーマを決めてもらい,それに関するキーワードを述べてもらう.その結果を第3者が評価を行い,そのキーワードが適切であるかどうかを判断し,どの手法を用いたいシステムを利用したときに十分に調べられているかを比較する.

今回は,情報処理学会創立45周年記念DVDの論文を使用し,論文データベースを作成した.作成したデータベースは登録論分数1000件,関連用語17万9332個となった.

被験者実験については,現在行っている最中であり,関連用語や類似論文の提示方法などを検討している段階である.

5. まとめ

本研究では,論文集合をクラスタリングすることにより類似論文を抽出し,その結果から関連用語の提示を行う論文検索支援システムを提案した.提案システムの有効性を検証するために,実際にシステムを構築した.現在,被験者実験を行っている最中であり,提案手法の有効性の検証と同時に,ユーザへの関連用語や類似論文の提示法について検討している.

参考文献

[1]国立情報科学研究所: CiNii-Nii論文情報ナビゲータ. http://ci.nii.ac.jp (アクセス日2011年1月22日).

[2]国立情報科学研究所: CiNii - 第２回 CiNii ウェブAPIコンテスト 実施要項. http://ci.nii.ac.jp/info/ja/web_api_contest

_2010.html（アクセス日 2011年1月22日).
[3]高久雅生，江草由佳: 簡易類似文書検索手法「ふわっと関連検索」の予備的評価と分析，情報処理学会第150回データベースシステム・第99回情報基礎とアクセス技術　合同研究発表会
[4]難波英嗣.論文間の引用情報を利用した関連用語の自動収集.言語処理学会 第11回年次大会.2005.
[5]中川裕志，森辰則，湯本紘彰.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, Vol.10, No.1, pp.27-45, 2003.
[6]横浜国立大学森研究所: 森研究室 - 専門用語自動抽出システム(配布).
http://www.forest.eis.ynu.ac.jp/Forest/ja/term-extraction.html(アクセス日 2011年1月22日).
[7]国立情報科学研究所: CiNii - 外部提供インターフェースについて.
http://ci.nii.ac.jp/info/ja/if_link_receive.html
(アクセス日 2011年1月22日).