2008－HCI－127 (13)
2008／2／1

リッチインターフェースを備えた グラフィカル論文検索支援システム

鈴木 雅人
早稲田大学 大学院 理工学研究科 情報・ネットワーク専攻

概要

論文を読む際、一つ興味のある論文に対して参考文献や類似論文などを検索し、それらの概要を確認して読みたいと思う論文を読むということを繰り返す。しかし現在多く用いられている、キーワード検索を主とした検索サービスでは、膨大な検索結果の中から目的の論文を見つけることが難しく、このような読み方に適していない。そこで本研究では一つの論文を入力として、それに類似・関連する論文を視覚的に検索・表示することができるインターフェースを提案する。関連する論文の概要などの情報を確認しながら検索することで、容易かつ効率的な論文検索が可能となる。これにより、専門的な単語の知識が無くても類似する論文を見つけることができる。

Graphical thesis retrieval support system that provided with rich interface

Masato SUZUKI
Department of Informatics and Network, Graduate School of Science and Engineering, Waseda University

abstract

It is repeated to read the thesis that wants to retrieve a reference literature and a similar thesis, etc. to a thesis interesting by one, to confirm those outlines, and to read when the thesis is read. However, it is difficult to find a target thesis, and it is not suitable for such a pronunciation in a lot of retrieval services that center on retrieval by keyword being used now from among a huge retrieval result. Then, one thesis is assumed to be an input in this research, and it proposes the interface that can retrieve and display the thesis that resembles and relates to it in the sight. An easy, efficient thesis retrieval becomes possible because it retrieves it while confirming information on the outline etc. of the relating thesis. As a result, a similar thesis can be found even if there is no knowledge of a special word.

1　はじめに

近年インターネットの発達に伴い、大量の論文が電子化データとして蓄積、提供されるようになり、学会のホームページで利用できる電子図書館やGoogleScholar、CiNiiといった論文検索サービスも登場してきている。我われ研究者は研究の過程においてそれらを利用して論文を探す機会が多くなってきた。しかし、電子化された文書データベースが大規模化するに伴って、目的の論文を見つけることが困難になってきている。検索者が欲しい文書を発見するためには文書検索技術が必要となってくる。一般的に多く使われている方法としてキーワード検索があるが、この手法は検索結果が膨大になる傾向があり、多くの必要としない情報が検索結果に含まれることが多い。また、ある研究分野における専門知識のある人ならば、検索に用いるキーワードを入力することはできるが、キーワードの想起自体が困難である事が多く、その領域に精通していない人にとっては困難である[1]。

われわれ研究者は研究過程において論文を読もうとする際、一つ興味のある論文に対して参考文献や類似論文などを検索し、それらの概要を確認して必要だと思う論文を集める、というプロセスをしばしば繰り返す場面。しかし、サービスは気ワード検索に頼るものがほとんどなので、特にこういった論文の探し方に適していないため、論文を探すことに時間がかかってしまう。

そういった背景から、検索支援技術の重要性が高まっている。そこで本研究では、キーワード検索などにおいて自分のニーズに合った論文を一つ見つけてあるという前提で、その論文を元に関連する論文を視覚的に検索・表示する、リッチインターフェースを備えた論文検索システムを提案する。

2　関連研究

論文検索支援としてはこれまでに、論文をトピックごとにクラスタリングする研究[2][3]や、情報を可視化する研究が行われてきた。特に可視化におけるこれまでのアプローチとしては、現在参照中のコンテンツのデータ構造全体における位置づけを可視化するものや、文献と関連キーワードの関係を可視化するシステム[4]、キーワード検索の結果を可視化し、新たに有効な検索キーワードを提示するシステム[5]などがある。

一つの論文から、関連性の高い論文だけをすぐに探したいという読み方をする場合、従来の手法では可視化された検索結果を元に新たに有効なキーワードを元に検索する必要があるため、目的の論文の内容を確認していくのに手間がかかる。

本研究では、論文自体を検索の入力とし、関連性の高い論文だけを必要な分だけを検索結果とし視覚化するため、効率的な検索支援を可能とする。

3　提案するシステムの概要

3．1　本システムの目的

一般に論文を検索する場合、公開されている論文のタイトルやabstractなどを読み、その論文が必要かどうか判断するというプロセスを経る。論文を探す時間が限られている研究者にとって、このプロセスに多くの時間をとられることは問題である。無料で本文を入手できる論文も多くあるが、最近では著作権により、タイトル・著者・abstractまでは無料で一般公開されているが、web上で本文を読む場合には料金が発生したり、会員登録をする必要がある場合が多い。

本研究では、一つのすでに読んだ論文を入力として（コア論文とする）、公開されている論文情報を用いて、コア論文と関連する論文だけを検索結果として返し、関連性の高い順に視覚的に表示し、上記の検索プロセスを短縮することで検索支援を行うものとする。

３．２　本システムの構成

本システムは、論文情報データベース、類似度データベース、論文情報 xml、GUI で構成される。

論文情報データベース：このデータベースには、論文のタイトル、著者、abstract、学会・収録誌情報、キーワード*、参考文献*、を入手することができる学会のホームページなどへのリンクといった論文情報を蓄積する（*：公開されている場合のみ）。現在のところ情報処理学会の日本語論文を無作為に約２５０本のデータを入れている。

類似度データベース：論文のタイトル、abstract を元に計算した、類似する論文のリストを類似性の高い順に得点付で保持する。

論文情報 xml：インターフェースに渡すし、表示させるための論文情報、類似度の情報を持つ。

GUI：ActionScript3.0 で製作したグラフィカルインターフェース。論文情報 xml を受け取り、可視化を行う。

３．３　類似度計算

類似度データベースに入れる論文の類似度の計算行うため、形態素解析を行う。次に、形態素解析の結果を元に専門用語抽出ツール「termmi」を用いて各論文ごとに専門用語を抽出し、それぞれに重み付けをする。そして、ベクトル空間法により、各文献の、他の文献との類似度を計算する。

以下、それぞれの過程についての詳細を述べる。

（１）　形態素解析

日本語の文章はそのままでは処理できないため、形態素解析を行い文章を分かち書きにし、する必要がある。論文情報データベースの全論文の abstract に対して日本語形態素解析ツール「茶筅」を用いて形態素解析を行った。

形態素解析の例：

「私は昨日カレーを食べました。」

→私/は/昨日/カレー/を/食べ/まし/た/。

（２）　専門用語抽出

類似度計算の精度を高めるため、形態素解析を行った abstract に対して専門用語抽出を行う。専門用語抽出に用いた「termmi」では、単語の隣接情報を用いて専門用語を抽出し、独自の単語の重要度を付ける。

（３）　ベクトル空間法による類似度計算

ベクトル空間法を用いて、論文間の類似度を計算する。類似度計算の結果と、類似度のランキング情報を類似度データベースに登録する。

３．４　GUI の機能・動作

機能：本システムの GUI は、キーワード検索、類似度検索、参考文献検索の機能をもつ

以下に動作の流れを示す。

１）すでに読んだ論文（コア論文）をデータベースからキーワードで検索し、入力とする論文を選択する。

２）コア論文が中心ノードとして表示され、その周りに類似性の高い論文が色違いの丸いノードとして表示される。類似性が高いほどコア論文のノードと近く表示される。一度に表示する数は設定可能。

３）それぞれの論文について、タイトル、著者、abstract がマウスオーバーすることで表示される。ノードをクリックすると、収録誌情報や参考文献、本文入手先へのリンクなどの詳細を閲覧することができる。(図１，２)

４）検索結果として表示された論文のノードをダブルクリックすることでコア論文となり、その論文を中心として類似性の高い論文を再検索して表示させることができる。

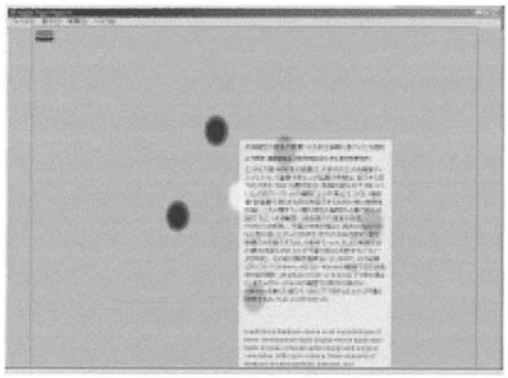


図.1:コア論文にマウスオーバーした例のイメージ

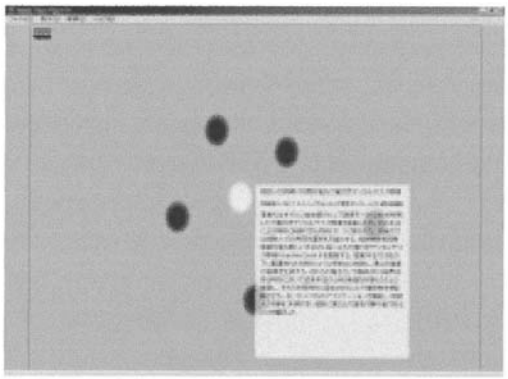


図.2:表示された類似論文の例のイメージ

4　システムの妥当性の評価

4．1　実験方法

一般的な PC 操作及びブラウザの操作に慣れている大学生及び大学院生６名を対象に実験を行った。GoogleScholar のようなキーワード検索システムを利用した場合と本システムを利用した場合での、一定数の目的論文達成までにかかった所要時間とチェックした論文の数（キーワード検索の場合はリンク先を見た回数）を比較した。まず被験者に、自分の精通している領域以外

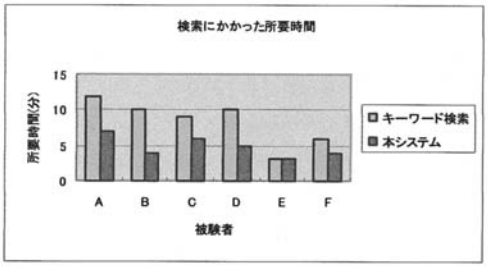
の日本語論文を一つ選んでもらう。次にその論文と類似性が高いと思う論文を５つ、abstract を読んで探してもらった。以下にそれぞれの場合の実験方法を示す。また、自由記述によるアンケートも行った。

- キーワード検索を利用した場合：論文のタイトルや abstract に含まれる単語,著者などを用いてキーワード検索して探す。
- 本システムを利用した場合：本システムのコア論文に指定し、類似検索を行い、類似性が高いとして出力されたノードの abstract を、類似性の高い順に読んでいき判断する。

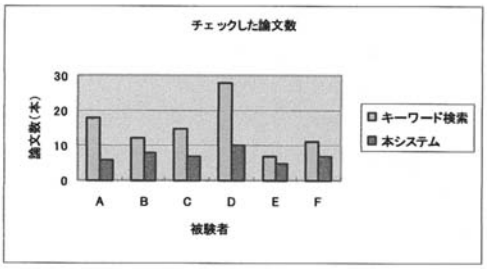
4．2　実験結果

実験から、以下のような結果が得られた。

図３に、類似論文を見つけるまでにかかった時間の比較を。図４にチェックした論文の数の比較を示す。



図．３　所要時間の比較



図．４　チェックした論文数の比較

結果を見ると、所要時間は E 以外の被験者に関しては全て短縮していた。E の被験者が見つけた類似論文はいくつかの論文において著者が

同じで、タイトルと abstract がかなり類似しているため、キーワード検索・本システム共に連続して出現したため、あまり差がでなかったと思われる。一方キーワード検索の場合では、検索に有効なキーワードの発見に時間がかかるためか、本システムに比べて全体的に時間がかかった。また、検索結果のリンク先に abstract が公開されていない場合や、検索の過程で類似性の高い論文が漏れてしまうことも原因として考えられる。

チェックした論文の数に関しては、本システムでは、abstract が類似しているとして出力された結果の精度が良かったのか、全ての被験者が１０本以内で検索を終えている。一方キーワード検索の場合では、検索結果のタイトルやハイライトされた abstract の一部を見て判断し、次にリンク先を見て内容をチェックするというプロセスを踏む。類似性の低い論文や、abstract が公開されていないといった検索におけるノイズが多く含まれてしまうため、チェックする回数が大幅に増えてしまったようだ。

また、アンケートからは、「自分で探す手間がだいぶ省けるので便利だった」、「類似度の精度が良かった」、「リンク先を見に行かなくても内容を確認できるのが良い」といった良好な意見が得られた。改善案として、「どういったところが類似しているのかをわかりやすく可視化してほしい」といった意見が得られた。

5　まとめ

本研究では、一つの論文を入力として、それに類似・関連する論文を視覚的に検索・表示することができる論文検索支援インターフェースを開発し、評価を行った。実験の結果から、本システムを用いることで、従来の論文検索に対して検索の手間を大幅に削減できるという点で有効性が示せた。

今後の課題としては、類似度計算に用いた専門用語の可視化やグラフィック面での改良など、よりユーザビリティを向上させていく必要があると考えられる。

6　参考文献

[1] R.N.Oddy : Information Retrieval through Man-Machine Dialogue, Journal of Documentation, 33(1), pp.1-14, (1997)

[2] 榊　剛史、松尾　豊、市瀬　龍太郎、武田　英明、石塚　満．論文データベースからの研究トピック抽出．人工知能学会全国大会,pp.1-4 (2005)

[3] 榊　剛史、松尾　豊、石塚　満．制約付きクラスタリングを用いた論文検索．人工知能学会全国大会,pp.1-4 (2006)

[4] 杉本　雅則、小山　照夫、堀　浩一、大須賀　節雄、絹川　博之、間瀬　久雄．文書間の関連性を可視化することによる文書検索システム．自然言語処理,112-3,pp.1-8 (1996)

[5] 野村　賢、河野博之、川原　稔．文書検索支援における可視化手法の提案とその評価．信学技報,DE2000-75,pp.1-8 (2000)