

非公開部分の多い論文データベースへのシソーラスの適用 Applying Thesaurus to the Database of Papers which has Many Closed Records

森川 由季子

Yukiko Morikawa

宮下 健輔[†]

Kensuke Miyashita

1. はじめに

近年、インターネットや高速な計算機の普及により、学術論文や卒業論文などのデータベース化が推進され、そのようなデータベースに WWW を用いてアクセスできるシステムも増えている。

しかし著作権保護などの理由から、会員限定でのアクセスしかできなかったり大学内での限定公開であったりするシステムが多い。また、そうでなくとも、一般の利用者が検索できるのは題名や梗概などの限られた項目のみで、本文に対しては会員登録が完了したり対価を支払ってからしか検索できないという状況が多く見られる。特に卒業論文の場合は、学内にはすべてを公開するが学外へは題名のみ公開している、または学外へは公開せず学内にも題名しか公開していないという状況がよく見られる。

このような論文データベースを検索することを考えるとき、一般的な全文検索では、非公開の項目は検索対象にならず、すべての項目が公開されている場合に比べて再現率が小さくなることが明らかである。

本論文ではこのような問題を少しでも解決するために、題名のみが公開されている論文について題名に使われている単語の同義語や関連語をシソーラスによって抽出し、それらも検索対象に加える論文データベース検索システムを卒業論文データベースに適用した。このような検索システムを利用することで、本文に対してごく短い文字数で構成されている題名が拡張され、再現率が大きくなることが期待できる。

以下ではこの論文検索システムの実装を述べつつ、概要と卒業論文データベースの検索実験を行なった結果を述べる。

2. 実装

シソーラスを利用した卒業論文データベース検索システムを以下のような環境に実装した。ここではその実装方法を説明しながらこのシステムの概要について述べる。

- OS : FreeBSD 4.10-RELEASE
- データベースサーバ : MySQL 4.0.20
- httpd : Apache httpd 1.3.33 (PHP 4.3.11)

このシステムは、PHP スクリプトで構成された WWW ページを利用し、ユーザから入力されたキーワードを用いて卒業論文データベースに対する全文検索を行ない、その結果を同じく WWW ページ上に表示するものである。データベースには卒業論文が題名、著者、本文などの項目ごとにレコードとして収められており、これらのうち非公開のものはレコードから除外されている。

[†] 京都女子大学現代社会学部, Kyoto Women's University, Faculty for the Study of Contemporary Society

2.1 全文検索

データベースサーバとして用いた MySQL には全文検索機能がある。しかしこれは単語間が空白で区切られていて1文字が1バイトで表される言語(英語など)でしか有効に機能しないので、これを日本語に適用するためにこの条件をみたす日本語の表記法を考えた。すなわち、検索対象となる文字列をデータベースに登録する際に形態素解析を行ない、文字列を分かち書きした後、さらに品詞間の空白を除く各文字の文字コードを16進数表記するという変換を行なうようにした。このようにすれば、単語間に空白が必要という条件をみたし、さらに1文字が1バイトで表されるという条件もみたす日本語の表記が得られる。この処理は [1] に記載されていたアイデアを参考に PHP で実現したものである。形態素解析には茶筌を利用した [2]。

さらに、キーワードとしてユーザが入力した検索語に対しても上記と同様の処理を行えば、日本語のレコードに対する全文検索が可能となる。

2.2 シソーラス検索

本システムでは題名を拡張するためにシソーラスを利用した。一般的にはデータベースに収められた論文と合致した専門分野の用語についてのシソーラスを利用する方が検索精度などの点でよいと思われる。しかしこのシステムは京都女子大学現代社会学部の卒業論文データベースに対して適用し、この学部には理系(物理、化学、情報、環境など)と文系(政治、経済、社会学、宗教、心理、マスコミなど)の研究室が混在することを考えると、そのようなシソーラスは選択し難い。そのため、ごく一般的なシソーラス [3] を利用した。なおシソーラスは論文をデータベースに登録するときのみ必要となり、検索するときには不要である。

例えば「うつ病と現代社会」というタイトルからは、まず「うつ病」というキーワードが抽出されシソーラスによって拡張される。その結果「鬱屈」、「心因性」、「ひきこもり」や「拒食症」などのキーワードがデータベースに登録され検索対象となる。

シソーラスによって拡張するキーワードは、上述した形態素解析によって題名から名詞のみを抽出して利用した。その際、単名詞だけでなく単名詞を複数並べることによってできる複合名詞もキーワードとして利用した。単名詞を利用するのは、複合名詞に比べて単名詞には非常に多くの同義語や関連語が存在するので、題名をより大きく拡張できるからである。しかし複合名詞の同義語や関連語はその部品となる単名詞に対するそれらよりも検索対象としての価値が高いと考えられるので、複合名詞も採り入れた。例えば「うつ病と現代社会」というタイトルに対して「うつ病」、「現代」、「社会」だけでなく「現代社会」もキーワードとして採用した。

表 1: 実験結果 (キーワードごとのヒット数)

公開率	精神	非行	児童虐待	法律	政治	市場	パソコン	産業	面接調査	新聞
100%	8/8	4/4	7/7	7/7	7/7	6/6	5/8	9/9	6/6	6/7
80%	6/7	4/4	5/5	6/6	6/6	5/5	5/9	8/8	6/6	5/6
60%	4/4	2/2	3/3	5/5	4/4	3/3	3/7	6/7	4/4	4/5
40%	2/5	1/1	3/4	2/2	4/4	1/1	2/8	3/3	3/3	3/5
20%	2/3	0/0	0/1	1/1	1/1	1/1	2/7	1/2	0/0	0/2
0%	0/3	0/0	0/1	0/0	0/0	0/0	0/6	0/1	0/0	0/2

3. 実験

前述の卒業論文データベース検索システムを利用して、シソーラスの有効性を確認する実験を行なった。

論文データベースの内容は、京都女子大学現代社会学部の2004年度卒業論文を利用した。データベースに登録された論文数は22であり、それぞれ題名、著者、本文が登録されている。題名の長さ本文の長さの平均はそれぞれ26文字と20,423文字であり、本文に比べて題名が圧倒的に短いことがわかる。

実験では、データベース上の全論文から無作為に抽出した論文について本文を非公開とし、それ以外の論文はすべてのレコードを公開とした。そして適当なキーワードを用いて論文検索を行ない、その結果として得られた論文の集合が、題名をシソーラスで拡張した場合とそうでない場合(全文検索のみ)とでどの程度異なるかを調べた。

3.1 実験結果

この実験では種々のキーワードを利用して検索を行なったが、そのうち公開率100%の場合に4個以上の論文がヒットするキーワードに絞って実験結果を表1に示す。

この表はデータベース中の論文における本文の公開率(左端の列)を変化させながら、最上行に列挙したキーワードについて、シソーラスを利用した場合とそうでない場合にヒットした論文の数を各セル内に記している。各セル内の数値は、スラッシュの左側がシソーラスを利用しなかった場合、右側がシソーラスを利用した場合の論文数を示している。

また、実際にヒットした論文は、シソーラスを利用しなかったときは公開されているレコード内に当該キーワードを含むものであり、シソーラスを利用したときには拡張された論文題名に当該キーワードを含むものであった。またそうでない論文はヒットしていなかった。

さらに、キーワードごとに各公開率においてヒットした論文の集合は次の2条件を常に満たしていた。

- シソーラスの利用に関わらず、公開率が $x\%$ のときにヒットした論文の集合は、公開率 $x + 20\%$ のときにヒットした論文の集合の部分集合である($x = 0, 20, 40, 60, 80$)。
- 公開率が等しいとき、シソーラスを利用しなかった場合にヒットした論文の集合は、シソーラスを利用した場合にヒットした論文の集合の部分集合である。

3.2 考察

シソーラスの利用に関わらず、公開率が小さくなるにしたがってヒットする論文数も少なくなっていることと、当該キーワードをどこにも含まないような論文がヒットしていないことから、この検索システムが期待通り動作していることがわかった。また、ヒットした論文の集合がそれぞれ前節後半の条件をみたしていることから、シソーラスの利用が論文データベースの検索について悪影響(検索精度や再現性を小さくする効果)は与えていないことが言える。

表1に記したキーワードのうち半数近くのものにおいては、公開率が40%以下となった場合、シソーラスを利用したときにヒットした論文数がシソーラスを利用しなかったときのそれを上回っている。その場合でも前節後半に述べた条件は満たされているので、これは、シソーラスを利用したときには、公開率100%のときにヒットした論文のうち数多くのものが低い公開率でもヒットし続けていることを示している。

4. まとめ

本論文では、非公開部分の多い論文データベースにおいて、シソーラスを利用して公開されているレコードを拡張することにより、より高い検索精度と再現性を得る論文データベース検索システムを実装した。この検索システムを本学部の卒業論文データベースに適用して実験したところ、本文の公開率が低い場合でもシソーラスを利用しない検索に比べて多くの論文がヒットしており、かつこれらの論文は公開率が高い場合にヒットしていた論文であることから、今回提案した論文データベース検索システムは有効に働いていることがわかる。

しかし今回登録された卒業論文の数が少ないため、データベースが巨大なものになった場合にも精度や再現率が維持できるかどうかを確認できていない。また、本学部の卒業論文では義務づけられていないが、論文に内容梗概が添付されていれば題名だけの場合に比べて検索精度が高くなることが期待できる。このようなことを考慮に入れ、さらに実装と実験を進めたいと考えている。

参考文献

- [1] 池内 淳, “MySQLについて 伍”, <http://www.daito.ac.jp/~ikeuchi/webdb/mysql.5.html>.
- [2] 松本 裕治, “ChaSen Wiki FrontPage”, <http://chasen.naist.jp/hiki/ChaSen/>.
- [3] 言語工学研究所, “デジタル類語辞典 第3版”, ジャングル, 2004.