テキストデータを用いた類似卒業論文検索システムの構築

著者　山田直正　　指導教員　斉藤徹

1.はじめに

本校での卒業研究論文の執筆には過去の卒業研究論文を参考にすることがとても多いと考えられる。しかし自分の研究に関係のある卒業研究論文を探すことに時間を取られて実際の卒業研究が進まなければ本末転倒であるので、この負担を減らすことが本校学生の卒業研究を支援することになるのではないかと考えた。そこで本研究では電子ファイルを用いた類似論文検索システムの構築を行うことで、この目的を達成しようとした。

これは論文を解析することで論文の特徴を表した単語を推定し、その単語の論文中での出現率を比較することで論文の一致度を計算するものである。

本研究の結果、電子情報工学科の平成 23 年、24 年の卒業研究発表会レジュメファイルと 25 年度の第一回中間発表レジュメファイルに対して、キーワードによる類似論文の検索と、論文による類似論文の検索ができる Web サービスの構築ができた。

2.システム概要

本研究では実際の論文検索システムへの足がかりとして卒業研究発表会レジュメファイルもしくは中間発表レジュメファイルを用いているが、5 章に記す問題点を改善すれば実際の論文ファイルを用いて運用することも十分に可能であると考えられる。そのため、これ以降は卒業研究発表会レジュメなどを論文と表記する。

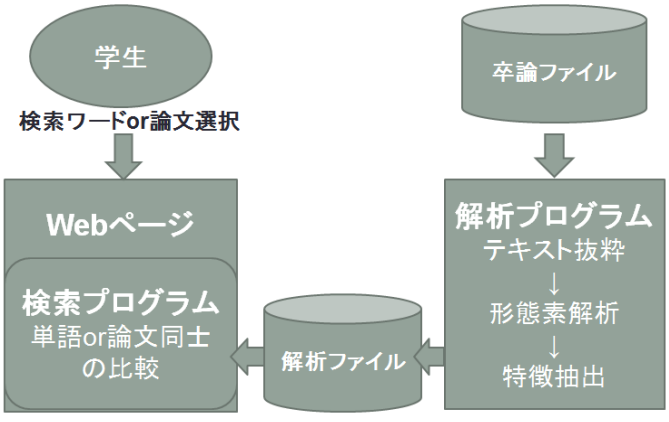

図1：システム構成図

このシステムは図 1 のように主に解析プログラムと検索プログラムの 2 つで構成されている。

1 つ目の解析プログラムでは次のようなことを行う。まず卒業論文発表会レジュメファイルから抜き出したテキストデータに対して形態素解析を行い、論文のメタデータとなりえない名詞以外の単語を取り除く。次に抽出された単語の論文内での出現回数をカウントする。このとき論文のタイトルや概要に出現した単語はよりその論文の特徴になると考えられるので、このような単語については他の単語よりもカウント時に倍率（重み）をかけて積算した。各単語の出現回数より論文中の単語の出現率を求めて、単語と出現率を解析ファイルに一論文一ファイルとして記録する。新しい論文が追加された場合はこのプログラムを実行することで、引数で指定したディレクトリ下のファイルが自動的に処理され、それらの論文を検索対象に含めることができる。

2つめの検索プログラムでは解析プログラムに作成された解析ファイルと、引数として与えられた単語またはある論文の解析ファイルを比較し一致度を計算、類似した論文として一致度が大きい順にファイル名のリストを出力する。

この検索プログラムをCGIプログラムから実行し、適宜処理をしてWebサービスとして構築した。

3.動作機能

解析ファイルで行っている処理の詳細は次のとおりである。まず pdf ファイルもしくは doc ファイルである論文ファイルから pdftotext、wvWare というソフトウェアを用いてテキストデータを抜き出した。そのままでは空白や句読点などが論文によってばらばらなので取り除いた。

そのテキストデータに対し MeCab を用いた形態素解析を行い、論文のメタデータとなりえない名詞以外の単語や、文字化けによって現れた可能性が非常に高い一文字の漢字や記号を取り除いた。

次に抽出された単語の論文内での出現回数をカウントした。出現回数のみでは重要な単語よりも何度も出現する一般的な単語が出現回数の上位になってしまう可能性があるので、確実に重要だと考えられる、論文のタイトルと概要に出現した単語を用いて重み付けを行った。重みとして通常は単語が出現する度に+1 とするところを、タイトルに出現した単語に対して+5、概要に出現した単語に対して+3 とカウントした。両方に出現する単語はより重要だと考えられるので+8 とカウントした。

各単語の出現回数を論文中に出現した単語の総出現回数で割った値をその単語の出現率として、その単語と出現率を解析ファイルとしてテキストファイルに記録した。

検索プログラムでは、キーワードもしくは論文の解析ファイルを比較して論文の一致度を求めた。キーワード検索の場合は次のような式を用いた。

$$\text{一致度} = \text{単語の一致数} + \sum_{\text{論文の特徴的な単語}} \text{単語の出現率}$$

類似論文検索の場合は次のような式を用いた。

$$\text{単語の出現率の比} = \left| \frac{\text{単語の出現率（小）}}{\text{単語の出現率（大）}} \right|$$

$$\text{一致度} = \text{単語の一致数} + \sum_{\text{論文の特徴的な単語}} \text{単語の出現率の比}$$

一致数を加算しているのは、出現率の大小よりも単語が一致したということを重要視したためである。

4.結果

実際に出来たWebサービスは次の図の通りである。ユーザーは、Webページのテキストボックスに検索したい単語を1つ以上入力し送信することで、その単語が特徴として含まれる論文の一覧が表示される（図2参照）。そこで表示された論文の中から一つを選び、「この論文で検索」ボタンをクリックすることで、その論文の内容が類似した論文の一覧が表示される（図3参照）。

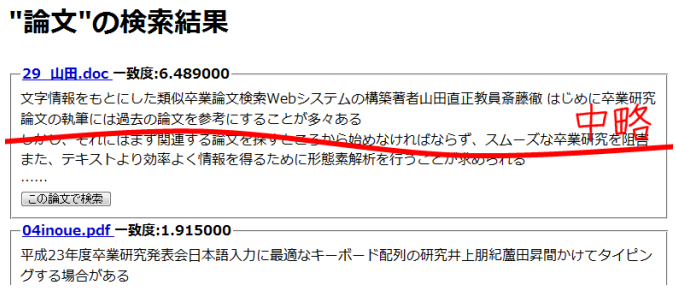

図 2：キーワード検索の例

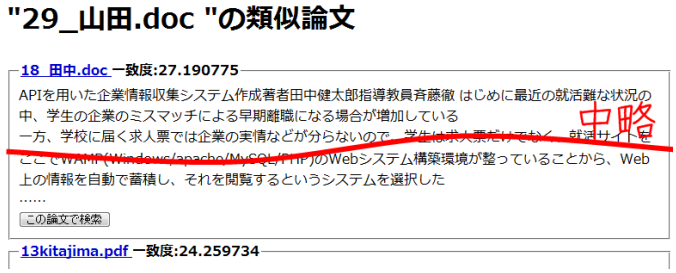

図 3：類似論文検索の例

実際にこのシステムで得られる類似論文については、例として類似卒業論文の検索によって得られる論文の例を 2 つ挙げる。得られた結果から上位 3 件を抜き出し、タイトルと一致度、そして著者と担当教員を記した。

タイトル	著者/指導教員	一致率
最適車間模型を用いたボトルネックによる渋滞流形成の数値解析	本山 野村	53.16
渋滞における発生と車両数の依存性	酒井 野村	49.36
MPS法を用いた水柱崩壊のシミュレーション	松井 下條	36.64

表 1：「後方参照を取り入れた最適速度模型による渋滞の数値解析」（著者：朝倉　指導：野村）

タイトル	著者/指導教員	一致率
連続母音におけるフォルマント周波数を利用した母音推定の研究	松山 西	60.00
マルチマーカを用いたARによるモデルビューアの開発	坂田 青山	33.32
FM一括変換方式における劣化要因シミュレーション	奥田 西	32.61

表 2：「音声スペクトログラムの画像解析による話者識別の研究」（著者：和田　指導：西）

表 1 では、同じ卒研室での前年度または後年度で継続して行われた研究であることがわかる。同じく表 2 では、年度で連続した研究ではないようだが、同じ卒研室で行われた音声に関する研究が 1 位となっていることがわかる。また、どちらも 2 位や 3 位には直接関係のないと考えられる研究が含まれているが、その分一致度の値の落差が大きくなっている。

5.まとめと今後の課題

一年を通して本研究を行い、これらのような結果を得ることができた。しかしまだ大きな課題がいくつか残っている。

1 つ目は、多くの論文に含まれるような一般的な単語によって関連性の低い論文が抽出される点である。現在は卒業研究発表会レジュメファイルを用いているので、出現した単語は多くて 30 件程度のデータ件数になっている。しかし、実際の論文を用いた場合はデータ件数が数百件程度まで増えるので、一般的な単語のフィルタリング方法を検証する必要があると考えられる。

2 つ目は、キーワード検索の場合にキーワードが一致しているかという判断しかできていないので、表記の揺れや形態素解析によって分割される「周波」「数」といった単語などによって必要としている論文を探すことが多少難しい点である。これは検索キーワードの形態素解析や類語辞書を用いた処理が必要になると考えられる。

以上の 2 点のような問題点もあるが、4 章で記した類似論文の検索例より、現状でもこのシステムは類似論文を探すために十分に有効なものであることが確認できた。