2007年度 修士論文

公開情報を用いた
類似論文検索精度の向上

早稲田大学大学院 理工学研究科
情報・ネットワーク専攻

鈴木　雅人

学籍番号：3606U050-5

提出：2008 年 2 月 4 日

指導：村岡 洋一 教授

目 次

表 目 次

第1章　はじめに

1.1　研究背景

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。われ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]。

特にわれわれ研究者は研究過程において、サーベイなどのために多くの論文を読む必要がある場面に遭遇することが多い。論文を読もうとする際、webを利用して検索し、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。しかし、先に述べた理由から、研究者、特に研究の初学者にとっては、類似する論文を探し出すことは困難であるといえる。
特に、一つ興味のある論文に対して、それに関連する論文や類似している論文を探そうとする際、多くの場合に論文のタイトルやアブストラクトを読み、それが類似性があるかどうか判断する。しかし、そのプロセスは時間の限られている研究者にとって大きな負担である。
そういった背景から、論文を検索する手助けとなる、検索支援技術の必要性が高まってきている。検索支援技術としては、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[4][5]、類似度を求める研究[6]などがあげられる。

1.2　研究目的

論文検索支援において、本研究では類似する論文を検索することに主眼を置き、検索対象を論文の日本語のアブストラクトに絞り、類似論文検索の精度を向上することを目的とする。

1.3　本論文の概要

本論文では、論文のタイトル、アブストラクト、キーワードといった公開されている論文情報を用いて類似性・関連性の高い論文を機械的に見つけ出す方法について述べる。論文のタイトルとキーワードには、論文を特徴付ける単語が含まれていることが多く、論文間の類似性を測る上で重要であると考えられる。本論文では、類似度を比較する論文集合のアブストラクトに対してそれぞれ専門用語抽出を行い、重要度を付けたリスト集合を作成し、ベクトル空間法により論文間の類似度を計算する。

提案する手法は、論文のタイトル、キーワードについても同様の操作をし、その重要度を、を先に得られたアブストラクトの専門用語リスト中の最も重要度が高い単語よりも強くし、新たに重要度をスコア付けされた専門用語リストにするというものである。従来手法である tf・idf 法を用いて単語の重み付けを行った場合、単純に専門用語だけを行った場合、提案手法を用いた場合とで、ベクトル空間法による類似度計算を行い、計算結果を比較する実験を行った。類似度計算結果について、本研究手法が他の２つの手法に比べ、精度の高い類似論文検索ができるか検証した。結果から、このリストに対して類似度計算を行うことで、従来よりも類似論文の検索精度が向上することがわかり、提案手法の有効性を示せた。

1.4　本論文の構成

本論文は 7 章から成る。

第 1 章　はじめに
本論文の目的，概要，構成について述べる。

第 2 章　既存の研究と問題点
既存の研究とその問題点について述べる。

第 3 章　本論文の提案手法
本論文で提案する手法について述べる。

第 4 章　解析の流れ
提案手法における、文章解析の流れについて述べる。

第 5 章　評価実験
提案手法についての評価を行うための実験方法と結果について述べる。

第 6 章　考察
実験結果についての考察を述べる。

第 7 章　結論
本論文のまとめと、今後の課題を述べる。

第２章　既存の研究と問題点

過去の研究では [6]（八太　2002）のように、本研究と同じく論文のアブストラクトを対象として類似度を計算し、論文検索支援を行った研究がある。従来の研究では、アブストラクトに対して形態素解析を行い、助詞や助動詞を除去しＴＦ・ＩＤＦ法を用いて単語の重み付けを行い、ベクトル空間法によって文献間の類似度を比較するという手法を用いていた。しかし、この方法では、論文を特徴づける専門用語以外の一般的な名詞（「研究」、「提案」など）が類似していない他の論文のアブストラクトの文中に多く現れた場合に検索ノイズとなり、類似していない論文が類似論文として出力されるといった問題がある。文書から専門用語を抽出する研究 [7] があり、多くの研究に利用されている。しかし文献内のアブストラクトから専門用語を抽出し、そのまま重要度を付けてベクトル空間法 [8] にによって類似度を計算しても、検索結果にノイズが発生したり、類似度の高い論文が検索結果から漏れてしまうことがある、といった問題がある。

第3章　本論文の提案手法

本研究の新規性としては、論文のタイトル、アブストラクト、キーワードを用いて、従来の研究に比べ、漏れなく類似度の高い論文を引っ張れるようにし、類似論文検索の精度を高める点にある。

論文を探す際に web を利用することが多いが、多くの場合、論文のタイトル、著者、アブストラクト、キーワード（書かれている場合)、収録誌情報までは一般に公開されている。しかし論文の本文全文に関しては、著作権の関係で公開されておらず、閲覧のためには課金が必要な場合も多い。そのため、公開されており、人間がある論文の類似論文を web を利用して検索する際に使う判断材料として考えられる、論文のタイトル、アブストラクト、キーワード用いて類似論文検索精度を向上させる手法を提案する。

論文のアブストラクトの文章中には、論文を特徴づける専門用語が多数現れる。また、論文のタイトルおよびキーワードには、特に重要度の高い専門用語が書かれていると考えられる。そのため、アブストラクトだけを解析して類似度を計算するのではなく、タイトル、キーワードに含まれる専門用語の重要度も考慮して解析することで、類似論文の精度があがると考えられる。

3.1　提案手法

1. アブストラクトの専門用語を抽出し、重要度を付けたリストを作成する。
2. タイトル、キーワードに関しても専門用語リストを作成する。
3. 2. で得られた単語の重要度を、1. の結果の中で最も重要度の高い単語の重要度より、高く設定する。
4. 3. で設定したリストを 1. のリストに追加する。
5. ベクトル空間法により、各論文間の類似度を求める。

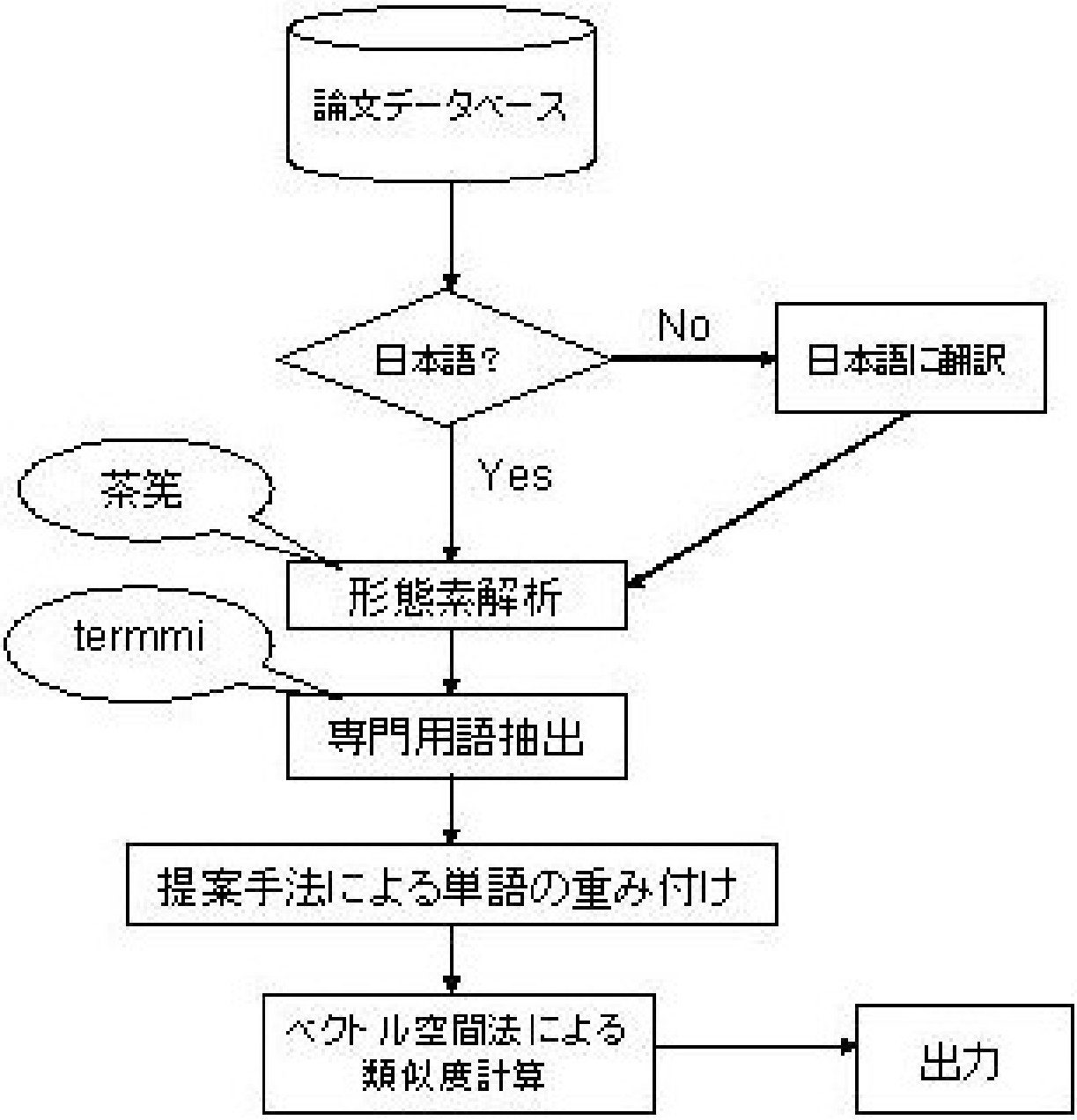
論文データベース
日本語？
No
日本語に翻訳
Yes
茶筅
形態素解析
termmi
専門用語抽出
提案手法による単語の重み付け
ベクトル空間法による
類似度計算
出力

図 3.1: 処理の流れ

以下に各過程についての詳細を述べる。

3.2 形態素解析

日本語の文章は区切りがなく、解析するに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」[11] を用いて形態素解析を行った。

形態素解析とは、文書を構成する文字列を単語に分割し、各単語に品詞や語形変化などの情報を与える処理である。　茶筅に日本語文章を入力すると、日本語文章を単語ごとに分けられたものが、単語の品詞や語形変化の情報とともに出力される。形態素解析の例：
入力文章：「私は昨日カレーを食べました。」
出力結果：

私	ワタシ	私	名詞-代名詞-一般
は	ハ	は	助詞-係助詞
昨日	キノウ	昨日	名詞-副詞可能
カレー	カレー	カレー	名詞-一般
を	ヲ	を	助詞-格助詞-一般
食べ	タベ	食べる	動詞-自立 一段 連用形
まし	マシ	ます	助動詞 特殊・マス 連用形
た	タ	た	助動詞 特殊・タ 基本形
。	。	。	記号-句点
EOS			

3.3 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アブストラクトに対して専門用語抽出ツール「termmi」を用いて専門用語抽出を行う。専門用語抽出を行う理由は、形態素解析の結果をそのまま専門用語として使うには次の 2 つの問題があるからである。

一つ目は、複合語に対応していないことである。専門用語の多くは単語を組み合わせて、複雑な概念を表すことが多くなる。特に本研究で使用している「茶筅」の場合、単語を品詞単位で細かく分割するため、そのまま使うには難がある。

二つ目は、どの用語が重要であるか判断する仕組みを持たないことである。

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。出力結果として、専門用語リストが重要度の高い順に重要度の値とともに得られる。なお、この過程で日本語の助詞や助動詞といった不要な品詞は除去する。

3.3.1 不要語の除去

日本語の助詞や助動詞は、日本語の文章中に頻繁に登場する単語であるため、これらの単語を使って検索しても、ほぼすべての文書が検索されてしまう。そのため、専門用語などのユニークな単語に比べ、類似度を計算する際に用いる単語として適当ではない。このような単語のことを不要語と呼び、類似度計算を行う際の単語リストから排除する必要がある。

3.3.2 専門用語の生成

「茶筅」で文章を単語に分割した際、単語の品詞情報も合わせて情報として出てくる。複合語は、語の並びと品詞情報を元に組み立てる。基本的には名詞が連続で出現した場合、それらを統合し複合語とする。

3.3.3 重要度計算の仕組み

「termmi」での専門用語は、単語そのものか複数の単語を組み合わせて作られる。この複合語を構成する最小単位の名詞を特に「単名詞」と呼び、この単名詞が他の単名詞と連結して複合語をなすことが多いほど、重要な概念を示すと考える。　簡単な例として、「情報科学技術」を考える。この語は次のとおり 3 つの単名詞に分割できる。この際、それぞれの単名詞が他の単名詞とどれだけ結びつくか統計的に分かっているとする。前の語に連結した回数を x, 後の語に連結した回数を y とすると

表 3.1:

単名詞	x	y
「情報」	1	2
「科学」	2	3
「技術」	1	1

複合語全体の重要度はこれらの６つ（単名詞数ｘ２）の数値の相乗平均から求める。「日本語マニュアル文における名詞間の連接情報を用いたハイパーテキストのための索引の抽出」[12] により、相乗平均がもっともよい結果になることがわかっているので、平均は相乗平均でとるようにしている。　また、相乗平均をとる際に、連結した回数が０回の単名詞に対応するため、各回数に１を加算した値を用いている。[10]

3.3.4　単名詞の連接情報

単名詞の連接情報は、連接した単語の種類をカウントする方法（異なり数）と、種類にかかわらず出現しただけカウントする方法（延べ数)、そして情報理論的に見た連接回数（パープレキシティ）の３つの方法がある。[10]

例えば、統計データで、「情報」という語が「科学」の前に２回、「技術」の前に３回連接したことが分かっているとする。この場合連接語の異なり数と延べ数は次のようになる。

異なり数　　２回　（「科学」　＋　「技術」」で２種）　　延べ数　　５回　（「科学」２回　＋　「技術」３回）

パープレキシティは、「情報理論的に見ていくつの単名詞が連接可能か」を示す。

本論文では、異なり数を使った。

3.4　単語重要度の改良

タイトルに含まれる単語がその論文にとって重要であるにもかかわらず、アブストラクト中には現れない、といった場合、類似論文検索時に類似している論文が検索結果から漏れてしまう可能性がある。これを解決するため、論文のタイトルおよびキーワードについても専門用語抽出を行う。

キーワードは元々単語にわけられているため、そのまま全て専門用語であるとする。タイトルに関しては、アブストラクトの時と同様に専門用語抽出を行う。ただし、タイトルは短い文章なので重要度は低い値で計算されるため、重要度の値に関しては値の大きさの順番のみ考慮するものとし、「研究」や「手法」といった、多くの論文に含まれるような一般的な単語に関しては専門用語とみなさずに除去する。

次に、先に得られたアブストラクトの専門用語リストの一番上に、タイトルと

キーワードから得られた専門用語リストの単語の重要度の値を、アブストラクトでのリストの最も重要度の高い単語の値より少し高くする。こうして新しい、重要度付き専門用語リストを作成する。
例)
タイトル：「物理的モデルによる 3 次元物体の乾燥によるひび割れ表現法」[9]
キーワード：「CG ひび割れ, 物理的モデル, ばねネットワークモデル, 含水率モデル」
アブストラクト：「CG (Computer Graphics) 研究の分野では, 自然物 自然現象の表現法の構築は重要な課題の一つてある. 本研究では CG において,3 次元物体に発生するひび割れをリアルに表現する方法を提案する. ひび割れは土壁, 田の表面, 陶磁器, 樹皮等に見られる身近な自然現象の一つてある 本研究では特に, 泥や粘土からなる物体が乾燥によって収縮する際に発生するひび割れの再現を目的とする. CG でひび割れを表現するには, 観察に基づくルールや簡易物理モデルに基づくアプローチが考えられるが, 本研究では材質や環境, 外力の変化への対応が容易に表現てきることから物理モデルによる方法を用いる. 具体的には粘土の収縮や弾性, 柔軟性を表現するためにはねネットワークモデルを, 乾燥による物体内部の水分移動を表現するために含水率モデルを導入し, これら二つのモデルを統合することにより, 乾燥によるひび割れ発生のメカニズムをシミュレーションする含水率を基本に, 使用する物理パラメータの実測方法も検討する最後に, 様々な形状をもつ 3 次元物体について実験を行い, 提案方法の有効性を確認した.」

表 3.2:

タイトル内の単語	重要度
物理的モデル	1.33
ひび割れ表現法	1.33
乾燥	1.00
物体	1.00

表 3.3:

キーワード
CG
ひび割れ
物理的モデル
ばねネットワークモデル
含水率モデル

重要度の数値が大きいほど、その単語は重要である。最低値は 1.00 であり、重要でない単語だとみなされる。

3.4.1 単語追加規則

このリストに、タイトル内の単語リストとキーワードリストを上位に追加し、全体の専門用語リストをつくる。上位に追加する理由は、タイトル内の専門用語は著者がタイトル文に含めているため重要な単語であると考えられ、キーワードについても論文中に特筆していることから同様に重要であると考えられるからである。専門用語リストにこれらの単語を追加する際、追加する単語が全てアブストラクト内の単語リストに存在し、かつ最上位にあった場合は、その部分の値はそのままにする。
いくつかは存在するが、全てが重要度が上位でない場合は、新しく上位に重要度を高くして追加した後、元の重複部分を削除する。
それ以外の場合は、アブストラクト内の単語 リストの最上位単語の上に、重要度をそれより高くして追加する。
キーワードとタイトルのどちらにも出現する単語は、タイトルでの追加規則に従う。

単語の重要度の値に関して、数値が大きくなるほど、その単語が論文集合全体における重要度が高くなるため、類似度計算への影響が大きくなる。例えば、各論文の最も高い単語の重要度の値が'3〜4' 程度であるとき、仮に論文A内の単語「サンプル」の重要度を'100' のように極端に大きい値に設定すると、「サンプル」を含んでる論文の論文Aに対する類似度が高く計算され、逆に「サンプル」を含まない論文の類似度が極端に低く計算されてしまうことがある。 本研究では、タイトル内の単語とキーワードが、アブストラクト内の最も重要度の高い単語と同等以上に重要だと仮定して値を設定する。そのため、追加する際の重要度の値は、キーワードと、タイトル内のリストで重要度が 1.00 の単語は、アブストラクト内の単語 リストの最上位単語より 1 高く設定する。同じく 1.33 のような値の単語は、さらに 0.33 増やす、というように差分だけ増やして追加していく。
このようにして、新しい単語リストを作成する。

3.5　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。ベクトル空間法 [8] は，共通したベクトル空間を用意することによって文書の類似性を判断するものである．各文書の内容は，使用された単語の評価値を基に，特徴ベクトルとして表現される．文書 d_i の特徴ベクトル $\vec{D_i}$ i は次式 のように定義される．

$$\vec{D_i} = [\boldsymbol{wa}(d_i, w_1)\ ,\boldsymbol{wa}(d_i, w_2)\ ,...,\boldsymbol{wa}(d_i, w_n)]$$

ここで n は単語の総数，$\boldsymbol{wa}(d_i, w_j)$ は文書 d_i 中の単語 w_j の出現頻度などのスコアである．ベクトル間の類似度ではよくコサイン尺度が用いられる。特徴ベクトル $\vec{D_i}$ と $\vec{D_j}$ のコサインを計算することで，文書 d_i と文書 d_j との類似度の評価が可能となる．類似度が 1 に近い値であるほど，d_i と d_j は類似した内容を持つ文書である．コサイン尺度は次式で定義される。

$$\cos(D_i, D_j) = \frac{D_i \times D_j}{|D_i||D_j|}$$

表 3.4:

アブストラクト内の単語	重要度
ひび割れ	2.57
研究	2.32
物体	2.25
方法	2.11
乾燥	2.00
CG	1.83
物理モデル	1.74
モデル	1.73
一つ	1.58
粘土	1.58
簡易物理モデル	1.58
自然物自然現象	1.55
物理パラメータ	1.50
ばねネットワークモデル	1.49
自然現象	1.47
含水率モデル	1.44
実測方法	1.39
提案方法	1.39
含水率	1.34
表現法	1.25
CG におい	1.25
水分移動	1.25
表現	1.25
有効性	1.25
ひび割れ発生	1.25
土壁	1.25
物体内部	1.25
分野	1.00
構築	1.00
環境	1.00
田	1.00
変化	1.00
樹皮	1.00

アブストラクト内の単語　重要度 (続き)	
柔軟性	1.00
再現	1.00
収縮	1.00
目的	1.00
泥	1.00
外力	1.00
観察	1.00
対応	1.00
最後	1.00
メカニズム	1.00
基本	1.00
陶磁器	1.00
課題	1.00
実験	1.00
アプローチ	1.00
二つ	1.00
ComputerGraphics	1.00
表面	1.00
形状	1.00
弾性	1.00
ルール	1.00
材質	1.00

表 3.5:

新たにできたリスト	重要度
物理的モデル	3.90
ひび割れ表現法	3.90
乾燥	3.57
物体	3.57
ひび割れ	3.57
ばねネットワークモデル	3.57
含水率モデル	3.57
研究	2.32
物体	2.25
方法	2.11
CG	1.83
物理モデル	1.74
モデル	1.73
一つ	1.58
粘土	1.58
簡易物理モデル	1.58
自然物自然現象	1.55
物理パラメータ	1.50
ばねネットワークモデル	1.49
自然現象	1.47
含水率モデル	1.44
実測方法	1.39
提案方法	1.39
含水率	1.34
表現法	1.25
CG におい	1.25
水分移動	1.25
表現	1.25

新たにできたリスト　重要度（続き）	
有効性	1.25
ひび割れ発生	1.25
土壁	1.25
物体内部	1.25
分野	1.00
構築	1.00
環境	1.00
田	1.00
変化	1.00
樹皮	1.00
柔軟性	1.00
再現	1.00
収縮	1.00
目的	1.00
泥	1.00
外力	1.00
観察	1.00
対応	1.00
最後	1.00
メカニズム	1.00
基本	1.00
陶磁器	1.00
課題	1.00
実験	1.00
アプローチ	1.00
二つ	1.00
ComputerGraphics	1.00
表面	1.00
形状	1.00
弾性	1.00
ルール	1.00
材質	1.00

第4章　評価実験

本研究の手法による類似論文検索精度を検証するための実験を行った。今回は、情報処理学会、電子情報通信学会の論文の中から、似た単語が登場する可能性が高い、同じ研究分野の中からより類似性の高い研究、トピックに関する論文を得られるか実験するために、バーチャルリアリティ分野の論文80本、自然言語処理分野の論文40本、その他の分野の論文を無作為に約200本を含めた論文データを対象に、再現率、適合率を求める実験を行った。

4.1　実験方法

1. tf・idf法による単語の重み付けを行った場合

2. アブストラクトから専門用語抽出だけを行った場合

3. 提案手法を用いてタイトル、キーワードの専門用語の重要度を付加した専門用語リストを用いた場合

の３つの方法に対し、ベクトル空間法で得られた類似度計算結果を比較する。
評価方法として、類似度計算された結果の研究分野やトピックスが同じであるなど、類似性・関連性が高いと考えられる論文なら正解、そうでないなら不正解として、正確性を評価するための尺度である適合率を求め、平均値を求めた。
また、一部主観的な評価になるが、共通していることやジャンルを分類し、類似性が高い順に5つまで並んだリストを正解データとして、完全性を評価するための尺度である再現率(recall)についても平均値を求めた。

正確性：検索質問に適合する文書だけを検索しているか？
適合率(precision)：正確性を評価するための尺度であり，検索された文書集合の

中で，検索質問に適合する文書の割合を示す．検索ノイズの少なさを示す尺度である。ここでは

$$\text{適合率} = \frac{\text{上位 5 位以内に現れた正解数}}{\text{システムが出した検索結果の数（= 5）}}$$

で定義される。

完全性：検索質問に適合する文書をもれなく検索しているか？

再現率 (recall)：検索対象となる文書集合の中の検索質問に適合する文書のうち，実際に検索された文書の割合を示す。検索漏れの少なさを示す尺度であり、ここでは

$$\text{再現率} = \frac{\text{上位 5 位以内に現れた正解数}}{\text{全ての正解数}}$$

で定義される。
正解かどうかの判定を行う基準として、次のことを考慮した。
・研究分野, 対象の絞込み
例えば、二つの論文がバーチャルリアリティの力覚提示分野論文なら、「力覚提示」というだけでは不正解。力覚の中でさらに「視覚と触覚」に関する研究なら正解。「インターフェース」だけなら不正解、両者とも「風圧インターフェース」なら正解。
結果は表記を簡略化するため、実験に用いた論文を「id50」のように、データベースに入れる際に登録した固有のキー番号で呼ぶ。専門用語抽出だけを行ったおよび提案手法の結果でえ得られる小数の値は、大きいほど類似性が高いことを示し、完全に同じ文章なら値は 1 となる。

4.2 実験結果：適合率 (precision)：

それぞれの方法での適合率の平均は以下のようになった。

結果は、提案手法が一番良い結果となり、tf・idf 法による重み付けの結果が一番悪かった。

表 4.1:

適合率平均値 (%)		
tf・idf 法	専門語抽出	提案手法
36.6	51.4	57.6

4.3 実験結果：再現率 (recall)

正解データのある論文における再現率の平均は以下のようになった。

表 4.2:

適合率平均値 (%)		
tf・idf 法	専門語抽出	提案手法
52.6	67.2	79.1

例として、3 種類の論文に対する再現率の計算過程および結果を示す。

例 1)id47 の類似論文を検索

表 4.3:

id47:正解データ	
論文番号	類似度
id48	類似性高い
id45	類似性高い
id57	類似性高い
id35	少し類似

表 4.4:

id47:tf・idf 法による結果		
論文番号	類似度	正解○/不正解×
id46	0.264	×
id48	0.259	○
id45	0.242	○
id35	0.169	○
id41	0.167	×
		再現率 0.75

表 4.5:

id47:専門用語抽出だけを行った結果		
論文番号	類似度	正解/不正解
id46	0.334	×
id48	0.333	○
id41	0.269	×
id45	0.264	○
id35	0.245	○
		再現率 0.750

表 4.6:

id47:提案手法での結果		
論文番号	類似度	正解/不正解
id48	0.328	○
id45	0.317	○
id46	0.292	×
id35	0.217	○
id57	0.189	○
		再現率 1.00

例 2)id50 の類似論文を検索

表 4.7:

id50:正解データ	
論文番号	類似度
id49	類似性高い
id38	類似性高い
id34	少し類似

表 4.8:

id50:tf・idf 法による結果		
論文番号	類似度	正解○/不正解×
id49	0.144	○
id34	0.088	○
id41	0.076	×
id58	0.074	×
id45	0.063	×
		再現率 0.666

表 4.9:

id50:専門用語抽出だけを行った結果		
論文番号	類似度	正解/不正解
id49	0.249	○
id34	0.161	○
id47	0.123	×
id37	0.114	×
id45	0.111	×
		再現率 0.666

表 4.10:

id50:提案手法での結果		
論文番号	類似度	正解/不正解
id38	0.141	○
id49	0.129	○
id40	0.109	×
id34	0.099	○
id56	0.078	×
		再現率 1.00

例 3)id53 の類似論文を検索

表 4.11:

id53:正解データ	
論文番号	類似度
id52	類似性高い
id59	類似性高い
id63	少し類似
id43	少し類似

表 4.12:

id53:tf・idf 法による結果		
論文番号	類似度	正解○/不正解×
id63	0.153	○
id52	0.135	○
id43	0.126	○
id37	0.093	×
id49	0.091	×
		再現率 0.750

表 4.13:

id53:専門用語抽出だけを行ったでの結果		
論文番号	類似度	正解/不正解
52	0.211	○
59	0.177	○
43	0.143	○
47	0.130	×
49	0.130	×
hline		再現率 0.750

表 4.14:

id53:提案手法での結果		
論文番号	類似度	正解/不正解
id52	0.272	○
id59	0.236	○
id43	0.166	○
id35	0.140	×
id49	0.101	×
		再現率 0.750

第5章　考察

全体的に、tf・idf 法による単語の重み付けを行った場合よりも、専門用語抽出だけを行った場合のほうが良い結果となり、、さらに提案手法を用いた場合のほうが良い結果となった。専門用語抽出に加えて論文のタイトル中の専門用語、キーワードを考慮した類似度検索を行うことで、類似論文の漏れが少なく、検索精度が高くなるという結果が得られた。提案手法を用いても検索漏れしてしまった論文については、人間が関連すると判断した重要な単語について、文字や表現が異なるため、専門用語抽出では現れなかったためだと考えられる。

第6章　結論

以上の実験結果から、本研究の手法を用いることによって、論文のタイトル、キーワード、アブストラクトの情報を用いて類似・関連論文の検索の精度を上げることができることがわかり、有効性を示せた。今後の課題としては、関連する専門用語間に関しての重み付けをするなどの方法で、より精度を上げて行く必要があると感じた。

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)

[6] 八太　絵美．文書間の類似度に基づく論文検索システムの開発と評価．(2002)

[7] 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)

[8] Salton, G.,“ The Vector Space Model, Automatic Text Processing.“ Addison Wesley Publishing, pp.312-325 (1985).

[9] 青木 公也, ゴ ハイ ドン, 金子 豊久．物理的モデルによる 3 次元物体の乾燥によるひび割れ表現法．電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理 No.12,pp. 1756-1764 (2003)

[10] 専門用語（キーワード）自動抽出用 Perl モジュール”TermExtract”の解説．http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html．2008 年 1 月 10 日

[11] 松本祐治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸：形態素解析システム茶筅 version 2.3.3 使用説明書, 2003.

[12] 中川裕志, 森辰則, 松崎知美, 川上大介．日本語マニュアル文における名詞間の連接情報を用いたハイパーテキストのための索引の抽出．情報処理学会論文,Vol.38 No.10,(1997 年 10 月）

謝辞

本論文を作成するにあたり、御指導をしてくださった村岡洋一教授に深く感謝します。
多くのアドバイスをしてくださった、秋岡先生に深く感謝をしています。
貴重なアドバイスをしてくださった村岡研究室のみなさまに深く感謝します。

付録

1月31日〜2月1日開催
第１２７回ヒューマンコンピュータインタラクション研究会
発表資料を添付する。

社会法人　情報処理学会　研究報告
IPSJ SIG Technical Report
2008-HCI-127 (13)
2008/2/1

収録誌:：情処研報 Vol.2008, No.11, pp. 87-91
ISSN 0919-6072