自然言語処理 128−19
（1998. 11. 6）

# 絞り込み用キーワードの抽出

塚田 政嘉，黒川 恭一

防衛大学校情報工学教室

本研究では、科学技術論文に対して類似論文検索を行う際に、検索者の所持する論文の中から検索に役に立ちそうなキーワードを自動的に抽出し、それらのキーワードを用いて類似論文を絞り込む手法について提案している。その手法は、科学技術論文の中から「既存の研究成果」等との共通点及び相違点を表現するキーワードを自動的に抽出し、それらのキーワードを用いることにより、類似論文をランキングするものである。科学技術論文70 件から、同じ分野と思われる論文 21 件を抽出し、それらをランキングした結果についても報告されている。

# Term extraction for keyword search

Masayoshi Tsukada and Takakazu Kurokawa
National Defense Academy

This paper proposes an automatic term extraction method from an original technical paper for keyword-based information retrieval and a ranking method of technical papers based on the extracted terms. Using such terms that express common part as well as different part among technical papers, the proposed method ranks papers according with their relevance to the original paper. The ranking result of 21 papers concerning with neural network parallel computing those are selected from 70 technical papers is also shown.

## 1　はじめに

近年、電子文書の発達により、多種多様な情報を容易に入手することができるようになった反面、それらの情報の中から有益な情報を効率良く検索することが困難になりつつある。このような状況の中で、文書検索に要する時間や作業量を軽減させるために、類似した文書を自動検索する研究［1］-［4］が行われている。

類似した文書の自動検索としては、文献［2］で文書の構造を利用して全文検索を行うシステムを試作している。ここでは、科学技術論文の中から「背景」「目的」「結果」「結論」等の情報を表現する文を抽出し、それらの文の意味役割を検索に反映させることにより、検索精度の向上を実現している。また、文献［3］においては、最初のキーワード検索の結果、得られた文書集合の中での出現頻度が、データベース全体の中での頻度と比べて相対的に大きい語を絞り込み検索の候補として抽出する手法を提案している。

本研究では、検索対象文書を科学技術論文に限定し、検索者の所有する論文（以後「検索元論文」と呼ぶ）中において記載された参考文献や他の論文（以後「既存の研究成果」と呼ぶ）との共通点及び相違点を抽出することにより、文書検索に役立つと思われる共通キーワード、プラスキーワード、マイナスキーワードの 3 種のキーワードを自動的に抽出する。そして、それらのキーワードを含む論文に対する点数の付与によってランキングを行い、類似論文の検索を行う方法を提案する。

## 2　キーワード検索

大量の文書の中から、興味のある文書や必要とする文書を探し出す手法の一つとして、キーワードを用いた検索がある。

キーワードを用いた検索は、処理の容易さなどがその利点として考えられる。しかし、使用するキーワードによって検索結果の量が大きく異なる［5］、検索者の検索要求とは関係のない文書を検索してしまう［2］、ヒット率の高い検索を実現するためには、検索要求式の作成が難しい［6］等といった問題も、これまでに指摘されている。

しかし、文書検索においては、文書中に出現する単語の出現頻度や出現位置、また文書の構造や表現の特徴等をもとに抽出したキーワードを使用することにより、検索精度の向上が実現されている。そのため、文書の特徴を利用したキーワード検索は、類似文書や関連の高い文書を検索する上で有効だと考えられる。

## 3　絞り込み検索

科学技術論文の検索としては、検索元論文の内容と類似した内容を持つ論文（以後「類似論文」と呼ぶ）を収集するための検索を行うことが多々ある。

しかし、科学技術論文の集合は巨大であるため、その集合の中から類似論文を検索するのは、容易ではない。そこで、類似論文を検索する際に、類似論文の候補を、科学技術論文の集合からまず抽出し、次にそれらの論文を、類似順にランキングして提示することが出来れば、類似論文の検索効率を上げることが出来ると考える。図 1 は科学技術論文の集合から類似論文の候補を抽出し、抽出した論文の集合をランキングするまでの流れを表した概念図である。

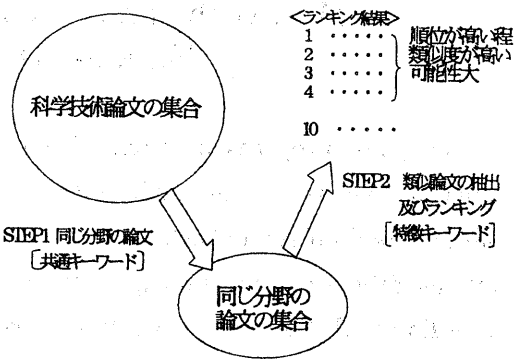



図1　類似論文を検索するまでの概念図

科学技術論文の場合における類似論文は、同じ分野や分冊に分類された中に出現しやすいと考えられる。従って、類似論文を検索したい場合は、同じ分野に分類された論文の中から検索を行えば、効果

的に類似論文を絞り込むことが出来ると考えられる。

図1のSTEP1は、科学技術論文の集合から、類似論文の候補を抽出する過程を表している。一方STEP2は、類似論文の候補の集合から、類似論文を抽出する過程を表している。なお、本研究では、類似論文を抽出するキーワードを「共通キーワード」、類似論文にランキングする働きを持つキーワードを「特徴キーワード」と呼ぶこととする。

## 4　共通キーワードによる類似論文の抽出

### 4．1　類似論文間の共通点

類似論文について考えた場合、これらの間には何らかの共通点が存在する。そこで、類似論文間に存在する共通点をキーワードとして、検索元論文から抽出することが出来れば、そのキーワードが共通キーワードとなる。

### 4．2　類似論文の抽出

今回、科学技術論文の集合から、類似論文の候補を抽出するための共通キーワードは、検索元論文の著者が、その論文に記載したキーワードを用いることとした。

## 5　特徴キーワードによる類似論文の抽出とランキング

### 5．1　類似論文の絞り込み

論文に記載されたキーワードを、共通キーワードとして類似論文の抽出を行った場合、以下のような不具合が生じることがある。

・十分な精度で類似論文を抽出できない。

・抽出された論文数が大量。

そこで、検索元論文中から、絞り込み検索に役立ちそうなキーワードを更に抽出すれば、より類似した論文を抽出することができる。本研究で、この絞り込みのために抽出するキーワードは、以下に示す3種である。まず1つめは、検索元論文の著者が、その論文に記載したキーワード以外のキーワードで、共通キーワードになるもの。もう1つは、検索元論文の特徴を強く表しているもの。残りの1つは、既存の研究成果の特徴であり、かつ検索元論文の内容とは関係の低いキーワードである。これら3種のキーワードの包含関係に従って、論文に対して点数付与を行い、その点数によってランキングを行い、上位のものを類似論文とすることとした。

### 5．2　共通キーワードの自動抽出

#### 5．2．1　共通キーワードと論文の類似度

4．で述べたように、類似論文の間には何らかの共通点が存在する。そして、論文間の共通点が多いほど、それらの論文は似ていると考えられる。

科学技術論文の場合は、検索元論文中で述べられている参考文献や他の論文（以後「既存の研究成果」と呼ぶ）が類似した論文、もしくは関連性を持った論文と考えられる。そこで、検索元論文から既存の研究成果との共通点を共通キーワードとして抽出することができれば、それらの共通キーワードを多く含む論文ほど、検索元論文に似た論文となる。例えば、検索元論文から抽出された共通キーワードが10種類だったとする。ここで、論文A,B,Cがそれぞれ3，8，5種類の共通キーワードを含んでいたとすると、類似順は論文Bが一番高く、以降C,Aの順となる。

#### 5．2．2　共通キーワードを含む文の抽出ルール

共通キーワードを抽出するためには、検索元論文の中から、既存の研究成果について記述されていると思われる文（以後「既存の研究成果に関係した文」と呼ぶ）と、検索元論文の内容に、特に強く関係すると思われる文（以後「検索元論文に強く関係した文」と呼ぶ）を抽出しなければならない。そのためのルールを、それぞれ表1と表2に示す。

表1　既存の研究成果に関係した文の抽出ルール

| | |
|---|---|
| ルール1 | 「従来」「既存」「今日まで」「これまで」「いままで」等の過去を表す表現が出現する文 |
| ルール2 | 「提案されている」「報告された」等の表現が出現する文 |

表1、表2の抽出ルールにおける基本的な考えは、既存の研究成果に関係した文と、検索元論文の内容

に強く関係する文の全てを検索元論文中から抽出するのではなく、単純なルールの中で、できるだけ適合率を高めようとするものである。

表2　検索元論文に強く関係した文の抽出ルール

| | |
|---|---|
| ルール1 | 「本論」「我々」「著者」等の表現が出現する文 |
| ルール2 | 「提案する」「報告する」等の表現が出現する文 |

### 5．2．3　実装と評価

UNIX 上において、上記のルールに従い、科学技術論文から既存の研究成果に関係した文と、検索元論文に強く関係した文を抽出するプログラムをC言語を用いて作成した。作成したプログラムを用いて、科学技術論文 21 件に対して、(上記のルールによって) 既存の研究成果に関係した文と、検索元論文に強く関係した文の抽出を行った。既存の研究成果に関係した文として抽出された文は147文で、検索元論文に強く関係した文として抽出された文は285 文であった。

抽出結果の評価は人手により行い、既存の研究成果に関係した文が 147 文中の 112 文で 76.2%、検索元論文に強く関係した文が 285 文中の 193 文で67.7%であった。

### 5．2．4　共通キーワード抽出のためのルール

次に、共通キーワードの抽出においては、既存の研究成果に関係する文と、検索元論文の内容に強く関係した文の両方に出現するキーワードを共通キーワードとして抽出することとした。ただし、抽出されたキーワードが、漢字だけで表記され、かつ 2 文字以下のものと、カタカナ、アルファベットだけで表記されたもので、1 文字以下のものは、共通キーワードとして抽出しないこととした。これは、キーワードとしては相応しくないと思われるものを排除するためである。

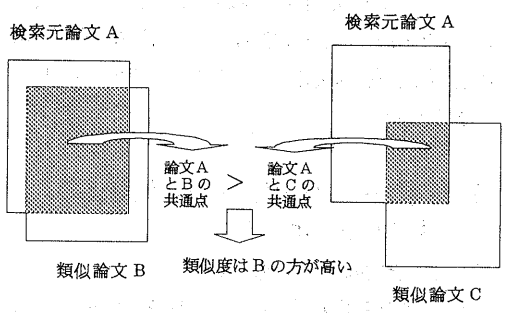
### 5．2．5　実装と評価

5．2．3と同様に、UNIX 上で上記のルールに従い、共通キーワードを抽出するプログラムを、C 言語を用いて作成した。前述の論文 21 件から共通キーワードを抽出したところ、152 個の共通キーワードが抽出された。なお、2 件の論文からは共通キーワードは抽出されなかった。

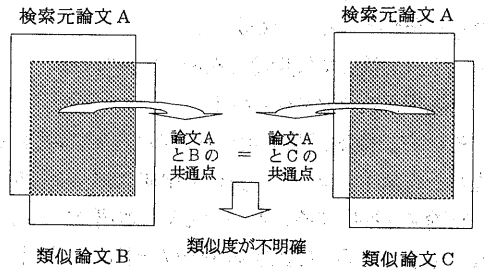
## 5．3　類似論文のランキング

5．1で述べた共通キーワードを用いることにより、それらを含む度合いに応じて、類似順のランキングを行えるが、検索元論文から抽出した共通キーワードが、類似論文の候補全般、もしくはその一部の論文に、同じように含まれる場合などには、類似順が明確にならない場合がある。

例えば、図2の(a) においては、論文Bと C に含まれる共通キーワードの種類が異なるために、類似順の違いが明確になるが、(b) のように共通キーワードの種類が同じ場合には、類似度の違いが明確にならない場合が考えられる。




(a) 共通キーワードが異なる場合




(b) 共通キーワードが同じ場合

図2　共通点の違いによる類似度の判定

そこで、共通キーワードと併用することで、より細かなランキングを行うために、図 1 に示した

特徴キーワードとして、以下に示す2種のキーワードを抽出することとした。それは、検索元論文の特徴を表したキーワード(以後「プラスキーワード」と呼ぶ) と、既存の研究成果の特徴であり、かつ検索元論文の内容とは関係の低いキーワード(以後「マイナスキーワード」と呼ぶ)である。これら共通キーワードとプラスキーワード、及びマイナスキーワードを用いて、類似論文の候補に対し点数を付与し、その点数によって類似論文をランキングする方法を提案するものである。

### 5．3．1　プラスキーワードとマイナスキーワードの抽出

プラスキーワードは、検索元論文に強く関係した文には出現するが、既存の研究成果に関係した文には出現しないキーワードとし、反対にマイナスキーワードは、既存の研究成果に関係した文には出現するが、検索元論文に強く関係した文には出現しないキーワードとすることとした。

プラスキーワードとマイナスキーワードは、共通キーワードと共に、ランキングを行う時点で用いるものであるので、4 章で用いた共通キーワードに比べて、より専門的なキーワードが適していると考える。そこで、より専門的な意味を表すキーワードを、プラスキーワード、マイナスキーワードとして抽出するために、文字種の組合せと、文字長に注目した。これは、単一の文字種で表記されたキーワードであれば、文字長が長いものほど専門的な意味を表す可能性が高く、また、単一の文字種のキーワードよりも、複数の文字種を含むキーワードほど、専門的な意味を表す語である可能性が高いであろうとの考えに基づくものである。プラスキーワードと、マイナスキーワードを抽出するためのルールを表 3 に示す。

なお、プラスキーワードとマイナスキーワードの抽出においても、共通キーワードの抽出の場合と同様に、キーワードとして適さないと思われるものを排除することとした。

表3　プラスキーワードとマイナスキーワードの抽出ルール

| | |
|---|---|
| ルール1 | 単一の文字種の場合、最も文字長の長いものを抽出。 |
| ルール2 | 複数の文字種のものは、文字長に関係なく、全て抽出。 |

### 5．3．2　実装と評価

UNIX 上で、上記のルールに従い、科学技術論文からプラスキーワードとマイナスキーワードを抽出するプログラムを、C 言語を用いて作成した。前述の論文 21 件から、プラスキーワードとマイナスキーワードを抽出したところ、抽出した結果は、プラスキーワードが 163 個、マイナスキーワードが 201 個であった。

## 6　実験と総合評価

前節までに示した方式に従い、科学技術論文から抽出した共通キーワード、プラスキーワード、マイナスキーワードにより、類似論文の候補にランキングを行うプログラムを、UNIX 上で C 言語を用いて作成し、類似度のランキングを行った。

なお、類似論文の候補に対するランキングは、点数付けによって行った。点数付けのルールは、1 種類の共通キーワードに対して 10 点を与え、プラスキーワード、マイナスキーワードに対しては、1 種類につきそれぞれ 1 点と−1 点を付与することとした。例えば、論文 A において、共通キーワードが 3 種類、プラスキーワードが 5 種類、マイナスキーワードが 4 種類出現したとすると、論文 A に与えられる点数は、30＋5－4＝31 点となる。

まず、任意に収得した科学技術論文 70 件から、共通キーワードとして「ニューラルネットワーク」を用いて、ニューラルネットワークに関係した論文 21 件を抽出した。ただし、抽出した 21 件の論文の中、検索元論文（文献 [7]）を 1 件含む。その論文から共通キーワード、プラスキーワード、マイナスキーワードを 5．2．3 及び 5．3．2 で示したプログラムを用いて抽出した。その結果、抽出された共通キーワードは 21 個、プラスキーワードは 11

個、マイナスキーワードは 6 個であった。表 4 に抽出されたキーワードを示す。

ランキングのために上述した、点数付けのルールによって、ニューラルネットワークに関係した論文 20 件に対して点数付けを行ったところ、ランキング結果は表 5 のようになった。

表4　抽出されたキーワード

| | |
|---|---|
| 共通キーワード | ニューラルネットワーク，バイナリニューロン，最小点，対称相互結合型，エネルギ関数，エネルギ最小化，Hopfield 型，制約条件，輸送問題，次形式，目的関数，最適化問題，Takeda，最適解，近似解，Hopfield，ニューラルネットワークモデル，ネットワーク全体，アルゴリズム，状態空間，検索能力 |
| プラスキーワード | 並列アルゴリズム，Hitchcock，対称相互結合型ニューラルネットワーク，線形計画問題，ijk 番目，入力バイアス，輸送コスト，ニューラルネットワーク表現，バイナリニューロンアレイ，シミュレーション結果，シンプレックス法 |
| マイナスキーワード | ソフトウェアシミュレーション，TSP，巡回セールスマン問題，A/D 変換器，2 値状態，線形プログラミング |

なお、このランキング結果に対する評価は、当事者が行い、その結果も表 5 に併せて示されている。ニューラルネットワークは、大きく学習型と、非学習型の 2 つに分けられるが、検索元論文は非学習型である相互結合型バイナリニューラルネットワークを用いたものである。このタイプのニューラルネットワークは、組合せ最適化問題に多く適用される。ランキングの 1~5 位の論文は、どれも相互結合型ニューラルネットワークを用いて、組合せ最適解を解いたものであるので、類似した論文と考えることが出来る。

しかし、類似論文との類似度が有ると判定された論文の内の 1 件のランキングが 14 位となり、その

表5　ランキング結果

| 順位 | 得点 | 論文のタイトル | 評価 |
|---|---|---|---|
| 1 | 66 | 多種フロー問題へのニューラルネットワークの適用に関する研究 | ◎ |
| 2 | 62 | ニューラルネットワークによる段階割り当て問題の並列解法 | ◎ |
| 3 | 60 | ニューラルネットワークによる周波数最適割当の解法 | ◎ |
| 4 | 51 | 集合被覆問題用ニューラルネットワークとその論理設計への応用 | ◎ |
| 5 | 41 | ECM 問題への出力更新間隔可変ニューラルネットワークの適用 | ◎ |
| 6 | 40 | ニューラルネットを用いたカラー画像における動的領域分割 | △ |
| 6 | 40 | ニューラルネットによる最小コスト経路設計方式の提案 | △ |
| 6 | 40 | 相互結合型バイナリニューラルネットワークのハードウェア化 | △ |
| 9 | 30 | ホップフィールド型ニューラルネットワークの連想メモリ効果を用いた超音波3次元イメージング系の後処理 | △ |
| 9 | 30 | 階層型ニューラルネットワークに基づく非線形自己回帰モデルによる高次ス | × |
| 11 | 21 | 仮想並列計算機システムによるニューラルネットワークシミュレーション | △ |
| 12 | 20 | 並列計算機上の誤差逆伝搬学習法の並列学習モデル | × |
| 13 | 19 | ホップフィールドニューラルネットワークと遺伝的アルゴリズムを組み合わせた最適化手法 | △ |
| 14 | 10 | ダブルトラックを用いたアレー再構成のニューラルネット解法 | ○ |
| 14 | 10 | 複数機能自己組織化ニューラルネットワーク | × |
| 14 | 10 | 回転物体の方位と形状を同時認識する回転拡散型ニューラルネットの認識特性 | × |
| 14 | 10 | 階層形ニューラルネットワークとホップフィールドネットワークを用いた画像復元 | △ |
| 18 | 0 | 拡散連想型ニューラルネットの拡散パターンによる物体位置と形状の同時認識 | × |
| 18 | 0 | ニューラルネットを用いた気象レーダ画像による降雨・降雪予測 | × |
| 18 | 0 | ニューラルネットを用いたソフトウェア信頼性予測モデル | × |

論文よりも、類似度が低いと判定された論文の順位の方が、高くなってしまっている。また、18 位にランキングされた論文は、学習型もしくはアナログのニューラルネットワークを用いたものなので、検索元論文との類似度は低いと判定された。従って、ランキング結果が低かったのは良いが、同じ分野の論文にもかかわらず、点数が付与されなかったという問題点もあった。。

今回提案した手法では、検索元論文に強く関係した文と、既存の研究成果に関係した文の抽出精度が、ランキングのための 3 種のキーワードの抽出結果に大きな影響を与える。従って、さらに精度の高い類似論文検索用キーワードの抽出ルールを考える必要がある。

## 7 結論

本研究では、検索元論文から類似論文の絞り込み検索に役立ちそうなキーワードを自動的に抽出し、それらのキーワードを用いて類似論文を抽出する手法を提案した。今回提案した手法によって、類似した論文の中でも、特に類似している論文は比較的高い精度で抽出をすることはできたが、それ以外の論文のランキングの精度が不安定であった。

今後は、実験対象とする科学技術論文の数を増やし、より詳細な検討を進めていく方針である。

## 参考文献


[1] 全裕里, 石間衛, 藤井敦, 石川徹也："ユーザの情報利用目的に基づく検索システム", 情報処理学会研究報告, NL127 - 5, pp.33 - 38, 1998.

[2] 三池誠司, 住田一男："文の意味役割解析に基づく全文検索", 情報処理学会研究報告 FI - 34 - 3, pp.17 - 24, 1994.

[3] 井上孝史, 林崎正之, 早川和広, 田中一男："絞り込み検索語候補の抽出に関する一検討", 第56 回情処全大, 3Y - 3, 1998.

[4] 篠原靖志："文書検索システム ExtractRequest における用語分析マップによるフィードバックの評価" 情報処理学会研究報告, 98 - DBS - 115, pp49 - 56, 1998.

[5] 長尾 真編：自然言語処理, 岩波書店, 1996.

[6] 芥子育雄, 乾隆夫, 石鞍謙一郎："大規模文書データベースからの連想検索", 信学技報, AI92 - 99, pp.73 - 80, Jan. 1993. 徳田克己, 塩見隆一, 青山昇一, 柿ケ原康二："分類パターンを用いた文書データの自動分類", 情報処理学会研究報告, NL123 - 9, pp.65 - 72, 1998.

[7] 土村将範, 狐塚茂樹, 黒川恭一："バイナリーニューロンによるニューラルネットワークを用いた輸送問題の並列解法", 電子情報通信学会 1992 年春季大会, pp.6 - 42.