類似論文ランキングの一手法

2U－7

塚田政嘉，　黒川恭一
防衛大学校情報工学教室

はじめに

電子文書の発達による情報量の飛躍的な増加の中で、効率的な文書検索の重要性が高まっている[1]。

本研究では、検索対象文書を科学技術論文に限定し、ランキングにより類似論文を検索する手法を提案する。その手法は、検索者の所有する論文（検索元論文）中から、類似論文のランキングに役立つと思われるキーワードを自動抽出し、それらのキーワードによって、他の類似論文をランキングするものである。本研究では、情報処理学会論文誌等に記載された100件程度の論文に対してランキングを行ったので、その結果についても報告する。

2　類似論文ランキングの過程

科学技術論文の検索は、検索元論文と類似した論文を収集するために行われることが多い。しかし、科学技術論文の集合は巨大なため、その全ての中から類似論文を検索する事は容易ではない。

そこで本研究では、類似論文を検索する際に、図1に示すように、まず類似論文の候補を科学技術論文の集合から抽出（STEP1）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(STEP2)ことにより、類似論文の検索効率を上げられると考えた。

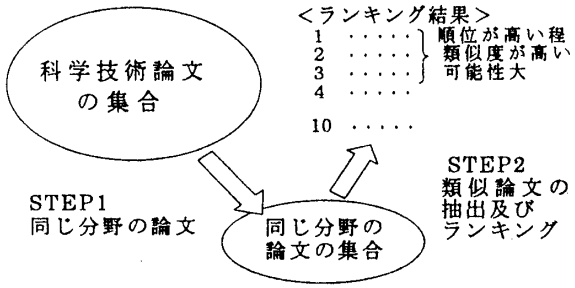

図1:ランキングまでの流れの概念図

ただし本研究では、図1のSTEP1のを研究の対象とはしておらず、STEP1で抽出された同じ分野の論文集合に対してキーワードによる点数付けを行い、その値を元にしたランキングを主として行った。

3　ランキング用キーワード

類似論文に対してランキングを行うためには、まず検索元論文からランキングのためのキーワードを決定し、抽出する必要がある。本研究では、類似論文のランキングを行うために、大きく2つのタイプのキーワードを抽出することとした。1つは検索元論文と類似した内容を持つ可能性がある論文を抽出し、大まかなランキングを行うためのキーワード(共通キーワード)であり、もう1つは、類似論文を細かくランキングするためのキーワード(特徴キーワード)である。これら2つのタイプのキーワードを併用することで、類似度順に細かくランキングを行うことができる。

3.1　共通キーワード

類似論文間には、何らかの共通点が存在するはずであり、検索元論文からその共通点を抽出し、それを共通キーワードとした。

3.1.1　2つのタイプの文の抽出

共通キーワードを抽出するために、まず既存の研究成果について記述されている文(既存の研究成果文)と、検索元論文の内容と強く関係した事項が記述されている文(検索元論文に強く関係した文)の2つのタイプの文を抽出した。

表1及び表2に、両タイプの文の抽出ルールを示す。

表1:既存の研究成果文の抽出ルール

ルール1	「従来」「古くから」「これまで」等の表現が出現する文
ルール2	「提案された」「報告された」等の表現が出現する文

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」「本研究」「我々」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

また、上記のルールに合致した文もしくは、その次の文の文頭が「そして」「また」等の接続詞で始まる場合には、ルールに合致した文の前もしくは次の1文も、対応するタイプの文とした。

3.1.2　共通キーワードの抽出

上記両タイプの文共に出現するキーワードを共通キーワードとすることとした[2]。

なお、英語のように語の区切りが明確でない日本語において、文書中からキーワードを抽出するためには、文章を語の単位に分割しなければならないが、今回語の分割については、日本語形態素解析システムChasen Ver.1.5[3]を用いて行った。

3.2　特徴キーワード

3.2.1　特徴キーワードの必要性

3.1で述べた共通キーワードは、類似論文間の共通点を表しているので、共通キーワードを多く含む論文ほど、類似度が高いと考えた(図2)。

しかし、いくつかの類似論文において、同じような頻度で共通キーワードが出現した場合には、類似度の大小関係が不明確になる可能性がある。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ものとして、特徴キーワードを抽出することにした。

A ranking method of similar technical papers
Masayoshi Tsukada and Takakazu Kurokawa
National Defense Academy, 1-10-20 Hashirimizu, Yokosuka 239-8686, Japan

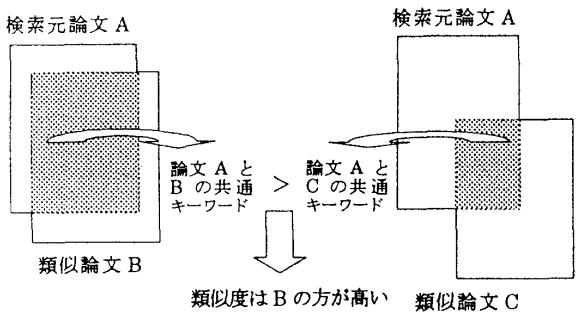

図2:共通キーワードと類似度

3.2.2 特徴キーワードの抽出

特徴キーワードは、2種のキーワードからなり、1つは検索元論文の特徴を強く表しているキーワード(プラスキーワード)であり、もう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(マイナスキーワード)である。これらのキーワードを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った。表3に、プラスキーワードとマイナスキーワードの抽出ルールを示す。

表3：プラス及びマイナスキーワードの抽出ルール

プラスキーワード	検索元論文に強く関係した文には出現するが、既存の研究成果文には出現しないキーワード
マイナスキーワード	既存の研究成果文において「不適当」「不十分」等の否定的な表現を含む文に出現するキーワードで、検索元論文に強く関係する文には出現しないキーワード

4 点数付け

ランキングを行うに当たり、各キーワードへの重み付けは、tf*idf[4]を用いた。今、Nを検索対象である文書の総数、tf_iを検索元論文におけるキーワードT_iの出現頻度、df_iをキーワードT_iを含む文書数とする。検索元論文から抽出されたキーワードT_iの重みW_iは以下の式によって与えた。

$$W_i = tf_i * \log (N/df_i) \qquad (1)$$

上記の式によって与えられた重みを各ランキング用キーワードの点数とし、それらのキーワードを含む論文に対して点数を付与した。

5 実 験

今回、科学技術論文として使用したのは、情報処理学会論文誌等に掲載された、情報工学分野の論文約100件である。まず、これまで述べてきたルールに従い、検索元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自動抽出し、それらのキーワードを元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た。そして、約100件の論文の中から、検索元論文として任意の論文を選び、上述のプログラムにより他の論文のランキングを行った。表4は、検索元論文をそれぞれニューラルネットワーク及びFPGAに関する論文とした時の、ランキング結果の上位10位までを示したものである。表4の結果から、上位にランキングされた論文は、右列最下段にランキングされた論文を除き、いずれも類似した論文と考えることが出来た。なお、ランキング結果に対する評価は、当事者が行った。また、ニューラルネットワークにおける9，10位にランキングされた論文は、タイトルからはニューラルネットワークと関連のない論文のようにも思えるが、その内容はニューラルネットワークに関連のあるものでり、提案した手法を用いることにより、論文の内容を反映した検索が実施可能であることが確認できた。

表4：ランキング結果

	論文タイトル		論文タイトル	
順位	検索元論文「ニューラルネットワークによる輸送問題の並列解法」	点数	検索元論文「多出力特性を利用したブロック統合によるFPGA回路最適化」	点数
1	ニューラルネットワークによる兵器割り当て問題の並列解法	520	ブロックからの複数出力と優先度付SFFDsを用いたFPGAブロック数最小化手法	822
2	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	475	エラー補償手続きに基づくFPGA回路最適化手法	725
3	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	360	許容関数に基づいた表参照型FPGAの最適化手法	400
4	ニューラルネットワークによる周波数最適割当の解法	346	FPGAを対象とした低消費電力指向配置・概略配線同時処理手法	348
5	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	316	テーブル参照型FPGAにおける論理ブロックの検査法	342
6	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	274	フレキシブル通信処理向けFPGA/MPU強結合システム	254
7	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	260	メモリバーンイン装置のテストパターン生成・制御回路のFPGAによる実現	253
8	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	214	通信向けFPGAおよび専用CADシステム	238
9	複数FPGAによるラピッドシステムプロトタイピング環境	193	FPGAおよびCADツールの同時評価システム	236
10	WSIを用いた自己組織化マップのフォールトトレランス	179	ATM網におけるデータトラヒックのEnd-to-Endの性能評価	232

6 むすび

本研究では、検索元論文から類似論文のランキングに役立つと思われるキーワードを自動抽出する手法を提案し、それらのキーワードの重複度を用いてランキングする手法を提案した。提案したランキング用キーワードの抽出ルールは単純なものであったが、類似した論文を上位にランキングすることが出来た。今後は、実験対象の論文数を増やし、より詳細な検討を行う方針である。

7 参考文献

[1] 塚田，黒川："マイナスキーワードの自動抽出"，第57回情報処理学会全国大会講演論文集，6R-8，1998.

[2] 塚田，黒川："絞り込み用キーワードの抽出"，情報処理学会研究報告，NL128-19，pp.135-141，1998.

[3] 松本，北内，山下，平野，今一，今村："日本語形態素解析システム『茶筌』version1.5使用説明書"，NAIST Technical Report, NAIST-IS-TR97007, July，1997.

[4] G. Salton and C. S. Yang："On the specification of term values in automatic indexing." Journal of Documentation, Vol. 29, No. 4, pp. 351-372, 1973.