4Q－10

技術文書標題からのキーワード抽出

望月 泰行, 藤井 洋一, 鈴木 克志, 丸山 冬樹

三菱電機(株) パーソナル情報機器開発研究所

1　はじめに

最近、全文検索システムの開発（参考文献[1]等）が盛んである。

検索に使用される可能性の高いフリーキーワード（名詞、サ変名詞、形容動詞）をテキストから自動抽出するシステムを試作した。抽出したフリーキーワードをキーとして索引を作成し、文書管理システムのキーワード全文検索（図1）において利用することを前提としている。フリーキーワードを字面の解析と辞書を用いた絞り込みよって抽出するのが特徴である。また、フリーキーワードを網羅的に抽出するため、フリーキーワード検索は疑似的なフルテキスト検索となり、索引を用いた検索はフルテキスト検索よりも高速である。

主な検索の対象を技術文書とする。技術文書やその標題には複合語が多い。フリーキーワードの抽出では複合語から有効な文字列を網羅的に抽出することを考えた。

社内技術文書の標題からフリーキーワードを抽出する実験を行ない、適合率、再現率を計算した。

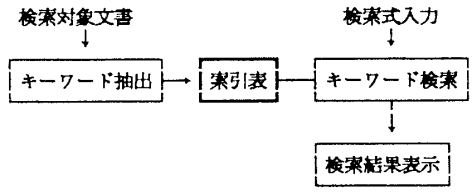

図1．キーワード全文検索

2　フリーキーワードの抽出アルゴリズムの概要

フリーキーワードの抽出アルゴリズムは、フリーキーワード候補の切り出し、フリーキーワード候補の部分文字列の抽出、フリーキーワードの絞り込みの3段階に分けることができる。

2.1　フリーキーワード候補の切り出し

字種によってテキストを分割することにより、フリーキーワード候補の切り出しを行なう。日本語のテキストを構成する文字の種類は、平仮名、片仮名、漢字、数字、アルファベット、記号からなる。このうち、平仮名と記号は文法的に機能的な役割を果たすことが多いので、これらを切り落とした残りの文字列をフリーキーワード候補とする。

2.2　部分文字列の抽出

一般にフリーキーワード候補は複合語を形成していると考えられる。複合語の部分文字列としてのフリーキーワードを抽出するために、複合語の切れ目となることが可能性な分割点を網羅的に設定し、その分割点から部分文字列切り出す。分割点を設定する方法は以下の3つである。

(1) 字種による分割　片仮名、漢字、数字、アルファベットの4つの字種の境目を分割点とする。ただし、例外的処理として、次の4点の処理を行なう。

・「々」を直前の文字で置き換える
・「・」（中黒）の前後が片仮名である場合は、これを削除する
・「－」「／」「&」は前後がアルファベットか数字のときは、これらをアルファベットとして扱う
・数字は、「4MDRAM」のように名詞の一部として用いられることがあるので、数字の前後にアルファベットまたは「－」「／」「&」があるときは、これをアルファベットとして扱う

(2) 接辞による分割　接頭語、接尾語の前後に分割点を設定する。ただし、部分文字列を切り出す際には、次のルールを適用する。

A Keyword Assignment Method for Technical Documents
Yasuyuki MOCHIZUKI,　Youichi FUJII,　Katsushi SUZUKI,　Fuyuki MARUYAMA
Mitsubishi Electric Corp.

・接頭語で終わる部分文字列は切り出さない
・接尾語で始まる部分文字列は切り出さない

(3) 辞書を用いた分割　基本語辞書を用いて分割点の設定を行なう。フリーキーワード候補の部分文字列が見出し語と一致したとき、その部分文字列の両側に分割点を設定する。

2.3　フリーキーワードの絞り込み

切り出した部分文字列は膨大な量となるため、抽出した部分文字列に対して、フリーキーワードとして相応しくないものや、検索に用いられることが少ないと予想されるものを削除する。

(1) 後接文字による品詞の推定　部分文字列に平仮名が後接している場合は、その部分文字列の品詞を推定することが可能である。フリーキーワードには、助詞、一部の助動詞、サ変の活用語尾、形容動詞の活用語尾が後接するので、それら以外の平仮名が後接している場合は削除する。

(2) 不要語の削除　部分文字列全体が基本語辞書の見出し語と一致する場合は削除する。基本語辞書の見出し語は一般的な語であるため、検索のキーとして用いられることが少ないためである。

(3) 文字数による絞り込み　長い文字列は検索に用いられることが少ないため、字数制限を設けることで、フリーキーワードを絞り込む。ただし、片仮名またはアルファベットの単一字種からなるフリーキーワードは字数制限の対象外とする。

切り出しから絞り込みまでの処理により、例えば、「機械」、「翻訳」が基本語辞書に存在する時、「マシン環境再確認」と「日英機械翻訳」からはそれぞれ次のフリーキーワードが抽出される。

「マシン環境再確認」
　マシン　　マシン環境　　環境
　環境再確認　　マシン環境再確認

「日英機械翻訳」
　日英　　日英機械
　機械翻訳　　日英機械翻訳

3　実験と評価

社内技術文書９２１１５件の標題から、フリーキーワードを抽出する実験を行なった。辞書は、基本語辞書から２文字漢字見出しだけを取り出したものを主記憶上に展開して使用した。また、文字数制限は１０文字以下とした。

抽出フリーキーワード数：*999,252*個

１標題あたり、約１０．８件のフリーキーワードが抽出される。抽出したフリーキーワードで検索用の索引を作成すると、巨大な索引となることが予想される。しかし、標題は本文と比較して接続詞や助詞などの平仮名の割合が低い。本文からフリーキーワードを抽出した場合、索引サイズの本文に対する比率は、上記の値から計算されるよりも小さな値となる。

また、社内技術文書の標題からランダムに抽出した１９３件に対して人手でフリーキーワードを抽出し、再現率と適合率を計算した。

再現率　*83.1%*，　適合率　*34.9%*

再現率を下げている最大の要因は、人手で抽出したフリーキーワードに、１１文字以上の文字列と平仮名混じりの文字列が含まれていることである。

また、フリーキーワードを網羅的に抽出しているため適合率が低いのは当然であり、これが疑似的なフルテキスト検索を実現している所以である。

4　おわりに

キーワード全文検索のための索引を自動作成するためのフリーキーワード抽出システムを試作した。本システムは字面の解析の後で基本語辞書の検索を行なうので、形態素解析よりも高速である。社内技術文書の標題を用いて抽出実験を行ない、フリーキーワードのサイズ、検索再現率、検索適合率を計算した。

現状のシステムの最大の問題点は、漢字平仮名混じりのフリーキーワードが抽出されないことである。これに対しては、必要語辞書を用意することを考えている。

参考文献

[1] 別所，他：テキストデータベースのためのキーワード抽出法，情報処理学会第４５回全国大会 3S-1(1993)