# 関連文書検索システムの開発 (3) － 複合語辞書 －

2U－5

梅基 宏†, 永峯 猛志†, 石飛 康浩†, 倉持 勉†, 倉橋政之†, 増市 博†‡, 舘野昌一†

†富士ゼロックス (株) IT 事業開発部 ‡スタンフォード大学 CSLI

## 1　はじめに

日本語の文書を対象にした全文検索の索引に、単語をキーワードとして登録する場合、テキストはあらかじめ分かち書きされておらず、ことばの切れ方が自明ではない。そこで、索引に登録されるキーワードは、形態素解析をテキストに施すことを通して抽出される。

したがって基本的に、形態素解析の辞書にない語はキーワードとして抽出されないので、ユーザが複合語をキーワードとして指定したときに検索できないことがある。

## 2　従来の索引構成と問題点

索引に登録されているキーおよび値とから、従来の索引構成を以下のように3つに大別することができる。

(1) キーとして単語、値として対応する文書集合が登録される

(2) キーとして単語、値として出現する文書中の位置情報が登録される

(3) キーとして単語と複合語、値として対応する文書集合が登録される

(1) では、複合語を構成する単語集合をキーワードとして、アンド検索を行うことができる。しかしその場合は、各単語が離れて出現する文書も多く得られてしまい、結果として検索の適合率が下がってしまう。

(2) では、検索時に複数の検索語とそれらの位置関係を入力とし、検索語が指定された位置関係にある文書を検索することができる。しかし、そのためには語に関する膨大な量の位置情報を準備する必要があり、処理にかかるコストが大きくなってしまう。

(3) では、ユーザが意図する検索を行うことができるが、単語のみを索引に登録する場合にくらべて、語を登録する索引の容量が遥かに大きくなってしまう。

以上のように、従来の索引構成 (1)(2)(3) にはそれぞれ問題がある。

## 3　提案する方法

ここで、複合語をキーワードとしたときでも、高い適合率で高速に検索でき、かつコンパクトな索引を用いる方法を提案する。本方法の骨子は次の4点である。

1. 単語が登録されている索引 (単語索引) と複合語が登録されている索引 (複合語索引) を併用する

2. 複合語を登録するときに、複合語を構成している各単語を、文字列そのものではなく、その単語が格納されている単語索引における位置で置き換えて表現する

3. 文書において連続して出現する語を複合語として登録する

4. 検索時にユーザの入力を解析し、複合語が入力されたときは複合語を構成する各単語の単語索引における位置を求め、複合語索引を検索する

---

Relevant Document Retrieval System (3) -- Compound Word Index --
Hiroshi UMEMOTO, Takeshi NAGAMINE, Yasuhiro ISHITOBI, Tsutomu KURAMOCHI,
Masayuki KURAHASHI, Hiroshi MASUICHI, Masakazu TATENO
IT Business Development, Fuji Xerox Co., Ltd.
CSLI, Stanford University

## 4　システム構成とアルゴリズム

本方法に基づくシステムは、大きく7つの要素から構成される。図1にシステム構成の概念図を示す。以降では、単語索引および複合語索引はともにトライ形式[1]と限定する。

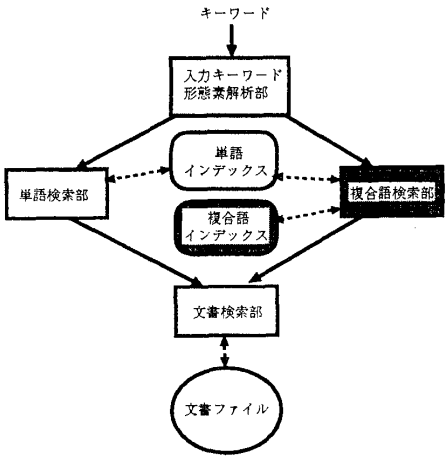


図 1: システム概念図

検索アルゴリズムのフローチャートを図2に示す。入力キーワードは形態素解析して、結果として得られる任意の2語の自立語のならびを入力キーワードの複合語とみなす。

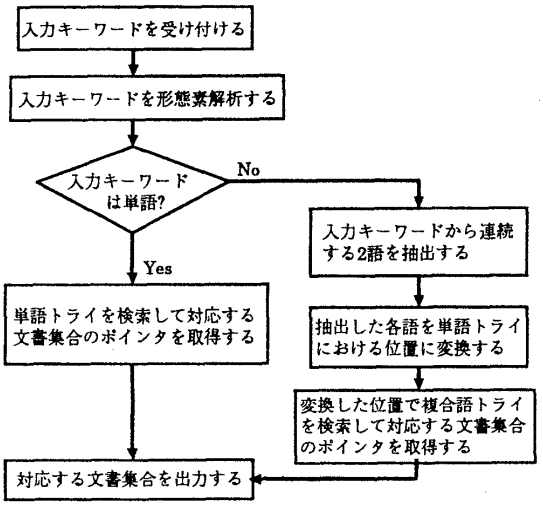


図 2: 検索アルゴリズムのフローチャート

## 5　評価実験とその結果

本方法で必要となる索引容量を評価するために、実際の文書を対象に評価実験を行った。対象文書は、特許公開公報CD-ROM1枚に収められている4,800 件(107 MB)の文書である。

結果として、従来の(3)の場合よりも48.6%の索引容量が削減できた。

また、複合語の語数は単語の語数の約4倍だが、複合語トライの容量は、単語トライの容量の77%であった。

| キーワードの種類 | 異なり語数 | トライの容量(バイト) | トライの容量比 |
|---|---|---|---|
| 単語 | 251,297 | 22,346,388 | 29.0 |
| 複合語 | 1,030,699 | 17,246,144 | 22.4 |
| 合計 | 1,281,996 | 39,592,532 | 51.4 |
| 単語＋複合語 | 1,202,967 | 76,946,780 | 100.0 |

表 1: 各索引の容量の比較

## 6　まとめ

本稿で提案した方法に基づいて、実際の文書(特許公開公報CD-ROM1枚分)を対象に評価実験を行い、キーワードを登録する索引に必要な容量が、従来よりも48.6%削減できた。また、複合語の語数は単語の語数の約4倍だが、複合語索引のトライの容量は、単語索引のトライの容量の77%であった。

## 参考文献

[1] 増市博, 山浦富久美, 小川剛弘, 舘野昌一, "形態素解析を用いた全文検索システムとその応用", 情報処理学会 自然言語処理 102-3, 1994.