An Approach to Extract Keywords Comparing Multiple Documents

Teruo KOYAMA

Abstract:

Today, most of the information retrieval systems are based on the keyword search method. Because extracting keywords from documents needs high expertise, it is costly and time consuming process. To improve this situation, automatic extraction of keywords from original documents (KWIC) is considered a promising method. In this paper, the author challenges to extract keywords based on statistical figures calculated from about 250 papers in information technology study. 50 to 70% of terms are confirmed to be proper keywords.

1.はじめに

現在の情報検索システムの多くはキーワード検索に基本をおいている。キーワード検索では特定の文献に対してどのようなキーワードを設定すれば効率の良い検索が可能となるか問題となるが、このようなキーワード設定を高い精度で行うには、専門家が選択されたキーワードに基づいて行うことが望ましい。しかしながらこのような形でのキーワード付与には大きなコストがかかり、専門家の人材の問題や、また、当該分野の最新の動向を反映したキーワード付与が困難な場合も存在する。そこで文献そのものに出現する用語を解析することにより、文献に対するキーワードを推定する試み（KWIC）が注目されている。

今回我々は、一部の文献についての形態統計結果に基づき、川原氏及び神戸に関する統計データに基づくキーワード推定の試みを行ったので、その結果について報告する。

2.検討に用いたデータ

今回用いたデータは、1994年度の情報処理学会論文誌に掲載された、日本語の単語を基とする語彙を含むデータである。解析にはこれらの論文を日本語形態解析ソフトウェアであるJUMANにより、ソフトウェアに付属する標準統計に基づいて形態統計を行った結果に基づいて行った。JUMANでは形態解析に当たって多義性のある場合の複雑のためにコストテーブルを用いるが、今回はJUMANに標準的な付属しているテーブルを用いている。また、多義性のある場合には、テーブルに基づき最も適当なと判定されたものを用いている。

3.検討の概要

3.1 基本的考え方

一つの文献に与えべきキーワードとして最も有力と考えられるのは、その文献で特徴的に出現する名詞である。ただし、専門性の高い文献の場合は、標準的な形態解析は適切でない場合や複合が数多くのキーワードとなることが多い。従って形態解析の結果に基づいてキーワードを推定するに当たって、一つの「語」を推定す
るものをいかに決定するか、および、これらの語の出現頻度をどのように取り扱うかが問題となる。

JUMANの新機能については、解析の結果得られた「新」との類縁、その機能に関する情報が含まれる。名詞を対象とする場合、語彙に基づいて名詞と判断されたものおよび語彙に存在しない未定語が統合の対象となる。

また、複合語も注目するという立場からは、文脈に名詞列は未定義語が統合して表現する頻度についても検討が必要である。

一般に一つの文を分離解釈した結果には、いくつかの名詞／未定義語の類縁が出現する。これらの類縁内、文脈、文脈または名詞／未定義語外の文脈で臨時区分されるものを以下「RUN」と呼ぶこととする。一つのRUNについて、その中の任意の部分列を類縁したものは、意味のある名詞を構成する可能性があると考えられる。したがって、複合語の性質からしてそのすべてが意味のある名詞であると考えることはできない。一方一つのRUNに含まれるすべての類縁を類縁したものは、独立した名詞として意味を持つ可能性がより高いと考えられる。

そこで、全文献から得られるすべてのRUNについて、それらから求められる類縁の類縁として類縁群を求めた上で、一つの文脈に類縁する「対象」として、その文脈に含まれるすべてのRUNについてそのすべての類縁列の類縁で、この類縁に含まれるものを考えることとした。ただし、今回は日本語キーワードについての検討を目的としたため、これらの対象内、日本語文書ではないアルファベットや特殊記号を含むものなどは対象の対象外としている。また、長さが1のもの、数字のみ、あるいは平仮名文字のみからなるものについても、有用なキーワードとなる可能性は低いと考え、検討の対象外とした。

このような処理の結果得られた、類縁の類縁を各類縁群について、各文脈に出現する類縁数を数え上げた上で、この結果に基づいてキーワードの候補を試みた。

3.2. 考慮すべき類縁数

一つの文脈に対するキーワードを、その文脈に特徴的に出現する「対象」であると考えるならば、文献ごとの出現類縁に偏りのある類縁に注意することが考えられる。この偏りを表す指標として、たとえばエイピスなどの統計値を用いることが考えられる。ただし、文献ごとに長さが異なり、出現する名詞列の数も異なっていることや、出現類縁の類縁列の少ない対象に対してマクロ的な統計的意味を持つかどうかという問題が存在する。一方、このような統計的意味があるのは、ある程度の類縁数の高い類縁についてのみである。しかしながら、キーワードの対象は、特定の文献に限って出現するため、出現類縁の類縁列数が大きくないものとも存在すると考えられる。

また、文献類縁に出現する類縁は、特にキーワードとなる可能性が大きいと考えて特別な扱いをすることも考えられる。

今回はこれらの問題を考慮して、
1. 文献に出現する全RUNの数を文献ごとに類縁する名詞列を代表するものと考え、名詞のこれらの文献をあたりの相関出数を対象として想定した上で、相関出数のT値を指標として採用する。
2. 頻度の特大なものは候補の対象としない。
3. 全類縁数が10に満たないものは候補解釈の対象としない。
4. そのかわり類縁を類縁する文脈数が10以下のものについては例外扱い。
5. 類縁に出現する類縁類縁満足に出現する対象は、基準を基本方針としている。

以上の基準に基づき、文献ごとに次のような数の候補を求める。
a: 出出現類縁のT値が大きいもの10語。
b: そのうち特にT値の大きい上位の10語。
c: 出現類縁数が10以下のものを、当該文献の出現数が3以上のもの

- 44 -
d: 当該文献に出現するcの語の内で特に出現数の多いもの上位20例。これからの合計からさらに、p = αc、およびq = b + dとする。また、表題に出現するcの語、pに含まれるものをrとし、k = q + rとする。（ここでは集計値を表す）

これからの集計の内kに含まれるものはキーワードとなる可能性が大きいと考え、キーワードの第一候補とした。このようなにして得られた集合kについて検査を行ったが、これらの語だけでは文献に対するキーワードとして十分とはいえないので、これに加え文献によって是証をみて、kはkに含まれる語の中で、3文字以上の長さを持つものと同一の文脈で出現するもので、pに含まれるもののうちで、3文字以上の長さを持つものの、を第2の候補とした。これで3文字以上の長さを持つもののみを対象としたのは、2文字語まで含めると相当多数のノイズが混入することが判明したことによる。

3.3. 検索結果

以上のようにして求められたキーワードの候補については検査するため、10文献をランダムサンプリングにより抽出し、これらの文献にてどのようなキーワードが選択されているかを調査した。まず、kのそれぞれについて、著者の判断でキーワードとして妥当なものk1、k2と妥当とは思えないものk0、k0に分類した。また、原文献を読みった結果、キーワードとして抽出したものが少ないものを除外k3とした。これらについて検査した結果の1例を次に示す。また、10の論文について、集計結果を表1に示す。

題名：ユーザインタフェースにおけるビデオ情報の構成

ki：インタラクション、カット、ビデオ、ビデオ情報、ポタン、マウス、ユーザインタフェース、ユーザインタフェース設計、映像、操作依存ビデオ情報

ko：画面、画面、映像、映像

ki：3次元モデル、アイコン、インタフェース構築、インタラクション設計、オブジェクト指向言語、カーネル、グラフィックス、通信等、映像、スクリーン、スクリーンバーや、テクストオブジェクト、ビデオカメラ、マルチメディア、メニューユー、ユーザインタフェース構築、操作依存ビデオ情報、時間依存ビデオ情報、操作依存ビデオ情報クラス、情報依存ビデオ情報

ko：アップ、アプローチ、イメージ、エンジン、オンライン、カメラ、クリック、コンピュータ、アプリケーション、タイミング、ダウン、ディスプレイ、トップ、ハードウェア、ハードディスク、メディア、ユーザ操作、ランダムアクセス、続き再、関係付け、構成変換、自動制、柔軟性、操作方法、対象物、同様に、内部状態、

ks：コミュニケーションチャネル、マウスカーソル

10文献の内、最後の文献についてはキーワード選択の効率を特に思わず、これは対象とする分野を分され特異なため、上記の推定方法が適当でないものと判断した。この文献を含むタイトルとキーワードの集計を示す。

題名：名作 שה_rollニ関する感性の効果的剖面

ki：王手、感性、評価、好意、将棋

ko：5手、5手目、5手目、手筋、コンテスト、コンテスト作品、作品、手筋、手筋問題、得点、名作

ki：ゲーム、移動移動、移動移動、感性、感性、評価、感性、評価、問題、将棋、将棋、将棋、将棋、将棋、将棋

ko：30題、5手目コンテスト、5手目問題、7手目コンテスト、7手目問題、オーバラップ、
この特異な1例を除けば、抽出された語の内、50%弱から70%弱がキーワードとして妥当なものと考えられる。また、キーワードとして抽出に失敗したもののがかなりの数にのぼっている。

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ki</td>
<td>6</td>
<td>6</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>ko</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>kš</td>
<td>7</td>
<td>10</td>
<td>19</td>
<td>16</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>22</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>kš</td>
<td>4</td>
<td>12</td>
<td>28</td>
<td>14</td>
<td>9</td>
<td>13</td>
<td>8</td>
<td>26</td>
<td>22</td>
<td>18</td>
</tr>
</tbody>
</table>

表1. サンプルされた文献に対する各集団の出現

4. 考察と今後の展望

このようなキーワードの抽出の仕方や抽出ノイズにどのような影響が現れるかについて、現在調査を進めている。文献集団全体についての新の出現の頻数、語の出現する文献の数、相出出現率の平均と標準偏差等について検討を進めているが、これまでのところはT値よりは減らし出現頻数について調整をした方が良いという感触を得ている。たとえば

1. 出現文書数の多い語は認識した方がよい。あるいは
2. 出現文書数が少ない語は認識した方がよい
3. 語の長い語については出現の頻度が少なくてもキーワードとして識別する、

等である。

今後はより多くの文書について調査を行うとともに、上記の要素を加味した上でどの程度の精度が可能であるかの検討を行っていくことを予定している。また、辞書の整備やストップワードの指示など、統制的というよりは語学的な考察によりどの程度の精度を上げることができるかについても検討を進めていきたいと考えている。