複数論文比較によるキーワード推定の試み

小山　照夫

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 considererd a promising method. In this paper, the author challange to extract keywors based on statistical figures calcurated from about 250 papers in information technology study. 50 to 70 % of terms are confermed 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が問題となるが、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、統制された語彙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るし、専門家の資質の問題や、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）が注目されてきている。

今回筆者は、一群の文献についての形態素解析結果に基づき、用語出現頻度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、日本語の題名を持つ約２５０の論文である。解析はこれらの論文を日本語形態素解析パブリックドメインソフトであるＪUMANにより、ソフトウェアに付属する標準の辞書に基づいて形態素解析を行った結果に基づいて行った。ＪUMANでは形態素解析に当たって多義性のある場合の解決のためにコストテーブルを用いるが、今回はＪUMANに標準的に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ついては、テーブルに基づき最も確からしいと判定されたもののみを用いている。

3. 検討の概要

3.1. 基本的考え方

一つの文献に付与すべきキーワードとして最も有力と考えられるのは、その文献に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、標準的な辞書には採録されていない語や複合語が重要なキーワードとなることが多い。従って形態素解析の結果基づいてキーワードを推定するに当たって、一つの「語」として認定す

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の解析結果には、解析の結果得られた「語」の候補と、その品詞に関する情報が含まれる。名詞を対象とする場合、辞書に基づいて名詞と判断されたものおよび辞書に存在しない未定義語が検討の対象となる。また、複合語にも注目するという観点からは、文の中に名詞ないしは未定義語が連続して出現する並びについても検討を加える必要がある。

一般に一つの文を形態素解析した結果には、いくつかの名詞／未定義語の並びが出現する。これらの並びの内、文頭、文末または名詞・未定義語以外の品詞で両側を区切られるものを以下「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連接したものは、意味のある名詞を構成する可能性があると考えられる。ただし、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語を連接したものは、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最長の接合としての語群を求めた上で、一つの文献に出現する「語」として、その文献に含まれるすべてのＲＵＮについてそのすべての部分列の連接で、この語群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、これらの語の内、日本語文字でないアルファベットや特殊記号を含むものは検討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、語の候補となる各文字列について、各文献に出現する頻度を数え上げた上で、この結果に基づいてキーワードの推定を試みた。

3.2. 考慮すべき統計量

一つの文献に対するキーワードを、その文献に特徴的に出現する「語」であると考えるならば、文献ごとの出現頻度に偏りのある語に注目することが考えられる。この偏りを表す標として、たとえばカイ２乗値などの統計値を用いることがまず考えられる。ただし、文献ごとに長さが異なり、出現する名詞の数も異なっていることや、出現頻度の絶対数の少ないものに対してマクロな統計値が意味を持つかどうかという問題が存在する。一方、このような統計的な値に意味があるのは、ある程度出現頻度の高い語についてのみである。しかしながら、キーワードの中には、特定の文献に限って出現するため、出現頻度の絶対数が大きくないものも存在すると考えられる。また、文献表題に出現する語は、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の数をその文献に出現する名詞数を代表するものと考え、各語のこれに対する文献あたりの相対出現数を尺度として想定した上で、相対出現数のＴ値を指標として採用する、

2.頻度の特に大きいものは検討の対象としない、

3.全出現頻度が１０に満たないものについては統計解析の対象としない、

4.そのかわり語が出現する文献数が１０以下のものについては別途取り扱う、

5.表題に出現する語は論文本体に出現する語よりも重視する、

ことを基本方針としている。

以上の方針に基づき、文献ごとに次のような語の集合を求める。

a: 出現頻度のＴ値が大きいもの１００語、

b: そのうち特にＴ値の大きい上位の１０語、

c: 出現文献数が１０以下の語で、当該文献の出現数が3以上のもの、

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2割。

これらの集合からさらに、p=a+c、およびq=b+dとする。また、表題に出現する語の内、pに含まれるものをrとし、k=q+rとする。（ここで+は集合和を表す）

これらの集合の内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。このようにして得られた集合kについて検討を行ったが、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、このほかの候補として各文献について、k': kに含まれる語の内、3文字以上の長さを持つものと同一の文内に出現するもので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、を第2の候補とした。ここで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. 検討結果

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検討するため、１０文献をランダムサンプリングにより抽出し、これらの文献につい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、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ki、k'iと妥当とは思われないものko、k'oに分類した。また、原文献を読んだ結果、キーワードとして追加した方がよいものを別途ksとした。

これらについて検討した結果の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部品の構成

ki：インタラクション、カット、ビデオ、ビデオ映像、ビデオ部品、ボタン、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、操作依存ビデオ部品

ko：画面上、蓄積、様子

k'i：3次元モデル、アイコン、インタフェース構築、インタラクション機能、オブジェクト指向言語、カーソル、グラフィクス、コミュニケーション、スクロール、スクロールバー、デスクトップ環境、ビデオカメラ、マルチメディア、メニュー、ユーザインタフェース構築、位置依存ビデオ部品、時間依存ビデオ部品、操作依存ビデオ部品クラス、蓄積型ビデオ映像

k'o：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、サブクラス、タイミング、ダウン、ディスプレイ、トップ、ハード、ハードディスク、メディア、ユーザ操作、ランダムアクセス、応用例、関連づけ、構成要素、自動車、柔軟性、操作方法、対象物、同期こ、内部状態、被写体

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推定の効率が特に悪いが、これは対象とする分野がかなり特殊なため、上記の推定法が適当でないものと判断した。この文献についてタイトルとキーワードの集合を示す。

題名：名作詰将棋における感性の定量的評価

ki：王手、感性、詰将棋、好感度、将棋

ko：5手、5手詰、7手、7手詰、コンテスト、コンテスト作品、作品、手詰、手詰問題、得点、名作

k'i：ゲーム、移動回数、移動駒、開放度、感性情報、詰将棋問題、玉移動、好感概念、好感度要因、相関係数、探索局面、探索局面数、特徴量、日本将棋連盟、評価モデル、不動駒

k'o：３０題、5手詰コンテスト、5手詰問題、7手詰コンテスト、7手詰問題、オーバラップ、

コンテスト作品群、コンテスト問題、シンボル、ヒント、作品群、市販問題集、集計値、創作問題、総合得点、平均得点、問題集、問題例

ks：感性情報処理、感性評価データ、詰将棋データベース、合い駒、手数、手順、余詰、早詰、変同、駒数、駒種、駒捨て、開き王手、両王手、紛れ

この特殊な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がかなりの数にのぼっている。

	1	2	3	4	5	6	7	8	9	10
ki	6	6	11	10	8	12	10	12	9	5
ko	6	6	3	5	3	3	2	9	1	11
ḱi	7	10	19	16	9	10	12	22	24	18
ḱo	4	12	28	14	9	13	8	26	22	18
ks	5	9	2	19	10	7	16	8	5	15

表1．サンプルされた文献に対する各集合の語数

4. 考察と今後の展望

このようなキーワードの抽出もれや抽出ノイズにどのような要因が関係しているかについて、現在調査を進めている。文献集合全体についての語の出現の絶対数、語の出現する文献の数、相対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値よりはむしろ出現頻度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.出現文献数の多い語は省略した方がよい、あるいは

1'.出現文献数によって、文献ごとの出現数の評価を変える方がよい、

2.出現頻度の大きい語は積極的に切り捨てた方がよい、

3.語長の長い語については出現の絶対頻度が少なくてもキーワードとして採用する、

等である。

今後はより多数の文献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推定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、統計的というよりは言語学的な考慮によりどの程度推定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。