シソーラスにおける語群を用いたキーワード自動抽出法の検討

3G−7

海老原 一郎　　岡田 義邦

電子技術総合研究所

1 はじめに

テキスト型データをデータベースに登録する際、少ない情報量で原文の特徴を登録する方法としてキーワードが一般に使用されている。キーワードを用いる方法は検索システムの記憶容量を節約でき、また、検索処理を迅速化・効率化するといった利点を持つ。そのため、キーワードは新聞記事や学術論文等での検索に広く使用されてきた。現在では、構文解析技術の発達や電子辞書の開発により、日本語文から名詞節を自動的に抽出することは高い精度で行えるようになった。しかし、抽出された全ての名詞をキーワードとすると原文の話題を特徴付けるために必要なキーワードの数倍程度不必要な名詞をキーワードとして抽出してしまうため、不必要な語を削除する研究 がなされている。

筆者らは、シソーラスにおける上位語／下位語・同義語などの関係から形成できる語のグループ化を使ってキーワード間の関係を解析し、それに文書中でのキーワード使用の状況の特徴を加味して不要キーワードを検出し削除する方法を検討中であり、その概要を報告する。

2 現在のキーワード自動抽出の問題点

自動抽出の場合のキーワードの冗長性は必要量の数倍といわれており、記憶容量・検索時の適合率・データの高度利用への応用の各観点から冗長性の改良を要求されている。キーワードの冗長性を減らす研究はいくつか行なわれているが、各キーワードを独立に扱っている場合が多い。ここでは、シソーラスを利用してキーワード候補語のグループ化を行ない、キーワードの冗長性を改善する方法を提案する。

3 キーワード候補語の抽出

まず、原文からキーワードの候補となる語をシソーラスを使用した統制コントロール方式により抽出し、その語に以下に示すように重みを与える。ここでは対象を特定分野（科学技術系学術論文）に限定して考える。

1. 日本語文を構文解析して単語単位に分かち書きにする。
2. キーワードは名詞について作成するのが一般的なので、各節の中から動詞・形容詞・助詞等を取り除き、名詞のみを抽出する。
3. 抽出された名詞の中から、対象分野のシソーラスに記載されている名詞を選び、キーワード候補語とする。
4. 上記候補語に重みを付加する。この時、単に文中での各語の出現回数を数えるだけでなく、
   (a) 文中の語の出現位置の特徴を加味して頻度を付ける。具体的には、題名・副題に出現する語に高い重みを与える。次に、「はじめに」・「概要」、「終りに」・「結論」等の言葉に続く節の中に現れる語に高い重みを与える。
   (b) 語の文章中での役割に応じて重みをつける。具体的には、並列表現の中に出現した候補語には低い重みを与える。連体修飾語の中に現れた候補語には低い重みを与える。主語、目的語の中に現れた語には高い重みを与える。

とする。以上の様にしてキーワード候補語とその重みが得られる。

4 不要語を削除するための方針

上記操作によって得られたキーワード候補語の中から、文書の主旨より遠いと判断される語を不要語として削除するための基本的な方針を説明する。この時、次の仮定を用いて不要語の削除処理を行なう。

仮定：ある文書がある主題について記述しているとき、その主題に意味の近い名詞が文中に多く出現する。

ここで主題とは文書の中で扱われる主要な話題を指し、一般には文章中に複数個存在する。候補語の取捨は、基本的にはその語の頻度に依存して行なうが、単に一単語の頻度のみによってその選択を決めるとすると、文書の話題に近い語でも出現回数が足らずに切り捨てられ、反対に、話題から離れた語でも文書中で引用される回数が多いと選択されてしまう可能性がある。これらの間違いを減らすために意味の近い単語どうしを集めてグループを作り、グループの重要さをそのグループが含む全ての候補語の頻度の総計で表して、グループ単位で取捨を行なうようにする。このようにすれば、頻度の低い単語でも文書の話題に適合したものは仮定より単語数の大き

An Automatic Indexing Method using Word Groups on Thesaurus

Ichirou EBIHARA, Yoshikuni OKADA

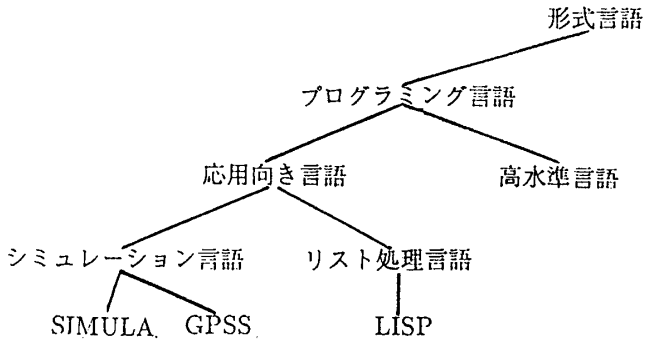
Electrotechnical Laboratory

なグループに属することができ、キーワードとして選ばれる可能性が高くなる。反対に頻度の高い単語でも文書の話題から外れた単語が属するグループは所有する単語数が少ないのでグループ全体の重要度は相対的に低くなり、この語がキーワードとして選択される可能性も低くなると期待される。

5 候補語のグループ化とキーワードの決定

上記アルゴリズムで問題となるのが語のグループ化の方法である。グループ化の方法はパターンのクラスタリングを行なうk-means法やward法などについて検討したが、グループ化を行なうか否かの判断基準を定めることが困難と思われたので、判断が簡単に行なえる以下の方法を検討中である。ここで用いるシソーラスでは、一つの語は上位語／下位語・同義語の記述を持つとする。このシソーラスを使って同義語を一つのノードで表し、上位語-下位語間を結ぶと、キーワード候補語間の関係が図1の様な木構造に類似したグラフで表される。

図1 キーワード木の例



上位語下位語の関係で連結したグラフはこれをキーワード木と呼ぶことにする。キーワード木の各ノードn(i)は重要度w(i)を持つ。重要度w(i)はその親のノードn(i-1)の重要度w(i-1)と親のノードn(i-1)がシソーラス上で持つ子の数p(i-1)により決定されるものとする。

$$w(i) = F(w(i-1), p(i-1)) \qquad ....(1)$$

この時、キーワード木の下の方が文章の話題の詳細を表していると考えられるから、

$$w(i) > w(i-1) \qquad ....(2)$$

となるように関数Fを定める必要がある。現在、上式のFを色々と変えて試行中である。グループ化のアルゴリズムは次のようになる。

1. グループ化の始めとして、木の最末端のノードの数だけグループを作る。ここで最末端とは選択された語のうちで選択された語を子として持たないノードを言う。
2. 一つのグループに着目して、一つ階層を遡った親のノードのところで他のグループと統合できないかを調べる。すなわち統合・分離の判断は親のノードの下に全ての枝が存在するとした時の重要度の総計に対して、実際に存在するノードの重要度の総計がある比率αを越えるならば統合し、そうでなければ分離したグループのままに残す。
3. グループ化されたものを一つのノードとみなしそのノードの重要度はそのグループに含まれる候補語の重要度の総計とし、親子のレイヤーを一段上にずらし、2）の操作を行なう。

この3.を繰り返してグループ化を行なう。このようにしてグループ化を行なった後、各グループについてその中に含まれる語について上記3節で求めた候補語の重さの総計を求める。重さの総計についてthresholdを設け、重さがそれ以下のグループを削除する。残ったグループの中に含まれる語全てをキーワードとする。上記仮定によれば、この手法によりキーワード候補語のなかの不必要な語の率の低減が期待される。キーワード候補語から不要な語を削除する際、候補語間の関係を木構造として処理を行なったが、実際にはシソーラスの上位語は各項目について一つとは限らないので、親ノードも意味が異なったものが二つ以上存在し得る。親が二つ以上存在する時は各々のノードを別の木と考えて処理を行なえばすむが、極低い確率であろうが輪を作ることもあり得るので、その時の処理は別に考えなければならない。

6 今後の課題

妥当な関数Fを実験から求める。また、クラスタリング法自体の再検討も必要かもしれない。この研究は、電子ニュースをターゲットの一つとしている。電子ニュースはディスクスペースの制約上短期間でデータが消去されてしまうので過去のデータが検索できない。それを解決するために、上記の方針に従って最終的にはニュースシステムのキーワード自動抽出システムを作成していく予定であるが、まずは文書の形式が比較的整っている学術論文用のシステムをシソーラスと合わせて試作している。試験システムが出来た後は、インデクサーと呼ばれる専門家が抽出した結果との比較を各種行なってみる予定である。また、問い合わせとの関係も考えてみたい。

参考文献

1. 木本晴夫：”日本語新聞記事からのキーワード自動抽出と重要度評価” 電子情報通信学会論文誌 D-I Vol.j74-D-I No.8 pp.556-565
2. 内山恵三：”重要キーワード抽出方式とその活用方法” 情処学会 データベース・システム 84-19 (1991. 7. 18)