遺伝的プログラミングを用いたデータマイニングアルゴリズムの組み合わせ手法

Combined Method of Data Mining Algorithms using Genetic Programming

新美　礼彦 1)
Ayahiko Niimi

1) 公立はこだて未来大学 システム情報科学部 情報アーキテクチャ学科
Department of Media Architecture, Future University-Hakodate

Abstract: Quality of keywords given to each document is important to search documents from a lot of document databases. It is necessary automatically extracting high quality keywords from a document to achieve a document retrieval with high efficiency. We proposed a keyword extraction approach with selection of extracting keyword method depended on document categories using genetic programming. This approach could select only one extracting keyword method. In this paper, we expand this approach to be able to select combination of some extracting methods. By our new proposed approach, we can construct more complex keyword extraction system with combination of some methods.

1　はじめに

現在、インターネットの爆発的は普及により、さまざまな情報が簡単に手に入るようになった。しかし、これらの情報の中から自分のほしい上を探すのは簡単ではない。多量の文献のなかから自分の欲しい文献を検索する時の効率は、各文献に付与されているキーワードの品質に大きく左右される。効率の高い文献検索を実現するためには、与えられた文献から高品質のキーワードを自動抽出する必要がある。今までにいくつかのキーワード抽出法が提案されているが、各キーワード抽出法は文献に応じて精度に違いがあり、パラメータチューニングなども大変である。

この問題に対して、文献をカテゴリーごとに分類し、遺伝的プログラミングを用いてカテゴリーごとにキーワード抽出法を自動選択し、キーワードの抽出を行うシステムを提案した。[1] 提案したシステムでは、1手法のみを用いたキーワード抽出しか行えなかった。そこで本研究では、それを複数のキーワード抽出法を同時に組み合わせてキーワード抽出が行えるように拡張する。これにより、提案手法では複雑なキーワード抽出アルゴリズムの組み合せが行えるシステムを構築可能になる。

提案した手法の検証のため、キーワード抽出実験を行い、その評価を行った。

2　遺伝的プログラミング

遺伝的プログラミング (Genetic Programming:GP) は、生物進化論の考えに基づいた学習法であり、そのアルゴリズムの流れは遺伝的アルゴリズム (Genetic Algorithm:GA) と同様である。[2] その特徴は染色体表現が GA と異なり、関数ノードと終端ノードを用い構造表現ができるように拡張してあることである。GP では、関数ノードと終端ノードを用いて LISP の S 式形式で個体を表現する。

GP では、個体評価に適応度関数を用いる。適応度関数には、個体の精度、大きさ、計算時間など複数の指標を総合して組み込むことが可能である。

3　キーワード抽出法

キーワード抽出法として、さまざまなものが提案されている。提案されているキーワード抽出法を大きく分けると、形態素解析を用いるもの、形態素解析を用いないもの、文章の構造をもとに解析するものなどがある。[3] 本論文では、主に形態素解析を用いるものについて検討した。

3.1 形態素解析

形態素解析とは、入力文を言語学的に意味をもつ最小単位である形態素に分割し、各形態素の品詞を決定するとともに、活用などの語変形化をしている形態素に対しては原形を割り当てることである。[4] 形態素解析で分割された単語を要素単語という。要素単語に分けることにより、頻度解析や特定品詞へのフィルタリングが行えるようになる。

3.2 出現頻度による抽出

形態素解析で分割された各要素単語の出現回数 (頻度) を調べる。出現頻度の高い要素単語をキーワードとして抽出する。出現頻度の高い要素単語をキーワードとして抽出するため、どんな文章からも最適なキーワードを抽出しやすい手法である。しかし、助詞などのキーワードとして適切でない語を抽出する傾向があるため、抽出後のフィルタリングが重要になる。単純な頻度を使わずに、$tf \cdot idf$ を用いることもできる。これは、以下の式で定義される。

$$\text{スコア} = tf \times idf \tag{1}$$

ただし、

tf : あるキーワードがその対象文章中に含まれる出現回数 (Term Frequency)

$idf = \log(N/n)$: (Inverse Document Frequency)

N: 全文章数

n: そのキーワードを含むファイル数

$tf \cdot idf$ 法を用いることにより、多数の文章に多く含まれる一般的なキーワードの重要度を下げ、特定の文章中に多く含まれるキーワードの重要度をあげることができる。

3.3 連続名詞の抽出

情報検索の世界では名詞概念をキーワードとして抽出する傾向が強い。[5] 一般的には、形態素解析を用いて名詞を抜粋し、キーワードの抽出をおこなう。

3.4 N-グラム

構文解析を行わない方法の 1 つとして、N-グラム (n-gram) 法がある。N-グラムは長い文字列から部分文字列を取り出す方法で、N には 2 や 3 などの数をとることができる。N-グラムのアルゴリズムでは 1 文字ずつずらしながら、連続する N 文字を取り出し、取り出した文字列の出現頻度を調べ、その集合の中で出現頻度の高い語をキーワードとして抽出するというものである。[5] あらかじめ文章に形態素解析による単語分けを行う必要がなく、任意の数の文字数を設定することができる。

しかし、単語分けを行わないで解析すると、単語の一部分を含んだ文字列を大量にキーワードとして抽出する恐れがある。これを改善するために、本論文では形態素解析を行い、要素単語に分けた後で、その要素単語の連続を調べる手法も検討した。

3.5 相関ルール

文章中に現れる文字や単語の相関から、キーワードを抽出することが考えられる。これを相関ルールと呼び、ルールはいくつかの文字 (または単語) からなり、どれだけ同時に現れやすいのか (相関があるか) が評価対象となる。相関ルールを高速に抽出する手法として、apriori アルゴリズムがある。[6] 相関ルールの探索では、支持度 (support value) と確信度 (confidence value) という 2 つの指標を用いて相関ルールを評価する。本論文では、相関ルールの支持度 (sup) は全データに対する構成要素が含まれる割合、確信度 ($conf$) はある構成要素が含まれたときに他の構成要素が含まれる割合の平均であると定義する。

相関ルール探索は、N-グラムを用いたアルゴリズムと同様に、形態素解析を行わなくてもキーワードを抽出することが可能である。しかしこれも、単語の一部分のみを抽出する可能性を減らすため、本論文では形態素解析を行った後に要素単語間の相関ルールからキーワードも作成することを考える。

4 GP によるキーワード抽出手法の組み合わせ

各キーワード抽出法には、対象文章に得意・不得意があると考えられる。構造化した文章には構造を解析しながらキーワードを抽出することができるが、あまり構造化されていない文章では同じ解析を行うことは難しい。メールなどの短く、あまり構造化されていな文章と、論文などのある程度の長さがあり、構造のはっきりした文章では、異なるキーワード抽出法を用いる方が効果的と考えられる。また、それぞれのキーワード抽出法において、パラメータを対象文章にあわせて、

表 1: AND ノードと OR ノード

関数ノード	定義
(AND A B)	A と B を評価し、両方に含まれているキーワードの割合を出力する
(OR A B)	A と B を評価し、少なくともどちらか一方に含まれているキーワードの割合を出力する

チューニングする必要もある。

そこで以前、GP を用いて、各情報カテゴリーをもとにして各キーワード抽出法を選択し、その時のキーワード抽出法の正答率を求め、正答率が一番高い情報カテゴリーとキーワード抽出法の組み合わせを見つける手法を提案した。この手法では、GP を用いることで情報カテゴリーに適したキーワード抽出法を自動選択し、キーワードの抽出を行うことが出来る。また、適応度関数の設計時に、キーワードの精度や数、抽出までの時間などを考慮することが可能となる。また、キーワード抽出法のパラメータも同時に学習させることが可能である。提案した定義では、関数ノードはどのカテゴリーの文章なのかの条件判断をあらわし、終端ノードはどのキーワード手法を用いるのかをあらわすようにした。

しかしこの定義では、選択る手法は 1 つになってしまう。そこで、本論文では、複数の手法が選択できるように、AND と OR の関数ノードの定義を追加した。(表 1 参照)

以前の定義では、

```
(if_news associate-w_key
   (if_editorial connect_noun_key
      (if_mail associate-w_key ngram-w_key)))
```

のような出力が得られたが、AND と OR を追加することにより、

```
(if_news associate-w_key
   (if_editorial
      (and connect_noun_key associate-w_key)
      (if_mail associate-w_key
               (or ngram-w_key connect_noun_key))))
```

のような出力が得られるようになる。

適応度は、以前と同様に GP の個体により情報カテゴリーからキーワード抽出法を選択し、そのキーワード抽出法によって得られてキーワードの正答率を求め、これをもとにした。これにより正答率が一番高い個体が適応度の高い個体となる。キーワードの抽出数や抽出時間なども適応度計算として定義することにした。

GP を用いたキーワード抽出システムの欠点として、実時間での学習が難しい点が考えられる。適応度をシステム利用者の評価により行う対話的なキーワード抽出システムも考えられる。しかし、GP の適応度計算が個体数やノード数に依存して増加してしまうので、対話的に学習をさせようとすると待ち時間が長くなってしまう。そこで、システム利用者からの評価入力待ち時間やシステムが利用されていない時間などを使って、評価と平行して学習するなどの工夫を行うことにより、実時間での学習に対応させることが可能であると考えられる。

表 2: GP のパラメータ

集団数	500
複製確率	0.1
交叉確率	0.8
突然変異確率	0.1
選択方式	トーナメント方式
関数ノード	表 3 の 7 種類
終端ノード	表 4 の 5 種類
訓練データ数	各カテゴリー 25 文章ずつ、合計 125 文章

提案手法で前提となるカテゴリー分けに関しても、以前と同様に、文章を自動的にカテゴリー分けする手法は含まず、カテゴリーは使用者により指定されるものとした。

5 検証実験

提案手法の有効性を検証するために、複数カテゴリの文章から複数手法を用いてキーワード抽出を行った。文章のカテゴリとして、論文、ニュース、社説、マニュアル、メールを用いた。まず、それぞれから手作業によりキーワードを抽出し、これを正解とした。キーワード抽出手法として、頻度解析、連続名詞の抽出、文字をもとにした N-グラム法、単語をもとにした N-グラム法、単語をもとにした相関ルールを用いた。

GP のパラメータは、以下のものを用いた。(表 2 参照) 適応度は、正答率から求めた。個体評価の際、毎回キーワード抽出を行うと時間がかかるので、実験ではあらかじめ各キーワード抽出法でキーワード抽出を行い、正答率を求めてから GP 学習を行った。以前の実験では正答率にあまり差がない場合にうまく学習が行えなかった。そこで、今回の実験では、正答率の差が適応度の大きく影響するように正答率に重み付けをおこなった。AND と OR に関して、あらかじめ個別の手法での正答率が得られているので、とりあえず表 3 のように定義した。

まず、AND と OR を含まない場合の GP を用いたカテゴリーを元にした各キーワード抽出法の学習では、

表 3: 関数ノード

表示	意味
and	引数 1 と引数 2 を評価し、評価値の小さい方を返す
or	引数 1 と引数 2 を評価し、評価値の大きい方を返す
if_paper	カテゴリーが論文なら引数 1 を、違うなら引数 2 を評価する
if_news	カテゴリーがニュースなら引数 1 を、違うなら引数 2 を評価する
if_editorial	カテゴリーが社説なら引数 1 を、違うなら引数 2 を評価する
if_manual	カテゴリーがマニュアルなら引数 1 を、違うなら引数 2 を評価する
if_mail	カテゴリーがメールなら引数 1 を、違うなら引数 2 を評価する

表 4: 終端ノード

表示	意味
frec_key	出現頻度による抽出法を用いる
connect_noun_key	連続名詞による抽出法を用いる
ngram-c_key	文字をもとにした N-gram による抽出法を用いる
ngram-w_key	単語をもとにした N-gram による抽出法を用いる
associate-w_key	単語をもとにした相関ルール抽出よる抽出法を用いる

以下の結果を得た。この場合の平均正答率は 0.65 であった。

```
(if_mail associate-w_key
   (if_news (if_news associate-w_key
                    ngram-c-key)
            (if_news connect_noun_key
                     connect_noun_key)))
```

つぎに、提案した AND と OR を含んだ GP を用いたカテゴリーを元にした各キーワード抽出法の学習では、以下の結果を得た。この場合の平均正答率は 0.656 であった。

```
(if_mail (if_manual ngram-c_key
                    (or associate-w_key
                        frec_key))
         (if_news associate-w_key
                  connect_noun_key))
```

得られたキーワード選択はほぼ同じであるが、カテゴリーがメールのときに相関ルール抽出と同時に頻度分析を用いるようになり、平均正答率がわずかではあるが上昇した。

6 おわりに

本論文では、以前提案した文献をカテゴリーごとに分類し、遺伝的プログラミングを用いてカテゴリーごとにキーワード抽出法を自動選択し、キーワードの抽出を行うシステムを、複数のキーワード抽出法を同時に組み合わせてキーワード抽出が行えるように拡張した。拡張のため、AND と OR を関数ノードとして定義した。提案した手法の有効性を検証するために、キーワード抽出実験を行い、その評価を行った。

その結果、AND と OR を用いた GP で、用いないものよりも高い正答率を得ることができた。

今後は、単に引数の最大、最小を返す実装になっている AND,OR によるキーワード選択時の正答率を、実際に複数手法でキーワードを抽出した時のキーワード数に応じたものになるように変更し、提案手法が実際に使えるかどうか検討する予定である。

参考文献

[1] 新美 礼彦、安信拓馬、田崎 栄一郎: 遺伝的プログラミングを用いたカテゴリーごとのキーワード抽出法選択. 第 18 回 ファジィシステムシンポジウム論文集:pp.303–306 (2002).

[2] Koza, J.R.: Genetic Programming. MIT Press (1992).

[3] 市村 由美、長谷川 隆明、渡部 勇、佐藤 光弘: テキストマイニング - 事例紹介, 人工知能学会誌 Vol.16 No.2,pp.192–200 (2001).

[4] 松本 裕治、北内 啓、山下 達雄、平野 善隆、松田 寛、浅原 正幸:日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版 (1999).

[5] 那須川　哲哉、河野　浩之、有村　博樹:テキストマイニング基盤技術,　人工知能学会誌 Vol.16,No.2,pp.201–211 (2001).

[6] Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, September 1994:32pages (1994).

[問い合わせ先]

〒 041–8655

北海道函館市亀田中野町 116–2

公立はこだて未来大学 システム情報科学部

情報アーキテクチャ学科

新美　礼彦

TEL:0138–34–6222　FAX:0138–34–6301

E-mail:niimi@fun.ac.jp