D-021

遺伝的プログラミングを用いたキー ワード抽出尺度の探索と進化

Exploring and Evolving Keyword Extraction Measures using Genetic Programming

冨坂 亮太† 相澤 彰子†‡
Ryota Tomisaka Akiko Aizawa

1. はじめに

本研究では、文書からキーワードを抽出する新しい手法を提案する。ここで言うキーワードとは、対象の文書の中に含まれているもので、その文書の内容を端的に表わす重要な語句のことを指す。対象の文書に含まれないようなキーワードも存在するが、本研究ではそういったキーワードは扱わない。

キーワードを抽出する研究は古くから行われており、語の特徴量を用いた TF-IDF 法などの抽出手法が有名である。語の特徴量としては、出現頻度、品詞、構成語数、語構成、共起語異なり数、文書中での出現位置など多くの手がかりが考えられる。どの特徴量がどのように有効であるかは問題依存である場合も多い。これより、語の特徴量を用いた手法では対象にする文書群によって適切な特徴量の組み合わせ方を考えねばならず、ただ、既存の抽出尺度を適用しただけではうまくいかないことがある。

そこで、本研究では、文書とその文書に対するキーワードをセットにした学習データから、遺伝的プログラミングにより、複数の特徴量から、対象の文書群に適したキーワード抽出尺度を自動的に構築する手法提案する。

2. 関連研究

キーワードの抽出の研究は、今までに多く存在する。前述の TF-IDF は古い有名な手法で、現在でも実用的な尺度として多くのアプリケーションで用いられている。TF は term frequency の略で対象の単語が対象の文書に含まれる個数を示している。対して DF は document frequency の略でコーパス中の対象の語句を含む文書の数を表し IDF はその逆数である。これらを掛け合わせたものが一般的な TF-IDF である。本稿では文献[1]で示されている TF×log(N/DF) (ここで N はコーパス中の文書数)の定義を用いる。

また、既存のキーワード抽出尺度を組み合わせて、対象の文書にあった新しいキーワード抽出尺度を構築しようとする研究も存在する。例えば、文献[2]では、キーワード候補を様々な抽出尺度で評価し、それらの値を整数計画問題に当てはめて解くことによりキーワード抽出を行っている。また、文献[3]では、統計的な手法を用いて既存の尺度を組み合わせる方法を提案している。

また、文献[4]においてはコロケーションを抽出する尺度を遺伝的プログラミングを用いて探索する手法を提案している。

† 東京大学 情報理工学系研究科 コンピュータ科学専攻

‡ 国立情報学研究所 コンテンツ科学研究系

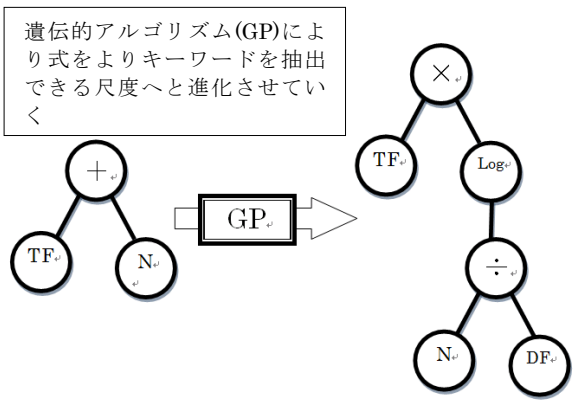
3. 遺伝的アルゴリズムによる抽出尺度の最適化

本節ではまず、本手法の流れを説明し、キーワード抽出の前処理として、キーワードの候補を選出する処理を説明した後に、学習に使うデータの説明と実際に遺伝的アルゴリズムを用いて抽出尺度を最適化する方法を述べる。

3.1 抽出尺度最適化の流れ

まず、キーワードがあらかじめ付与された文書を学習データとする。学習のターゲットとなるのは、図 1 に示すように木構造で表現された抽出尺度関数である。学習では、遺伝的プログラミングの各遺伝子によって表わされるキーワード抽出尺度により選出されるキーワードがあらかじめ付与されたキーワードに近づくように尺度を進化させていく。

図 1:GP によるキーワード抽出尺度の進化



3.2 キーワード候補の選出

まずキーワードの候補を選出する。これは、動詞などの、キーワードとはなりにくい品詞をふるい落とす処理で、実際には以下の品詞をもった単語の連続のみを取り出した。

- 未知語
- 名詞(非自立を除く)
- 接頭詞-名詞接続
- 接頭詞-数接続
- 記号-一般
- 記号-アルファベット

この、品詞の選び方は文献[2]の研究に従った。これにより、「非線形の」(‘の’が名詞-非自立)や「述べる」(動詞)などといった語が候補から除外され、「整数計画問題」や「データベース I」といった語が候補として残ることになる。

3.3 遺伝的プログラミングの適用

遺伝的プログラミングは進化論的計算手法の一種である。木構造で表わされたひとつの式をひとつの遺伝子とし、いくつかの遺伝子を交叉および突然変異させて進化させていき、目的の評価関数に対して良い結果を得られるような式を構築する手法である。

本研究では、抽出尺度の計算式をひとつの遺伝子に対応させる。式のオペランドとしては、TF、IDF と、語の文書に対する出現位置の情報(PC)、コーパスの文書数(一定)を用いた。この 4 つの特徴量のみをオペランドとしたのは、本手法による特徴量の組み合わせ方の優位性を、一般の TF-IDF の定義によるキーワード抽出と比べることによって示すためである。もちろん扱う特徴量を増やすことにより、得られる結果の精度は向上する。

また、用いた演算子は＋、－、×、÷、平方根、二乗、log の七つである。計算の単純化の為に平方根、二乗を用いたが、これらを累乗にしても問題ない。

これらから、遺伝子は、図 1 のような木構造で表わされる。

3.4 適合度の計算法

ここでは、本手法で用いた遺伝的プログラミングにおいて、各遺伝子の適合度を評価するための評価関数について述べる。

まず、学習用の各文書から前述の方法によりキーワードの候補を選出する。その候補それぞれを各遺伝子によって表わされた抽出尺度により評価し、それを評価の高い順にランキング付けする。評価の高い順に 3 つ取り出し、その取り出された三つを、文書に与えられているキーワードと比べ、F 値を算出する。

F 値は情報検索の精度を表すための一般的な尺度で、適合率と再現率の相加平均で表わされる。適合率は全検索結果(この場合 3 つ)の中の検索要求を満たす検索結果(文書に与えられたキーワードと一致するもの)の割合である。再現率は検索要求を満たす全キーワード(あらかじめ文書に与えられているキーワード)のうちの、検索要求を満たす検索結果の割合である。この値を F3 とする。さらに、評価の高い順に 5 つ取り出して同じ様に F 値を算出したものを F5 とする。

また、遺伝子の木の大きさに制限がないといくらでも、式は大きくなれるため(例えば、TF/TF はいくら掛けても式の結果は変わらないため遺伝子はいくらでも大きくなれる)、評価関数に－w×tl という評価値を加える。ここで tl は遺伝子の木のノード数を表し、w は任意の値である。w の値を操作することによって、最終的に導き出される尺度の式の大きさを操作することができ、w の値が小さければ小さいほど式の大きさは大きくなる。

前述の F3 と F5 の平均値に－w×tl を足し合わせたものを各トレーニング用の文書毎に算出し、すべての文書での平均値を最終的な遺伝子の評価値とする。

4. 実験

以上の方法によって作り出された、キーワード抽出尺度の方が、既存の一般的な定義の抽出尺度よりもより対象の文書群に適したキーワードを抽出できることを示すために以下のような実験を行った。

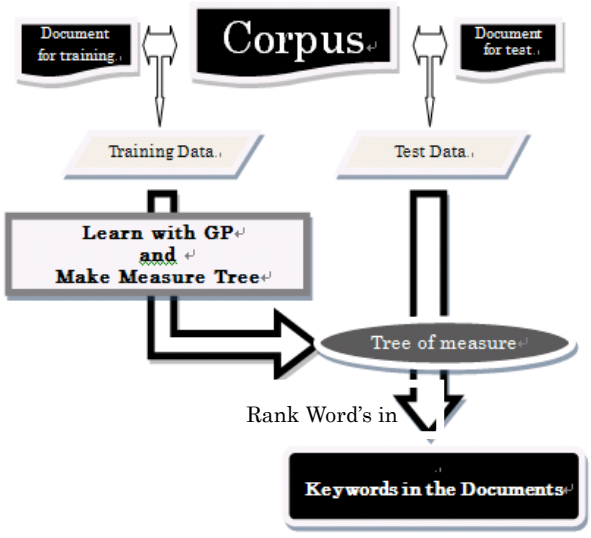
4.1 実験の概要

実験は図 2 のような流れで行う。まず、トレーニング用の文書とコーパスからトレーニングデータを作る。ここでトレーニングデータとは、単に、トレーニング用のデータからキーワードの候補を選出し、それらの特徴量の値をあらかじめ計算しておいたものである。同様にテスト用の文書群からもテストデータを作っておく。

その後、トレーニングデータを使って遺伝的プログラミング(図中 GP)により、キーワード抽出尺度を構築する。

構築された尺度により、テストデータからキーワードを抽出する。最後に、テストデータに対して F 値を計算して各手法の性能値とする。

図 2.キーワード抽出のフローチャート



4.2 使用データ

実験には NTCIR-2 の情報検索用のテストコレクションを用いた[1]。このコレクションには約 40 万件の学会発表等の著者抄録が含まれ、抄録それぞれに著者によるキーワードが 1 つから 5 つ付与されている。この抄録の中から 100 件ずつトレーニング用とテスト用に取り出し、残りをコーパスとして用いた。実験は 4 回行い毎回トレーニング用とテスト用の文書を変えて行った。文書は、トレーニング用とテスト用のデータが同じ分野の抄録になるように選択して取り出した。

[1] http://research.nii.ac.jp/ntcir/

4.3 実験方法

上述のデータから、以下のそれぞれの手法によりキーワード抽出を行い結果を比較した。

・提案手法によるキーワード抽出
・TF-IDF によるキーワード抽出
・TF-IDF に出現位置情報を掛けたもの(以下 TF-IDF-PC)によるキーワード抽出
・SVM によるキーワード抽出

SVM 用の素性は、遺伝的アルゴリズムのオペランドのために用いたものと同じもの(この場合 TF と DF と出現位置情報とコーパスの文書数)を用いた。

3.4 で定義したパラメタ w は、0.000001 と 0.001 と 0.005 の三つの値でそれぞれ遺伝的プログラミングを適用した。

4.4 結果の評価値

それぞれのキーワード抽出結果の評価は、遺伝的プログラミングの各遺伝子の評価と同じように、各尺度によりキーワード候補をランキング付けし、評価の高い順に 3 つ取り出し、F 値を算出した。それをすべての文書に対して行い、平均をとったものを F3 とする。

同様に、評価の高い順に 5 つを取り出し F 値を算出したものを F5 とする。

SVM でのキーワード抽出では、キーワードと判別されたものすべてから F 値を算出した。

5. 実験結果

この実験による実験結果を表 1 に示す。

5.1 TF-IDF・TF-IDF-PC との比較

今回行った実験では、出現位置情報が大きな影響を及ぼしており、TF-IDF-PC によるキーワード抽出の精度の方が TF-IDF によるキーワード抽出の精度を大きく上回った。

提案手法と TF-IDF-PC を見比べると、w が 0.000001 の時と 0.001 の時は、多少下回る時があるが全体的に提案手法の方が、TF-IDF-PC の結果を上回っているといえる。

今回の実験では TF-IDF-PC と提案手法で手がかりとする素性が共通になるように素性数を 4 つのみとしたため、遺伝的プログラミングの探索空間が限定されたものとなり、劇的な F 値の上昇は見込めなかった。しかしながら、著者があらかじめ与えておいたキーワードを抽出するという難しいタスクでほぼ安定して既存の手法を上回る結果を得られたことは価値があるといえる。

5.2 SVM との比較

今回の実験では素性数が 4 個とあまりにも少なく、また一般語の削除といった前処理も行っていないため、SVM によるキーワード抽出はうまくいかず、F 値はすべてにおいて 0 となってしまった。また、今回のキーワード抽出では、正解キーワードの部分文字列で素性の値が全く変わらない不正解キーワードや、キーワードらしいワードだが、著者によってキーワード指定されていないワードが存在したりしたため、SVM ではうまく学習することができなかったと考えられる。

表 1:実験結果

一回目	F3	F5
GP(w=0.000001)	0.209	0.209
GP(w=0.001)	0.232	0.212
GP(w=0.005)	0.177	0.192
TF-IDF	0.138	0.155
TF-IDF-PC	0.193	0.192
SVM	0	
二回目		
GP(w=0.000001)	0.143	0.127
GP(w=0.001)	0.130	0.130
GP(w=0.005)	0.134	0.121
TF-IDF	0.113	0.118
TF-IDF-PC	0.130	0.127
SVM	0	
三回目		
GP(w=0.000001)	0.125	0.160
GP(w=0.001)	0.141	0.155
GP(w=0.005)	0.129	0.155
TF-IDF	0.106	0.124
TF-IDF-PC	0.149	0.155
SVM	0	
四回目		
GP(w=0.000001)	0.228	0.240
GP(w=0.001)	0.240	0.214
GP(w=0.005)	0.192	0.208
TF-IDF	0.180	0.197
TF-IDF-PC	0.216	0.217
SVM	0	

5.3 パラメータ w に対する考察

今回、w の値として 3 つの値を用いたが w=0.005 の時は式の大きさが TF-IDF と同程度になり結果は TF-IDF と TF-IDF-PC の中間程度のものとなった。しかし、この条件でも TF-IDF より十分よい結果を出しているといえる。また、w=0.000001 の時と w=0.001 の時を比べると、w=0.001 の時の方が安定して良い結果を出すことがわかた。これは w=0.000001 の時は、式がトレーニングデータに特化しす

ぎてしまい、違う文書群のテストデータに対してうまく働かなかったのではないかと考えられる。

5.4 遺伝的プログラミングによって得られた尺度の考察

w=0.005 のとき得られた尺度はせいぜいオペランドが三つ程度なので人にとって直観的で分かりやすく、対象の文書群において、どのような特徴量が有効かということを判断する材料となる。たとえば今回の実験だと、3 回目の文書群に対しては、logTF/(DF+√n)といった式が得られた。これは、TF-IDF の変種と見ることができる。四回目の文書群に対してもまったく同じ式が得られている。これから、今回用いたような文書群に対してはこのような形の TF-IDF が有効であることが推測される。

また、一回目の文書群に対して得られた式は PC/(DF+N)であり、ここでは、PC がより有効な素性であると推測される。これは、一回目においては TF-IDF-PC の結果が TF-IDF の結果を他よりも大きく上回っていることから納得できる。二回目の文書に対しては PC^2×logTF/(DF+√n)で 3(4)回目の文書群に対して得られたものに PC の二乗をかけたものとなった。このように、遺伝的プログラミングを用いた手法ではトレーニング用文書セットに応じたキーワード抽出尺度を選択できているといえる。

6. 考察

今回用いた素性は 4 個と少なく(さらにそのうち一つは定数)、機械学習によってキーワードを抽出するのは難しいタスクであり単に SVM を用いてキーワード判別を行っただけでは全く結果が出なかった。このような場合でも、遺伝的プログラミングを用いることで、既存の抽出尺度よりも良いキーワード抽出手法を学習することができた。

もちろん、素性数を増やせばさらに結果の向上が見込め、現に単語の長さなど若干素性を増やした場合 TF-IDF-PC をはるかに上回る結果を出すことに成功している(表 2)。また、このときは SVM でのキーワード抽出による F 値も 0 ではなかった。

また、本手法では、得られるものは、単純な式であるため 5.4 のように出来上がった式を見れば、対象の文書群に対してどのような素性がきき、どのように組み合わせるのがよいかといった知見を得ることができる。

ただし、本手法は学習速度が遅いといった欠点がある。それは、今回用いた遺伝的プログラミング用の評価関数のせいであるが、これをもっと軽い評価関数にすれば、学習速度も向上し使いやすいシステムにしていけると考えられる。

このような、適切な評価関数の考案と、多くの素性を投入して、さらにいろいろなキーワード抽出尺度との比較をし、本手法の有効性を確かめていくことが今後の課題である。

表 2: 一回目に用いた文書群と同じ文書群を用いて実験したもの

	F3	F5
GP(w=0.000001)	0.238	0.264
TF-IDF	0.138	0.155
TF-IDF-PC	0.193	0.192
SVM	0.049	

参考文献

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schuetze, “Introduction to Information Retrieval”

[2] 乾孝司,橋本泰一,高村大也,内海和夫,石川正道, “キーワード抽出の整数計画問題としての定式化”, 情報処理学会研究報告, 自然言語処理研究会, 2008-NL-188, pp. 29-36 (2008)

[3] 内山将夫,井佐原均“複数尺度の統計的統合法とその専門用語抽出への応用”, IPSJ SIG NotesVol.96, No.87(19960912) pp. 115-122 (2003)

[4] Jan Snaider, Bojana Dalbelo Basic, Sasa Petrovic, Ivan Sikiric , “Evolving new lexical association measures using genetic programming”,ACL,HLT,ShortPapers,pp. *181*-184 (2008)