社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－171 (6)
2006／1／12

Web資源からの決定木学習によるWebページに対するキーワード付与

小磯拓也† 但馬康宏†
藤本浩司†,☆ 小谷義行†

Webページの内容を上観的に表すキーワードをインターネットをコーパスとして自動的に付与するシステムを構築した．キーワードを付与するためにインターネット上のWebページをコーパスとして使用してキーワード候補語とその属性値を求め，決定木を用いてキーワード候補からキーワードを選択する．すでにキーワードを持つWebページに対して付与処理を行った結果，精度15 %，再現率27 %を得た．

A Keyword Assigning System using Tree Learning with the Web

TAKUYA KOISO,† YASUHIRO TAJIMA,† KOJI FUJIMOTO†,☆
and YOSHIYUKI KOTANI†

We developed a keyword assigning system for web pages, using decision tree. It assigns keywords to web page with searched web pages from internet with search engine. First, the system picks up some words from input webpage for searching internet, then it searches internet using this words, next it picks candidated keyword from input web page and searched web pages. Last, it inferences keywords from these using the decision tree. The decision tree is learned from several web pages that have been already assigned keyword by authors. We evaluated this system, precision is 15 % and recall is 27 %.

1. はじめに

RSSに代表されるように，昨今インターネットのWebページに対してメタ情報を付与することで，インターネットの閲覧をさらに便利にする技術が注目されている．これまでもWebページに対してそのページの著者がメタ情報を付与することはなされていたが，その手順の煩雑さから全てのページに付けられていたとは言い難い情報である．RSSはそうした情報を自動的に付与することによっておおいに普及したが，日付情報やタイトルといった簡単に付けられる属性しか持っていないことが多い．そこで，Webページの内容を表すようなメタ情報を自動的に付与することによって，こうしたメタ情報の普及を目指すものである．

本稿ではメタ情報としてWebページの内容を端的に表現する単語（キーワード）の集合を，Webページに対して付与することを目的とし，それを行うシステムを作成した．そのための手法として，文書中の情報のみならず，それをもとに検索して得られたWebの情報も付与に際して使用する．

2. キーワードとは

キーワードとはWebページに対してページの著者が明示的に付与する単語の集合であり，そのWebページの内容を簡潔に表すものである．キーワードには，製品名，一般名詞，複合名詞や動詞など，ページの著者が適合すると判断した，さまざまな種類の単語が考えられる．そこで本稿におけるキーワードを以下の通り定義する．

- Webページの内容を表す単語であり，ページの作成者が主観的に付与するものである．
- 単語はアルファベットと数字で構成される文字列である．
- 複合名詞も含む．

あるWebページの内容を表現するためには，キーワードの集合を用いて表現する．Webページに使われているマークアップ言語であるHTMLでは，こうしたキーワードを明示するための属性値が定義されている．すなわち，METAタグの中のkeywords属性値に主にカンマ区切りで入力されている単語がキーワードだと判断することができる．以下はWebページに含まれているキーワードの例である．

† 東京農工大学工学教育部電子情報工学専攻
Tokyo University of Agriculture and Technology
☆ 現在，金融エンジニアリング・グループ
Presently with Financial Enginnering Group Inc.

```
<meta name="Keywords" content=
"programming, web hosting, keyword">
```

3. 関連研究

キーワードを抽出することに関する研究はこれまでも頻繁に行われてきている．文書中に含まれているキーワードを抽出するもっとも頻繁に用いられるアルゴリズムは tfidf である．ある文書の内容を表すキーワードとなる単語はその文書に対して集中的に現れるという仮説のもとに，単語の頻度とその単語が現れた文書の割合から，キーワードを推定するアルゴリズムである．

キーワードを推定するというシステムには，Keygraph[3] もある．このシステムは「文書は著者独自の考えを主張するために書かれるという仮説」([3] より引用）をもとに，文書を建物に例えている．その上で単語の共起によるネットワーク構造を作ってキーワードの抽出を行っている．共起を用いるアルゴリズムとして[4] もある．

また，インターネット，Web 資源を用いて何らかの情報を得るシステムとして[5] がある．これは研究者間の協働関係ネットワークを作成するアルゴリズムであるが，その内容として検索エンジンを用いて Web 上のファイルを解析して，ネットワークを構築している．

本稿は Web ページに対するキーワードを付与するためのシステムであるため，関連研究[3] と[4] と目的は近い．ただしこれらが文書中に含まれる単語をキーワードと見なすのに比べると，本稿のシステムは文書中に現れないキーワードを Web 上のファイルを検索することによって得ようとするところが異なる点である．例えば文書に含まれないキーワードとして，例えば果物の名前を列挙しているが「果物」という単語がでない文書が考えられる．そういった問題に対して対処できるシステムの構築が目的である．

4. 処理対象の Web ページ

本稿で対象とする Web ページはディレクトリが指定されていないドメインネームのみの URL で現されるもの，すなわちトップページとする．そうした理由として，こうしたページは Web サイトの入口であり，こうしたサイトのキーワードが，それ以下のディレクトリに用いられている例が多いためである．

5. キーワード付与システムの概要

本システムは，基礎知識データベースとインターネットを用いて入力した Web ページに対するキーワードを付与するシステムである．基礎知識データベースとは，あらかじめインターネットを広く浅く検索して得られたページを解析して得られたさまざまな値である．

本システムは入力として，キーワードを付与したい Web ページ（入力 Web ページとする）と，付与したいキーワードの個数を与えられると，入力 Web ページの内容を基礎知識データベースを用いて解析し，まず検索クエリーを決定する．そして，この検索クエリーをインターネット上の検索エンジンに入力し，関係する Web ページの集合（検索 Web ページ集合とする）を取得する．入力 Web ページと得られた検索 Web ページ集合を用いて解析を行い，キーワードを出力する．

5.1 問題の簡略化

入力 Web ページからキーワードを付与するというタスクを以下の通り変更する．入力 Web ページを単位とすると扱いづらいので，本システムでは「ある単語がキーワードかどうか」ということを求める処理とする．キーワード付与システムはキーワードの候補になりうる単語の属性値を調べ，それがキーワードであるかどうかを何らかの基準を用いて決定する．簡略化することにより，後述する判断基準として決定木やさまざまな分類アルゴリズムを用いることができるようになる．

5.2 システムの付与処理の概要

5.2.1 基礎知識データベースの構築

本システムの方針は，インターネットから関連するジャンルのコーパスを集めてそれを用いてキーワードを付与するというものであるが，これだけではすべてのジャンルに均等に現れる単語について判断することはできない．そこで，ジャンルに依存しない単語と，その属性値を調べて，基礎知識データベースを構築する．まず，あらかじめインターネットから広いジャンルのページを収集する．そしてその中に含まれる単語の属性値を取得する．本稿では属性値として idf 値を調べている．

広いジャンルのページを収集するために，インターネットの検索エンジンを使用した．KDD Cup 2005 に含まれるジャンル名をそのままインターネットの検索エンジンに入力して，得られたページの集合を基礎知識データベースを構築するためのページ集合とした．

こうして得られた基礎知識データベースはキーワードの付与の決定に使用するだけでなく，検索クエリーを決定するためにも用いられる．よって，この基礎知識データベースはシステムが最初に 1 回作成し，基本的には更新しないものとする．

5.2.2 システムの学習

キーワードである META タグを持った Web ページ（学習 Web ページとする）を学習用のデータとして，それを用いてシステムの学習を行う処理である．まず，学習 Web ページを解析する．本システムでは頻出した単語を選択している．そして，これをあらかじめ作成した基礎知識データベースの idf 値と掛け合わせて tfidf 値を計算し，それの上位 n 個を検索クエ

リーとする．こうして得られた検索クエリーを用いて検索エンジンで検索を行う．こうして検索 Web ページを得た上で集合に含まれている単語を調べ，属性値を計算する．すでにキーワードは分かっているため，この属性値とキーワードであるかどうかを組にして学習データを作成し，システムの学習を行う．

5.2.3 キーワード付与処理

システムの学習を行ったら，キーワードの付与を実行する．キーワードの付与方法は，システムの学習と同様に，処理対象の Web ページを解析して検索クエリを決定し，検索クエリを用いて検索エンジンで検索 Web ページ集合を取得．それを解析してキーワード候補語とその属性値を決定する．すでに学習がなされているので，システムは学習された判断基準に基づいてキーワード候補語の中からキーワードを決定し，それをシステムの出力として決定する．

5.2.4 判断基準

キーワード候補語は，原理的にはすべての単語でありうるが，システムはすべての単語を処理することができないために，本システムでは以下の基準でキーワード候補語を決定する．

- 入力 Web ページのソースに含まれている．
- 検索 Web ページ集合に含まれる Web ページのソースに含まれている．

入力 Web ページもしくは検索された Web ページのソースに含まれている単語をすべて対象とするが，タグやコメントになっているものは含まないものとする．基本的には Web ブラウザで表示したときにユーザの目に触れることのできる文字列がキーワードの候補となる．

6. 決定木によるキーワードの決定

キーワードの決定には属性値から判断できるものであれば，判断方法として何を用いても構わない．例えば SVM 等をシステムの判断方法として用いることもできる．本稿では決定木を使用する．これは作成された決定木を分析することにより人間が大きく分類に寄与している属性値を判断することができるためである．使用する決定木は c4.5[1] とする．c4.5 は ID3 の発展系であり，よく用いられている決定木作成アルゴリズムである．

また，決定木やその他のアルゴリズムにおいて，キーワードであるかどうかを判断するためには，属性値を用いる必要がある．属性値の内容を分析することによって適切な判断基準を作成する．本システムで使用する属性値は以下の通りである．a〜i は，生起数等のその属性値単体で用いられるものであり，j〜m は，a から i の値を組み合わせたものである．

6.1 a. 入力 Web ページの単語生起数

入力 Web ページに含まれている単語の生起数である．含まれている場合は 0 より大きな値になり，入力 Web ページに含まれていない，すなわち検索 Web ページ集合にのみ含まれるキーワード候補語に関しては 0 になる値である．

6.2 b. 検索 Web ページ集合での単語生起数

検索 Web ページ集合に含まれる全てのページに現れる単語の総数である．検索 Web ページには複数の Web ページが含まれているが，それらを全て一つのページにまとめたときに現れる単語の生起数を指す．検索 Web ページに現れていない単語に関しては 0 とする．

6.3 c. 検索 Web ページのタイトルでの単語生起数

検索 Web ページ集合に含まれるページのタイトルに含まれる単語の生起数である．これも全ての Web ページを一つのページと見なした上でそこに含まれる単語の生起数を求めたものである．タイトルに含まれていないものに関しては 0 である．

6.4 d. 基礎知識データベースの idf 値

基礎知識データベースにおける idf 値をそのまま使用する．その単語がインターネットにおいて一般的であるかどうか，そういった問題について対処するためのものである．基礎知識データベースに値が存在しない場合，1 回存在した場合と同様と見なす．0 回にすると無限大になってしまうためである．

6.5 e. 検索 Web ページ集合における idf 値

検索 Web ページ集合のみを対象として idf 値を求める．基礎知識データベースの値と異なり，関連するジャンルにおいて頻出する単語を判断するために用いる．値が存在しない，すなわちキーワードに対応する単語が見つからない場合，d と同じように 1 回現れた場合と同じ値にする．

6.6 f. タイトルにおける idf 値

検索 Web ページ集合のタイトルのみを対象として idf 値を求める．タイトルでの頻出語をこれで判断する．これも同様に，キーワードに対応する単語が見つからない場合，1 回現れたと見なす．

6.7 g. 検索 Web ページ集合の単語生起数

検索 Web ページ集合の中にはキーワードを持つものページも存在する．そういった情報は非常に有用だと判断してこういったデータを取得することにした．検索 Web ページ集合を一つの Web ページとして見なした場合，その中に含まれるキーワードの生起数である．キーワードとして存在しなかった場合は 0 になる．

6.8 h.description の単語生起数

Web ページにおいて MATA タグの一種に description が存在する．これはページの作成者が主観的に作成した要約を現したものである．従ってそこに含まれる単語は他の単語に比べて重要であると考えられる．検索 Web ページ集合のページを一つのページと見なし，そこに含まれる description の生起数を求めたも

のである．

6.9 i. キーワードの個数

本システムは入力 Web ページと同時に付与したいキーワードの個数も同時に入力する．入力 Web ページのキーワードの個数が全ての単語に対して与えられる．この値が大きい場合と小さい場合とで，キーワードの内容が変化すると考えられる．

6.10 j. 入力 Web ページの単語生起数の tfidf 値

a と d の積によって tfidf 値を現すことができる．入力 Web ページに現れない単語，もしくは idf 値が存在しないものに関しては 0 になる．

6.11 k. 検索 Web ページ集合の単語生起数の tfidf 値

b と d の積によって現される tfidf 値である．検索 Web ページ集合に含まれる単語に対しても tfidf 値を付けるものとする．

6.12 l. タイトルに表れる単語総生起数の tfidf 値

f と d の積によって現される tfidf 値である．どちらかが 0 の場合は 0 になる．

6.13 m.description の単語生起数の tfidf 値

h と d の積によって現される tfidf 値である．どちらかが 0 の場合は 0 になる．

表 1 本システムで使用する属性値の一覧

a. 入力 Web ページの単語生起数
b. 検索 Web ページ集合での単語生起数
c. 検索 Web ページのタイトルでの単語生起数
d. 基礎知識データベースの idf 値
e. 検索 Web ページ集合における idf 値
f. タイトルにおける idf 値
g. 検索 Web ページ集合の単語生起数
h.description の単語生起数
i. キーワードの個数
j. 入力 Web ページの単語生起数の tfidf 値
k. 検索 Web ページ集合の単語生起数の tfidf 値
l. タイトルに表れる単語総生起数の tfidf 値
m.description の単語生起数の tfidf 値

7. 評価実験

本システムの性能を調べるために以下の条件で実験を行った．なお，URL はインターネットからランダムに取得してきたもので，さらにランダムに並び替えを行うことによって内容の均質化を試みている．また，大量のキーワードが存在するページは処理対象としてふさわしくないと考えて，キーワードが 40 個（複合名詞を含んだ数）よりも少ないもののみを対象とした．これでほぼ 90 %の URL が含まれることとなる．

- 検索クエリー：上位 2 位までの値を検索クエリーとして使用する
- 検索エンジン：米 Yahoo![2)]
- 検索 Web ページの個数：最大 30 ページ．
- 基礎知識データベースの規模：14514 単語
- 使用する決定木作成システム：C4.5[1)]
- 668 件を学習データとして使用し，残り 202 件を用いて評価を行った．

実験用の URL に含まれているキーワードの内訳は表 2 の通りである．「システム認識数」は，本システムが処理可能な単語の数，すなわち入力 Web ページと検索 Web ページ集合に含まれている単語の個数である．「単語のみ」はキーワードとして含まれている単語の数であり，「全て」は複合名詞も含んだ全体のキーワードの数である．

c4.5 に関しては[1)] 標準の決定木作成と，末端の葉のノードが 100 以下の場合に木の成長を止める場合の 2 通りの決定木を作成した．[1)] 標準では木が細かくなる傾向にあるためである．

表 2 データのキーワード数内訳

	システム認識数	単語のみ	全て
学習データ	3427	4491	8989
テストデータ	1830	2304	4166

学習用データのレコード数は約 673 万，テストデータのレコード数は約 330 万である．一つの Web ページにつき約 15000 個のキーワード候補語がある．

8. 実験結果

実験結果は以下の通りである．表 3 は[1)] の標準の決定木での分類結果である．この例は再現率よりも適合率を重視しているので，再現率を重視した場合の結果を一つ示す．再現率の値が 2 個あるが，これはシステムが把握しているキーワード数 (1) と，本来処理できる単語の個数 (2) のそれぞれで再現率を求めたものである．また，表 4 は再現率が高くなるように決定木で得られたルールから選択した一例である．

表 3 c4.5 標準設定での実験結果

再現率 (1)	96 / 1830	5.2 %
再現率 (2)	96 / 2304	4.1 %
適合率	96 / 149	39.2 %

表 4 再現率が上昇するようにルールを選択した結果

再現率 (1)	611 / 1830	33.3 %
再現率 (2)	611 / 2304	26.5 %
適合率	611 / 4080	15.0 %

また，作成された決定木を図 2 以下に示す．0 の場合はキーワードではなく，1 の場合はキーワードであると判断したノードである．括弧内は，0 の場合は左が 0 の出現数，1 の場合は 1 の出現数である．例えば

「0 (3332.0/440.0)」はキーワードではないとシステムが判断し，そのうちの約 3332 個がキーワードではなく，約 440 個がキーワードであるということになる．なお，c4.5 では一度過剰に木を生成した後に簡略化を行うようになっているが，図 2 は簡略化する前の決定木である．簡略化した決定木は図 1 である．

```
tfidf_normalize <= 0.00382 : 0 (6670598.0/1457.9)
tfidf_normalize > 0.00382 :
    tfidf_keyword_in_result <= 2.72965 : 0 (59795.0/859.7)
    tfidf_keyword_in_result > 2.72965 :
    |   keyword_in_result <= 3 : 0 (3332.0/440.0)
    |   keyword_in_result > 3 :
    |   |   numofkeywords <= 14 :
    |   |   |   title_tf_all > 0 : 0 (63.0/14.8)
    |   |   |   title_tf_all <= 0 :
    |   |   |   |   freq <= 13 : 0 (135.0/35.0)
    |   |   |   |   freq > 13 : 1 (13.0/5.7)
    |   |   numofkeywords > 14 :
    |   |   |   title_idf > 4.62815 : 0 (28.0/11.3)
    |   |   |   title_idf <= 4.62815 :
    |   |   |   |   title_tf_all > 0 : 0 (48.0/23.9)
    |   |   |   |   title_tf_all <= 0 :
    |   |   |   |   |   tfidf_normalize > 0.22934 : 1 (10.0/1.3)
    |   |   |   |   |   tfidf_normalize <= 0.22934 :
    |   |   |   |   |   |   tfidf_keyword_in_result <= 34.3399 : 1
    |   |   |   |   |   |   tfidf_keyword_in_result > 34.3399 : 0
```

図 1　作成された決定木 (簡略化されたもの)

9. 考　　察

9.1　決定木に関する考察

決定木の傾向として顕著なのは，tfidf_normalize の値である．これが 0.00382 以下の場合にほぼ全てのキーワード候補語が含まれており，これらは全てキーワードではないと見なされている．図 2 を見るとこの条件に約 1000 件のキーワードが含まれていることがわかる．学習データのキーワードは 3427 個なので，約 1/3 以上のキーワードが破棄されてしまっていることになる．図 2 は簡略化する前のものなのだが，簡略化した場合には約 1500 個のデータが捨てられてしまっていることになる．

これにより，全体のキーワードのうちの大多数が捨てられてしまっていることになり，また，入力 Web ページに含まれていない単語もキーワードとして付与することができるという当初の目標を達成してるとはいいづらい．しかしながら，認識はしているため，将来的に適切な属性値があればこの中からキーワードだけを取り出すための条件が設定できることになる．従って，キーワードを取り出すための属性値をさらに追加することが今後の目標である．

9.2　性能に関する考察

再現率と精度の結果から，キーワードを付与するシステムとして用いることは難しいと考えられる．現段階では，Web ページに付与されるキーワードがページあたり平均 10 前後存在するため，おおむね 3 個程度が見つけられている計算になる．ただし，キーワードの定義が後述するようにかなり厳密なものであるた

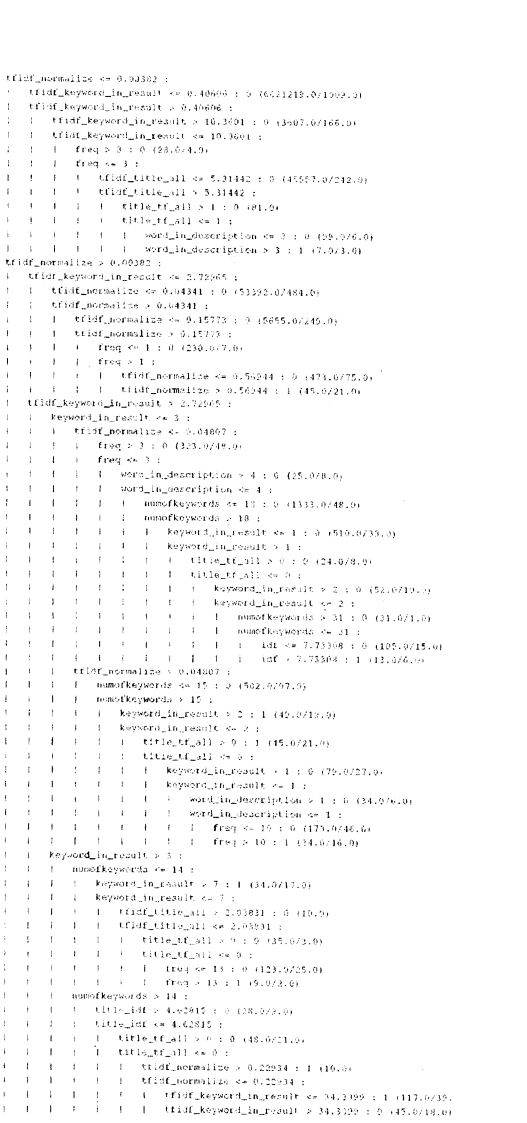

図 2　作成された決定木 (簡略化されていないもの)

めに精度や再現率が低くなる傾向にはあると考えられる．これらの結果の理由としては以下のものが考えられる．これらの問題の解消は今後の目標となる．

9.2.1　キーワードの定義が厳密である

ある Web ページに対してそのページの著者が付与するキーワードはその著者の主観に定義される．そして主観的には複数の候補が存在していても最終的に数個の単語がキーワードとして集約されることになる．本システムではそういう曖昧な状態を正解としておらず，そうした単語を検索しても不正解と見なすようになっている．

9.2.2　適切な属性値が発見できていない

本稿の属性値としては主に tfidf 値を重視した属性

値となっている．これはこの tfidf 値がこうしたキーワード抽出の分野ではよく用いられるためである．しかしながら実際のキーワードは tfidf にそっていない場合もあるため，それらを分離するための属性値が欠落している場合適切な結果が得られない．本稿ではそういう状況が起こっていると考えられる．

9.2.3 キーワード候補の中にキーワードがない

キーワード候補の中に全てのキーワードが網羅されていないため，その時点で再現率が低下する要因となっている．テストデータの場合半分以上が候補に含まれていないため，本システムが最高の性能を出したとしても再現率に限界がある．また，複合名詞は本稿では扱っていないのであるが，複合名詞はキーワードとしてはよく現れることがデータからは判断できる．したがって，これまでに見付けられていない単語を発見し，複合名詞に対応することができれば，再現率の向上が期待できる．

10. 今後の予定

前節の通り，キーワード候補語に現れていないキーワードが 30 %以上の割合で存在する．こうしたキーワードを発見することで，再現率を高めることができるため，今後の予定としてはこうしたキーワードを発見するための手法の改良が考えられる．またそうして得られたキーワード候補語の中から適切なキーワードを選び出せる属性値の発見も行う必要がある．また前述した通り，本システムでは複合名詞を全く考慮していないため，こうした複合名詞の付与を行えるようにする予定である．

11. おわりに

本稿では，インターネットからコーパスを収集し，それを用いて入力した Web ページに対してキーワードを付与するシステムについて述べた．本システムは入力した Web ページや関連する Web ページに含まれるキーワード候補語の中からキーワードに成りうる単語を選択して出力する．判断基準として学習データにより作成した決定木を用いる．すでにキーワードを持つ Web ページに対して付与処理を行った結果，精度 15 %，再現率 27 %の正解率を得た．

参 考 文 献

1) Ross Quinlan: C4.5 , http://www.rulequest.com/Personal/ , 2004
2) Yahoo! , http://www.yahoo.co.jp/ , 2005
3) 大沢, ネルス, 谷内田, KeyGraph:語の共起グラフの分割・統合によるキーワード抽出, 電気情報通信学会論文誌, D-I, Vol.J82-D-I No.2, pp.391-400, 1999
4) Small World 構造に基づく文書からのキーワード抽出, 松尾 豊, 大澤 幸生, 石塚 満 , 情報処理学会論文誌, Vol.43, No.6, pp.1825-1833
5) 浅田, 松尾, 石塚 , Web からの研究者ネットワーク抽出の大規模化, 人工知能学会論文誌, Vol.20, No.6, pp.370-378 2005