検索エンジンを利用したキーワードの特徴抽出手法

藪　達也† 遠山　元道††

† †† 慶應義塾大学理工学部情報工学科　〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
E-mail: †yabu@db.ics.keio.ac.jp, ††toyama@ics.keio.ac.jp

あらまし　キーワードの意味概念や特徴を定量的にとらえることは様々なテキスト文書を解析するために必要不可欠である. そこで現在までキーワード間の意味的距離や意味関係等をとりあつかうために大規模な辞書やコーパスなど静的なデータオブジェクトの利用や, 文書集合中でキーワード間の共起関係の利用等が行われてきた.

しかしながらこれらの情報は予め辞書にキーワードの情報を登録しておく必要があったり大規模な文書集合を解析する必要があるため, 新出語や固有名詞などを扱う事が困難であった.

そこで本研究では 1000 語程度の連想語と検索エンジンを利用することで新出語や固有名詞などのキーワードに対しても有効な特徴ベクトルを生成する手法を提案する.

キーワード　検索エンジン, キーワードマイニング

Features extraction technique of keyword with web search engine

Tatsuya YABU† and Motomichi TOYAMA††

† ††Department of Information and Computer Science, Faculty of Science and Technology,
Keio University
Hiyoshi3–14–1, Kouhoku-ku, Yokohama-shi, Kanagawa, 223–8522 Japan
E-mail: †yabu@db.ics.keio.ac.jp, ††toyama@ics.keio.ac.jp

Abstract　It is essential to understand quantitatively the general meaning and the characteristics of the keywords in order to analyse various text documents. Up to now, co-occurence between keywords from the document and static data such as large-scale dictionary or corpus have been used to calculate semantic distance and relationship between keywords to deal with this problem. However, using these, it is difficult to handle new words when they appear or proper nouns since information on the keyword have to be registered in a dictionary beforehand and it is necessary to analyse a large-scale document set. Therefore, we propose in this paper a new method of generating an effective features vector for these problematic keywords using association between around 1000 words and a search engine.

Key words　Web search engine, keyword mining

1. はじめに

現在テキストマイニングなどの文書解析を行う場合, テキストをまず個々の語に分解し, それぞれの語の意味や共起関係などを辞書や大規模なコーパスを用いて解析を行う, といった方法が一般的である. これらの分野の多くの研究では語というものをその語に関連する様々な別の語を用いて表現している. 固有名詞の例で言えば「坂東英二」という語を「野球選手」「タレント」といったように表すのである. しかしこれは語の特徴を定性的にとらえたものであり, では「坂東英二」はどの程度「野球選手」としての話題性があるのか,「タレント」としての話題性があるのかといったことまでは表せていない.

また, 普通名詞や形容詞などは辞書を用いればほとんどの語の意味や関連語, あるいは類似語などが容易に得られるが, 新しい語が日々生まれ続けている現代においては全ての語を辞書に登録して利用するのは限界があるし, ましてや固有名詞の解析に至っては辞書だけでは対応しきれないのが現状である. また, 辞書に無い語の意味を大規模なコーパスを解析することでとらえようとする場合コストがかかりすぎてしまうという問題がある.

そこで本稿ではとある語 (特に固有名詞) の特徴を定量的に表すための特徴ベクトルを, 辞書を利用せずに検索エンジンの検索結果のみを用いて生成する手法を提案する.

以下に, 本稿の構成を示す. まず 2 章で関連研究について述

べる．次に 3 章で特徴ベクトル生成手法の詳細について述べる．さらに 4 章で実装について述べる．そして 5 章で実験と評価を行い, それについての考察を述べ, 6 章で結論およびまとめを述べる．

2. 関連研究

キーワードの特徴をベクトルにする web サービスとして kizasi.jp [1] が挙げられる．この web サービスは毎月 25 万件以上ものブログ記事を時系列毎に要約し, ユーザが与えた一つのキーワードによってブログ記事を検索することでそのキーワードとともに用いられている関連語や, そのキーワードとともに語られている記事の話題を成分としたベクトルをユーザに返すというサービスである．ブログ記事を時系列毎に要約しているため, ホットなキーワードについては非常に多くの関連語が得られたり話題の特徴を表すベクトルを得ることができる．例えば年末の時期に「紅白歌合戦」というキーワードを入力すれば「大晦日」や「年越し」といった関連語が出力されるし, 話題の成分として「TV」と「ワクワクする」といった成分が高い特徴ベクトルが得られる．しかし, あまりホットではないキーワード, 例えば数年前に流行ったおもちゃの名前だとか事件の名前等を入力してもほとんど関連語は出てこない．また全てのブログにおいて一切語られていないキーワードは出力結果さえ返ってこない．kizasi.jp は本来, ある時期におけるホットな話題や関心を抽出するのが目的であるからあらゆるキーワードに対応しているわけではないということになる．

また web 検索エンジンを利用して知識抽出を行う研究については大島らの研究 [2] が挙げられる．この研究ではユーザが与えた 1 語のクエリに対して web 検索エンジンが持つ情報のみから同位語とそのコンテキストを発見する手法を提案している．この手法では web 検索エンジンの検索結果の情報のみから知識抽出を行うことを web 検索エンジンインデックスのクイックマイニングと呼んでおり, 知識取得のために必要な追加コストが低いことと適応分野が広いことを大きな利点としている．本稿で提案する手法もこの研究と同様に web 検索エンジンの検索結果のみから知識取得を行っており, その利点としても本研究はこの研究と同様のものを挙げているため, 類似性の高い研究といえる．しかしながら本稿で我々が提案する手法はキーワード自体のもつ特徴を抽出することが目的となるため, この研究のとあるキーワードに対する同位語とそのコンテキストの発見とは目的が異なる．

さらに野田らの研究 [3] では, 従来 web 検索の結果をクラスタリングする際にキーワードの意味内容までは考慮していないということから, クラスタリングにはキーワードと意味的に密接な関係をもつ語を発見する必要があると考え, そのような語を発見するためにキーワードに付帯する「親概念」と「話題語」という 2 つの補助概念を提案している．親概念というのは「そのキーワードとは何者なのか」という概念であり, 話題語というのは web 検索における話題, つまり求める情報の方向性を特徴付ける語のことである．この研究では web 検索エンジンを用いて質問キーワードに対する「話題語」を自動抽出することを行っているが, 「親概念」については概念を提案しただけでその取得方法には言及していない．ここで, 本稿で提案するものはまさに「そのキーワードとは何者なのか」という知識を抽出するものであり, そのために情報の方向性を特徴付ける語 (本稿では「連想語」と呼ぶ) を用いて解析を行っている．そのため本稿は, この研究で抽出しようとしている「話題語」のようなものは既に取得できているという前提の上で, それを使うことで「キーワードが何者であるのか」つまりキーワードの「親概念」のようなものを特徴ベクトルとして取得しようとする研究, というとらえ方もできる．

3. キーワードの特徴ベクトル

本稿で提案するキーワードの特徴ベクトルは, キーワードがどのカテゴリにどのくらい属するのか, という度合いを定量的に表すものとする．そのためベクトルの各成分は各カテゴリについてキーワードが属している度合いということになる．この度合いを本稿では「連想語」という指標を用いて定義している．そこで, まず以下でカテゴリ及び連想語についての詳細を述べてから特徴ベクトルの生成手法について述べる．

3.1 カテゴリ

本稿ではキーワードが何者なのかを表す特徴ベクトルの成分をカテゴリと定義する．カテゴリは図 1 のようになっており, 本稿ではこの 21 カテゴリをルートカテゴリと呼ぶ．各カテゴリには図 2 のように下の階層に詳細カテゴリが存在しており, それによってルートカテゴリだけでは表現できない詳細情報を表現できるようになっている．

なお, カテゴリは Yahoo!カテゴリ [4] を参考に筆者が独自に作成したものである．ただし本稿で提案するのは特徴ベクトルの生成手法についてであり, カテゴライズはここで挙げたようなものでなくても構わない．あくまでここに挙げたカテゴリは本手法を実装するにあたって比較的広い分野の固有名詞を扱えるようにしたものであり, 例えば何かの分野について特化した特徴ベクトルを得たかったら, その分野についての詳細なカテゴリ定義を行えば良い．

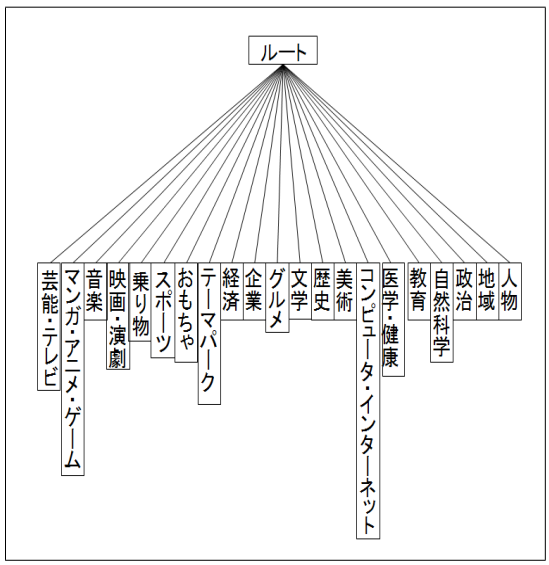

図 1　ルートカテゴリ

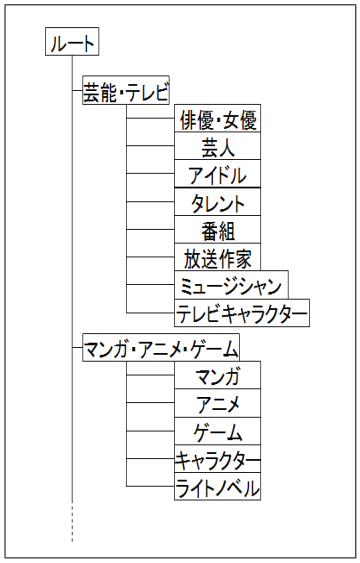

図 2　カテゴリの階層構造

3.2　連　想　語

本稿では, 前節で定義したようなあるカテゴリ階層内で, 各カテゴリを連想させる語を連想語と呼ぶ. 例えば「芸能・テレビ」カテゴリの連想語としては図 3 のようなものが挙げられる. ここに挙げたものは「芸能・テレビ」というカテゴリだけを連想させる語であり, 他のカテゴリを連想させるものではない, というのが前提である.

しかしながら「人物」カテゴリだけは特別なカテゴリとなっている. これは与えられた固有名詞が人名なのかそうではないのか, を判断するための重要なカテゴリであり, 例えば「タレント」という連想語を挙げてみると, これは「芸能・テレビ」カテゴリと「人物」カテゴリの両方を連想させる語である. このように連想語は基本的には一つのカテゴリを連想させる語であるが, 「人物」カテゴリだけは他のカテゴリとともに同時に連想しても良いということにしている.

また, 連想語は特徴ベクトル生成時に検索結果の情報を形態素解析したものとマッチングを行うために使用するので, 連想語自体が形態素であり, あらかじめデータベースに格納しておくものとする.

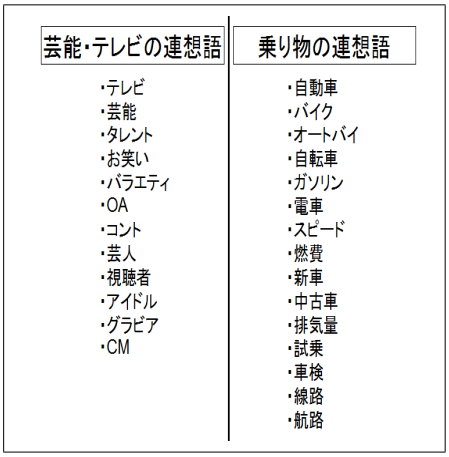

図 3　連想語の例

3.3　特徴ベクトル生成手法

本稿で提案する特徴ベクトルは以下の手順で生成する.

（1） キーワードをクエリとして web 検索を行い検索結果のうち上位 50 件のページタイトルとスニペットを取得する.

（2） 取得した検索結果に対して形態素解析を行い, 名詞だけを取り出してそれぞれの名詞の出現回数をカウントする.

（3） あらかじめデータベースに格納してある連想語とこれらの名詞とのキーワードマッチングを行う.

（4） マッチした名詞の出現回数を連想語に割り当てられているカテゴリの成分に足していく.

（5） 全ての足し合わせが終了した時点で得られたベクトルを正規化し, クエリとして入力されたキーワードの特徴ベクトルとする.

まずは特徴ベクトルを生成したいキーワードを入力する. この時, 入力するキーワードは一語でなくても良い. 例えば「ヤマハ」というキーワードを入力すると音楽教室やバイクや楽器など様々な内容のページの検索結果が取得されることになる. ここで仮に特徴ベクトルとして取得したい対象がバイクについてであったとすると「ヤマハ バイク」といったようにスペースを空けてキーワードを複数入力することで AND 検索がなされ, 検索結果として取得されるページを絞り込むことができる.

次に取得した検索結果 50 件のページタイトルとスニペットに対し形態素解析を行う. ここで名詞以外の品詞は全て排除するとともに, 名詞の出現回数をカウントする. 例えば

バイク・スクーターホームページ
スポーツバイク, **スクーターの製品情報**, **試乗車検索**, **販売店検索等**. ... **バイク・スクーター**. **スポーツ バイク**, **スクーターを中心とした**, **ヤマハのバイクを紹介します**. ... **ヤマハオートバイ** (**新車**) **初回点検** (1**ヶ月目**) **は無料**. **新車** ...

という文章があったとする. ここで形態素解析を行って名詞のみをとりだし, その回数をカウントすると, 以下のようになる.

バイク：5 , **スクーター**：4 , **スポーツ**：2 , **検索**：2
ヤマハ：2 , **新車**：2 , **ホーム**：1 , **ページ**：1
製品：1 , **情報**：1 , **試乗車**：1 , **販売店**：1
中心：1 , **紹介**：1 , **オートバイ**：1 , **初回**：1
点検：1 , 1：1 , **目**:1 , **無料**：1

さらに, これらの名詞に対して連想語とのキーワードマッチングを行う. その結果,

バイク：5 , **スポーツ**：2 , **新車**：2 , **オートバイ**：1

という 4 つの連想語が抽出される.

それぞれの連想語について, 割り当てられているカテゴリのベクトル成分に出現回数を足し合わせていく. 上の例では,「バイク」「新車」「オートバイ」の 3 語は全て「乗り物」のカテゴ

リに属し, 「スポーツ」は「スポーツ」カテゴリに属すことになる. そのため, この例からは「乗り物」カテゴリの成分が 8 であり,「スポーツ」の成分が 2 であるようなベクトルが得られる. 最終的にはベクトルを正規化し, そのキーワードの特徴ベクトルとする.

4. 実　　装

4.1 web 検索と形態素解析

キーワードをクエリとして web 検索を行い検索結果の形態素解析を行う際に, 本研究では Yahoo!検索及び Yahoo!形態素解析を利用した. なお検索結果として利用したのは上位 50 件のページタイトルとスニペットのみである.

4.2 実装の概要

実装は PHP を用いて web アプリケーションとして作成した. このシステムはまずユーザからキーワードをクエリとして受け取りそれをもとに web 検索を行う. そして 3.3 節で示した手順でキーワードの特徴ベクトルを生成し, ベクトルの各成分を図 4 や図 5 に示したようなグラフにして出力する. なお, クエリ入力毎に毎回検索と形態素解析を行ってベクトル生成をしているのではなく, 一度入力されたキーワードの特徴ベクトルは二次利用のためにデータベースに格納している. また, こうすることで再度同じキーワードのベクトルを生成する際にかかる時間を短縮することができるとともに検索エンジン側に過負荷をかけないよう配慮している.

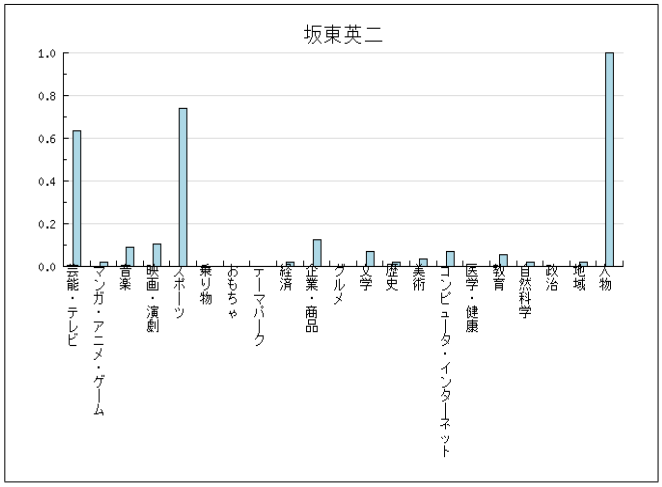

図 4　生成されたグラフ (坂東英二)

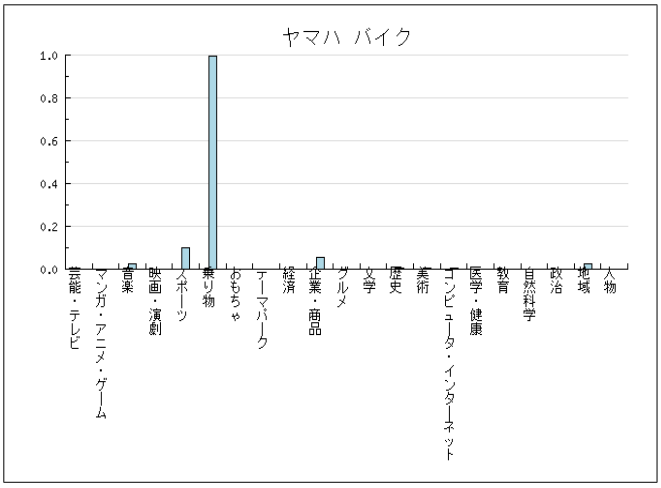

図 5　生成されたグラフ (ヤマハ　バイク)

5. 実験とその結果及び考察

キーワードの特徴ベクトルの妥当性の評価のために, 以下の手順で評価用の正解ベクトルを作成した.

（1） 一般のユーザ 10 名に表 1 に示した 22 語のキーワードについての wikipedia を一読してもらい, その語に対するある程度の知識を得てもらう.

（2） 全てのキーワードについて, ルートカテゴリ 21 次元の各々が, 強く当てはまるなら 2, 少し当てはまるなら 1, 全く関係がないなら 0 を成分としたベクトルを作成する.

（3） あるキーワードの正解ベクトルの各成分の値を 10 名のユーザの回答結果の各成分の和とする.

（4） 正解ベクトルを正規化する.

ここで, 上記の手法によって得られた正解ベクトルと本稿で提案した手法を用いて作成した特徴ベクトルとの類似度を測る. また, 比較のために上記 (1) を行ってユーザがキーワードに対する深い知識を得る前に, おおよその知識だけで上記 (2)〜(4) に示したようなキーワードの評価を行ったユーザベクトルを生成し, 正解ベクトルとユーザベクトルとの類似度も測る. もし正解ベクトルと本稿で提案した特徴ベクトルとの類似度が, 正解ベクトルとユーザベクトルとの類似度よりも高ければ, それは少なくともユーザが漠然ともっているキーワードへのイメージよりも本稿で提案する手法のほうがより正確にキーワードのもつ概念をとらえられたことになる.

正解ベクトルと本稿で提案した特徴ベクトルとの類似度は cos 類似度を用いる. 2 つのベクトル $\boldsymbol{v1}$ と $\boldsymbol{v2}$ の cos 類似度は以下のように表される.

$$\boldsymbol{Sim(v1, v2)} = \frac{\boldsymbol{v1 \cdot v2}}{|\boldsymbol{v1}||\boldsymbol{v2}|}$$

ここで,cos 類似度 $\boldsymbol{Sim(v1, v2)}$ は 2 つのベクトル $\boldsymbol{v1}$ と $\boldsymbol{v2}$ のなす角とされ, 最も類似している状態, すなわち 2 つのベクトルが全く同じ場合に最大値 1 をとる.

表 1　評価に用いた全てのキーワード

大乱闘スマッシュブラザーズ X ，東京マラソン ，名古屋国際女子マラソン ，タイタニック ，ラストフレンズ ，ROOKIES ，川田亜子 ，中国四川省 ，岩手・宮城内陸地震 ，植田辰哉 ，シバトラ ，崖の上のポニョ ，北京オリンピック ，谷亮子 ，麻生内閣 ，リーマン・ブラザーズ ，セレブと貧乏太郎 ，緒方拳 ，筑紫哲也 ，レッドクリフ ，飯島愛 ，NON STYLE

なお, 表 1 に示したキーワードは 2008 年 2 月から 2008 年 12 月までの 11ヶ月間において,kizasi.jp のきざしランキングアーカイブの月間ランキングの上位から固有名詞であり且つ wikipedia に記事が掲載されているもののみを 2 件ずつ取り出したものである. きざしランキングアーカイブでは日毎, 週毎, 月毎にブログで語られている件数の多いキーワードのランキングを得ることができる. ただし, ブログに出現する語としては正式名称というより略語で現れることが多いため, その場合は正式名称に直した上で wikipedia の記事を参照し, 評価した. 実験の結果を表 2 に示す.

表 2　正解ベクトルとの cos 類似度

キーワード	本手法	ユーザベクトル
大乱闘スマッシュブラザーズ X	0.991	0.695
東京マラソン	0.895	0.892
名古屋国際女子マラソン	0.952	0.943
タイタニック	0.741	0.658
ラストフレンズ	0.968	0.859
ROOKIES	0.700	0.682
川田亜子	0.961	0.792
中国四川省	0.688	0.875
岩手・宮城内陸地震	0.564	0.889
植田辰哉	0.996	0.598
シバトラ	0.969	0.821
崖の上のポニョ	0.729	0.837
北京オリンピック	0.916	0.890
谷亮子	0.932	0.857
麻生内閣	0.980	0.873
リーマン・ブラザーズ	0.795	0.680
セレブと貧乏太郎	0.866	0.590
緒方拳	0.966	0.559
筑紫哲也	0.904	0.744
レッドクリフ	0.950	0.598
飯島愛	0.858	0.678
NON STYLE	0.926	0.882

表 2 のとおり, ほとんどのキーワードにおいて, 本手法を用いた特徴ベクトルと正解ベクトルとの類似度の方がユーザベクトルと正解ベクトルとの類似度よりも高い値となっている. ユーザベクトルと正解ベクトル間の類似度を見てみると, どちらもユーザのイメージに基づくベクトルなのに高い値や低い値があることがわかる. この値が低いということは, ユーザが漠然と思い描いていたキーワードに対するイメージと, ある程度正確な知識を得た後に思い描くイメージにかなり差があったということになる. これは, キーワードの知名度がもともと低かったり, キーワードについての一般的にはあまり知られていなかった意外な一面を知ることでイメージが変化したということが考えられる. また, この値が高いということは逆にイメージの変化があまりなかったということが考えられる. 例えば「東京マラソン」や「岩手・宮城内陸地震」など, キーワード自体がカテゴリを明確に表している場合は高い値となっていることがわかる.

次に本手法と正解ベクトルとの類似度について見てみると, 全体的に高い値がでていることがわかる. しかし「中国四川省」や「ROOKIES」「崖の上のポニョ」などで低い値となっている. 「中国四川省」については 2008 年 5 月 12 日に大地震が発生しているが, ユーザとしては「地震」という自然科学カテゴリのイメージの他にも「四川料理」というグルメカテゴリのイメージや「都市」という地域カテゴリのイメージなど様々なものがあった. しかし本手法を用いた場合は web 検索を行うため, 大地震に関するページが上位に多く含まれたことによって自然科学カテゴリの成分が他のカテゴリの成分に比べはるかに大きくなってしまった. そのために正解ベクトルとの類似度が低下したことが分かった. また,「ROOKIES」や「崖の上のポニョ」についても同様に, 本手法を用いた場合, ある一つのカテゴリの成分が飛びぬけて高い値をとることがあるために類似度の低下が起こっていることが分かった.

しかし, このようなことは決して悪いことではない. キーワードのもつ意味概念というものは時間とともに変化することも多々あるからである. 例えばとあるお笑い芸人がハリウッド映画に出演したということで世間で話題になったとする. そういった話題が出る前のその人物の概念というものはただの「お笑い芸人」であったかもしれない. しかしひとたびそういった話題がでると「映画俳優」としての概念が形成されることであろう. 本手法では, こういった時事とともに変化するキーワードのもつ意味概念を web 検索を用いることによってとらえることができる. そのため,wikipedia などで得られるキーワードの広い概念とは類似していないとしても, 本手法を用いた時期における, そのキーワードの「話題性」のようなものをとらえることができるということで有用である.

6. ま と め

本研究では辞書だけでは対応できなかった固有名詞をはじめとするキーワードの意味概念や特徴を定量的にとらえるために, 検索エンジンと 1000 語程度の連想語を用いることでキーワードの特徴ベクトルの生成を行った. 従来研究においては巨大なコーパスを解析することでキーワードの意味概念を解析して

いたが, 我々の手法では web 検索によって即座にキーワードの意味概念を得ることができる. 更に, 実験によって web 検索エンジンから得られる情報から, キーワードの現在語られている話題性のようなものをうまく取得できることが分かった. そういった意味では本手法によって生成されるキーワードの特徴ベクトルとは, キーワード自体の正確な意味を表しているのではなく, 現時点におけるキーワードの話題性を表すベクトルということができる.

また, 本稿で定義したカテゴリや連想語についてはまだ曖昧なところが多いのも事実である. 今後の課題としては, より有用な特徴ベクトルを得るためのカテゴリと連想語の定義の厳密化を進めるとともにこの特徴ベクトルを応用したテキストマイニングなどの諸手法について検討を行っていく.

文　　献

[1] kizasi.jp: http://kizasi.jp/

[2] 大島 裕明, 小山 聡, 田中 克己: 『web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見』, 論文誌 (トランザクション) - 論文誌 データベース（TOD）Vol.47 No,SIG19, 2006 年 12 月

[3] 野田 武史, 大島 裕明, 手塚 太郎, 小山 聡, 田中 克己: 『web 検索結果のクラスタリングに用いる話題語の質問キーワードからの自動抽出』, 電子情報通信学会第 17 回データ工学ワークショップ (DEWS2006),2006

[4] Yahoo!カテゴリ: http://dir.yahoo.co.jp/