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

野田　武史† 　大島　裕明† 　手塚　太郎† 　小山　聡† 　田中　克己†

† 京都大学大学院 情報学研究科 社会情報学専攻 〒 606-8501 京都市左京区吉田本町
E-mail: †{noda,ohshima,tezuka,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　現在，Web 検索におけるユーザインターフェースとして，検索結果をクラスタリングして表示する手法が提案されているが，これらはキーワードの意味内容までは考慮しておらず，ユーザが検索結果を効率的に探索するためのツールとして必ずしも十分な利便性をもっているとはいえない．効果的なクラスタリングを行うためには，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．本研究では，このような語を抽出するために，キーワードに付帯するものとして「親概念」と「話題語」という 2 つの補助概念を提案する.「親概念」とは，キーワードがそもそも「何であるか」について表現する概念であり,「話題語」とはある親概念に結びつけられた，それについて語られる「話題」の類型である．これらの概念を用いることで，キーワードが指す対象の曖昧性を回避しながら，話題の意味内容に則したクラスタリングを行うことが可能となると考えられる．今回は，ユーザが入力したキーワードからこれらの補助概念を生成する手法の開発に主眼を置いて研究を行った.

キーワード　情報検索，Web とインターネット，データマイニング

Automatic Extraction of Topic Terms for Web Search Result Clustering

Takeshi NODA†, Hiroaki OHSHIMA†, Taro TEZUKA†, Satoshi OYAMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshidahonmachi, Sakyo, Kyoto 606-8501 Japan
E-mail: †{noda,ohshima,tezuka,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　Meta-search engines that clusterize search results into index-labeled categories have recently been proposed and are gaining popularity. Although categorized views of these search engines are extremely useful, the labels of clusters are not satisfactory in many cases, because most of these do not consider the semantic relationship between the query word and the cluster label. We propose two complementary concepts, “super class concepts” and “topic terms” of keywords. The super class concept represents the class that the keyword belongs to and the topic term is a typology of topics about the keyword, which depends on the keyword’s super class concept. Using these complementary concepts, topic aware Web page clustering can be achieved. We describe a methodology in this paper for extracting these concepts based on Web meta-searching.

Key words　Information Retrieval, Web and Internet, Datamining

1. はじめに

近年，Web ページ検索技術の進歩は非常に速く，日々新たな技術が開発され，公表されている．これらの中の 1 つとして，ユーザが入力したキーワードを用いて他の検索エンジンで検索を行い，その結果をいくつかのクラスタに分けて表示するものがある．このような検索エンジンが登場する背景には，旧来の検索エンジンにおいて検索結果が単なるリストとして表示されることへのユーザの不満があると考えられる．この表示方法においては，似通った内容のページ群を近い位置に配置することで視認性を高め，ユーザが検索結果を順に閲覧していく作業を効率化してくれるという利点があるため，多くのユーザによっ

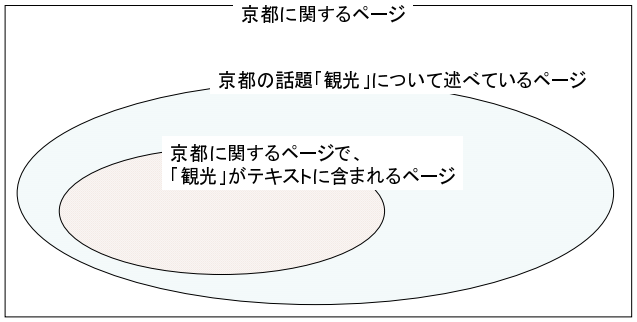

図 1　話題適合ページと話題語を含むページ

て支持されている．しかし，現在の Web ページクラスタリングは，質問キーワードの近傍に現れる単語列や結果ページ群に頻出する単語を元にクラスタリングを行うものであり，ラベルとして用いられる語が質問キーワードとどのような意味的つながりをもつかについては考慮していない．このため，しばしば入力した質問キーワードと意味的な関連性の薄い語がクラスタのラベルに用いられることがある．

また，そもそも Web 検索を行う際にはユーザが求める情報は漠然としてではあれ，ある程度の対象範囲をもっていることが多い．たとえば，同じ「京都」というキーワードで検索を行う場合でも，京都の観光名所をリストアップしたい場合と，特に京都の歴史に関する情報のみを知りたい場合とでは求めるページは異なっているはずである．Web 検索の結果をクラスタリングするのであれば，このような求める情報の「話題」に即した方法で分類が行われていることが望ましいが，現在のクラスタリング手法ではこれは不可能である．なぜなら，京都の観光名所を紹介するページにおいて，必ずしも「観光」というテキストが含まれているとは限らないためである (図 1)．

本研究では，これまでの Web ページクラスタリングとは別のアプローチによって Web ページをクラスタリングすることを考える．すなわち，ユーザが入力する「京都」のような質問キーワードを Web 検索の主題と捉え,「観光」や「歴史」,「グルメ」など，その主題に関する各々の話題によって Web ページをクラスタリングする手法である．このような Web 検索における話題，求める情報の方向性を特徴づける語を本研究では「話題語」と呼ぶ．

本稿では，キーワードに関する話題語をどのように検出するかについて焦点をあて，考察と実験を行った．2 章で関連研究について述べ，3 章で本研究で利用する諸概念について，4 章で提案する話題語抽出の手法について，5 章で実装について，6 章で実際に行った実験結果について述べ，7 章でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 クラスタリングを行う検索エンジン

与えられたキーワードの話題語を検索できる Web 検索エンジンとして，Clusty [1] が挙げられる．Clusty では，検索結果をクラスタリングし，それぞれを特徴づける語を提示する機能が実装されている．Clusty は Vivisimo 社のクラスタリングエンジンを用いたメタ検索エンジンであるが，そのクラスタリング手法について詳しくは明らかにされていない．Clusty はユーザの入力を元にメタサーチを行い，その結果をクラスタ化して表示するが，クラスタのラベルの選定はあくまで Web ページのクラスタリングから導き出されたものであり，入力キーワード自身との関係性は考慮していない [2]．

KartOO [3] は，検索結果を地図を模したグラフィカルな表現によって提示する検索エンジンで，サイトを都市，サイト間の関連性を道として表現している．評価の高いサイトを大きく表示するなどの工夫が凝らされているが，表示されるものは本研究が対象としてるような話題ではなく，キーワードに関連するあらゆるものを対象としている．

2.2 関連語の抽出

小山らは，Web ページのタイトルと本文という内部構造に着目して，ある主題とそれを詳細化する話題を抽出する手法を提案している [4]．彼らは Web ページのタイトル部分にはそのページの主題となる語が含まれやすく，本文にはそれを詳細化するような語が含まれていると考え，統計的手法を用いて有意性を判定し，詳細語の抽出を行っている．この手法では高い精度で主題を詳細化する語を抽出することができるが，ここでいう詳細語は本研究の提案する話題語とは必ずしも同じではなく，広く質問キーワードに関連をもつ語を対象としている．

山本らは，NTCIR の学術文書データ・毎日新聞記事データ・中国語による新聞記事データを対象に，文章中における単語の使用形態でスコアリングを行い，関連語シソーラスの自動構築を行った [5]．関連の種別の判別は行っていないため，本研究で論じる話題語以外の関係性を持つ関連語も多数抽出されている．

佐藤らは，クラスタリングされたニュース文書を対象に，クラスタを代表する語句を話題語と定義した [6]．クラスタ内の文書における頻度をもとに語句話題度を定義し，その上位を話題語として抽出した．この手法において特定の単語に対する話題語ではなく，文書集合に対する話題語が取得されている．

検索エンジン Google が提供する Google Suggest [7] では，ユーザが入力したキーワードに対し，過去にそれと共に複数キーワード検索された単語が表示される．しかし，表示される単語がユーザのキーワードに対してどのような関係を持つのかは必ずしも明確でない．本研究で扱われる話題語も含まれるが，それ以外の関係性を持つ単語も多く取得されてしまう．

3. 質問キーワードの話題語と親概念

3.1 質問キーワードの話題語

Web 検索においては，多くのユーザが多数の語からなる質問キーワードよりも，語数の少いより単純なキーワードを使用する傾向があることが知られている [8] [9]．これは，いきなり多数の語を入力してピンポイントに検索を行うより，まず求める情報の核として絶対に外すことのできない語のみを用いて大雑把に検索を行い，その結果を見ながら徐々に入力語を増やして結果を絞り込んでいく，という検索スタイルが好まれているためだと考えられる．

なぜこのような検索スタイルが好まれるのか．それは，たと

えば京都の観光に関する情報を得たいときなど，いきなり「京都 観光」と入力して検索を行うと，検索結果に含まれてほしいページが除外されてしまう可能性があるからだと考えられる．すなわち，明示的に指定された「観光」という語が，京都の観光に関して述べている全てのページに含まれているとは限らないためである．もし，祇園祭について詳しく解説を行っているページがあった場合，このページは「京都の観光」というユーザの検索対象範囲に含まれるべきであろう．しかし，このページ内に「観光」という語がテキストとして明示されているとは限らない (図 1)．このため，ユーザは必要以上に多くの語を質問キーワードとして入力することを避けようとする．

このように，Web 検索を行う際，ユーザは求める情報の範囲を暗黙的に想定していながら，大雑把なキーワードのみを入力して検索を行うことがある．本研究では，ある検索対象において絶対に外すことのできない，文字通りキーとなる語のことをキーワードと呼び，キーワードに関連する特定の情報の範囲のことをそのキーワードの「話題」，話題を言葉で表現したものを「話題語」と呼ぶ．この他にも，想定している話題をうまく言語化できないなどの理由で仕方なくキーワードのみが入力される場合が多々ある．この結果，ユーザは小数の質問キーワードによって検索を行い，得られた膨大な量の検索結果を 1 つ 1 つチェックする必要に迫られることとなるが，この作業は非常に繁雑なものであるため，結局上位の数十件をチェックするだけで諦めてしまう場合が少なくない．

このように，現在の検索エンジンには検索対象を絞り込む際，残ってほしい Web ページがふるい落とされてしまうという問題が存在する．これは，キーワードとその話題語という本来異なる階層の語を同等に扱おうとしているために起こる問題だと考えることができる．

このことから，本研究では Web 検索において，キーワードとその話題語を区別して扱うことを提案する．話題語の例としては，京都の場合は他に「グルメ」や「写真」,「歳時記」などが考えられる．

3.2 話題語の性質

質問キーワードの話題語となる語は，日本語による表現ではしばしば「京都の観光」や「京都の歴史」,「京都の歳時記」といったように,「質問キーワード の—」という形で用いることができる．このことに注目すると,「質問キーワード の—」という表現が含まれるページは，質問キーワードの—という話題に言及していると期待できる．もちろん，先に指摘したように全ての適合ページがこの方法で検出できるわけではないが，少くとも話題語が存在するということは知ることができる．このような「名詞 の 名詞」という形で用いられる「の」は，言語処理の分野では連体助詞と呼ばれる [10].

3.3 キーワードの親概念

キーワードとその話題語との関係について更に考察すると，キーワードと話題語との間には一定の制約があることがわかる．たとえば，キーワード「京都」の話題語として「観光」や「歴史」を考えることができるが，キーワード「小泉純一郎」の話題語として「歴史」を考えることはできても,「観光」を考えることは通常できない．これは，京都と小泉純一郎が互いに異なったものであること，すなわち京都は「都市」であり，小泉純一郎は「人間」であるということに起因すると考えることができる．つまり，あるキーワードがどのような話題語を持ち得るかは，そのキーワードがどのような親概念をもつかによって決まる．キーワードとその話題語との関係を考える際，そのキーワード自体が「何ものであるか」ということがわかっていれば，話題語の候補を発見する際に利用できる．本研究ではこれを「キーワードの親概念」と呼ぶ．キーワードとその親概念との関係は図 2 のようになる．

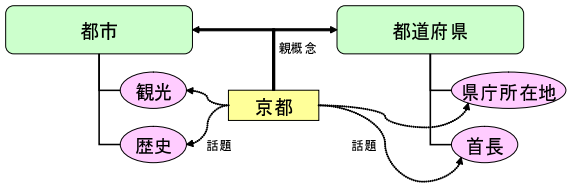

図 2　キーワードの親概念と話題語

4. 質問キーワードからの話題語の抽出

本研究の最終目的は，ユーザが入力した質問キーワードに従ってその話題語を発見し，その話題について言及しているページをその話題語の下にクラスタリングして表示することである．本稿ではその前段階として，キーワードの話題語としてふさわしいものを抽出することを目的とし，以下のような手順で抽出を試みた．

4.1 Step1: 話題語の抽出

質問キーワードを k，話題語を t とすると,「k の t」というフレーズが多くの場合に成立するということに着目し,「k の」という文字列をクエリとして検索エンジンに送り，検索結果として Web ページのサマリ群 S を得る．これらのサマリ群に形態素解析を適用し,「の」以降に続く普通名詞を抽出する．このようして抽出された名詞群を，キーワードに対する話題語の候補 T_c とする．

ここで抽出対象を名詞に限定したのは，たとえば「京都の古い町並み」のように，話題語の前に挿入されている修飾語を抽出しないようにするためである．また,「京都の嵐山」や「京都の清水寺」といった固有名詞も除外することにした．なぜなら，このような固有名詞は京都の話題の 1 つという見方もできる一方,「京都の嵐山の紅葉」というように，それ自体が更に深い話題をもつキーワードであるとも見ることができるからである．このように,「の」は何度もつなげて用いることができる語であるため，本研究では「の」の後に続く話題語の候補としては，普通名詞のみを用いることとした．

4.2 Step2: 話題語のランキング

抽出された話題語の候補には，例えば「情報」や「ホームページ」のような，Web 上で広汎に使用されどのようなキーワードにおいても話題となり得る語が多数含まれている．このような語は，キーワードに特徴的な話題語に比して重要性が低くなるようにしたい．

これらの語を取り除く方法としては，既存のシソーラス等を用いた辞書的な方法で検証することも考えられるが，常に新語が生まれ続けている Web 上での情報検索において，内容が固定された辞書を用いることには限界がある．また，このような語が全てのキーワードに対して話題語としての価値をもたないとは言い切れないため，完全に取り除いてしまうのは問題である．そのため，今回の実験では辞書的な手段を用いない代わりに，情報検索において広く用いられている TF-IDF 法 [11] を応用した尺度として KTPF(Keyword-Topic Phrase Frequency) を定義し，利用する．

KTPF はキーワードを k，話題語候補を $t \in T_c$ とおいたとき以下の式で定義される．

$$ktpf(k,t) = \frac{count(k,t) \cdot df(k)}{df(t)},$$

ここで，$count(k,t)$ は Step1 で得られたサマリ群 S 中での t の出現回数の数え上げであり，$df(k)$ は “k の” が出現する文書数，$df(t)$ は “の t” が出現する文書数である．

このように定義すると，$ktpf(k,t)$ はキーワード k とその話題語 t との関連の強さを表す尺度として利用できる．なぜなら，$count(k,t)$ は話題語候補 t とキーワード k との親密さを表し，$df(t)$ で割ることで一般性の高い語の値が相対的に低くなるためである．この手法の優れている点は，キーワードと話題語を結ぶ「の」という語を用いた検索だけでキーワードと話題語との関連性を測ることができるところにある．KTPF を用いてランキングを行うことで，キーワードにより関連性の深い順に話題語候補を整列することができる．

5. 実　　装

5.1 Web 検索および検索結果の取得

キーワードから話題語を求める作業および各フレーズの出現頻度を求める作業では，Web ページ検索エンジンとして Google [12] を用いた．検索結果にはページのタイトルや URL，ページサマリ，該当ページ件数などが含まれるが，本研究ではこのページサマリに対して形態素解析を行い，該当ページ件数を DF として利用した．

5.2 形態素解析

連体助詞「の」を用いて前後の語句を抽出するために，奈良先端科学技術大学院大学情報科学研究科の工藤拓氏が開発した Mecab [13] を用いた．Mecab は，同大学自然言語処理学講座の開発する ChaSen を基に開発された高速な形態素解析器である．

5.3 実装の概要

実装はスクリプト言語 Ruby を用いて行った．このシステムはまず，キーワード k を入力として受け取り「k の」をクエリとして Google による Web 検索を行って，Web ページ 500 件のページサマリと $DF(k)$ を得る．次にこのページサマリの全てに対して Mecab による形態素解析を適用し，話題語候補群 T_c と $count(k,t)$ を作成する．そして，全ての $t \in T_c$ に対して再び「の t」をクエリとして Google で検索を行い，$DF(t)$ の値を取得し，$ktpf(k,t)$ を求め, 表示する．

6. 実験とその結果／考察

6.1 実 験 内 容

実験では，前章で提案した手法を用い，具体的なキーワードで実際に検索を行い，再現率—精度グラフを描いて本手法の妥当性の検証を行った．

6.2 抽出された話題語の検証

まず，いくつかのキーワードで実際に検索を行った．Google での取得ページ数を 500，ランキング数を上位 20 件とし，質問キーワードを「京都」と「コーヒー」として検索した結果をそれぞれ表 1，表 2 に示す．

これらの表はそれぞれ，検索された話題語の候補とその $count(k,t)$，$df(t)$ および KTPF の値の組を $count(k,t)$ の降順で並べたものである．表から分かるように，$count(k,t)$ と $ktpf(k,t)$ の値は食い違っているものが多く見受けられる．これは，京都の表に現れている「情報」や「ホームページ」などの一般性の高い語は，$df(t)$ の値が高くなる傾向にあるためである．これは，質問キーワードと強い関連性をもつ話題語を重視するという本研究の目的に合致している．コーヒーの表においても,「味」や「香り」,「鮮度」などのコーヒーと深い関りがあると考えられる話題語の KTPF 値は高く保持されており，本手法は十分実用的な機能を有していると推測される．

その一方で，京都の「マン」やコーヒーの「クロ」など，明かに話題語として不適切な語でありながら高い KTPF 値をもつものも存在する．これらの語は，実際には「マン喫 (マンガ喫茶)」と「クロロゲン酸」という名詞の一部であり，形態素解析に使用した Mecab の内部辞書に登録されていなかったために誤って抽出されたものであると考えられる．

今回の手法を用いる限り，このようなノイズが候補に含まれることが避けられないが，このような語が含まれないようにするには，他のものに較べて極端に大きな値をもつものを外れ値として無視することや，Mecab の辞書を増強することなどによって対応が可能である．また，このような工夫を行うにあたっては，コーヒーの表中の「カフェイン」のように，本当の意味でコーヒーと強い関連性を持つ語を除いてしまわないような配慮も必要となると考えられる．

6.3 再現率—精度グラフ

次に，本手法において KTPF を用いることで候補のランキングがどのように変化するかを検証するため，話題語の評価尺度として単純な数え上げ ($count(k,t)$) を用いた場合と，KTPF を用いた場合について質問キーワードからの話題語抽出の再現率—精度グラフ (図 3) を描き，比較した．今回は 20 件の候補を抽出したので，これらの 20 件から上位 n 件に含まれれる話題語候補の集合に対して再現率と精度を計算し，n を 1 から 20 まで変化させることによってグラフを描いた．

話題語としての適切性の判断基準としては，質問キーワードを k，話題語候補を t として,「k の t」という表現に違和感を覚えない候補を適合候補とし，7 人がそれぞれ適合／不適合を判断して 4 人以上が適合と判断した候補を最終的な適合候補とした．

表 1　京都 (df=1750000)

話題語の候補 t	$count$(京都, t)	$df(t)$	$ktpf$(京都, t)
情報	9	29500000	0.534
旅館	7	1430000	8.566
不動産	7	1690000	7.249
魅力	6	5190000	2.023
風景	6	2820000	3.723
中心	6	3520000	2.983
グルメ	5	3120000	2.804
文化	5	3440000	2.544
ホームページ	5	15700000	0.557
老舗	5	3300000	2.652
街	5	3570000	2.451
大学	4	3990000	1.754
マン	4	155000	45.161
伝統	4	2590000	2.703
天気	4	2800000	2.500
文化財	4	634000	11.041
ホテル	3	2670000	1.966
格安	3	3170000	1.656
写真	3	8670000	0.606
カフェ	3	3220000	1.630

表 2　コーヒー (df=2560000)

話題語の候補	$count$(コーヒー, t)	$df(t)$	$ktpf$(コーヒー, t)
味	28	2910000	24.632
香り	25	2340000	27.350
歴史	11	8350000	3.372
実	10	1890000	13.545
量	9	2470000	9.328
豆	9	2640000	8.727
粉	8	1320000	15.515
風味	8	3740000	5.476
基礎	6	6630000	2.317
種類	6	7450000	2.062
専門	6	3420000	4.491
成分	5	2820000	4.539
楽しみ	5	1860000	6.882
通販	5	1830000	6.995
木	5	2350000	5.447
鮮度	4	416000	24.615
効能	4	932000	10.987
カフェイン	4	59800	171.237
味わい	4	2300000	4.452
クロ	4	165000	62.061

また，検証には以下のキーワード群を利用し，各上位 n 件の再現率および精度を平均したものを用いた．

京都，コーヒー，大阪，ジャガー，ディスプレイ，マッコウクジラ，聖徳太子，アルゴリズム，電話，鉛筆

これらのキーワードの選択理由は恣意的なものであり，何らかの公平なテストコレクションに基づくものではないが，様々な特徴をもつキーワードを選択する目的で以下のように考え，選択した．

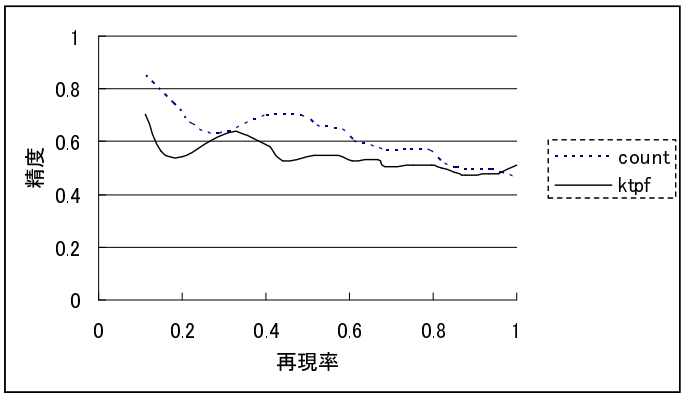

図 3　再現率—精度グラフ

多義語の例として知られる「ジャガー」や「ディスプレイ」，逆に生物としての「マッコウクジラ」，人物としての「聖徳太子」のような親概念が固定できるもの，話題が広がりやすいと考えられる「アルゴリズム」や「電話」,「コーヒー」，逆に広がりにくいと考えられる「鉛筆」と都市としての「京都」,「大阪」である．

グラフをみると，KTPF を用いた場合のほうが全体的に精度が低くなっていることから，本研究の期待とは逆に KTPF を利用することによって話題語の抽出精度を低下させてしまうことがわかった．

これは，話題語の適切性の判断基準に曖昧性が残っていることが原因と考えられる．本研究では，話題語はあるキーワードで Web 検索をする際にユーザが暗黙的に考慮している対象範囲だと定義した．しかし，実際に列挙された話題語候補のなかから本当に話題語としてふさわしいものを厳密に区別するのは難しく，今回の実験でも被験者 7 人で意見が分かれるものが多数存在した．たとえば，マッコウクジラの「胃」や「腸」，聖徳太子の「没年」などは「の」による接続に違和感は存在しないものの，話題としての広がりが乏しいという意見があり，このようなものを話題語として認めるかどうかについては各個人によって判断が分かれた．このように，日本語の「の」は非常に広い範囲で利用される語であるため，候補の評価を行うためには話題語のより厳密な定義が必要となることがわかった．

また，今回の話題語の適切性判断の際には，KTPF で順位を低下させている「情報」や「ホームページ」などの語も適合／不適合の区別では適合と判断されるため，精度の計算において逆効果をもたらしていると考えられる．直観的な満足度に比して評価結果が低くなってしまったことから，話題語の評価には適合／不適合以外の評価基準を用いる必要があるのかもしれない．

7.　まとめと今後の課題

本研究では，Web 検索の結果を質問キーワードに関連の深い話題語を用いてクラスタリングすることを最終目標とし，本稿ではそのために用いる話題語を発見する手法について日本語の

連体助詞による修飾関係を用いることを提案し，検証した．

実験の結果から，本稿で提案する話題語抽出手法によって概ね満足できる話題語が発見できていることがわかった．しかし，その一方で KTPF を用いたランキング手法が話題語の抽出精度の向上に役立つとはいえないということもわかった．更に，本稿で提案した話題語の定義には曖昧な部分が多分に残っており，人手による評価を行うにはより厳密に定義を固定する必要があることがわかった．

今後の課題として，話題語検出の際の単語検出精度の向上と話題語の定義の厳密化を進めるとともに，最終目標であるクラスタリング型検索エンジンの実装へ向けた諸手法について検討を行っていく．

謝　　辞

本研究の一部は，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己），および，平成 17 年度科研費特定領域研究 (2) 「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己），および，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Clusty
http://www.clusty.com/

[2] How the Vivisimo Clustering Engine Works
http://vivisimo.com/docs/howitworks.pdf

[3] KartOO
http://www.kartoo.com/

[4] 小山聡, 田中克己: “文常構造を利用した web からの話題発見”, 電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003), 2003.

[5] 山本英子, 梅村恭司: “辞書を用いない関連語リストの構築方法”, 情報処理学会研究報告-自然言語処理, 2002-NL-148(12), pp. 81-88, 2002.

[6] 佐藤吉秀, 川島晴美, 佐々木努, 大久保雅且: “文書の類似度と新鮮度に基づく話題語抽出”, 情報処理学会研究報告-自然言語処理, 2005-NL-165(5), pp. 29-35, 2005.

[7] Google Suggest
http://www.google.com/webhp?complete=1&hl=ja/

[8] D. Butler: “Never trust a human”, Nature, nol.405, p.115, 2000.

[9] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz: “Analysis of a very large altavista query log”, SRC Technical Note 1998-014, DEC Systems Research Center, 1998.

[10] 美野秀弥, 橋本泰一, 徳永健伸, 田中穂積: “日本語の連体修飾関係に関する研究”, 言語処理学会第 10 年次大会発表論文集, 2004.

[11] 徳永健伸: “情報検索と言語処理 (言語と計算 5)”, 東京大学出版会 (1999).

[12] Google
http://www.google.co.jp/

[13] Mecab
http://chasen.org/~taku/software/mecab/