主題語からの話題語自動抽出とこれに基づく Web 情報検索

Web Information Retrieval by Extraction of Topic Terms from Subject Terms

野田 武史♥ 大島 裕明♦
小山 聡♠ 田島 敬史♠ 田中 克己♠

Takeshi NODA Hiroaki OHSHIMA
Satoshi OYAMA Keishi TAJIMA
Katsumi TANAKA

われわれが検索の対象とする語には，それについて関連のある話題を表すような別の語を考えることができる．たとえば，京都という語について考えた場合，その話題として「観光」や「グルメ」，「写真」，「ホテル」などが考えられる．本研究では最初に対象とした語を主題語，それに関連する話題を表す語を話題語と呼び，Web 上の情報を利用して主題語に関連する話題語を自動抽出するとともに，その話題に関する情報を提供している Web ページを検索する手法について考察する．

When we search the Web, we choose some keywords to verbalize it. Generally, for every keyword we can think of other keywords which represent a "typical topic" of them. For example, if we want to know something about Kyoto and choose the word "Kyoto" as initial keyword, then we can think of other keywords like "sightseeing", "gourmet", "photo" or "hotels" which describe some specific aspects of Kyoto as a "typical topic".

In this research we call the initial keywords "theme terms" and latter associated keywords "topic terms"' We discussed how we could find topic terms for a given theme term from the Web and tried to find appropriate Web pages that have the theme term and topic terms.

1. はじめに

現在広く利用されている検索エンジンにおいては，ユーザが検索を行うためのキーワードを自分で入力することが前提とされている．しかし，ユーザにとって自らの検索対象を過不足なく適切にキーワード化することは容易ではなく，自由に入力可能なキーワードによる検索には限界があるといえる．このような問題に対処するためには，ユーザが入力した少しの手がかりを用いてユーザが求める情報を予測し，絞り込みに用いるべき適切な語を提示することができればよい．このような試みとしてGoogle Suggest[1]があるが，これは単に候補のリストを表示するものにすぎず，候補の語との意味的つながりを知ることができない．ユーザが入力するキーワードをある「概念」ととらえれば，それに関連する上位/下位概念，兄弟概念など，入力したキーワードとそれを補うべきキーワードとの間にはいくつかの異なる関連性のありかたが存在すると考えられ，これを意識した提示のしかたが行われるべきである．

ディレクトリ型検索エンジンでは，概念の上下関係に基づいてWebページが分類されているため，ユーザは比較的容易に求める情報へ到達できるという特徴がある．このような特徴を，ロボット型の検索エンジンにおいても実現できないだろうか．すなわち，入力したキーワードと関連のあるキーワードの候補をその関連のありかたごとに分類して提示することで，ユーザの絞り込み作業を支援することができないかと考える．提示された候補を選択することによって適切な絞り込み検索が行われれば，ユーザは求める情報のキーワード化という繁雑な作業から開放される．

本研究では，検索に用いられるキーワードに対してその意味内容をより詳細化し分化させるような別のキーワードを「話題語」と呼び，分化の元となったキーワードを「主題語」と呼ぶ．本稿ではある主題語に対する話題語を発見し，それに適合するWebページを検索する手法について考察した．本研究で目標とするシステムのイメージを図1に示す．このインターフェースでは画面上部に話題語のリストを表示し，選択された話題語に沿って画面下部に適合するWebページのリンクを表示する．話題語の右のリストは後ほど説明する親概念および兄弟語である．

2. 諸概念

2.1 話題語

一般的に，あるキーワードkについて考えたとき，kのある一部分をより詳細に取り上げるような意味をもつ別の語k'が存在する．たとえば，「京都」というキーワードを考えると，この語が意味する内容は非常に大きく漠然としており，含まれている情報，すなわちWebページの数も多い．ここで，「観光」という語を新たに考えると，京都と観光それぞれがもつ情報の共通部分，すなわち「京都に関する情報で，かつ観光に関するもの」が得られる．これは，「観光」という語によって「京都」という語がもつ情報を詳細化し，絞り込んだと考えることができる．このような，あるキーワードkに付随して考えることのできる語k'を本研究ではkの話題語

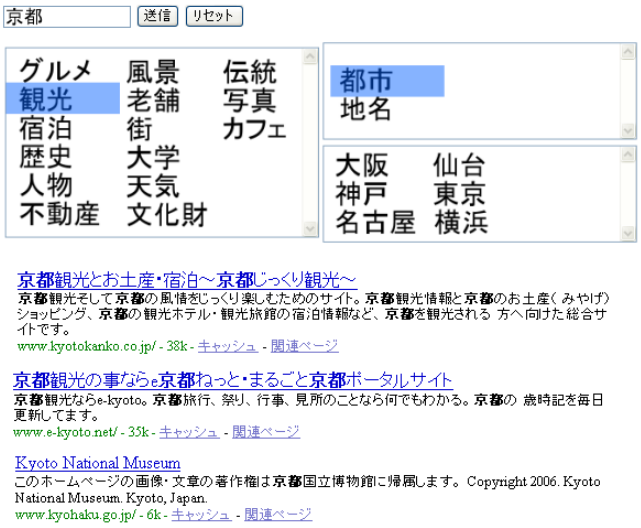

図 1. システムイメージ
Figure 1. System Image

♥ 学生会員 京都大学大学院情報学研究科博士前期課程
noda@dl.kuis.kyoto-u.ac.jp
♦ 学生会員 京都大学大学院情報学研究科博士後期課程
ohshima@dl.kuis.kyoto-u.ac.jp
♠ 正会員 京都大学大学院情報学研究科
{oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

と呼ぶ．話題語は元のキーワードと関連のある語であるが，元のキーワードと関連のある語全てが話題語であるわけではない．たとえば「京都」に対する「大阪」は，互いに県庁所在地名，都市名であるという関連性をもつものの，大阪によって絞り込まれる京都の情報が意味的に明確ではないため，話題語ではない．

2.2 主題語

あるキーワードkに対してその話題語k'を考えたとき，逆に話題語k'に対して元のキーワードkをk'の主題語と呼ぶ．主題語は，話題語を考えることのできるキーワードであれば，どのようなものでもなることができる．

2.3 親概念

多くの語は，字面上は全く同じ語であっても，それが使用される文脈によって異なる意味をもつことがある．たとえば「Ruby」という語を例にとれば，これは宝石の 1 種類として用いられる場合もあれば，組版の文字を指す場合も，プログラミング言語の一つを意味する場合もある．さらに，どの意味で用いられるかにより，その「Ruby」にふさわしい話題語も変化する．「硬度」という語は言語としての「Ruby」の話題としては適切ではない．このように，ある語が与えられただけではその語に対する話題語を求めることが難しいため，本研究ではある語 k が実際に「何であるか」を表す語 k”を，k の親概念と呼び，必要に応じてこれを使用することで話題語の発見を補助することを考えている．現在のところ，親概念を取得する適当な手段が得られていないため，実際には利用していない．

2.4 兄弟語

Ruby に対する Perl や Python など，プログラミング言語という親概念を共有するキーワードどうしは互いに対等な関係にあると考えられ，互いに兄弟語関係にあるとする．1 つのキーワードからだけでは親概念が決定しにくい場合に，複数の兄弟語を知ることができればそれらに共通な関連語として親概念を検索することなどが可能であると考えられるが，これも現在のところ実際には使用していない．以上の諸概念について，Ruby を主題語として図 2 に例示した．

2.5 話題語の性質

ある主題語 p の話題となる語 t 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」といったように，「p の t」という形で用いることができる．このことに着目すると，「p の t」というパターンを Web ページから抽出することで，与えられた主題語に対する話題語を発見することができる．また，このような表現が含まれるページは，「p の t」という話題に言及しているということも期待できる．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[2]．

3. 主題語からの話題語の抽出

本研究では，ユーザが入力した主題語からその話題語を発見し，各話題について言及している Web ページをその話題語の下に分類して表示することを目標としている．まず，ユーザが入力した主題語からその話題語を発見する手法について説明する．以下のような手順で抽出を試みた．

3.1 話題語の抽出

主題語を p，話題語を t とすると，「p 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，「”p 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

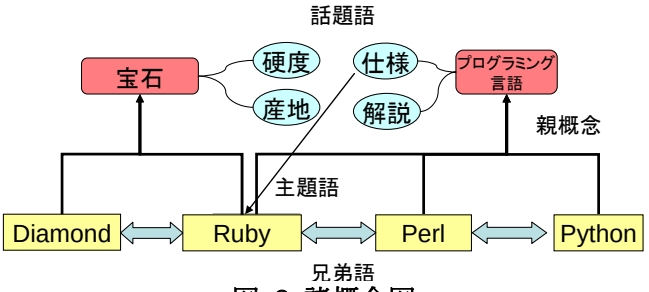

図 2 諸概念図
Figure 2. Concepts

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」以降に続く名詞を抽出する．このようして抽出された名詞群を，主題語 p に対する話題語の候補群 Tt^p とする．

ここで，「京都の観光の動向」や「京都の嵐山の紅葉」といったように「の」が連続して使用される場合については，「の」を介してつなげられる語の全てを話題語の候補とした．日本語の文法としては，「の」で接続される語の順序関係は可換ではなく，意味上の上下関係が含まれていることは明らかであるが，本研究において各話題語はたんに主題語に関わる情報を絞り込んでいく機能をもつ語であると定義しているので，簡単化のため話題語間での修飾関係は考慮しないこととした．

3.2 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主題語においても話題となり得る一般性の高い語が含まれている．このような語は，その主題語に特徴的な話題語に比して重要性が低いと考えられるため，話題語の候補を何らかの手法で評価し，評価値の高いものから順に表示することが必要となる．このために利用する尺度として，本稿ではある主題語 p に対する話題語 t の重要度の評価を以下のように考えた．

主題語 p に対する話題語 t の重要度は，主題語側からみた話題語の重要度，すなわちその話題語 t が主題語 p に関する全ての話題のうちでどれだけの割合を占めるかということと，逆に話題語側からみた主題語の重要度，すなわちその話題語 t が登場する場合にその主題語が p である割合との積で表される．

まず，主題語側からみた話題語の重要度 $imp_p(p,t)$ と話題語側からみた主題語の重要度 $imp_t(p,t)$ を以下のように定義する．

$$imp_p(p,t) = \frac{df("p\text{の}t")}{df("p\text{の}")} \qquad imp_t(p,t) = \frac{df("p\text{の}t")}{df("\text{の}t")}$$

すると，これらの値を用いて主題語 p に対する話題語 t の総合重要度 $imp(p,t)$ は以下のように定義できる．

$$imp(p,t) = imp_p(p,t) \cdot imp_t(p,t)$$

この評価値を用いて，各話題語候補 $t \in T$ についてその評価値が高かった話題語候補の上位 10 件を p に対する話題語 T^p とする．

4. 話題語に適合する Web ページの取得

次に,各話題語$t \in T^p$に適合するWebページを取得する．主題語 pおよび話題語 tが与えられたとき，これに適合するWeb ページは以下のようなものであるとする．すなわち，そのページが pに関わる内容について言及しており，かつ tに関わる内容についても言及していることである．ここで重要な条件として，pや tがページ内に明示的に現れなくともよいものとする．これは，Web ページの適合判定があくまでそのページの意味内容であるということを考えれば自然な条件である．たとえば，京都の観光について言及しているページを考えると，それらのページ全てに必ずしも「京都」や「観光」という語が含まれているとは限らない．祇園祭りや嵐山の紅葉について詳細に述べたページは「京都」と「観光」の両方に関連すると考えられるものの，それらのページの記述において「京都」や「観光」という語が必須であるとは限らないためである．

主題語や話題語が文字列として含まれないページも適合させるためには，主題語や話題語を用いて検索を行うだけでは不十分である．そこで，本研究では各ページの特徴ベクトルを利用して適合ページを検索する手法をとった．具体的には，以下の手順に従って行った．なお，今回は各ページの特徴ベクトルを作成するにあたって，そのページ自体を取得するのではなく,Google による検索結果中のページサマリを用いて行った．

4.1 Step1: 見本 Web ページの取得

まず，ある主題語と話題語に適合させる Web ページの見本となるものを取得する．今回は主題語 pから話題語 tを抽出する際に検索された，「p の t」という表現が含まれる Web ページ群ををを見本ページ P_t^p として利用する．

4.2 Step2: 特徴ベクトルの作成

次に，見本 Web ページ P_t^p を形態素解析し，それに含まれるそれぞれの語の出現回数を用いて特徴ベクトルを作成し，それを主題語 p, 話題語 tに適合するページの特徴ベクトル $vec(p,t)$とする．

4.3 Step3: 適合ページの取得

主題語のみをクエリとして新たに Web 検索を行い，得られた Web ページ群の各 Web ページの特徴ベクトルと $vec(p,t)$ とを比較し，類似度の高いもの上位 5 件を適合ページとして表示する．

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では，Web ページ検索エンジンとして Google を用いた．Google の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った結果を用いて話題語の候補を求め，該当ページ件数を df として用いることで各候補の重要度を計算した．

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自然言語処理学講座の開発する ChaSen[3]を用いた．

6. 実験とその結果／考察

6.1 実験内容

実験では前章の実装を用いていくつかの具体的な主題語に対して話題語及び適合 Web ページの検索を行った．

6.2 話題語の抽出

まず，主題語を「京都」として実際に検索を行った．Google での取得ページ数を 500 件として検索した結果の一部を表 1 に示す．

この表は，抽出された話題語の候補とその Google 検索結果 500 件中での総出現回数，*df("京都の t")*, *df("の t")*, 京都からみた各候補の重要度，各候補からみた京都の重要度およびそれらの積である総合重要度と，各候補を総合重要度に従って並べた順位を示している．この表は結果のごく一部であるが，いくつかの特徴的な語を含んでいる．たとえば，上位にランキングされた「伏見」や「清水寺」は，地理的に京都の一部分を表す固有名詞である．京都の地名を表す固有名詞が上位にランキングされているのは，これらの語からみた京都の重要度が大きいからであり，京都という語との関連の深さからみても，高い重要度を得ていることは妥当であると考えられる．「人」，「桜」，「老舗」などはこの順で京都の側からみた重要度が大きいものであるが，最終的な重要度は「人」よりも「老舗」のほうが上位となっており，この点が興味深い．古い歴史をもつ都市である京都にとって，「人」という一般性の高い語よりも，「老舗」のほうが重視されるという結果は好ましいものである．このほか，「情報」や「ホームページ」といった一般性の高い語は概して京都からみた重要度が低いため，順位が低く抑えられており，この点も直観的な評価と符合している．

このようにしてみると，主題語「京都」対する話題語に関しては概ね良好な結果が得られているといえる．改めて上位 20 件を表示すると，以下のようになる．

「京都」の話題(上位 20 件) = 和菓子,伏見,清水寺,町家,町屋,祇園,紅葉,寺社,老舗,四季,観光,芸妓,桜,町,きもの,文化財,花街,伝統,街,歳時記

特に問題があると思われるような語は含まれていない．実際には「もの」や「こと」といった話題語としてふさわしくない語も抽出されていたが，重要度の計算によってこれらの語は排除されていた．

次に，主題語を「大阪」，「iPod」として同様の実験を行った.これらの上位20件の結果は以下に示すとおりである．

「大阪」の話題(上位 20 件) = 濃厚,おば,梅田,賃貸,下町,貸家,高槻,ホテル,風俗,街,御堂筋,探偵,税理士,天気,夜景,味,岩盤,中心地,分譲,人

「iPod」の話題(上位 20 件) = 充電,ＣＭ,小道具,バッテリー,車載,バッテリ,付属,液晶,リモコン,音量,音楽,動画,ケース,操作,再生,すべて,曲,箱,進化,成功

この内容をみると，大阪の場合はやはり地名が多く含まれる傾向があることがわかる．大阪の「おば」は Chasen が「おばちゃん」を認識できなかったことによるものと思われる他，iPod の「バッテリー」と「バッテリ」の表記ゆれなどの問題がみられるものの，ほぼ話題語として問題ない語が抽出できているといえる．以上から，主題語と話題語の双方か

表 1 京都の話題語(抜粋, df(“京都の)= 3460000 Table1. Topic terms of Kyoto(extract)

話題語の候補 t	出現回数	df(″京都の t″)	df(″の t″)	$imp_{京都}$(京都 ,t)	imp_{t}(京都 ,t)	imp(京都 ,t)	順位(/162)
和菓子	1	137000	1110000	0.123423423	0.020726173	0.002558095	1
伏見	2	47000	155000	0.303225807	0.007110439	0.002156069	2
清水寺	1	36000	99600	0.361445783	0.005446294	0.00196854	3
紅葉	6	138000	2640000	0.052272727	0.020877458	0.001091322	7
寺社	1	24500	172000	0.142441861	0.003706505	0.000527962	8
老舗	5	149000	7070000	0.021074965	0.022541604	0.000475064	9
観光	29	115000	5250000	0.021904762	0.017397882	0.000381097	11
桜	4	149000	10600000	0.014056604	0.022541604	0.000316858	13
人	2	183000	200000000	0.000915	0.027685325	0.000025332	48
情報	5	53200	273000000	0.000194872	0.008048412	0.000001568	89
ホームページ	8	18500	84600000	0.000218676	0.00279879	0.000000612	100

らの重要度を用いた今回の評価手法は，話題語の発見において有効であったと考えられる．

6.3 適合ページの検索

次に，主題語「京都」，話題語を「観光」としてこの話題に適合する Web ページの検索を行った．

これらのページを実際に閲覧してみると，その内容は確かに京都の観光に関わる情報を多く含んでいた．本来は「観光」という語が含まれないページも含まれるはずであるが，今回は全てのページに「観光」という語が含まれていた．これは，上位 5 件のみを選んだことも関係していると考えられる．

7. まとめと今後の課題

本稿では，主題語と話題語という概念を用い，ユーザに絞り込み候補となる話題語およびそれに適合する Web ページを提示できるようなシステムの構築を目指し，主題語からの話題語の抽出と，適合ページの検索に関する実験を行った．

実験の結果，話題語の抽出においては概ね良好な結果が得られたが，これらの評価は個人的なものでしかなく，全ての人々にとってこれらの結果が受け入れられるものであるか否かについては定かではない．Web ページの検索では多くの話題語において同じページが上位に表示されてしまい，話題の相違による Web ページの選択が十分に行えていなかった．

今後はページの適切性を定量的に評価できるような手法について今後考察していくべきであると考えられる．また，これ以外にも今後の課題として，さらに話題語の定義を明確化し，あらゆる主題語に対して話題語としてふさわしくない語のリストを作成するなどして話題語の誤抽出の確率を低下させること，適合ページの検索手法について見直しを進めていくこと，概念の設定のみに止まっている親概念および兄弟語の有効活用などが挙げられる．

[謝辞]

本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14～18 年度），文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，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（研究代表者：田中克己），および文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関 する研究」（研究代表者：田中克己，A01-00-02，課題番号：18049041）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] Google Suggest
http://www.google.com/webhp?complete=1&hl=ja/
[2] 美野秀弥，橘本泰一，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.
[3] Chasen
http://chasen.naist.jp/hiki/ChaSen/

野田 武史 Takeshi NODA
京都大学大学院情報学研究科博士前期課程在学中．2005 年京都大学総合人間学部自然環境学科卒業．日本データベース学会学生会員．

大島 裕明 Hiroaki OHSHIMA
京都大学大学院情報学研究科博士後期課程在学中．2004年神戸大学大学院自然科学研究科博士前期課程修了．Web環境におけるパーソナライゼーションの研究に従事．情報処理学会，日本データベース学会，ACM各学生会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002年京都大学大学院情報学研究科博士後期課程修了．博士(情報学)．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI各会員．

田島 敬史 Keishi TAJIMA
京都大学大学院情報学研究科社会情報学専攻助教授．1996年東京大学理学系研究科情報科学専攻博士課程修了．博士(理学)．主にデータベースシステム，Web検索の研究に従事．ACM，情報処理学会，日本データベース学会等各会員．

田中克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．工学博士．主にデータベース，マルチメディアコンテンツの処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．