クエリトピックに対する一般的知識範囲の推定

小紫　弘貴†　　田島　敬史††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †komurasaki@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　Web 上には膨大な情報が存在する．そのため，あるトピックについて調べる際に，そのトピックに関する Web 上の全情報ではなく，一般的な知識として知られている範囲内の情報のみを取得したいという場合がある．そこで，本研究では，与えられたトピックに関する Web 上の情報のうち，一般的な知識にあたるものと，そうでないものとを判別する手法を提案する．本研究で提案する手法では，まず，与えられたトピックに関する Web ページ群からそのトピックを構成する要素知識の集合を抽出し，各要素知識がどの程度一般的なものであるかを判定する．次に，各 Web ページがどの程度の一般性の情報を提供するためのものであるかを，そのページ中の要素知識の一般性に基づいて判定する．全ての関連 Web ページについて，このようにして一般性を判定し，その結果から，どの程度の一般性のページがどの程度の数存在するかの分布を取得する．そして，この分布から一般的知識の範囲を推定する．また，このようにして得られた一般的知識の範囲の情報を過不足無く含むような Web ページ集合を求めて，これを結果としてユーザに返す手法についても提案する．本稿では実験によりこれらの提案手法の有用性を検証する．

キーワード　情報集出，知識抽出，トピック抽出，情報詳細度，情報要約

1. はじめに

近年，インターネットの普及に伴い，膨大な情報の中からユーザーにとって必要な情報を取捨選択する機会が増大しており，Web 閲覧の際の情報選択を支援する技術の重要性が高まっている．

あるトピックについての情報を Web 検索を用いて収集しようとする時，それぞれのトピックごと，またユーザーごとに必要とする情報の量には差があると考えられる．例えば専門家になるのであれば Web 上にあるそのトピックに関連する情報はほとんど全て必要となるであろうが，そうでないユーザーがそのトピックに興味を持ち概要・要点だけを知りたいような場合には，そのトピックに関する情報の中でも一般的，常識的とされる情報だけを取得することで十分であろう．しかし，調べたいトピックに関連する大量の Web ページ群からそういった情報だけを人手で選別するのは大変な手間がかかり，また，そもそも，専門家でないユーザにとっては，どの情報が一般的，常識的とされる情報であり，どの情報がそうでないかの判断が困難である場合も考えられる．

例えば，あるユーザが「徳川家康」について調べたいとする．この時,「江戸幕府」や「関ヶ原の戦い」については広く一般的に知られていることであると考えられるが，それに対して正室が「築山殿」という人物であったことや元服してから「次郎三郎元信」と名乗っていたこと等は一般的ではないと考えられる．よって,「徳川家康」について知らないあるユーザーが,「徳川家康」について一般的に知られている程度のことを知りたいという場合であれば，上に挙げた四つの情報の範囲内で考えれば,「江戸幕府」と「関ヶ原の戦い」について書かれた Web ページを見るだけで十分であると考えられる．しかし,「徳川家康」という語で Web 検索を行った場合，大量のページが検索結果として返されるので，これらのページ全てに目を通すのは大変な手間がかかる．実際には，広く一般的に知られていることのみを知りたいのであれば，そのうちの上位のある件数のみを見れば十分であると考えられるが，上位何件まで見れば十分なのかを判断するのは容易ではない．また,「徳川家康」について詳しくないユーザの場合,「関ヶ原の戦い」は有名であるが,「築山殿」は有名ではないという判断がつかない場合も考えられる．

よって，Web 上に存在する膨大な知識の中から，一般的，常識的であると考えられる知識と，一般的とはいえない専門的な知識とを自動で判別することが出来れば，調べたいトピックについて一般的に知られていること，言い換えれば一般的なユーザでも知っておくべきことのみを無駄なく取得でき，Web 検索による情報収集の作業の効率化につながると考えられる．そこで，本研究では，一般的な Web ユーザーが知っていると推測される範囲，これを一般的知識範囲と呼び，調べたいトピックに関連する Web ページ集合から，そのトピックについての一般的知識範囲を推定し，その範囲内に存在する知識を過不足無く含む程度の Web ページ集合を結果として返す手法を提案する．

本研究において，あるトピックの持つ情報を構成する要素を要素知識と呼ぶことにする．ここで，ある要素知識が「一般的」であるとは，その要素知識が多くのユーザに知られているということとして定義される．しかし，ある要素知識がどの程度の数のユーザに知られているかを直接知ることは困難である．そこで，本研究では，ある要素知識が「一般的」であるかどうかを，その要素知識が，Web 上の多くのページで記述されているかどうかで近似することを考える．これは，多くの Web ページで記述されている要素知識は，多くのユーザの目に触れ，多くのユーザに知られているはずであり，一方，多くのユーザに

知られている要素知識は，知っているユーザ数に比例して，より多くのユーザによって Web ページに記述されるはずであるという仮定に基づいている．

また，ある要素知識が「一般的」であるか「専門的」であるかは，二値的に決定されるものではなく，ある要素知識の「一般性」は 0 から 1 までの連続的な値をとり，その値が 0 に近いほど「専門的」，1 に近いほど「一般的」であるとする．

なお，本研究の最終目標としては，Web 上のみならず社会における一般的知識範囲の推定を目指しているが，本論文では Web 上での情報に焦点を当て,「Web 上での，Web ユーザにとっての，一般的な知識範囲」の推定を目標とする．

本研究で提案する手法は以下の四つのステップからなる．

まず，与えられたクエリトピックについての情報を含む Web ページを収集し，この Web ページ群から，そのクエリトピックが持つ情報を構成する要素知識の集合を抽出する．本研究では，これらの各要素知識は対応する語によって表現できると仮定する．例えば，先に挙げた「徳川家康」の例で言えば,「江戸幕府」や「築山殿」などが，要素知識に対応する．

次に，各要素知識について，それがどの程度一般的なものであるかを，その知識を表す語の，関連 Web ページ集合中における DF 値から求める．要素知識を表す語の DF 値が高い，言い換えれば関連 Web ページ集合内に多く出現する要素知識は一般的であると推定できる．つまり一般的である語は出現頻度が高く，クエリトピックとの関連度も高いと考えられる．その反対に，関連 Web ページ集合内の少ないページにしか出現せず，DF 値が低ければその要素知識は専門的であるといえる．同様にクエリトピックとの関連度も，一般的な要素知識と比べて低くなると考えられる．

また，各要素知識の一般性の推定結果をもとに，前記の Web ページ群の各 Web ページについても，そのページが，どの程度の一般性の情報を提供することを意図して記述されたページであるかの推定を行う．

ここで，要素知識がどの程度一般的であるかを判定を行ったとしても，その要素知識を説明する際に一般的でない要素知識を含んでいる文章は目的の達成に不適であると言える．つまり，一般的な情報を一般的なレベルで説明している文章を抽出する必要がある．そのためにそれぞれの要素知識の一般性だけでなく，それぞれの Web ページが持つ一般性を判定することによって，一般的な要素知識をそれに見合ったレベルで記述しているページを発見することが出来ると考えられる．

続いて，三つ目のステップとして，前記の各 Web ページに対する推定結果をもとに，どの程度の一般性の Web ページが，どの程度のページ数，Web 上に存在するかの分布を求める．図 1. に，そのような分布をグラフで表した例を示す．このような分布においては，一般性の高い情報，例えば「徳川家康」の例で言うと,「江戸幕府」のような情報について書いている Web ページは数多く存在するが，より一般性の低い情報，例えば「築山殿」のような情報について書いている Web ページは数が少なくなる傾向があると予想される．そして，この分布中で，情報の一般性がある点より下がると，急にそのような情報を含む Web ページ数が減少するような点を発見し，そのような点を一般的知識範囲の境界であると推定する．

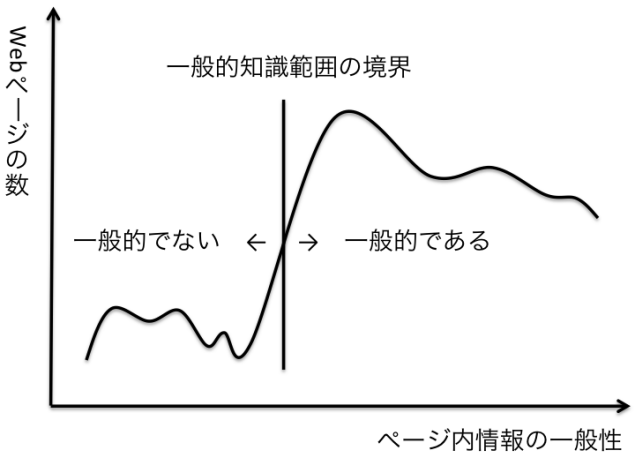

図 1　Web ページの持つ情報の一般性による分布

最後に，それぞれの Web ページ内にどの要素知識が存在しているかということを利用して，求まった一般的知識範囲内の一般性を持つ要素知識を網羅するような，なるべく小さな Web ページ集合を各ページの持つ情報の類似性と差異を利用して抽出する．

本稿では，上記の四つのステップのうちの一つ目のステップであるトピックを構成する要素知識の抽出の部分について，具体的な手法を提案し，また，実験により提案手法の妥当性を検証する．

2. 関 連 研 究

本研究では，Web ページ集合から一般的な要素知識のみを抽出することを目的としているが，これに近い既存技術として，文書群の自動要約の技術が挙げられる．自動要約に関しては既に多くの既存研究がある．例えば，桜井らはユーザーが求める情報であり，かつ，文書に特徴的である情報を含む複数文書の要約を作成するために，クエリとの共起関係と関連度の高さを用い，これらに加え，他の文書との違いを明確にするために，関連度の低さを利用する手法を提案した [1]．クエリとの関連度の高さだけでなく低さも利用するという点で，本研究が要素知識の一般性の判定にクエリとの関連度の高さと低さを利用する点と関連がある．

また，複数文書からの情報要約については，既存研究として Carbonell らによる研究 [2] と Goldstein らの研究 [3] がある．Carbonell ら [2] は Maximum Marginal Relevance（MMR）によって発見されたクエリの多様性を用いた文章の選択による，単一，あるいは複数文書の要約について研究を行った．また，複数文書に絞った自動要約の研究として，Goldstein ら [3] は，単一文書の自動要約に比べて要約の速さ，冗長性の排除，文章の選択方法が重要であると考え，ドメイン非依存の速度のある統計処理で冗長性を減らしつつ多様性を最大化するように文章を選択する技術を提案した．

文書の自動要約によってユーザーが情報を取捨選択する労力

を省くことが可能だが，自動要約における情報の取捨選択においては，その情報が，自動要約の対象となる文書または文書群の持つトピックに対して重要であるかどうかが基準であるという点で，一般的知識範囲を基準として情報の取捨選択を行う本研究とは大きな違いがある．

ここで，重要性と一般性の違いについてであるが，クエリトピックについて重要な情報とはクエリトピックに関連する情報の中でも，クエリトピックについて知る上で欠かせない情報と言い換えることが可能な情報であるが，誰でも知っている情報であるとは限らない情報である．一般的な情報とは多くの人は知っているが，だからと言ってクエリトピックに対して欠かせない情報ではなく，取るに足らない情報である可能性もある．例えば,「徳川家康」について重要な情報として「関ヶ原の戦いにおける小早川秀秋の裏切り」が挙げられるが一般的な情報であるとは言えない．反対の例として徳川家康は「戦国大名であった」という情報は一般的であるが重要であるとは言えない．

本研究では，クエリトピックに関する要素知識の集合を，クエリトピックに関連する Web ページ群から抽出する．このような，Web ページ集合からのトピック抽出については，多くの関連研究がある．例えば，Zeng ら [4] や Khoo ら [5] の研究が挙げられる．Zeng らは，トピックを階層構造で考え，単語の文書頻度を混合ガウス分布で表し，その分布から EM アルゴリズムを用いて階層的なトピックを抽出する手法を提案した．これは単語の文書頻度の分布からトピック抽出を行うという点で本研究と関係があるが，抽出する対象がトピックであり，情報をトピックとしてまとめて扱っている点で，本研究が情報を語単位とページ単位とで扱う点と異なる．また Khoo らは異なる手法として TF*PDF アルゴリズムを用いてニュース記事から話題となったトピックの抽出を行った．複数文書からトピックを抽出するという点で，本研究の要素知識集合の抽出と関連がある．しかし抽出に TF*PDF アルゴリズムを用いるという点で，一般性による抽出を行う本研究と異なる．

本研究で，それぞれの語がどの程度一般的なものであるかを判定する際に，関連 Web ページ中において各語が出現するページ数を利用するという点の関連研究として，山本ら [6] の研究が挙げられる．山本らは語やフレーズの信憑性の判定に Web 上でどの程度それらが出現しているかを利用している．

また，本論文の提案手法においては，一般的知識範囲内を推定した後，そのような要素知識のみを含むような Web ページ集合を発見してくる技術も必要となる．Dai らはクエリトピックに関連する文書群から，(1) 新規性に基づいた方法，(2) 文書のクラスタリングを利用した方法，(3) サブトピック抽出を利用した方法の 3 種類のランキングアルゴリズムを用いて，サブトピックとなる要素を含み，かつ，重複が最小であるような文書群を取得する方法を提案し，またそれらを比較する研究を行った [7]．文書群から必要な文書を抽出するという点で本研究で関係があるが，本研究では一般的知識範囲内に存在する要素知識を内包するような文書を必要としており，Dai らはクエリトピックに関連する文書群全体からサブトピックとなる要素を含む文書を取得するという点で異なる．

3. クエリトピックに関する要素知識語の抽出

本章では，ユーザーによって与えられたクエリトピックを構成する要素知識に対応する要素知識語の集合を関連 Web ページ群から抽出する手法について述べる．

3.1 要 素 知 識

本研究では，あるトピックの持つ情報を構成する要素を要素知識と呼ぶことにした．他にトピックを構成する要素の呼称としては「サブトピック」という言葉が挙げられるが，ある主トピックについてのサブトピックとは，その主トピックを構成する情報を階層的に分類した際に現れる，主トピックの部分集合となるようなトピックのことである．

一方，本研究で言う要素知識を表現する語は，主トピックを構成する情報を階層的に分類した場合に，必ずしも主トピックの下位に位置づけられるような語であるとは限らない．そこで，本研究では,「サブトピック」という語は用いず,「要素知識」という語を用いる．

例えば,「徳川家康」というトピックを構成する要素知識としては「江戸幕府を創設した」という情報が挙げられる．しかし，この知識を表す「江戸幕府」という語自体は，トピックの階層構造において，必ずしも「徳川家康」というトピックより下位に位置づけられる情報であるとは言えないため,「徳川家康」のサブトピックと呼ぶのは適切ではない．

本研究の目標は，与えられたトピックに対して，一般の人が知っているであろう情報を求めることであるので,「徳川家康」というトピックが与えられた時に,「江戸幕府」という情報が，知っているべき情報に含まれないということは考えられない．しかし，上述のように,「江戸幕府」は必ずしも「徳川家康」のサブトピックではない．このように，われわれの目的においては，与えられたトピックの，いわゆるサブトピックのみを考えるのでは，不十分である．

また，要素知識は文章で表されるものであり，無限に存在してしまうため，対応する語を「要素知識語」とし，情報を語単位で扱うことする．

トピック「徳川家康」についての,「要素知識語」の例を図 2 に示す．

3.2 要素知識語が持つ性質

ここでは，先ほど定義した「要素知識語」がどのような性質を持っているかについて述べる．要素知識語とは，あるクエリトピックに関する情報を構成する要素であるから，クエリトピックとの関連が深い語で表されるようなものであると考えられる．また，クエリトピックにとって重要な語は要素知識語になるといえる．なぜなら，重要な語とはクエリトピックを知る上で欠かせない情報と言い換えることが可能であり，それはクエリトピックを構成する情報，つまり要素知識語であると考えられるからである．

この分析に基づき，本研究では，このような要素知識語を抽出するために，各語に関する以下の指標を用いる．

（1） クエリトピック語との共起率が高い．

（2） 関連 Web ページ集合内における出現回数が多い．

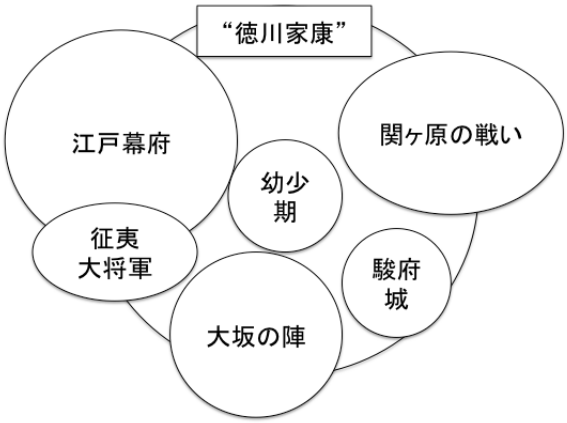

図 2　トピックに関する要素知識語の例

（3） 各 Web ページ内においてページの上部，或いは全体に渡って出現している．

（4） ページタイトルやリンクテキストに出現している．

（5） 固有名詞である．

まず，(1) のクエリトピック語との共起率について，詳しく説明する．ある語に関連する Web ページ群の中に，クエリトピック語を含む Web ページが多く現れる場合，その語が持つ情報のうち，多くがクエリトピックに関連する情報を表していると推測できる．従って，そのような語はクエリトピックとの関連が強いと言える．また，その逆に，クエリトピックに関連する Web ページ群の中に，ある語を含む Web ページが多く現れる場合，クエリトピックの持つ情報の中で，その語に関連する情報が占める割合が大きいと考えることができる．従って，その要素知識語はクエリトピックについて知ろうとする時に重要であると推測される．

そこで，クエリトピック語を Q とし，Q が出現する Web ページの集合中に出現する各語 t について $CoOccurrence(t)$ を，t が出現する Web ページ集合中の，クエリトピック語 Q が出現するページが占める比率，すなわち，

$$CoOcurrence(t) = \frac{|Doc(Q \wedge t)|}{|Doc(t)|} \tag{1}$$

で与える．同様に，$Ratio_t$ を，クエリトピック語 Q に関連する Web ページ集合中の，t が出現する Web ページが占める比率，すなわち，

$$Ratio(t) = \frac{|Doc(Q \wedge t)|}{|Doc(Q)|} \tag{2}$$

で与える．ただし，ここで，

$Doc(t)$：語 t が出現する Web ページの集合

$Doc(Q)$：語 Q が出現する Web ページの集合

$Doc(Q \wedge t)$：語 Q と語 t がともに出現する Web ページの集合

である．そしてこれら 2 つの値を用いて語 t の Q に対する共起度による要素知識としての価値 $Value(t)$ を，

$$Value(t) = CoOccurence(t) \times Ratio(t) \tag{3}$$

によって与える．

次に (2) の出現回数について述べる．この場合の出現回数とは出現ページ数ではなく，全ページ中で何回その語が出現しているか，ということを表す．前述のように本研究では，要素知識はクエリトピックに関連する Web ページ内に登場する語で表される．このことから，関連 Web ページ集合内での出現回数が多ければ多いほど Web ページとの関連度が強くなり，従ってクエリトピックとの関連も強くなると考えられる．

また，ある語の出現位置が Web ページの上部であったり，ページの全体に渡っている場合，その語はそのページとの関連が強いと推測され，同様にクエリトピックとも関連が強いと言える．これが (3) についてである．それぞれのページ中の全ての語が何番目の語であるかを出現位置と定める．

次に (4) について述べる．クエリトピックに関連する Web ページのタイトルや，そのページへのリンクテキストに登場している語は，その Web ページの内容をよく表している語であると推測できる．このことから，関連 Web ページとの関連が強いそのような語は同様にクエリトピックとも関連が強いと考えられる．これらの語について重みをおく．

最後に (5) についてであるが，固有名詞は一般的な名詞と異なり文章の記述や説明に用いられることはほとんど無い．従って，固有名詞が Web ページ中に出現しているということは，その固有名詞はそれ自体で文章中において意味がある語であると考えられるため，一般的な名詞よりも重要である．固有名詞に対して普通名詞よりも重みをおく．

3.3　要素知識語集合の抽出

要素知識語は，第 3.2 章で述べたような性質を持つ．そこで，本研究では，その性質を利用して関連 Web ページ集合から要素知識語集合を抽出する手法を提案する．

まず，クエリトピック語を一般の Web 検索エンジン(注1)に投入し，その検索結果から上位 100 件の Web ページ集合とそれぞれのページタイトルを取得する．そして取得した 100 件の Web ページ内にある名詞を抽出する．その際に，例えば「徳川家康」などの固有表現や複合名詞が形態素解析器(注2)によって「徳川」と「家康」という 2 つの名詞に分割されてしまうという問題があるので，名詞が連続して現れる場合には，個別の名詞と，それらを連結したものの双方を要素知識の候補とする語として抽出する．よって，偶然名詞が連続しているだけで，意味のある連続ではない場合や,「徳川家康」では出現頻度が低いが「徳川」だけで見た場合には出現頻度が高いというような場合にも，個別の語と連結された語のうち，頻度が低いものは以降の手順において除外され，頻度が高い方のみが使用される．

また，これらの名詞を抽出する際に，各語の Web ページ中での出現位置や Web ページ集合内での出現回数を取得する．出現位置は Web ページ内で何番目の語であるかを取得するものとする．

次に抽出された各名詞について，$CoOccurence(t)$，$Ratio(t)$，

(注1)：Google 検索：http:www.google.co.jp

(注2)：本研究では MeCab を用いた．

$Value(t)$ の値を求める．

上で述べたこれらの値を用いて関連 Web ページ集合内の全ての語についてスコアリングを行い，そのランキング上位の語を要素知識語であるとする．

4. ページが持つ情報の一般性判定

本章では，取得した Web ページが持つ情報がクエリトピックに関してどれだけ一般的であるかどうかを判定する手法について述べる．

4.1 情報の一般性

ここではまず，Web ページに含まれる情報が一般的である，ということがどういう性質を持つのかを説明する．

まず，前述のように，本研究では，あるトピックに関連する要素知識語の中で一般的なものとは，関連する Web ページ集合内において多くのページに出現する要素知識語のことであり，一般的でない，つまり専門的な要素知識語とは関連 Web ページ集合内でもあまり多くの Web ページに出現しない要素知識語のことであると仮定する．

ここで，本研究では，要素知識語の一般性は，その語が関連 Web ページ集合内においてどの程度の数の Web ページ内に出現するか，つまりその語が持つ DF 値によって求められる．従って，要素知識語集合内の各語について DF 値を計算し，その値によってランキングを行うことによってそれぞれの要素知識語の，知識集合内での相対的な一般性を求められる．

「徳川家康」の場合で例を挙げると，徳川家康の正室が「築山殿」であったという情報を含んでいる Web ページは,「徳川家康」で Web 検索して得られた Web ページ内には 0 では無いがあまり得られず,「江戸幕府」についての情報を含むページよりは少ない．したがって「築山殿」という語が持つ DF 値は低く,「徳川家康」についての要素知識語として専門的であると言え,「江戸幕府」はそれに比べて DF 値が高く，一般的な要素知識語であると言える．

4.2 ページ内情報の一般性

クエリトピックに関連する Web ページが持っているクエリトピックについての情報がどの程度一般的であるか，ということをページ内情報の一般性，と呼ぶ．

関連 Web ページ内にはクエリトピックについての情報が一つ以上存在しており，一般的な情報と専門的な情報が混在していることもありうる．図 3 のように，一般的な情報が多く，専門的な情報が少ないページをここではページ内情報の一般性が高いとし，そうでない専門的な情報が多いページをページ内情報の一般性が低いとする．つまり，第 3.3 章で述べた手法によって抽出した要素知識語集合内に存在する語で，ある Web ページ中にどの程度の一般性を持つ語がどの程度出現しているか，によってそのページ内情報の一般性が判定できる．その際に，要素知識語集合内の語がどの程度出現しているか，つまり一般性に関わらず要素知識語がどれだけ多く出現しているかという側面については，これはページ内情報の一般性ではなく，その情報量を表していると考え，一般性を求めるためには用いない．

上で述べたように，ページ内情報の一般性は，ページ内に存在する要素知識語がどの程度専門的なものがあるか，ということによって決定される．従って，一般的知識範囲を求める際にはページ内に一般的な要素知識語，つまり DF 値の高い要素知識語を含んでいるかどうかも考える必要がある．なぜなら一般的知識範囲内の情報，言い換えればあまり専門的でない情報を抽出することを目的としているからである．したがって誰でも知っているような，一般性の非常に高い要素知識語，つまり DF 値が非常に高い要素知識語ばかり含んでいるページであっても目的を満たすために必要となる．

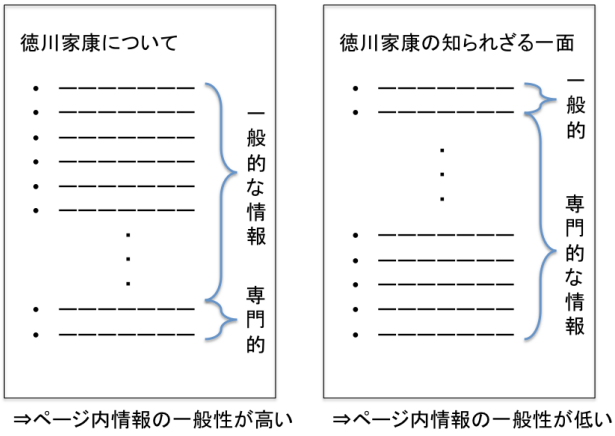

図 3 ページ内情報の一般性

例を挙げると,「江戸幕府」という語は「徳川家康」について説明する際によく登場するため多くの Web ページ内に出現する．したがって DF 値は高くなるが,「徳川家康」について知りたい人にとっては欠かせない情報であると言え,「江戸幕府」について説明してあるページは必要であると言える．

5. 一般的知識範囲の推定

本章では，上で求めた Web ページ内情報の一般性についての Web ページ集合の分布から，一般的知識範囲を推定する手法について述べる．

5.1 Web ページ内情報の一般性による分布

第 4.2 章で求めた各 Web ページ内情報の一般性にもとづき，どの程度の一般性の Web ページがどの程度のページ数存在するのかついての分布を，クエリトピックの関連 Web ページ集合から得る．図 1. は，そのような分布の例である．この図では，x 軸が「ページ内情報の一般性」，y 軸が「ページの数」を表す．

5.2 分布から一般的知識範囲を推定する

上で作成した分布から一般的知識範囲を推定する手法について述べる．

一般的知識範囲とは，クエリトピックの要素知識について多くの一般的な Web ページに記述されているような情報の範囲と考えることができる．前述の分布から一般的知識範囲を抽出するにあたり，本研究では，一般的な要素知識について書いてあるページ数と一般的でない専門的な要素知識について書いてあるページ数には差があり，またそのページ数の関係はページ

表 1　パターンセット

	Value(t)	出現回数	出現位置	タイトル語	固有名詞
a		○			
b	○				
c		○	○		
d		○	○	○	○
e	○	○	○	○	○

内情報の一般性が低くなるにつれて，多少の増減はあるとしても全体では減少していき，かつある点において急激に減少するという仮説を置く．なぜならある点までは多くのページがその程度の一般性を持っていのに，その点から一般性を少し下げた場合にページ数が急に減少するということは，その程度の情報の一般性からが，専門的なであると考えられるからである．

よって，クエリトピックの関連ページに対して一般性の分布を求めた際に，一般性がある値となるところで，Web ページの数が急激に減少するような点があれば，その点の持つ一般性が一般的知識範囲かどうかの閾値となると推測し，その閾値よりも一般性の高い要素知識語を一般的知識範囲内の情報として扱う．

こうして推定された一般的知識範囲を満たすような Web ページ集合を発見し，結果として返すことで問題を解決する．

6. 実　　験

本論文では，提案手法の一つ目のステップである，関連 Web ページ中に存在する名詞の集合から要素知識語集合を抽出する部分について，提案手法の有効性を実験によって検証する．

6.1 手　　法

ここでは，3.2 節で提案した 5 つの指標，すなわち，(1) クエリトピック語との共起度から求められる $Value(t)$ ，(2) 出現回数，(3) 出現位置，(4) タイトル中に出現するかどうか，(5) 固有名詞であるかどうか，を用いてクエリトピックの要素知識語の候補となる語の抽出を行い，これらの指標の有効性を比較する．

実験の手順を以下に述べる．まず，ランキング手法として表 6.1 にあるような (a) から (e) の 5 つのパターンを用意する．クエリトピック語を含む Web ページ群から抽出した名詞を，これら 5 つのパターンでそれぞれランキングを行い，その上位 100 語からなる語集合を抽出する．そして，これらの 4 つの語集合に対して，適合率による評価と，各語の DF 値の平均値による比較を行った．この際に，評価基準として，(a) によるランキング，すなわち，語の出現回数のみによるランキング結果を用いることとする．

正解となる語，つまりクエリトピックについての要素知識語としてふさわしい語は人手で判定を行った．全ての正解語を発見することは不可能であるので，再現率は考慮しない．また，ページ内情報の一般性を求める際に語の持つ DF 値が重要となる．その参考として，それぞれの指標によるランキングが，どの程度知識の一般性に影響を与えるかを調べるために DF 値の平均値による比較を行った．

表 2　適合率と DF の平均値

パターン	a	b	c	d	e
適合率	7%	7%	12%	18%	23%
DF の平均値	0.3134	0.0286	0.0173	0.0220	0.0200

今回の実験では，$CoOccure(t)$ と $Ratio(t)$ の値を求める際に Web 検索エンジンの検索結果数を用いており，検索を行うタイミング等によって得られる結果が変化してしまう．そのため，たまたま検索結果数が多くなるといったことが生じる可能性があり，これにより $CoOccure(t)$ や $Ratio(t)$ の値が期待される程度の値より高くなることがある．つまり， $Value(t)$ の値が高くなり，クエリトピックとあまり関連が深くないと推定される語が誤って要素知識語として判定され，また，逆に関連が深いと推定される語の $Value(t)$ が期待される値よりも低くなり，要素知識語でないと判定される，という問題がある．

そこで，そのような不確定な要素を排除したランキングを比較対象として行うために，上の 5 つのパターンの選択において，(b) として $Value(t)$ のみによるランキングを選択している．

また，(c) として，出現位置によるランキングを選択している理由は，タイトル語と固有名詞は文章中において説明に用いられない語に重みをつけるための指標であり，そういった意図とは異なる，出現位置による語のページに対する関連度についての指標を別に評価するためである．

「説明に用いられる語」とは，クエリトピックについての情報を記述する際の語であり，またクエリトピック自体とは関係が無いが，記述に用いられるため関連 Web ページ中に出現する語のことを指し，以降では説明語と呼ぶ．従って，説明に用いられない語とはクエリトピック自体と関係があり，関連 Web ページ中に出現する語のことである．

6.2 データセット

クエリトピックの例として「徳川家康」を用意し，知識集合を抽出する提案手法の有用性を検証した．今回は一般のウェブ検索エンジン(注3)に「徳川家康」をクエリとして投入し，その検索結果上位 100 件の Web ページ中に存在する 13962 語の名詞を抽出した．ただしこの中には第 3.3 章で述べたように連続する名詞を連結させた名詞も含んでいる．

6.3 結　　果

(a) から (e) の 5 パターンでのランキングによる適合度の結果を表 6.3 に示す．

6.4 考　　察

評価基準として用意した (a) の適合率が 7%であり，提案する指標を全て用いた (e) の適合率が 23%であることから，提案する指標による適合率の向上が見られた．以下で詳しく述べる．

まず (b) のパターンで適合率が向上しなかった理由は，Web 検索エンジンの検索結果ページからプログラムによって取得した検索結果数と，実際に検索して得られる数値との間に大きな差があり，共起度が高くなると推定される語であっても非常に低い数値となっていることがあり，期待される効果が得られな

(注3)：Google 検索：http://www.google.co.jp

かったから，と考えられる．

次に (d) と (e) の結果を見ると，(a) に比べて文章中において説明に用いられるような語の数が少なかった．つまり，クエリトピックと関連がある，あるいは重要であると判定されるような語が多く見られたということであり，これは説明が非常に少ないタイトルや，説明文にあまり用いられない固有名詞に重みをつけたことによる適合率の向上であると推測される．

説明文に用いられないことを表す指標として出現位置のみによるランキングを行った (c) は適合率の向上は (d) と (e) より小さかったが，語の DF 値が全てのパターン内で最小となっており，一般性の低い語を抽出する際に効果的であることが明らかとなった．

また (d) と (e) との間にも差異があり，(d) の方が語の DF 値は低く，$Value(t)$ を用いるとクエリトピックとの関係がある語を多く抽出することが可能であるが，その反面，多くのページに出現している説明文に用いられるような語を抽出するため，語の一般性は低くなる傾向がある．

7. おわりに

本論文では，あるトピックに関して専門的でない，一般的な情報を抽出するための手法を 5 段階に分けて述べた．

まず調べたいトピックの持つ情報を構成する要素知識語集合を抽出する．それによってクエリトピックに関連する Web ページの持つ情報の一般性を判定する．そして，どのような一般性のページがどの程度存在するかの分布を作成し，分布されるページ数が大きく変化する点から一般的知識範囲を推定し，その範囲内にある情報を抽出する手法について提案した．

要素知識語集合の抽出において，提案する指標が妥当であることが一定程度示されたが，$Value(t)$ の値が期待されるものと差が大きいため，今後原因を究明し，適合率の上昇を目指す．そして，抽出した要素知識語集合を用いて一般的知識範囲の推定に向けて提案手法を実装し，その有用性を検証していく．

謝辞　本研究の一部は科研費 (23650048) の助成を受けたものである．

文　献

[1] 桜井俊彦, 内海彰. “情報検索のためのクエリに基づく文書自動要約”. 言語処理学会第 10 回年次大会発表論文集, pp. 265–268, 2004.

[2] Jaime Carbonell and Jade Goldstein. “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. *In SIGIR*, pp. 335–336, 1998.

[3] Jade Goldstein Vibhu Mittal, Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. “Multi-document Summarization By Sentence Extraction”. *In Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, pp. 40–48, 2000.

[4] Jianping Zeng, Chengrong Wu, and Wei Wang. “Multi-grain hierarchical topic extraction algorithm for text mining”. *Expert Systems with Applications*, pp. 3202–3208, 2010.

[5] Khoo Khyou Bun and Mitsuru Ishizuka. “Topic Extraction from News Archive Using TF*PDF Algorith”. *WISE'02*, pp. 73–82, 2002.

[6] Yusuke Yamamoto, Taro Tezuka, Adam Jatowt, and Katsumi Tanaka. “Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis”. Vol. 4505, pp. 253–264, 2007.

[7] Wei Dai and Rohini K. Srihari. “Minimal document set retrieval”. *CIKM*, pp. 752–759, 2005.