# 文書構造を利用したWebからの話題発見

小山　聡†　田中　克己†

† 京都大学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: †{oyama,ktanaka}@i.kyoto-u.ac.jp

**あらまし**　本研究では，ユーザから大雑把な話題 (broad topic) についての質問が与えられたときに，それを詳細化するような話題を Web から抽出し，ユーザに提示する手法を提案する．従来のページ単位での単語の共起に基づいて関連キーワードを抽出する手法では，必ずしも主題と関係のないキーワードが多く抽出されてしまうという問題があった．提案手法では，単語の共起を測る際に，キーワードが Web ページの中のタイトルに現れる場合とそれ以外に現れる場合を区別することで，主題を詳細化するキーワードを精度良く抽出することができる．これらのキーワードをユーザに提示することで，ユーザは Web の中にどのような話題があるかを把握し，より詳細な情報の検索を行うことが可能となる．

**キーワード**　話題抽出，話題構造，インターネットと Web，情報検索，知識発見，データマイニング

# Using Document Structure for Extracting Topics from the Web

Satoshi OYAMA† and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida–Honmachi, Sakyo, Kyoto 606–8501, Japan
E-mail: †{oyama,ktanaka}@i.kyoto-u.ac.jp

**Abstract**　In this paper we propose a method for extracting keywords that detail the broad topic given by the user from the Web. Existing methods for extracting related keywords are based on term co-occurrence in each page and they extract many keywords irrelevant to the topic. Our method can precisely extract detailing topic keywords by considering positions (title or text) of keywords when they co-occur in documents. Using these keywords, the user can have a grasp of the topics in the Web and use them for more detailed search.

**Key words**　Topic distillation, Topic structure, Internet and Web, Information retrieval, Knowledge discovery, Data mining

## 1. はじめに

今日では，我々が情報を収集する際にまず最初に利用するのは Web である場合が多い．すなわち，Web を情報の探索の出発点として用いている．Web には，様々な分野の大量の情報が提供されており，Web ページ自身に有益な情報が掲載されている場合もあれば，外部の情報源（出版物やニュース）へのポインタが示されていることも多い．まず Web にアクセスすることで，未知の話題についても何らかの手がかりが得ることができる．

このように，Web を情報の探索の出発点として用いる場合，どのような情報が欲しいかあらかじめ具体的に分かっている場合（例えば，お店の住所や電車の出発時刻を調べるなど）もあるが，そうではない場合も多い．例えば，旅行先での観光プランなどを探すとき．初めて訪問する場所であればユーザはどのような観光スポットがあるのかすら知らないことがある．このようなとき，とりあえず大雑把な質問（ハンガリーの観光情報であれば “ハンガリー” というキーワード）を検索エンジンに投入し，検索結果のページを見て，どのような情報があるのかを確認するであろう．

しかしながら Web では，初期の大雑把な質問を検索エンジンに投入した場合，非常に多くの検索結果が返される．これらのページを１つ１つ確認し，中に含まれる話題について調べることは非常に労力のかかる作業である．また，キーワードを追加して検索結果を絞り込む場合でも，どのようなキーワードを追加すればよいかは，ある程度の数のページを見て初めて分かる場合が多い．

このような，大雑把な話題（broad topic）の検索は，Web ページの数が増大するに従って，ますます困難になっている．一般のユーザが，多くの検索キーワードからなる複雑な質問を

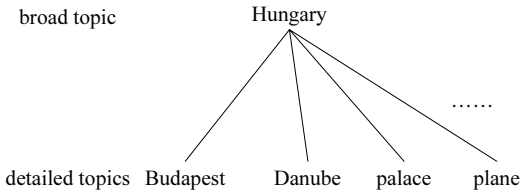


図 1　話題の構造.
Fig. 1　Topic structure.

投入できないことはよく知られている [1] [2]．ほとんどのキーワードに対して，何千，何万のページが存在することは，潜在的には Web の中から必要な情報を取得する可能性を高めている．しかし現実は，ユーザの多くは検索エンジンの上位のページの幾つかを見るだけであきらめてしまい，必要な情報までたどり着けないことになる．

このような問題を解決するために，話題を階層的に分類した Yahoo!(注1)や Open Directory(注2)などの Web ディレクトリの開発が行われて来た．しかし，ユーザの多様な興味にうまく合致したディレクトリが存在するとは限らない．ユーザが入力した多様な話題に対して，その検索を支援できるような手法が求められている．

そこで本研究では，ユーザから大雑把な話題についての質問が与えられたときに，それを詳細化した話題を表すキーワードを Web から発見し，ユーザに提示することで情報検索を支援する方式を提案する．

例えば，ユーザがハンガリーについて調べたい場合を想定する．“Hungary” というキーワードを検索エンジンに投入した場合，検索結果のページ集合には，図 1 のように，ブダペストやドナウ川，平原や王宮など，ハンガリーという主題を詳細化する様々な話題が含まれるであろう．そこで我々の手法では，このように主題を詳細化する話題を表すキーワードを，Web から抽出してユーザに提示する．これにより，ユーザが Web ページ集合の中にどのような話題が含まれているのかを把握し，さらにこれらのキーワードを用いて詳細な検索を行うことを支援することが可能となる．

従来，キーワードの関連を抽出するために，連想ルールがしばしば用いられてきた [3] [4]．これらの研究での連想ルールは，ページ単位での語の共起度に基づいている．しかし，ある話題を表すキーワード A が与えられたとき，Web においてこれと共起するキーワード B が必ずしも，かならずしも A を詳細化するために用いられているとは限らない．キーワードの間の共起だけでなく，それらが「どのような関係で」用いられているかを考慮する必要がある．そこで，Web ページの内部構造に着目し，キーワードがページのタイトルに現れるのか，それ以外の部分に現れるのかを考慮してキーワードの間の共起度を測ることで，キーワード A を詳細化するキーワード B を抽出することを提案する．

以下では，2 章で Web ページの文書構造を利用した話題の抽出法について述べ，3 章で実際の抽出例を示す．4 章で関連研究を示し，5 章で本論文の結論を述べる．

(注1)：http://www.yahoo.com/
(注2)：http://dmoz.org/

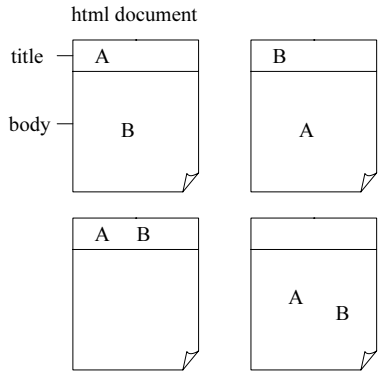


図 2　同じキーワードを異なった場所に含む Web ページ.
Fig. 2　Web pages that have same keywords but in different positions.

## 2. Web ページの文書構造を利用した関連キーワードの抽出

### 2.1 連想ルール

連想ルール (association rule) [5] はもともと，データマイニングの分野で POS データの解析などに利用されているものである．情報検索に用いる場合には，キーワード集合の間の連想ルールを用いる．例えば，キーワード A とキーワード B の間の連想ルール $\mathrm{A} \Longrightarrow \mathrm{B}$ を考えたとき，このルールは以下に示す確信度 (confidence) と支持度 (support) で評価される．

$$\mathrm{confidence}(\mathrm{A} \Longrightarrow \mathrm{B}) = \frac{DF(\mathrm{A} \wedge \mathrm{B})}{DF(\mathrm{A})}$$

$$\mathrm{support}(\mathrm{A} \Longrightarrow \mathrm{B}) = \frac{DF(\mathrm{A} \wedge \mathrm{B})}{N}$$

$DF(\mathrm{A})$ はキーワード A を含むページの文書頻度 (document frequency)，$DF(\mathrm{A} \wedge \mathrm{B})$ はキーワード $A$ および $B$ を含むページの文書頻度，$N$ は全体のページ数である．

例えば，

$$\mathrm{confidence}(\text{ハンガリー} \Longrightarrow \text{ブダペスト}) = 0.12$$

$$\mathrm{support}(\text{ハンガリー} \Longrightarrow \text{ブダペスト}) = 0.01$$

であれば，“ハンガリー” というキーワードを含む Web ページが与えられた場合，そのうち “ブダペスト” というキーワードを含むものが 12%あり，“ハンガリー” と “ブダペスト” というキーワードを含むページが全体のページの集合の中に 1%あることを示している．

### 2.2 文書構造を反映した連想ルール

従来の連想ルールを情報検索に用いる研究 [3] [4] では，ページ内にキーワードが現れるか否かにのみ基づいている．しかし，図 2 に示すように，たとえば二つのキーワードがページ内で共起する場合であっても，キーワードがどこに現れるかによって，キーワードがページの話題を記述する上で果たす役割は異なっている．

通常は，ページのタイトル部分に現れるキーワードはページ

表 1　2x2 分割表
Table 1　2x2 Contingency Table

| | Class $B_1$ | Class $B_2$ | Total |
|---|---|---|---|
| Class $A_1$ | $x_{11}$ | $x_{12}$ | $a_1$ |
| Class $A_2$ | $x_{21}$ | $x_{22}$ | $a_2$ |
| Total | $b_1$ | $b_2$ | $N$ |

の主題を表し，ページの本文に現れるキーワードは，主題に関して詳細な説明を行うために用いられると考えられる．このような前提に立てば，キーワード A が与えられたとき，これをタイトルに含むページ内に頻繁に現れるキーワード B が見つかれば，B は A を詳細化した話題を表している可能性が高いと言えるであろう．

もちろん，個々のページを見れば，タイトルが正確にページの主題を表していない場合もある．しかし，検索エンジンの膨大なインデックスにおける，キーワードの出現頻度を参照することで，このような例外の影響を低減することができる．

以下では，Google(注3) の検索式の記法に従い，キーワード A をページのタイトル部分に含むという条件を，intitle(A) と記述する．ここで，

$$\text{intitle(A)} \Longrightarrow \text{B} \tag{1}$$

の連想ルールを考える．このルールの確信度と支持度はそれぞれ，

$$\text{confidence(intitle(A)} \Longrightarrow \text{B)} = \frac{DF(\text{intitle(A)} \wedge \text{B})}{DF(\text{intitle(A))}} \tag{2}$$

$$\text{support(intitle(A)} \Longrightarrow \text{B)} = \frac{DF(\text{intitle(A)} \wedge \text{B})}{N} \tag{3}$$

で与えられる．$DF$(intitle(A)) はタイトルにキーワード A を含むページの数，$DF$(intitle(A)$\wedge$B) はタイトルにキーワード A をふくみ，かつページのどこかにキーワード B を含むページの数である．このルールの確信度が高いことは，キーワード A がタイトルに含まれるとき，高い割合でキーワード B がページ内に含まれることを意味する．

キーワード A が与えられたとき，これに対して高い確信度でルール (1) が成り立つキーワード B が発見されれば，それは A を詳細化するキーワードである可能性が高い．しかし，B がもともと出現頻度の高い語であれば，ルール (1) の確信度が高かったとしても，必ずしもキーワード A を詳細化しているとはいえない．例えば，“情報” や “ページ” というキーワードは Web において頻繁に現れ，ルール (1) の確信度も高くなるが，このようなキーワードを抽出しても有用ではない．そこで，キーワードの配置を考慮しない単純な連想ルール

$$\text{A} \Longrightarrow \text{B} \tag{4}$$

を考え，ルール (1) とルール (4) の確信度を比較し，前者が高ければ，B をキーワードとして抽出することを行う．

ここで注意する必要があるのは，例え前者が後者よりも大き

(注3)：http://www.google.com

$$\begin{aligned}
N &= DF(\text{A}) \\
a_1 &= DF(\text{intitle(A)}) \\
a_2 &= N - a_1 \\
b_1 &= DF(\text{A} \wedge \text{B}) \\
b_2 &= N - b_1 \\
x_{11} &= DF(\text{intitle(A)} \wedge \text{B}) \\
x_{12} &= a_1 - x_{11} \\
x_{21} &= b_1 - x_{11} \\
x_{22} &= N - a_1 - b_1 + x_{11}
\end{aligned}$$

図 3　変数の定義．
Fig. 3　Definitions of variables.

な値をとったとしても，偶然による場合もあり，それだけで意味のある違いをいうことはできないということである．そこで，二つのルールの確信度の違いが偶然によるものなのか，意味のあるものなのかを測定するため，統計的検定の手法を利用する．具体的には，以下のような $\chi^2$ 検定を行う．

表 1 の分割表のように，母集団が 2 つの性質 $A, B$ の両方において，互いに背反なクラス $A_1, A_2$ と $B_1, B_2$ に分けられており，大きさ $N$ の標本における，各クラスの観測度数が，$a_1, a_2$ および $b_1, b_2$ であるとする．ここで，クラス「$A_i$ かつ $B_j$」の観測度数が $x_{ij}(i = 1, 2; j = 1, 2)$ であったとする．（ここで，$\sum_{j=1}^{2} x_{ij} = a_i, \sum_{i=1}^{2} x_{ij} = b_j$ である．）このとき

$$\chi_0^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(x_{ij} - a_i b_j / N)^2}{a_i b_j / N} \tag{5}$$

の値がは自由度 1 の $\chi^2$ 分布に従う．ここで，有意水準を $\alpha$ に設定したとすると，$\chi_0^2 > \alpha$ であれば，2 つの性質 $A, B$ の間に関連があるということができる．

例えば，2 つのルール intitle(ハンガリー) $\Longrightarrow$ ブダペスト と ハンガリー $\Longrightarrow$ ブダペスト の確信度を比較する場合は，表 2 のような分割表をつくり，$\chi_0^2$ 値を計算する．危険率 $\alpha = 0.05$ とすれば，自由度 1 の場合の $\chi^2$ 分布表 [6] によると上側確率 5%点の値は 3.84 であるので，$\chi_0^2 = 752.858 > 3.84$ より，“ハンガリー” がタイトルに含まれるという事象と，“ブダペスト” がページ内に含まれるという事象に統計的な有意な関係があるということができる．分割表の各セルの値は全て，図 3 の式より，4 つの質問（A，intitle(A)，A∧B，intitle(A)∧B）に対する検索結果の数から求めることができる．

$\chi_0^2$ 値は二つの性質の間に正の依存関係がある場合だけでなく，負の関係がある場合にも大きな値をとる．これは，confidence(intitle(A) $\Longrightarrow$ B) が confidence(A $\Longrightarrow$ B) よりも有意に小さくなる場合に対応する．このようなキーワード B を抽出することは目的でないため，

$$\frac{DF(\text{intitle(A)} \wedge \text{B})}{DF(\text{intitle(A))}} > \frac{DF(\text{A} \wedge \text{B})}{DF(\text{A})}$$

という条件を課すことで，このようなキーワードを除く．

### 2.3　アルゴリズム

上で示した指標を用いて，ユーザから話題を表すキーワードが与えられたとき，これを詳細化するキーワードを Web から抽

表 2　主題キーワード “ハンガリー” と詳細化キーワードの候補 “ブダペスト” の分割表.
Table 2　Contingency Table for the subject keyword “Hungary” and the candidate of the detailing keyword “Budapest”.

| | “ブダペスト” を含む | “ブダペスト” を含まない | Total |
|---|---|---|---|
| “ハンガリー” をタイトルに含む | 979 | 2741 | 3720 |
| “ハンガリー” をタイトルに含まない | 11121 | 86159 | 97280 |
| Total | 12100 | 88900 | 101000 |

// $A$ = ユーザの入力キーワード
// $N$ = サンプルページ数
// $p$ = サンプルページにおける支持度の閾値
// $M$ = ユーザに提示する詳細化キーワード数
// $\alpha$ = 検定の有意水準
（1） 質問 A で検索を行い，$DF$(A) の値を取得する．
（2） 質問 intitle(A) で検索を行い，$DF$(intitle(A)) の値を取得する．
（3） (2) の検索結果の中から，A をタイトルに含むページ $N$ 件をサンプルとして抽出する．
（4） サンプルページ集合において一定割合 $p$ 以上のページに現れるキーワードを，候補として選択する．
（5） (4) で抽出したそれぞれの候補 B について，質問 A∧B, intitle(A)∧B で検索を行い，$DF$(A ∧ $B$)，$DF$(intitle(A) ∧ $B$) を求め，図 3 および式（5）に従って $\chi_0^2$ 値を求める．
（6）

$$\frac{DF(\text{intitle(A)} \wedge \text{B})}{DF(\text{intitle(A)})} > \frac{DF(\text{A} \wedge \text{B})}{DF(\text{A})}$$

および $\chi_0^2 > \alpha$ 満たすキーワード B の中から，確信度が大きなキーワード $M$ 個を話題を詳細化するキーワードとしてユーザに提示する．

図 4　話題を詳細化するキーワードを発見するアルゴリズム.
Fig. 4　Algorithm for extracting keywords that describe the topic in detail.

出するアルゴリズムを図 4 に示す．ユーザの投入したキーワードと，Web における全てのキーワードの組み合わせに対して，$DF$(A ∧ $B$) および $DF$(intitle(A) ∧ $B$) の値を求めることは，時間がかかる処理である．そこで，質問 intitle(A) で検索した結果の中からサンプルページを収集し，その中で出現割合が一定の閾値 $p$ 以上のキーワードに対してだけ，実際に表 3 の値を求める．これはサンプルページの数を $N$ としたときの式 (3) の支持度が $p$ 未満のルールを枝狩りすることに対応している．

## 3. 抽　出　例

ネットワークを通して検索エンジンに複数の質問を投入することは，時間のかかる処理であり，提案手法は，検索エンジンの側のサービスとして実現する方が適していると考えている．しかし，本論文では，十分な数の Web ページ集合に対して評価を行うため，Google が提供している Web サービスである Google Web APIs [7] を用いて実験を行った．一種のメタ検索の形で，クライアント側からサーバ側に複数の質問を投入して質問に合致する検索結果数を取得している．また，ここでは $N = 50$ として， intitle(A) に対する検索結果の上位 50 ページをサンプルとして Web からダウンロードした．これらのページの中から日本語形態素解析システム茶筅 [8] を用いて名詞を抽出した．$p = 0.1$ として，サンプルページ集合内でこれ以上の出現頻度を持つキーワード B について，質問 A∧B, intitle(A)∧B に対する Google の検索結果数を元に，図 3 の値を求め，$\chi_0^2$ の値を計算した．有意水準 $\alpha = 0.05$ として検定を行い，有意性がみられたルールを確信度の高い順に示した．

$A$ = ハンガリー として，抽出した結果が表 3 である．ハンガリーの地名（ブダペスト，ドナウ）や名所（王宮），通貨（フォリント），民族名（マジャール）など，もっぱらハンガリーについての話題を詳細化するキーワードが多く抽出されていることが分かる．次に，$A$ = ブダペスト として抽出した結果を表 4 示す．$A$ = ハンガリー の場合と共通するキーワード（ドナウ, 王宮）などの他に，ブダペストについての場所を表すキーワード（橋，広場）も抽出されている．

$A$ = 環境 とした場合の結果が，表 5 である．地球，リサイクル，保全といった，環境についての話題を詳細化するキーワードが現れている．

比較のために，表 6, 7, 8 に従来のページ単位の共起に基づく連想ルールで抽出したキーワードを示す．“ハンガリー” に対しては，他の国名等，関連性はあるが対等な関係にあると思われるキーワードが抽出されている．“ブダペスト” に対しても，“ハンガリー”，“ヨーロッパ” などの，より広い話題を表すキーワードが抽出されている．“環境” に対しても，かなり一般的なキーワードが抽出されていることが分かる．

これらの実験結果から，ページ内でのキーワードの配置を考慮することで，抽出されるキーワードを，話題を詳細化するものに限定することが可能になったと言える．

## 4. 関 連 研 究

連想ルールを用いてキーワードの関連を抽出した研究としては，Mondou [3] がある．これは，連想ルールを用いてユーザに関連キーワードを提示し，ユーザがその中からキーワードを選択をして検索の絞り込みを行う検索エンジンである．

Sanderson と Croft [9] は，連想ルールと同様の指標を用いて，文書から概念階層を抽出することを行っている．彼らは，キーワードを含むページの間に包含関係が存在する場合，すなわち，

$$p(\text{A}|\text{B}) = 1, p(\text{B}|\text{A}) < 1$$

が成り立つ場合，A を B の親とする概念階層を抽出し，ユーザに提示することを提案している．($p$(A|B) はページにキーワード B が含まれるとき，A が含まれる割合であり，confidence(B $\Longrightarrow$ A) に対応する．) これは，図 5 のように，キーワード B を含むページの集合が，キーワード A を含むペー

表 3 “ハンガリー” を詳細化するキーワード

Table 3 Keywords detailing “Hungary”.

| B | confidence(intitle(A) ⇒ B) | confidence(A ⇒ B) | $\chi_0^2$ |
|---|---|---|---|
| ブダペスト | 0.263172 | 0.119802 | 752.858432 |
| ハン | 0.137634 | 0.093168 | 90.386535 |
| ドナウ | 0.110484 | 0.055941 | 217.568895 |
| フォリント | 0.086828 | 0.015545 | 1282.458293 |
| ブタペスト | 0.073118 | 0.042376 | 89.947452 |
| ブダ | 0.059409 | 0.013069 | 642.984328 |
| 王宮 | 0.058065 | 0.029307 | 112.277652 |
| マジャール | 0.043011 | 0.018713 | 124.176963 |
| マーチャーシュ | 0.026075 | 0.009822 | 104.913207 |
| 建国 | 0.025806 | 0.020099 | 6.388029 |

A=ハンガリー

表 4 “ブダペスト” を詳細化するキーワード

Table 4 Keywords detailing “Budapest”.

| B | confidence(intitle(A) ⇒ B) | confidence(A ⇒ B) | $\chi_0^2$ |
|---|---|---|---|
| 旅 | 0.343939 | 0.263158 | 23.010772 |
| ドナウ | 0.292424 | 0.147895 | 113.335319 |
| 王宮 | 0.242424 | 0.097895 | 161.731524 |
| ブダ | 0.227273 | 0.065789 | 290.102866 |
| ガイド | 0.224242 | 0.107895 | 96.160642 |
| 広場 | 0.219697 | 0.155263 | 21.643924 |
| 橋 | 0.218182 | 0.130526 | 46.291585 |
| 観光 | 0.218182 | 0.172105 | 10.188 |
| 予約 | 0.210606 | 0.119474 | 53.979518 |
| 川 | 0.207576 | 0.166842 | 8.161537 |

A=ブダペスト

表 5 “環境” を詳細化するキーワード

Table 5 Keywords detailing “environment”.

| B | confidence(intitle(A) ⇒ B) | confidence(A ⇒ B) | $\chi_0^2$ |
|---|---|---|---|
| 地球 | 0.323782 | 0.202259 | 34408.92826 |
| リサイクル | 0.182808 | 0.08193 | 50861.80528 |
| 物質 | 0.17937 | 0.096715 | 29399.61912 |
| エネルギー | 0.163897 | 0.135113 | 2665.404813 |
| 化学 | 0.130372 | 0.106366 | 2279.473625 |
| 保全 | 0.100287 | 0.068789 | 5822.674902 |
| 用語 | 0.096275 | 0.077618 | 1827.805433 |
| 大気 | 0.093983 | 0.031417 | 48361.17722 |
| 排出 | 0.086533 | 0.045175 | 14908.36125 |
| 温暖 | 0.084814 | 0.062218 | 3289.81083 |

A=環境

ジの集合に包含されている場合である.

このような包含関係を用いる際の問題点として，実際にこのような包含関係が成り立つ例が稀であるということである. Sanderson らも実際にはノイズの影響を避けるため,

$$p(\mathrm{A}|\mathrm{B}) \geqq 0.8, p(\mathrm{B}|\mathrm{A}) < 1$$

の指標を用いており，これは，confidence(B $\Longrightarrow$ A) $\geqq 0.8$ を満たすキーワード B を抽出することに対応する.

しかし例えば，A = ハンガリー と B = ブダペスト のように，包含関係が予想される場合であっても，$DF$(ブダペスト) = 19000, $DF$(ブダペスト ハンガリー) = 12100 であり，$p$(A|B) = 0.637 にしかならず，抽出からもれてしまう. また，逆に $p$(A|B) の閾値を下げると，必ずしも包含関係にあるとはいえないキーワードまで抽出されてしまうという問題がある.

表 9,10, 11 に，第 3. 章と同じサンプル集合に対して，Sanderson らの基準を適用して抽出したキーワードを示す. (前章の表

表 6 “ハンガリー” についてページ単位の共起で抽出したキーワード
Table 6 Keywords extracted for “Hungary” by using co-occurrences in pages

| B | confidence(A⟹B) |
|---|---|
| 日本 | 0.719802 |
| 中 | 0.709901 |
| 人 | 0.70099 |
| ドイツ | 0.694059 |
| フランス | 0.652475 |
| イタリア | 0.611881 |
| 国 | 0.60198 |
| 世界 | 0.548515 |
| スペイン | 0.539604 |
| アメリカ | 0.521782 |

A=ハンガリー

表 7 “ブダペスト” についてページ単位の共起で抽出したキーワード
Table 7 Keywords extracted for “Budapest” by using co-occurrences in pages

| B | confidence(A⟹B) |
|---|---|
| ハンガリー | 0.636842 |
| 日本 | 0.521053 |
| ドイツ | 0.432632 |
| ヨーロッパ | 0.391053 |
| ウィーン | 0.338421 |
| フランス | 0.304211 |
| プラハ | 0.269474 |
| 旅 | 0.263158 |
| ホテル | 0.254737 |
| オーストリア | 0.252632 |

A=ブダペスト

とは，連想ルールの向きが逆になっていることに注意して欲しい.）“環境” に対しては，我々の手法と同様の良い抽出結果が得られているが，“ハンガリー” に対しては，閾値を 0.8 にした場合，わずか 2 つのキーワード（表の太線より上の部分）しか抽出されていない．また，閾値を下げて行けば，包含関係にあるとはいえないキーワードが抽出されてしまうことが分かる．“ブダペスト” については，条件を満たすキーワードは抽出されていない.

これは，“環境” というキーワードは Web での出現頻度が非常に大きい（4,900,000 件）であるが，“ハンガリー” や “ブダペスト” は比較的小さい（101,000 件および 19,000 件）ことが影響していると考えられる．即ち，包含関係を用いた場合，非常に一般的なキーワードに対しては多くの詳細化キーワードが抽出できるが，出現頻度が比較的少ない（それでも，数万以上の出現回数がある）キーワードについては，詳細化キーワードを抽出できない可能性が高い.

同様の研究として，あるページ集合が与えられた時に，その集合自身を記述するキーワード (self)，上位概念を記述するキーワード (parent)，下位概念を記述するキーワード (child) を抽出するものがある [10]．この研究においては，対象クラスタと

表 8 “環境” についてページ単位の共起で抽出したキーワード
Table 8 Keywords extracted for “environment” by using co-occurrences in pages

| B | confidence(A⟹B) |
|---|---|
| 情報 | 0.595918 |
| 利用 | 0.522449 |
| ページ | 0.477551 |
| システム | 0.444898 |
| 日本 | 0.410204 |
| 機能 | 0.397959 |
| サイト | 0.393878 |
| 技術 | 0.393878 |
| 開発 | 0.367347 |
| 製品 | 0.310204 |

A=環境

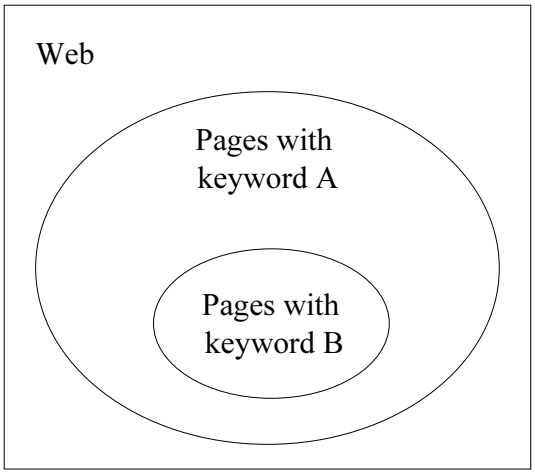


図 5 キーワード A がキーワード B を包含する例.
Fig. 5 Example of keyword A subsuming keyword B.

表 9 “ハンガリー” に対する包含関係で抽出したキーワード
Table 9 Keywords extracted by using subsumption under “Hungary”

| B | confidence(B⟹A) |
|---|---|
| フォリント | 0.945783 |
| マーチャーシュ | 0.901818 |
| マジャール | 0.713208 |
| ブダ | 0.694737 |
| ブダペスト | 0.636842 |
| ブタペスト | 0.600281 |
| ブルガリア | 0.413613 |
| ポーランド | 0.319130 |
| チェコ | 0.313725 |
| オーストリア | 0.300741 |

A=ハンガリー

全体のページ集合におけるキーワードの出現頻度を比較し，クラスタ内で頻出するが，全体では比較的稀なキーワードを self，クラスタ内でも全体でもよく現れるキーワードを parent，クラスタ内で良く現れるが，全体では稀なキーワードを child として抽出している．彼らの研究では，Open Directory のページ

表 10 “ブダペスト” に対する包含関係で抽出したキーワード
Table 10 Keywords extracted by using subsumption under “Budapest”

| B | confidence(B⟹A) |
|---|---|
| マーチャーシュ | 0.795455 |
| ブダ | 0.657895 |
| 漁夫 | 0.136729 |
| ドナウ | 0.121121 |
| ハンガリー | 0.119802 |
| プラハ | 0.115056 |
| ウィーン | 0.063663 |
| 東欧 | 0.049153 |
| ペスト | 0.037397 |
| チェコ | 0.037059 |

A=ブダペスト

表 11 “環境” に対する包含関係で抽出したキーワード
Table 11 Keywords extracted by using subsumption under “environment”

| B | confidence(B⟹A) |
|---|---|
| 排出 | 0.920502 |
| ダイオキシン | 0.918182 |
| 大気 | 0.910714 |
| 保全 | 0.900538 |
| 汚染 | 0.889868 |
| アセスメント | 0.888372 |
| 土壌 | 0.879581 |
| 公害 | 0.866667 |
| 水質 | 0.846154 |
| エコ | 0.844538 |

A=環境

群を例題として用い，これらの 3 種類のキーワードを区別する閾値を設定している．彼らの研究における問題も，適切な閾値を設定することであり，Open Directory を参照して抽出した閾値が，必ずしも多くの話題について適切であるとは限らない．

これに対して我々の手法では，キーワードのページ単位での共起関係だけでなく，キーワードのページ内での配置という新しい視点を用いることで，ロバストに話題の詳細関係の抽出を可能にしていると考えている．

自然言語処理の技術を用いて，下位語 (hyponym) を抽出した研究としては, [11] が存在する．彼らの研究では，A, such as B... といった，下位語を示すような語彙構文 (lexico-syntactic) パターンを用いて，大量の文書から上位語-下位語の組を抽出している. [10] の著者が述べているように，この研究はシソーラスを作成するのには適しているが，実際の Web ページの集合における出現頻度を反映していないため，検索に役立つとは限らない．

Web ページの内部構造を検索に利用した研究としては, [12] [13] がある．ここでは，同じ複数のキーワードからなる質問であっても，そのうちどのキーワードをタイトルに指定し，どのキーワードを本文に指定したかによって，検索されるページの内容が異なるという性質を利用している．ユーザが入力した複数のキーワードを，自動的にタイトルやテキストに振り分けることで多数の異なった質問を生成して検索エンジンに投入し，検索結果 URL の重なりの少ない質問の組を発見して，それらの検索結果を比較のためにユーザに並べて提示するシステムを提案している．これにより，同じキーワードの組を含むページ集合の中から，異なった話題をもつページ集合を分離することで，ユーザが必要とするページを探し出すことを支援している．

## 5. おわりに

本論文では，ユーザから大雑把な話題についての質問が与えられたときに，それを詳細化するような話題を表すキーワードを Web から抽出し，ユーザに提示する手法を提案した．これらのキーワードをユーザに提示することで，ユーザは Web の中にどのような話題があるかを把握し，より詳細な情報の検索を行うことが可能となる．

従来のページ単位でのキーワードの共起に基づいて関連キーワードを抽出する手法では，必ずしも主題と関係のないキーワードが多く抽出されてしまうという問題があった．また，キーワードの包含関係を用いて概念階層を抽出する手法では，上位概念となるキーワードの出現頻度に依存して，抽出される下位概念のキーワード数が不足してしまうという問題が生じていた．これに対して提案手法では，単語の共起を測る際に，キーワードが Web ページの中のタイトルに現れる場合とそれ以外に現れる場合を区別することで，主題を詳細化するキーワードを精度良く抽出することができる．

表 3 において現れるキーワード，例えば “フォリント” などは，それだけがユーザに対して提示されたとしても，その単語を始めて見たユーザにとっては，もとの話題との関連性が理解できない場合がある．特に固有名詞の場合には，例えば，フォリントがハンガリーの “通貨” であることが分かるように，もとの話題との関係を表す，一般的なキーワードを呈示する必要がある．また，逆に表 3 に現れている “橋” や “広場” は一見非常に一般的な話題を表すキーワードが抽出されているように見える．しかし，実際に “ブダペスト 橋” や “ブダペスト 広場” という質問を検索エンジンに投入してみると，これらは，“鎖橋” や “英雄広場” といった特定の地名を表していることが分かる．本論文では，詳細化キーワードは一つの単語として扱ってきたが，今後は複数のキーワードの組合せを考慮して抽出を行う予定である．

## 謝　辞

本研究の一部は，平成 14 年度科学研究費補助金特定領域研究 (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 14019048，研究代表者 田中克己) による．ここに記して謝意を表します．

## 文　献

[1] D. Butler, “Never trust a human,” Nature, vol.405, p.115, 2000.
[2] 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.

[3] H. Kawano, and T. Hasegawa, "The structure of mondou – web search engine with textual data mining," Proceedings of the 12th International Conference on Systems Engineering, pp.373–378, 1998.

[4] 小山聡，石田亨，"情報ナビゲーションへの連想ルールの適用," 電子情報通信学会論文誌，vol.J84-D-I, no.8，pp.1266-1274，2001.

[5] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, September 1994.

[6] G.W. Snedecor, and W.G. Cochran, Statistical Methods, Iowa State University Press, 1989.

[7] Google, Inc., Google Web APIs Reference, , beta 2 edition, August 2002, http://www.google.com/apis/.

[8] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，浅原正幸，"日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版," Technical Report NAIST-IS-TR99012，奈良先端科学技術大学院大学，December 1999.

[9] M. Sanderson, and B. Croft, "Deriving concept hierarchies from text," Proceedings of the 22nd ACM SIGIR Conference (SIGIR'99), pp.206–213, 1999.

[10] E. Glover, D.M. Pennock, S. Lawrence, and R. Krovetz, "Inferring hierarchical descriptions," Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02), pp.507–514, 2002.

[11] M.A. Hearst, "Automatic acquisition of hyponyms from large text corpora," Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), pp.539–545, 1992.

[12] 小山聡，田中克己，"質問の階層的構造化を用いた web 検索手法の提案," 日本データベース学会 Letters，vol.1, no.1，pp.63–66，2002.

[13] S. Oyama, and K. Tanaka, "Exploiting document structures for comparing and exploring topics on the web," The 12th International World Wide Web Conference (WWW2003), Poster Session, 2003, to appear.