修 士 論 文

ポータルサイト自動作成のための用語説明獲得

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

菅井 俊介

2004 年 3 月

修 士 論 文

ポータルサイト自動作成のための用語説明獲得

指導教官　白井 清昭

審査委員主査　白井清昭 助教授
審査委員　島津明 教授
審査委員　東条敏 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

210046 菅井 俊介

提出年月: 2004 年 2 月

概 要

ポータルサイトとは、ユーザーがあるテーマに関連した事柄を調べるときに最初に訪れることを目的に作られ、そのテーマに関する様々な情報を集約したWebサイトである。その主な構成要素は、検索エンジン、リンク集、用語集などである。現在、ポータルサイトの多くは人手により作成されているため、多くの労力を要する。そのため、与えられたテーマに沿ったポータルサイトを自動作成する技術の確立が期待される。

そこで、本研究はポータルサイトのうち用語集の作成を目的とし、用語の説明文を自動的に獲得する。用語の説明はWeb文書から動的に獲得する。なぜならWeb上には既存の辞典や辞書には掲載されていない専門性の高い用語や、造語、新語等の説明文や定義文の存在が期待されるからである。また、Webは辞書や辞典に比べ、頻繁に更新されるという特徴を持っているため利用価値が高い。一般に、ユーザーが知りたい用語を調べたり、用語集を作る場合、検索エンジンで用語説明を得ようとすることがある。しかし、膨大な検索結果の中から用語の説明が記述されているページを見つけることは困難である。これに対し、本研究はWebページから用語説明を自動的に獲得し、ユーザーに提示することを目指す。

また、同じ語でも分野により意味（語義）が異なる場合がある。そのため、目標となるポータルサイトのテーマに合致する語義を有する説明文を自動的に選別し、ユーザーにスコアの高い順に提示するシステムを提案する。具体的には、ポータルサイトのテーマを表わす分野に特有の名詞を多く含む用語説明文を見つけることにより、ポータルサイトのテーマと関連のない語義に対応した説明文を排除する。また、ポータルサイトのテーマは動的に変化するため、分野コーパスも動的に獲得し、用語説明の関連度を求める点に本研究の特徴がある。

本システムは、ポータルサイトのテーマと用語を入力とし、その用語の説明をWebから獲得する。これは以下の手続きから成る。

1. 用語の説明があると思われる候補ページの取得

   用語集に掲載する用語Xを与える。システムは検索クエリを「Xとは」、「Xは」に分け、これを検索エンジンgooに入力し、Webページを取得する。網羅性を重視し、最大2000ページのWebページを収集する。同時に文字コードをEUCに統一する。

2. 候補ページからの不必要なタグの除去、ページ整形

   文字装飾タグやスタイルタグ等のリストを用意し、１で取得した候補ページから除去する。次に、HTMLタグとそれ以外の文章が一行毎に交互に現れるように、タグの前後に改行を入れ、タグ同士が続く場合は間に空行を入れ、ページ最後に終端記号を付与する。

3. HTMLタグを利用した用語説明箇所の抽出

HTML タグを利用し、２で整形した Web ページから用語説明の書かれている箇所を抽出する。ここでは、用語が見出し語となっていて、その後に用語説明が書かれている場合と、用語が文自体に内包されている場合を考慮し、それぞれについて用語説明箇所抽出アルゴリズムを作成した。また、形態素解析を行い、後続する文が連体詞や接続詞で始まる場合、その文も用語説明として抽出する。また、抽出の際、用語説明箇所を複数の文型パターンに分類する。

4. ふさわしくない用語説明の除去

特定の文型パターン、長すぎる説明文、用語が複合語の一部となっている場合の 3 つについて、ふさわしくない用語説明を除去する。

5. 用語説明のスコア付け

ポータルサイトのテーマとの関連の深さを表わすスコアの付与。用語説明 E へのスコア付けは以下の式で行う。

$$Score(E) = \frac{1}{|E|} \sum_{n \in E} score(n) \tag{1}$$

ここで E は用語説明文、n は説明文中の名詞、$score(n)$ は名詞 n のスコア（テーマとの関連度）を表わす。$score(n)$ を求めるため、テーマをクエリとして検索を行い、テーマに関する Web 文書群を収集する。$score(n)$ として、単語の頻度や文書頻度をもとにした TF-IDF と、文書頻度をもとにした RDF の 2 通りの定義を考えた。いずれにせよ、そのテーマに特有の名詞に対し高いスコアを与えることを目的とする。最終的に、用語説明を $score(E)$ の降順に並べ、その上位の用語説明を出力する。

本システムの有効性を測るため、評価実験を行った。実験では、求めたい用語に対し、語義が１つの用語、複数の語義をもつ用語の 2 種類について用語説明を獲得した。また、それぞれの用語に対し、TF-IDF によるスコア付けと RDF によるスコア付けの両方を試した。得られた用語説明のスコア上位 10 件について、正しい用語説明が得られたかどうかを評価した。実験の結果、ほぼ全ての用語で、スコア上位 10 件にふさわしい用語説明が含まれていた。また、スコア付けは TF-IDF より RDF のほうが優れていることがわかった。語義を複数持つ用語の場合、語義が 1 つの用語に比べ、ふさわしい用語説明を抽出することが困難であった。これは多くの語義に対応した説明が多数抽出されたことや、用語説明とテーマとの関連度を測る手法がうまく働かなかったことが原因として考えられる。これを解決するため、本システムの手続き３で得られた文型パターンをスコア付けに反映させる方法が考えられる。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究の背景と目的

ポータルサイトとはインターネットの入り口となるWebサイトの事である。例として、Yahoo!、Exite、gooなどの検索エンジン系サイト、Netscape、Microsoftなどの Webブラウザメーカー系サイト、AOL、リクルートなどのコンテンツプロバイダ系サイト、So-net、BIGLOBEなどのネットワークプロバイダ系サイトなどがある。

しかし、本研究で取り扱うポータルサイトとは、ユーザーがあるテーマに関連した事柄を調べるときに最初に訪れる事を目的に作られ、そのテーマに関する様々な情報を集約したWebサイトを指す。その主な構成要素には検索エンジン、リンク集、用語集などがある。現在、ポータルサイトの多くは人手により作成されているため、多くの労力を要する。そのため、与えられたテーマに添ったポータルサイトを自動作成する技術の確立が期待される。

本研究はポータルサイトのテーマをユーザーが与えると、ポータルサイトを自動生成をするシステムを作成する。特に、ポータルサイトにおける様々なコンテンツのうち、用語集の自動作成を目的とする。用語集作成には次の二つのプロセスが必要である。

1. 用語の選別
2. 用語説明の獲得

このうち、本研究では2のプロセスの実現を目的とする。すなわち、ポータルサイトのテーマと用語集にのせる用語を入力とし、その説明を獲得する。特に、用語説明はWeb文書から動的に獲得する。なぜならWeb上には既存の辞書や辞典には掲載されていない非常に専門性の高い用語の説明文の存在が期待されるからである。また、Webは頻繁に更新されるという特性を持っているため、造語、新語等の説明文や定義文の存在も期待される。さらに、Webから用語説明を抽出することで、文の特徴だけでなく、HTMLタグやレイアウト等の情報から用語説明箇所を判定することもできる。

ある用語の意味をインターネットで調べる際、一般に検索エンジンを使用し、説明文を探すことが行われている。しかし、以下の点で困難な場合がある。

1. 用語の説明箇所自体を見つける事が困難

   一般に、数多くの検索結果の中から、適切な説明文を見つけることはユーザーにとって困難である。

2. 用語が多義であるとき、ポータルサイトのテーマに沿った用語説明を獲得するのが困難

   テーマに沿った用語説明を得るためには、ポータルサイトのテーマと検索したい語を And 検索することが考えられる。しかし、検索された文書集合に目的とする説明文や定義文が含まれていないが、Web 上には存在する場合がある。すなわち、テーマとなるキーワードが存在しないページにも、テーマに沿った用語説明が存在することがある。この時、テーマと用語との And 検索ではそのような用語説明は獲得できない。

本研究ではこれを解決し、目標となるポータルサイトのテーマに合致する語義に対する説明文を自動的に選別し、ユーザーに提示するシステムを構築する。

1.2 本論文の構成

2 章では、関連研究における用語説明抽出と多義語の取り扱いについて説明し比較することで、本研究の特色を明らかにする。3 章では、本研究の用語説明抽出システムの各モジュールについて説明する。4 章では、システムの評価実験結果を示し、提案手法と従来法との比較や考察を述べる。5 章では結論と今後の課題を述べる。

第2章　関連研究

Web を辞書や事典として利用し、Web ページ群から用語に関する説明情報を収集することで、自動的に用語の意味を調べる研究が近年盛んに行われている。これは、Web が有益無益に関わらず既存の辞書や事典を遥かにしのぐほどの膨大な情報を有し、かつ頻繁に更新される特性があるためである。

ここでは、まず、用語集の自動構築や用語あるいは用語説明に関連する研究を概観する。

Kaji らは、テキストコーパスから用語及び用語間の意味的関係を自動的に抽出し、シソーラスを生成する方法を述べ、情報検索に対しそのアプリケーションのデモを行っている [?]。シソーラスの生成は、コーパスから用語や共起データを抽出し、統計的に用語間の相互関係を分析することにより構成する。また、用語抽出のための手掛かりとなる構成要素である複合名詞の構造を解析するための新しい方法も提案している。Youngja らは、大きな文書集合から特定の分野の用語解説を自動的に抽出する方法について述べている [?]。また、桜井らは、与えられた用語に対して、その用語を説明する文を Web から収集し、それらを編集してユーザーに提示するシステムを提案している [?][?]。藤井らは、Web を事典的に利用することを目的とし、質量ともに優れた事典知識情報を生成するためのシステムを提案している [?][?][?]。山田らは、ニュース原稿を解析し、作成された用語説明を抽出するパターンを Web から用語説明を自動抽出する方法へ応用している [?][?]。

本章では、特に本研究と関連性の高い桜井ら、藤井ら、山田らによる研究を用語抽出、多義語の取り扱いといった観点から紹介し、本研究との違いを明確にする。

2.1　用語説明箇所の抽出

用語集作成における最も重要な技術として、用語説明箇所の抽出が挙げられる。これは目的となる用語を説明する文章（用語説明）を、Web から収集するということである。ここでいう用語説明とは、辞書や事典と同じく、それぞれの語や項目に対してそれらを説明する文章の事を指す。

用語説明箇所の抽出には、次の 2 つのタスクに分けることができる。

- 候補となる Web ページの取得
- 用語説明箇所の抽出

2.1.1 候補となる Web ページの取得

ある用語がシステムに入力されたならば、システムは、その用語説明が掲載されている可能性があるページ（候補ページ）をまず収集する。

桜井らは、2 つの検索エンジン[1]に対して、次の 4 種類の検索質問を入力している。

「X」,「X とは」,「X という」,「X は」

ここで、X は入力された用語を表す。それぞれのクエリに対して、最大 50 ページを収集するため、最大 200 件のページを取得する。さらに、取得したページ中に存在するアンカーに着目し、アンカーテキスト ⟨A⟩ と ⟨/A⟩ に囲まれた文字列に入力された用語が含まれていた場合、そのアンカーのリンク先ページも候補ページとして取得している。

山田らは、検索エンジン google[2]を利用し、対象用語自体をクエリとして検索した上位 10 個の Web ページから、次の 4 パターンを含むページを取得している [?]。

- 定義型リストに対象用語があるページ、もしくは、対象用語が見出し化されているページ
- 対象用語にページ内アンカーが付加されているページ
- 対象用語を主題とする説明文を含むページ
- 対象用語が連体修飾を受けている文を含むページ（連体修飾節が用語の説明となっているとき）

対象用語として、2002 年 11 月のニュース原稿に出現した最新時事用語 71 語を利用したところ、上記パターンで十分な説明が得られた割合は 60.6%であった。

また、山田らは、NHK の放送用読み原稿として利用されるニュース記事をコーパスと見立て、対象用語が説明されている記事の抽出も行っている [?]。

藤井らは、検索エンジン google を用い、対象用語をクエリとし、対象用語を含む Web ページを取得している。

2.1.2 用語説明箇所の絞り込み抽出

2.1.1 項では、候補となる Web ページの取得について触れた。しかし、取得した膨大な Web ページ群のどこに目的とする用語説明が書かれているかを絞り込まなければならない。そのため取得したそれぞれのページを調べ、用語説明が書かれた箇所を抽出する必要がある。

桜井らは、収集したそれぞれのページを調べ、用語説明が書かれた段落を抽出している。実際に抽出しているのは、(1) 内包的定義文を含む 1 段落、(2) 用語が見出しとなって

[1] goo (www.goo.ne.jp) と Inforseek Japan(www.infoseek.co.jp)

[2] www.google.co.jp

いる場合の見出しに続く 1 段落の 2 種類の段落である。ここでいう内包的定義文とは、長尾による用語定義文の分類に準拠しており、上位語を用いて用語を定義しているパターンを指す [?]。桜井らの詳細分類では、被定義語がトピックとなっている場合を直接的内包、被定義語が連体修飾句を構成している場合を間接的内包とした 2 種類に分類している。これは、用語説明は内包的定義文で始まることが多いという見識から考察された。さらに、パターンをあらかじめ用意し、内包的定義文かどうかの判定を行っている。以下に判定に用いる内包的定義文パターン例を示す。

直接的内包

<X:*><*:(名詞|未定義語)>*<と:助詞>?<は:助詞><*:*><*:(名詞|未定義語):E>+<だ。:*>

間接的内包

<X:*><*:(名詞|未定義語)>*<と:助詞><(い|言)う:*>+<*:(名詞|未定義語):E>+<*:助詞>

ここで <X:*>:被定義語,<*:*>:任意の形態素,?:あってもなくてもよい,*:任意個数,+:1 個以上,E:上位語 とする。

また、見出しで始まる段落を抽出するのは、Web 上に用語集のような形式で用語に対する説明を記述しているページを想定しているためである。

次に、以下の (1)(2) で行う処理を簡単にするため、前もって 14 種類の文字装飾タグの削除を行う。その他のタグの前後に改行を挿入し、独立行とする。以降の処理は、これにより整形された行単位で行われる。

(1) 内包的定義文を含む段落の抽出

- 内包的定義文の文型パターンにマッチする行を見つける
- 次の行が空行またはタグであれば、見つかった行のみを抽出
- それ以外の場合は、その行を段落開始行とする
- 段落開始行から逆方向に行を調べ、最初に見つかった ⟨br⟩ 以外のタグを段落開始タグとして記憶する
- 段落開始行から順方向に行を調べ、以下のいずれかの条件を満たす行を段落終了行とする
  - 段落開始タグと同じタグが存在する行
  - ⟨br⟩ タグが存在する行
  - そこまでの全文の長さが 200 バイトを超える行
- 段落開始行から段落終了行までを用語説明として抽出

(2) 見出しで始まる段落の抽出

- 対象用語が含まれている行を見つける
- その行の文字数が、用語+5バイト以内の場合、見出しと判定する。これにより、「X」、「X とは」、「X は?」などが見出し語と判定される
- 見出し語の直前にある ⟨br⟩ 以外のタグを見出しタグとして記憶する
- 見出しの行以降の最初のテキスト行を、段落開始行とする。その行の直前のタグを段落開始タグとする
- 段落開始行から順方向に行を調べ、以下のいずれかの条件を満たす行を見つける。これを段落終了行とする
  - 段落開始タグと同じタグが存在する行
  - 見出しタグと同じタグが存在する行
  - ⟨br⟩ タグが存在する行
  - そこまでの全文の長さが 200 バイトを超える行
- 見出しと段落を対にして用語説明として抽出する

本研究においてもこの方法に準じて用語説明の抽出を行う。さらに、上記以外で絞り込みに有効なパターンや HTML タグ情報を独自に調査する。

藤井らは、文章表現と HTML タグを手掛かりとし、用語説明箇所を絞り込んでいる [?]。文章表現に関する手掛かりには、X とは Y、X という、などの文章表現テンプレートを 18 種定義し、絞り込みを行っている。具体的には、テンプレートにマッチした文を含む領域や、見出し・リンク先に続く一定の領域のうち、以下の条件に合う領域を抽出している。

- 対象用語が用語タグ ⟨DT⟩ でマークされている場合は、用語説明タグ ⟨DD⟩ でマークされた領域
- 段落タグ ⟨P⟩ でマークされた領域
- 箇条書きタグ ⟨UL⟩ でマークされた領域
- 抽出開始箇所から 3 文

また、HTML タグに関する手掛かりでは、⟨DT⟩、⟨B⟩、⟨H1⟩ などを用いた用語を見出しとみなし、後続する段落をその見出しに対する説明としている。

山田らは、やはり ⟨P⟩ 等の HTML タグを手掛かりとして利用している [?]。また、文の区切りは句点を手掛かりとして判定している。次に用語が含まれる文に対し、構文解析を行い、文中に「X は Y です」というパターンがある場合、Y の部分を用語 X の説明として抽出している。用語にかかる連体節がある場合は、連体節の部分を用語説明として抽出している。ただし、十分な情報量を持つ説明を抽出するため、動詞が含まれている連体節

のみを処理対象としている。

以上に共通するのは、いずれの方法においても HTML タグによる用語説明文の抽出を行う点、不必要なタグを除去している点である。これは、あらかじめ用意されたテキストや新聞記事等のコーパスとは大きく異なる点である。HTML タグは Web 特有の情報であり、用語説明抽出の手掛かりとしての利用価値が期待できる。

2.2 多義語の取り扱い

用語説明の収集によって、一般に、複数の用語説明が得られる。また、対象用語が多義語であった場合、異なる内容の用語説明が得られることもある。この場合、それぞれの用語説明は別のものとして捉えなければならない。

桜井らは、同じ内容を表す段落をグループ化し、それぞれのグループに対して、代表となる用語説明と上位語を決定している

まず、グループ化において、2 つの用語説明が同じ内容を表している（同義）かどうかを判定するために、上位語が一致しているか、内容語に重複が見られるか、を基準としている。ここで上位語として抽出しているのは、「名詞、未定義語の列」、「動詞＋こと」、「サ変名詞＋すること」のいずれかの構造を持つものである。

次に、代表語の選抜のために次の基準 1〜4 を設け、これらを順に適用し、用語説明が 1 つに絞り込めた時点で、それを代表としている。

1. 内包的定義文を含むものを選ぶ
2. 1 のうち「X は Y である」というタイプの内包的定義文を含むものを選ぶ
3. 2 のうち用語定義文を多く持つものを選ぶ
4. 3 のうち定義の種類を多く持つものを選ぶ

最後に上位語を確定する。ここでは、代表となった用語説明に基づいて上位語を決定するのではなく、グループ全体を考慮し、次の方法で上位語を決定している。

1. それぞれの用語説明から抽出された上位語のリストの中で、グループ内の用語説明中最も出現数の多いものを核語とする
2. 1 の上位語リストから、核語を文字列として含むものを取り出し、それらの中で、用語説明中の出現度が最も高いものを新たな核語とする。

例えば、「Java」の場合、核語が「言語」→「プログラミング言語」→「オブジェクト指向プログラミング言語」と変化し、詳細かつ適切な上位語を得ることができる。

藤井らは、用語説明を分野ごとに分類し、各語義に対して最適な用語説明を選択し、出力している。この目的は、関連分野 c に対して最適な用語説明 d を選択することにある。実際の処理では、まずすべての専門分野に対して $P(d|c)$ を計算し、$P(d|c)$ の値があるしきい値以上の用語説明だけを選択する。ここで $P(d|c)$ は分野 c に用語説明 d が出現する

確率であり、単語 n-gram によって推定される。その結果、対象用語が関連する分野と、分野にとって適切な用語説明を同時に特定することができる。ここで、関連分野 c は「専門語辞書」に基づき、以下の 19 の専門分野で構成されている。

> 航空・宇宙、バイオテクノロジー、ビジネス、化学、コンピュータ、土木・建築、防衛、地球環境、電気・電子、原子力・エネルギー、金融、法律、数学・物理、機械工学、医療・医学、金属、海洋・船舶、プラント、貿易

藤井らの手法は、特定の分野と関連度が高く、かつそれ自身が用語説明としてふさわしい候補を最終的な事典情報として出力する方法であり、用語説明は与えられた分野について分類している。

2.3 本研究の特色

本研究は、ポータルサイト自動作成のための用語説明を自動的に Web から獲得する手法について提案する。そのため、関連研究と多くの共通点があるが、以下の点で異なる。

- 候補となる Web ページの取得数
  - 桜井らは 1 クエリに対し最大 50 ページ、山田らは、対象用語に対し上位 10 ページを取得してくるのに対し、本研究では、1 クエリに対し、最大 1000 ページの取得を行う。これは、関連研究が応答性に優れたシステムを目指すのに対し、本研究ではより網羅性の高い説明文抽出を目指しているためである。
- 定義文の分類とパターン化
  - 藤井らは、事典の説明文から定義文のパターンを自動的に生成している。桜井らは、定義文に 13 種類のパターンを設定している。また、山田らは、説明文に 5W1H の役割を与え役割判定している。このような定義文の分類を行うと、分類パターンにマッチしない説明文は取得できない。本研究では用語説明取得のカバレッジを広げるため、定義文の分類やパターン化を行わない。ただし、用語説明としてふさわしくない文のパターンを設定し、不適切な用語説明を除去している。
- 見出し語に対する用語説明の抽出
  - 藤井らは、見出し語の後、連続する 3 文を用語説明として抽出している。桜井らは、形式段落を判定し、段落単位で抽出している。本研究では、さらに、複数の段落で用語説明が行われている場合に対しても、その用語説明全体を抽出するため、用語説明が次段落に続くかどうかを判定し、その結果によっては次段落も用語説明として抽出する。これにより、より詳細な用語説明を取得できる。

- 多義語に関する取り扱い
  - 桜井らは、多義語に対し、グループ化および上位語の選定という作業を 行う事で、一つの用語をそれぞれの語義に分けて出力している。また、藤井 らは、19 の専門分野にそれぞれの用語を振り分ける事で語義ごとに分類出 力している。山田らは多義語に関しては考慮していない。本研究では、ポータルサイトに掲載する用語説明文を抽出することが目的であるため、全ての語義を出力する必要はない。しかし、ポータルサイトのテーマに合う語義を正確に出力することが求められる。また、専門用語のコーパスをあらかじめ用意することもできない。そのため、本研究では動的に分野コーパスのようなものを作成し、ユーザーが与えるポータルサイトのテーマと説明文の類似度を測ることで、多義語へ対応している。これにより、あらかじめ分野を規定することなく、用語説明として最適な用語を選択することができる。また、分野コーパスを利用して、ある用語説明の語義がその分野に合致しているかどうかを判定する手法も藤井らの手法とは異なる。藤井らは事前に大量の分野コーパスを用意し、それから学習した単語 n-gram モデルを用いて、分野と用語説明の類似性を判定している。これに対し、本研究では分野コーパスを動的に作成するため、n-gram モデルを推定できるほど大量のコーパスを用意できない場合もある。そのため単語の文書頻度に基づいて類似性を判定する。

- システム形態
  - 桜井らは、Web を事典や辞書の代わりとして使うことが目的であるため、応答時間を考慮している。本研究では、応答時間をそれほど考慮せず、網羅性に重点をおくことで、より多くの用語説明文を抽出することを目指す。

第3章　用語説明抽出システム

3.1　システムの構成

本システムの処理の流れを図3.1に示す。

本システムでは、ポータルサイトのテーマ(単語)と用語集に掲載される対象用語が与えられる。それぞれを検索エンジンにかけ、候補ページを取得する。用語を検索した結果、得られた候補ページから、用語説明が記載されている箇所を抽出し、それらを選別する。一方、テーマに関する候補ページから、名詞群を取り出し、それぞれに対してテーマとの関連の深さを表すスコアを付ける。さらに、これらに基づいて用語説明文の候補に対するスコア付けを行う。用語が複数の語義を持ち、語義毎に異なる用語説明が得られた場合でもテーマに添った用語説明文に大きなスコアが与えられ、ポータルサイトにふさわしい用語説明が得られる。

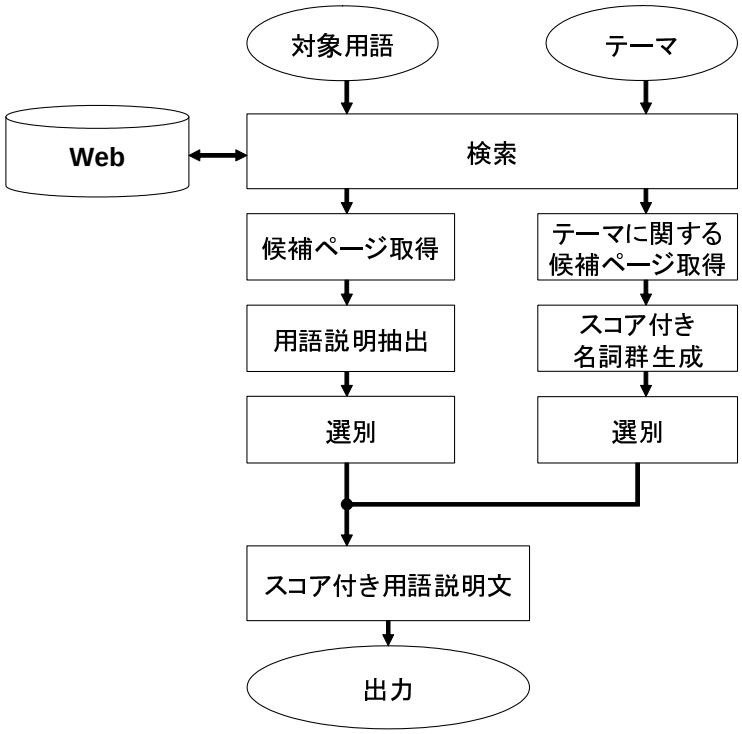

図 3.1: システムの処理の流れ

本システムは以下の 3 つのタスクからなる。

- 候補ページの取得
- 用語説明抽出
- 用語説明選抜

それぞれのタスクついて説明する。

3.2 候補ページの取得

対象となる用語と、それに続く助詞を一まとめにしたクエリを検索エンジンへ入力することにより、クエリが記載されていると思われる候補ページを Web 上から自動的に獲得する。ここでクエリとしたのは「X とは」、「X は」の 2 種類である（X は対象となる用語）。なぜなら、対象用語 X のみを検索エンジンにかける場合に比べ、X に対する説明の取得が期待できるからである。検索エンジンには goo を使用した。以下に候補ページ取得アルゴリズムを示す。

1. 各クエリを検索エンジンに入力し、検索ヒット数と URL リストを取得
2. 検索ヒット数が 1000 を超えた場合、最大取得 URL 数を 1000 とする
3. URL リストに基づき、各 Web ページを取得
4. Web ページの文字コードを統一するため、取得した Web ページを全て EUC コードに変換

3.3 用語説明抽出

3.3.1 前処理

Web ページの中の記述は HTML タグにより装飾されていることがある。そのため、用語説明箇所抽出の前段階として文字装飾タグやスタイルシート等を除く必要がある。しかし全てのタグを除去しプレーンテキストにするわけではない。HTML タグは用語説明箇所を特定するため有効に使える。除去するタグの一覧を表 3.1 に示す。pre タグは入力テキストをそのまま表示させるタグであるため、どこまでが用語の説明であるのか判断する事が困難である。そのため除去対象とした。なお、本研究で用いた HTML タグは HTML4.0 に準拠する。

不必要な装飾タグを除去したのち、各 Web ページに対し次の操作を行う。まず全ての文およびタグを一行に整形し、スペースやタブによる空白を除去する。次に、タグの前後に改行を、タグが続く場合にはタグ同士の間に空行を挿入する。最後の行に::last::を加える。これらの規則を用いたページ整形を図 3.2 に示す。

表 3.1: 除去するタグ一覧

装飾タグ類	⟨a⟩ ⟨b⟩ ⟨i⟩ ⟨s⟩ ⟨u⟩ ⟨tt⟩ ⟨cite⟩ ⟨font⟩ ⟨small⟩ ⟨big⟩ ⟨sub⟩ ⟨sup⟩ ⟨strong⟩ ⟨strike⟩ ⟨span⟩ ⟨em⟩ ⟨blink⟩ ⟨ruby⟩ ⟨rt⟩ ⟨rp⟩ ⟨marquee⟩ ⟨blockquote⟩ ⟨img⟩ ⟨inputtype=⟩
特殊記号	< > " & &#1〜¦
スタイルシート	⟨style⟩ 〜 ⟨/style⟩
pre タグ	⟨pre⟩ 〜 ⟨/pre⟩

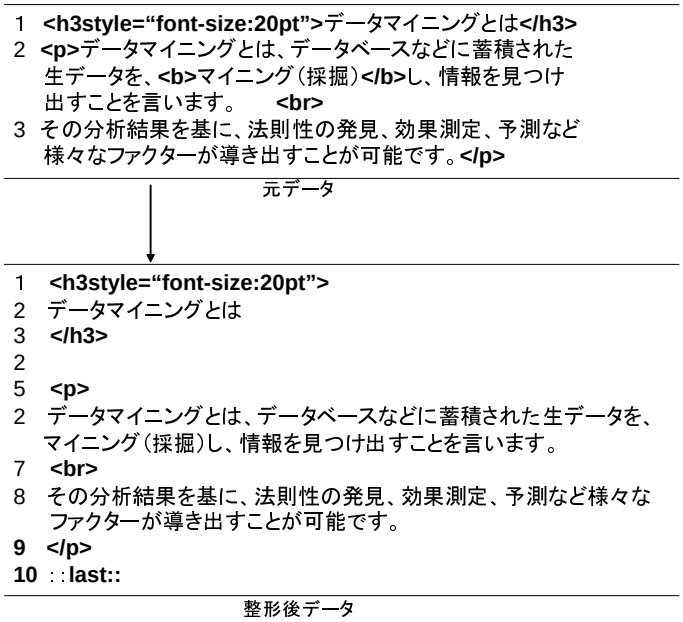

図 3.2: ページ整形の実行例

3.3.2 用語説明抽出

ここでは3.2節で取得したWebページから、用語説明箇所を抽出する。なお、以下の処理は前述の前処理によって、図3.2のように整形し、標準化された行単位で行う。図3.3にクエリ「Xとは」で検索・収集されたWebページからの用語説明箇所抽出アルゴリズムをフローチャート形式で示す。以下、図3.3のアルゴリズムを説明する。

まず始めに、行を入力する。
次にクエリ「Xとは」を探す。「Xとは」の場合、次の2つのパターンがある。

1. 見出し語パターン

   ⟨H3⟩Xとは？⟨/H3⟩
   ⟨P⟩– である。⟨/P⟩

2. 基本パターン

   ⟨P⟩Xとは – – – – – – です。⟨/BR⟩ その – – – – – –。⟨/P⟩

見出し語パターンとは、⟨H⟩タグ等で用語が見出しのような形で出現し、その下に説明が続く場合を指す。一方、基本パターンとは、用語の後に続けて説明文が現れる場合をさす。以下、それぞれのパターンに対する用語説明の抽出方法について述べる。なお、出力は1つのクエリに対し、1つのテキストファイルとして保存される。

見出し語パターン

入力された行がクエリ＋4文字以下の文かどうかを判定する。クエリ＋4文字以下の文には、「Xとは何か？」や「Xとは…」と言った見出し語であるもの、またページタイトルであるものなどがある。そのため、クエリ＋4文字以下の文であった場合、一つ前のタグを調べ、⟨title⟩、⟨dt⟩、その他、の3つの場合に分ける。以下でそれぞれの場合について説明する。

- **⟨title⟩の場合**
  次の行から最後の行（:: last::）までを調べ、「X」もしくは、クエリ「Xとは」があるかどうかを調べる。あれば次の行を開始地点としてアルゴリズムを最初から適用し直す。無い場合にはそのWebページには目的とする用語説明が無いと考え、次のWebページに対象を移す。

- **⟨dt⟩の場合**
  ⟨dt⟩は⟨dd⟩とセットで用いられ、語の定義を表す。そのため、次の行から順に見ていき、初めに現れる⟨dd⟩の次の行を用語の定義部として出力する。

- **その他**
  ⟨title⟩⟨dt⟩ 以外のタグであった場合、一つ後のタグを調べ、⟨br⟩ または一つ前と同じタグかどうかを判定する。違った場合には次の行へ進む。同じだった場合、この行を見出し語と判定する。このとき、解説文がすぐ後に来る可能性が高い。そのため、次の行から最大 100 行までを調べ、最初に出現した 5 文字以上の文を見出し語に対する解説文として出力する。さらに次の行を茶筌 [?] によって形態素解析する。これは、続く文が用語に対する説明であるかどうかを判定するためである。茶筌をかけた結果、次の行が接続詞もしくは連体詞で始まっている文であった場合、その行を見出し語に対する解説文の続きとして出力する。

基本パターン

入力された行がクエリ＋5 文字以上で構成される文であった場合、前後の行のタグを調べる。前後のタグは、同一である、後のタグが ⟨br⟩ である、その他、の 3 つの場合に分かれる。以下でそれぞれの場合について説明する。なお、クエリ自体が文の中に内包され、用語説明を構成する文をここでは用語内包文と呼ぶ。

- **同一の場合**
  用語内包文として出力する。

- **一つ後のタグが ⟨br⟩ の場合**
  この文を用語内包文 2 として出力する。さらに次の行を茶筌によって形態素解析する。茶筌をかけた結果、次の行が接続詞もしくは連体詞で始まっている文であった場合、その行を用語内包文 2 の続きとして出力する。

- **その他の場合**
  この文を用語内包文 3 として出力する。上記と同様に、茶筌を用い次の行が説明の続きかどうかを判定する。

本研究では、従来の研究とは異なり、1 つの段落もしくは 3 文のみといったように、あらかじめ用語説明の長さを固定するのではなく、接続詞や連体詞でつながる文も抽出することが出来るようにした。また、この時点で、獲得された用語説明がタイトル・見出し語・見出し語に続く解説文・用語内包文・用語内包文 2 ＋さらに続く文・用語内包文 3 ＋さらに続く文、のいずれでのパターンで取得されたかを判定できる。この分類は現時点では後続の処理に利用してはいないが、獲得された用語説明を詳細に分類する際の手掛かりとなる可能性がある。上記の分類の利用方法を検討することは今後の課題である。

ここでは検索クエリが「X とは」の場合の説明をしたが、「X は」の場合には、見出し語パターンの処理を行わない。なぜなら「X は」は「X とは」に比べ、見出し語として Web 上で利用されている例が少ない。しかし、「X は」は「X とは」に比べ、獲得できる Web ページが多いため基本パターンだけでも十分な結果が得られる。その他の処理は「X とは」と同一である。

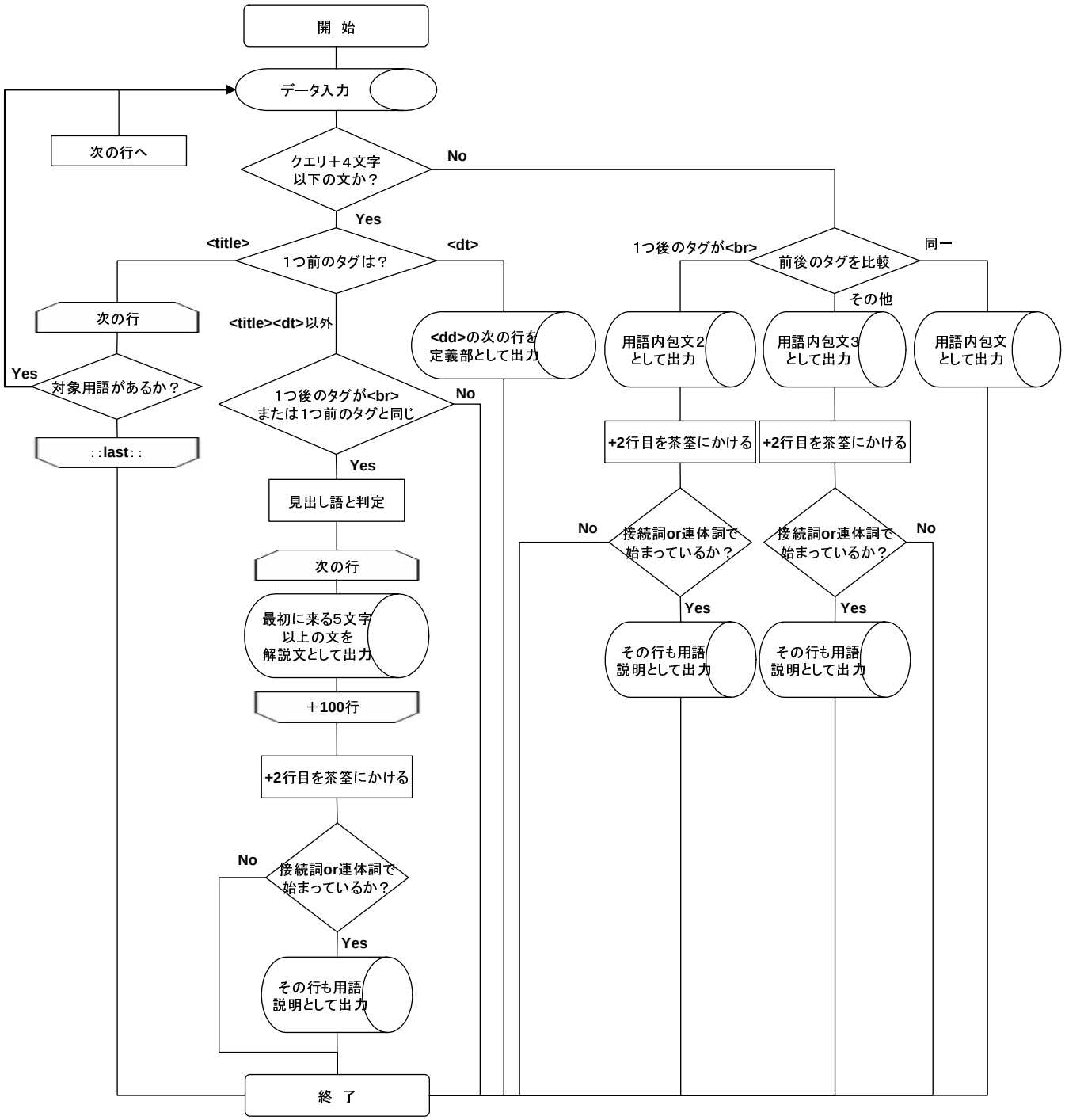

図 3.3: 用語説明箇所抽出アルゴリズム

3.4 用語説明選抜

3.3 で抽出した用語説明候補は、Web から得られた説明に簡単な選抜をしたに過ぎない。そこで、本節ではより詳細な選抜を行う。これにより、多義語に対してもテーマに添った用語説明を抽出する。用語説明選抜は大きく分けて、ふさわしくない用語説明の除去、用語説明へのスコア付けの 2 つのタスクから成る。

3.4.1 ふさわしくない用語説明

ふさわしくない用語説明には次の 3 つのパターンがある。

1. ふさわしくない文型パターン
2. 用語説明文の文字数の制限
3. 複合語の制限

以下でそれぞれを説明する。

ふさわしくない文型パターン

3.2 節での用語説明候補抽出では、明らかに用語説明としてふさわしくない文が含まれることがわかった。このような文のパターンを 78 種類特定した。そしてこのパターンに合致するものを候補から除くことにより、用語説明の絞り込みを行った。表 3.2 にふさわしくない文型パターンを示す。また、その一部について、特徴と例を以下に示す。

X とは。	(とは で終わる文)
X とは（一体）（何｜なに｜なん）〜ですか？	(X に対する疑問文)
X とは異なり〜	(X との比較文)
X とは …	(感嘆詞・記号等で終わる文)
アルファベットや記号のみ	(明らかに説明文になっていない文)
Y と X とは〜で〜ある。	(特殊な文型)

しかし、表 3.2 の文型パターンでフィルタリングを行った場合、次のような用語説明もマッチする。

> バイパスとは、交通渋滞の激しい道路の混雑を解消するために、その区間を迂回してつくる道路です。ではバイパスが交通渋滞緩和に本当に役立っているのでしょうか？

上記の例は用語説明としてふさわしいため、除去の対象から外す必要がある。この文の特徴は、「X とは」〜「でしょうか？」の間に句点があることである。このような場合を救済文型パターンとし、除去対象に含めないことにした。表 3.3 に救済文型パターンの一覧を示す。

表 3.2: 用語説明にふさわしくない文型パターン

X とは思えない	X とはオカシイ
X とは逆	X とは判断されていない
X とは何ぞや	X とは判断されない
X とは言っても	X とは縁がない
X とはいっても	X とは縁がなく
X とは言えない	X とは無関係
X とはいえない	X とは関係ありま
X とは言いません	X とは関係ない
X とはいいません	X とは関係なく
X とは言わなくても	X とは冗談きつい
X とはいわなくても	X とはじょうだんきつい
X とは見なさない	X とは呼ばない
X とは見なされない	X とは呼ばれない
X とはしないで	X とは違った
X とはならない	X とはちがった
X とはいかない	X とは違って
X とはよく言ったもの	X とはちがって
X とはよく言ったもん	X とは違う
X とは別方向	X とはちがう
X とは別の	X とは違い
X とは別です	X とはちがい
X とは雲泥の差	X とは異なり
X とは認めず	X とは異なる
X とは限らない	X とは異なって
X とは …	X とは似ている
X とは...	X とはにている
X とはどういうものですか	X とは似た
X とはどういう物ですか	X とは一体？
X とはどういう意味ですか	X とはいったい？
X とはどういうことですか	X とは～かな？
X とはどういう事ですか	X とは～ですか？
X とは？	X とは～でしょうか？
X とは?	X とは～でしょう？
X とは！	X とはどんなもの～？
X とは!	X とはどんなこと～？
X とは。	X とはどんな事～？
X とは (スペース)	X とは～どのような～？
X とはから始まり	X とは～という意味ではありません
X とはおかしい	X とは～という意味ではない

表 3.3: 救済文型パターン

X とは〜。〜ですか？
X とは〜。〜かな？
X とは〜。〜でしょうか？
X とは〜。〜でしょう？
X とはどんなもの〜。〜？
X とはどんなこと〜。〜？
X とはどんな事〜。〜？
X とは〜どのような〜。〜？
X とは〜。〜という意味ではありません
X とは〜。〜という意味ではない

用語説明文の文字数の制限

文字数の制限は 800 文字を制限とし、800 文字以上の文は説明文とみなさい。800 文字以上（原稿用紙 2 枚分）の説明は、あまりにも長いため、用語説明としてふさわしくない場合がほとんどであった。そのため、候補から除いた。

用語の複合語を制限

目的とする用語またはクエリの直前の語を調べ、それが名詞もしくは接頭詞だった場合、その文は目的とする用語ではなく、用語を含む別の複合語について説明している文であると判断し、除去する。品詞の解析には茶筌を用いた。

例として用語が「バイパス」の場合、以下のように除去するかどうかを判定する。

○ バイパスとは、交通渋滞の激しい道路の混雑を解消するために、その区間を迂回してつくる道路です。

× 西環状バイパスと北回りバイパスとは、いわゆる二環状八放射の外環状道路を形成するものです。

× 美幌バイパス

1 番目の説明文は直前の語が無いため除去対象とならない。2 番目の説明文は、「環状」の品詞が名詞-一般、「北回り」の品詞が名詞-サ変接続であるため除去対象となる。なぜなら「西環状バイパス」や「北回りバイパス」は、バイパスの具体例であり、バイパス自体の説明は得られないからである。3 番目も「美幌」の品詞が名詞-固有名詞-一般であるため除去される。このように、複合語はより細かな制限された概念を表すことが多いため、除去すべきである。また、目的とする用語自体が複合語で構成されている場合にも対応した。例えば、仮に「バイパス」が「バイ」と「パス」の 2 つの語の複合語であると解析さ

れても、「バイパス」として検出する。

3.4.2 テーマとの関連度による用語説明へのスコア付け

3.4.1 項の前処理を経た用語説明文に対し、テーマと用語説明の関連度を調べるため、用語説明文中の名詞に着目した。用語説明文 (E) に対し、個々の名詞とテーマの関連度を式 (3.1) で計算しスコア付けを行う。

$$Score(E) = \frac{1}{|E|} \sum_{n \in E} score(n) \tag{3.1}$$

ここで E は用語説明文、n は用語説明文に含まれる名詞、$score(n)$ は語 n に対するスコア、$|E|$ は E 中の名詞の数を表す。score(n) は、名詞 n とポータルサイトのテーマとの関連の深さを表わすスコアである。

すなわち、用語説明文に含まれる全ての名詞について、そのスコアの平均を求め、用語説明文全体のスコアとする。そして、式 (3.1) で得られたスコア付き用語説明文を、スコアの高い順に出力する。

これにより、あらかじめ特定の分野コーパスを用意すること無く、動的にテーマに関するスコア付き名詞群を作成し、ポータルサイトのテーマに沿った用語説明文を抽出できる。

次に、$score(n)$ の計算方法について説明する。テーマに関しても同様にして候補ページを取得するが、クエリとして与えるのはテーマのみであり、助詞を付属しない。なぜならテーマに関する説明文を取得する必要はなく、テーマ自体が記載された Web ページを取得することで、テーマと関連が深いページを取得するためである。 次に、テーマに関する候補ページに特徴的な単語を探す。テーマと関連の深い単語には様々なものがあるが、ここでは名詞および未知語を対象とする。具体的には、候補ページ中の文章を茶筌で形態素解析し、表 3.4 の品詞を持つ単語を抽出する。ただし、半角文字だけからなる文は除き、それ以外の半角文字は全角文字に修正する。また、未知語には罫線文字や記号文字が多く含まれるため、これを取り除かなければならない。取り除く未知語を表 3.5 に示す。ただし、表 3.5 中のｎｂｓｐは半角スペースを意味する HTML のメタ文字である。

次に、これらの名詞群に対して、二通りのスコア付けを行った。TF-IDF スコア付けと RDF スコア付けである。

TF-IDF スコア付け

TF の大きい語、すなわち文書中に繰り返し生起する語はテーマに特有の語であると考える。しかし、汎用的な語は分野に特有の語ではないため、IDF の大きい語、すなわち語が出現する文書数の小さい語に対して、大きいスコアを与える。

TF(term frequency) は、対象となる文書群において、ある語がどれくらいの頻度で出現するかを表したものであり、以下の式で表される。

$$TF = tf(d, t) \tag{3.2}$$

IDF(inverse document frequency) は、ある語が出現する文書数が少なければ、特徴のある語であるとする考えで、以下の式で表される。

$$IDF = \log_2 \frac{N}{df(n)} \tag{3.3}$$

対象となる語 n に対するスコアは次式で表される。

$$Score(n) = TF \times IDF \tag{3.4}$$

ここで、d はある文書、t はある語、N は文書群、n は対象となる語、$df(n)$ はある語 t が出現する文書数、をそれぞれ表している。

RDF スコア付け

RDF(relative document frequency) は、用語を含む文書が全体の文書の中でどれくらいあるのかを表す相対文書頻度のことである。RDF は式 3.5 で表される。

$$RDF = \frac{df(n)}{N} \tag{3.5}$$

対象となる語 n に対するスコアは RDF であるため、次式で表される。

$$Score(n) = RDF \tag{3.6}$$

ここで、N は文書群、n は対象となる語、$df(n)$ はある語 t が出現する文書数、をそれぞれ表している。

表 3.4: テーマ候補ページのスコア付けを行う品詞

名詞
名詞-一般
名詞-固有名詞
名詞-固有名詞-一般
名詞-固有名詞-人名
名詞-固有名詞-人名-一般
名詞-固有名詞-人名-姓
名詞-固有名詞-人名-名
名詞-固有名詞-組織
名詞-固有名詞-地域
名詞-固有名詞-地域-一般
名詞-固有名詞-地域-国
名詞-副詞可能
名詞-サ変接続
名詞-形容動詞語幹
名詞-ナイ形容詞語幹
未知語

表 3.5: 取り除く未知語

┃ ━ │ ┌ ┐ ┘ └ ├ ┬ ┤ ┴ ┼ ━ ┃ ┌ ┐ ┘ └ ├ ┬ ┤ ┴
┼ ├ ┬ ┤ ┴ ┼ ├ ┬ ┤ ┴ ┼ 、 。 ， ． ・ ： ； ？ ！ ゛ ° ´
` ¨ ^ ¯ ＿ ヽ ヾ ゝ ゞ 〃 仝 々 〆 〇 ー ― ‐ ／ ＼ ～ ‖ ｜ …
‥ ‘ ’ “ ” （ ） 〔 〕 ［ ］ ｛ ｝ 〈 〉 《 》 「 」 『 』 【 】
＋ － ± × ÷ ＝ ≠ ＜ ＞ ≦ ≧ ∞ ∴ ♂ ♀ ° ′ ″ ℃ ¥ ＄ ¢ £
％ ＃ ＆ ＊ ＠ § ☆ ★ ○ ● ≡ ∇ ∂ ⌒ ⊥ ∠ ∬ ∫ ∵ ∝ ∞ √ ≫
≪ Å ‰ ♯ ♭ ♪ † ‡ ¶ ◯ ◎ ◇ ◆ □ ■ △ ▲ ▽ ▼ ※ 〒 →
← ↑ ↓ 〓 ・ ∈ ∋ ⊆ ⊇ ⊂ ⊃ ∪ ∩ ∧ ∨ ¬ ⇒ ⇔ ∀ ∃ Α Β Γ
Δ Ε Ζ Η Θ Ι Κ Λ Μ Ν Ξ Ο Π Ρ Σ Τ Υ Φ Χ Ψ Ω α β
γ δ ε ζ η θ ι κ λ μ ν ξ ο π ρ σ τ υ φ χ ψ ω А
Б В Г Д Е Ё Ж З И Й К Л М Н О П Р С Т У Ф Х Ц
Ч Ш Щ Ъ Ы Ь Э Ю Я а б в г д е ё ж з и й к л м
н о п р с т у ф х ц ч ш щ ъ ы ь э ю я ≡ ∫ √
⊥ ∠ ∵ ∩ ∪ ＿ 彡 凵 n b s p

テーマを「暗号化技術」とした時のスコア付き名詞群生成の過程を図 3.4 に示す。図中のスコアは TF-IDF である。また、わかりやすさのため名詞群をスコアの降順に並べた。

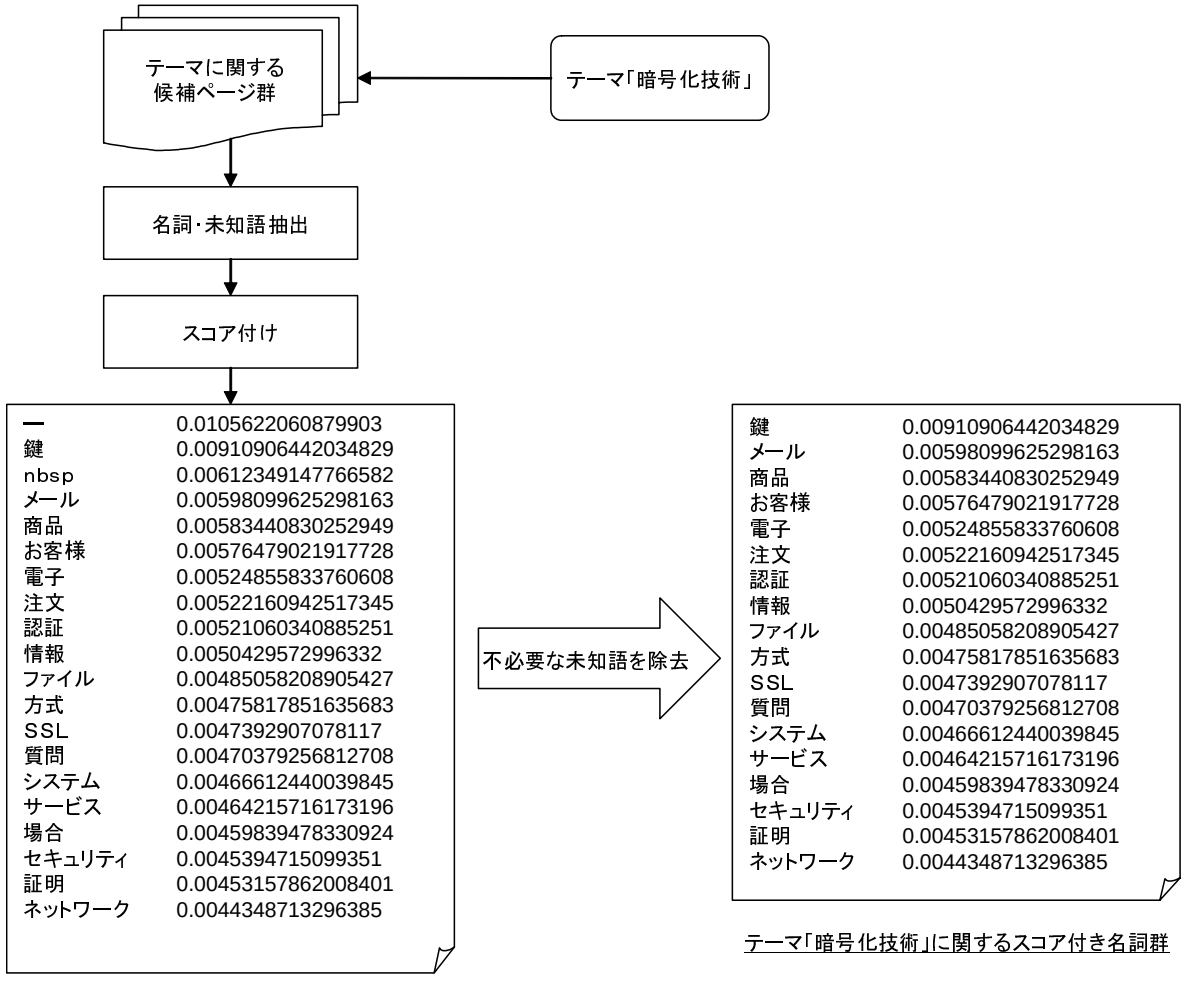

図 3.4: テーマ「暗号化技術」に関するスコア付き名詞群生成

このように、本研究では動的にポータルサイトのテーマに関連した文書を獲得し、その文書に含まれる名詞について、テーマとの関連度を定量化している。また、式 (3.1) のように、説明文中に含まれる個々の名詞の関連度スコアの平均によって、用語説明とポータルサイトのテーマとの関連度を測っている。

第4章　評価実験

4.1　実験方法

提案手法の評価実験を行った。実験では、ポータルサイトの「テーマ」と用語集に掲載する「用語」を与える。用語はそれぞれ「X とは」「X は」の 2 つのクエリに分ける。それぞれのクエリを検索し、取得した用語説明を 1 つに統合し、スコア付けを行う。語義が 1 つの用語と、語義が複数ある用語を実験した。スコアの上位 10 個の用語説明を出力した。また、それぞれについて、用語説明としてふさわしいかどうかを判定し、正解率を求めた。

実験に使用した「テーマ」と「用語」の一覧を表 4.1 表 4.2 に示す。表 4.1 は語義が 1 つの用語で使用したテーマと用語の一覧を示す。表 4.2 は語義が複数あるときの使用したテーマと用語の一覧を示す。

表 4.1: 語義が 1 つの用語の「テーマ」と「用語」一覧

テーマ	用語
動物	ヒポポタマス
	カピバラ
	エミュー
	インプリンティング
	グルーミング
	ワシントン条約
	特別天然記念物
	反芻
	ジュゴン
	ペリット
証券	青天井
	エコファンド
	外貨建て MMF
	空売り
	元本
	ディスクロージャー
	店頭有価証券
	額面割れ
	株主優待
	裁定取引

表 4.2: 語義が複数ある用語の「テーマ」と「用語」一覧

テーマ	用語
医療	キメラ
神話	キメラ
医療	バイパス
交通	バイパス
ヘリコプター	アパッチ
民族	アパッチ
情報技術	エージェント
プロ野球	エージェント
Perl	ハッシュ
暗号化技術	ハッシュ

表 4.2 のキメラの場合、テーマが医療のときは、「キメラは 1 個体の生物の中に遺伝的に異なる 2 つの細胞系が存在することである。」のような説明が、神話のときは、「キメラとは、頭がライオン、体が山羊、尻尾が蛇というギリシャ神話の中に出てくる怪獣の名前です」のような説明が取得されるのが望ましい。

4.2 実験結果

実験結果は、語義が 1 つのときと、語義が複数あるときに分けて出力した。また、それぞれの場合について、獲得 Web ページ数を示す表と、用語説明の評価表、出力された用語説明の一例を示す。

4.2.1 語義が 1 つのとき

表 4.1 のテーマに関する用語の説明文を取得した。表 4.3 で各クエリに対し取得した Web ページの数を示す。X は用語である。表からわかるように、クエリ「X+とは」に比べ「X+は」により取得した Web ページが多い。また、各クエリとクエリの合計で取得した Web ページの数を数え、平均を出した。なお、テーマ「動物」に関する候補ページは 727 ページ、テーマ「証券」では 716 ページ獲得した。この Web ページは、3.4.2 項で述べたように、獲得した用語説明文とテーマとの関連度を測るために用いる。

表 4.4 で用語説明の評価を示す。用語説明総数は、表 4.3 の「X とは」「X は」を 1 つに統合し、ふさわしくない用語説明を除いた後の用語説明の数である。正解数は、TF-IDF と RDF それぞれでスコア付けをした上位 10 件のうち、テーマに沿った用語説明が獲得できた数を表わす。そのため、分母は 10 である。一方、順位は、テーマに沿ったふさわしい用語説明が現れた順位を表わしている。

例えば、テーマが「動物」、用語が「カピバラ」のとき、得られた用語説明は 101 個で

あり、TF-IDF スコア付けでは、上位 10 件中 5 件が、RDF スコア付けでは 6 件が「カピバラ」の説明をしている。また、TF-IDF スコア付けで最初にふさわしい用語説明が現れたのは、スコア上位から 3 番目であり、RDF スコア付けでは一番ふさわしい用語説明として現れたことを意味している。

表 4.4 の一番下の行にそれぞれの平均を示す。正解数の平均は、数が大きいほどふさわしい用語説明を含むことを意味し、順位の平均は、数が小さいほど、ふさわしい用語説明が出力の上位に現れることを意味する。

表 4.3: 取得した Web ページ数：語義が 1 つ

テーマ	用語	X+とは	X+は	合計
動物	ヒポポタマス	9	27	36
	カピバラ	15	114	129
	エミュー	38	266	304
	インプリンティング	40	49	89
	グルーミング	40	269	309
	ワシントン条約	78	186	264
	特別天然記念物	7	40	47
	反芻	12	408	420
	ジュゴン	68	480	548
	ペリット	14	65	79
証券	青天井	12	42	54
	エコファンド	40	109	149
	外貨建て MMF	358	543	901
	空売り	67	290	357
	元本	38	746	784
	ディスクロージャー	91	295	386
	店頭有価証券	0	1	1
	額面割れ	7	24	31
	株主優待	55	307	362
	裁定取引	46	83	129
平均		51.75	217.2	268.95

表 4.5 は、獲得した用語説明上位 10 件の例である。これは、スコア付けに RDF を用い、テーマを「動物」、用語を「ワシントン条約」とした例である。用語説明の左上の数字は、その用語説明に与えられたスコアを表わし、降順に並べ出力している。用語説明の前の○×はその用語説明がふさわしいものであるかどうかを人手で判定した結果である。

表 4.4: 用語説明の評価

テーマ	用語	用語説明総数	正解数 TF-IDF	正解数 RDF	順位 TF-IDF	順位 RDF
動物	ヒポポタマス	14	5	6	1	1
	カピバラ	101	6	8	3	1
	エミュー	304	2	8	8	2
	インプリンティング	31	2	2	1	2
	グルーミング	223	5	6	2	1
	ワシントン条約	151	8	9	2	1
	特別天然記念物	19	7	7	3	2
	反芻	66	3	4	7	2
	ジュゴン	564	9	10	1	1
	ペリット	60	1	3	6	1
証券	青天井	26	1	2	10	6
	エコファンド	76	10	10	1	1
	外貨建て MMF	15	7	6	3	1
	空売り	297	6	6	2	1
	元本	594	0	0	0	0
	ディスクロージャー	214	0	5	0	4
	店頭有価証券	1	1	1	1	1
	額面割れ	16	6	7	2	1
	株主優待	276	5	7	5	2
	裁定取引	77	2	5	5	1
平均		156.25	4.3	5.6	3.15	1.6

表 4.5: テーマ「動物」・用語「ワシントン条約」を与えた時の RDF スコア付き用語説明

0.152
○　ワシントン条約は絶滅の危機に瀕している動物のみを保護するものだ。動物の福祉と保護という面で言えば、ＥＵ協定 36 条とＧＡＴＴ 20 条が、動物と人間の健康と生命を守る方策を採ることを定めている。
0.135
○　動物の保護に関係する法律としては、前記の動管法のほかに環境庁が所管する鳥獣保護法、通産省が所管するワシントン条約があるが、動管法は、「人が占有する動物」を対象としているのに対し、鳥獣保護法は「国内の野生の鳥類やほ乳類」を対象としており、また、同じ野生動物であってもワシントン条約は、「国際取引される希少動物」を対象としているように、対象動物によって所管省庁が異なっている。
0.123
○ 野生動物の保護のためにワシントン条約に基づく輸入規制が行われている
0.118
○ １９７２年にスェーデンのストックホルムにおいて開催された「国連人間環境会議」にて、地球上の貴重な野生動物を保護し、絶滅から護ろうという提案が示されました。
0.0962
○　＊ワシントン条約は動物が好きな人なら、ほとんどの人が知っていると思うし、そうでない人でも、ニュースで耳にしたことくらいはあると思う。世界の国々で生き物が滅びないように守ろうと作った条約です。まだ作られて 30 年しかたっていません。実際、人間が原因となって絶滅してしまった動植物は、どのくらいいるんでしょうか？
0.094
○　ワシントン条約は、野生動物の国際取引を規制して、絶滅の恐れのある野生動植物を保護することを目的とした国際条約です。
0.078
○　ワシントン条約は、世界中でほとんどいなくなったり、少なくなったりしている動物や植物をむやみに売ったり買ったりはできないようにして、守っていくためにできた約束です。少なくなっている野生の動植物を守っていくために、たくさんの国が協力しているのです。今年９月で、１５８か国がこの条約を結んでいます。
0.073
○ 毛皮の敷物、コート、象牙などの細工品など…。この部屋は滅びゆく野生動物で飾られています。ワシントン条約は絶滅の恐れのある野生動植物の国際取引を規制しています。
0.070
○　ワシントン条約は、世界中で絶（ぜつ）めつが心配されている野生の動物や植物を守るために、海外へ輸出（ゆしゅつ）したり輸入したりする時のルールを決めた国際的（こくさいてき）な約束（やくそく）ごとです。「絶めつ」は、ほろびるという意味です。
0.068
× 週末にジャッジャックウイークエンドマーケットのペットコーナーへ行った。これまた色々な犬や猫が、いや小動物から小鳥、鴨、鶏そして鰐や綺麗な色をした蛇、大蛇まで売られている。此れってワシントン条約は存在するのって感じだ。

4.2.2 語義が複数あるとき

表 4.2 のテーマに関する用語の説明文を取得した。表 4.6 は、各クエリに対し取得した Web ページの数である。また、各用語の検索クエリの合計と、クエリ別の平均を出した。また、テーマをクエリとしたときの獲得 Web ページを表 4.7 に示す。この Web ページは、3.4.2 項で述べたように、獲得した用語説明文とテーマとの関連度を測るために用いる。

表 4.6: 取得した Web ページ数：語義が複数

用語	用語＋とは	用語＋は	合計
キメラ	81	439	520
バイパス	81	804	885
アパッチ	26	374	400
エージェント	596	667	1263
ハッシュ	101	671	772
平均	177	591	768

表 4.7: テーマの獲得 Web ページ数

テーマ	獲得ページ数
医療	174
神話	625
交通	324
ヘリコプター	633
民族	662
情報技術	462
プロ野球	445
Perl	459
暗号化技術	613

表 4.8 に用語説明の評価を示す。表 4.4 と同様、用語説明総数は、表 4.6 のクエリを用語毎に 1 つに統合し、ふさわしくない用語説明を除去した数を表わす。表 4.8 で用語説明の評価を行った結果、スコア付けは TF-IDF より RDF の方がよい結果が出るため、ここでは RDF で実験を行った。RDF 欄の数字は、スコア上位 10 件中に含まれるふさわしい用語説明の数を表わす。また、正解の順序 (A) は表 4.8 の順位と同様、ふさわしい用語説明が上から何件目に初めて出力されたかを表わしている。不正解の順序 (B) は、与えられたテーマと異なる用語説明が上から何件目に初めて出現したかを表わしている。文が途中で切れていたり、形態素解析が不十分であるため出現する複合語など、用語説明としてふ

さわしくなかったり、説明文の形を成していなかったりする場合もある。そのため、表中の A と B の出現する順序を A<B かどうか、つまり A が B より上位に出現するかを調べることで評価を行う。すなわち、A<B であるとき、複数の意味の中からテーマに合った意味の用語説明をうまく選別できたことを表わす。また、表 4.4 と同様に、それぞれの数値の平均値を求める。正解の順序 (A) は値が小さいほど上位にふさわしい用語説明が来ていることを示し、不正解の順序 (B) の値が大きいほど、異なる分野の用語説明が抽出されないことを示している。さらに、A<B であれば、スコア上位に現れるものがふさわしい用語説明であると言える。

表 4.8: 用語説明の評価

テーマ	用語	用語説明総数	RDF	正解の順序 (A)	不正解の順序 (B)	A<B か？
医療	キメラ	376	2	7	6	×
神話	キメラ	376	4	1	5	○
医療	バイパス	219	1	1	2	○
交通	バイパス	219	3	1	0	○
ヘリコプター	アパッチ	228	1	9	6	×
民族	アパッチ	228	2	4	10	○
情報技術	エージェント	804	3	2	6	○
プロ野球	エージェント	804	4	3	1	×
Perl	ハッシュ	444	3	7	4	×
暗号化技術	ハッシュ	444	6	2	7	○
平均		413.3	3	3.5	5.2	

また表 4.9 に、獲得した多義語の用語説明上位 10 件の例を示す。表 4.9 は、スコア付けに RDF を用い、テーマ「プロ野球」、用語「エージェント」とした例である。用語説明の左上の数字が、用語説明に与えられたスコアを表わしている。また、○×はその用語説明がふさわしいものであるかどうかを人手で判定した結果である。ここで、語義が「情報技術」のエージェントの説明は 7 位にランクされている。1 位は「エージェント」の説明ではないが、「フリー・エージェント」の説明であり、テーマの「プロ野球」に関連していることは確かである。これにより、テーマに合った説明の選別がうまくいっていることがわかる。

表 4.9: テーマ「プロ野球」・用語「エージェント」を与えた時の RDF スコア付き用語説明

0.303 × プロ野球選手が所属球団の拘束なしに外国も含めた他球団と交渉、契約できる制度 0.274 × MLB：エージェントはまず選手捜しから 0.253 ○ 詰まるところ、選手とチームの契約や選手と企業の CM 契約を取り持つ仕事。 0.250 ○ 尚、日本プロ野球界で頻繁に使われるフリー・エージェントとは、プロで 9 年の実働実績を積んで他球団への移籍する権利を得た選手のことを指している。移籍が成立するには獲得したチームが FA 権を得た選手が所属していた球団に対し、前年度の年俸の 1.5 倍の補償金を払うか、人的補償として選手を移籍させなければならない。 0.240 ○ 代理人＝エージェントはビジネスとしては、スポンサーは複数に渡るため単価及びリスクは、プロ（球団などに対してする契約）に比べるとネガティブにならざるを得ない。だが、日本で代理人として認められているスポーツ界は前記でも言ったように、野球とサッカーだけである。野球に限っては暫定的で代理人を認め、更に「弁護士」のみである。 0.201 × パーソナルプレースエージェントはユーザーから個人情報を受け付け、モーバイルエージェントを生成して、情報を与える。 0.172 × プロ・エージェントは人材紹介業務のためのアプリケーションサービスです 0.155 ○ アメリカから生れた代理人＝エージェントは主に野球・バスケットボール・ホッケー・アメフト選手の契約代理が仕事であった。つまり、選手と競技のフィールドを提供する球団との間を主なビジネス範囲としていた。 0.148 × エージェントとは有名どころだとＪ○Ｂとか近○日○ツーリストとか日本○行とかです。 0.139 × あなたの希望条件に合う企業情報、仕事情報をあなたに代わって「エージェント」が収集し、マッチした情報を E メールでお知らせします。一度検索した条件をそのまま「エージェント」として保存することも可能です。

4.3 考察

実験の結果、ほぼ全ての用語で、スコア上位 10 件にふさわしい用語説明が含まれていた。

表 4.4 で評価した結果、スコア付けの平均正解数、順位ともに RDF スコア付けが TF-IDF スコア付けを大きく上回った。これにより、テーマに関するスコア付けは RDF のほうが優れた結果が出ることがわかった。また、RDF スコア付けによる平均順位 1.6 は、ポータルサイト自動作成のための用語説明獲得に十分応用できる。

表 4.8 では、多義語に対する用語説明の評価を行った。その結果、実験の規模は小さいが、A<B 判定により 10 件中 6 件において、テーマに関連した用語説明がその他の分野に関連した説明より上位にくることがわかった。

また、表 4.5 の「ワシントン条約」の例のように、条約が出来た年号や、関連条約、管轄省庁に関する説明を抽出し、予想以上にバラエティに富んだ用語説明を得ることができた。

語義を複数持つ用語の場合、語義が 1 つの用語に比べ、ふさわしい用語説明を抽出することが困難であった。これは多くの語義に対応した説明が多数抽出されたことや、用語説明とテーマとの関連度を測る手法がうまく働かなかったことが原因として考えられる。これを解決するため、本システムの手続き 3 で得られた文型パターンをスコア付けに反映させる方法が考えられる。文型パターンは、「定義部」、「見出し語に対する解説文」、「見出し語に対する解説文＋続く解説文」、「用語内包文」、「用語内包文 2」、「用語内包文 2 ＋続く用語説明」、「用語内包文 3」、「用語内包文 3 ＋続く用語説明」がある。これらのパターンのうち、例えば「用語内包文」や「定義部」の文型パターンで抽出された文章は用語説明としてふさわしいものが多く、一方、「見出し語に対する解説文」のパターンで抽出された文章は用語説明ではないものが多かった。したがって、用語説明として精度の低い「見出し語に対する解説文」には低いスコアを与え、また、「用語内包文」や「定義部」には高いスコアを与えるようにスコア付けの定義を修正をする方法が有効な改良案となる。

第5章　おわりに

本研究では、ポータルサイトのテーマと用語集に掲載する用語を入力すると、その用語の説明を Web から探し出し、それらをテーマと関連のありそうな順に整理して出力するシステムを実現した。本システムの特徴は、次の通りである。

- 1 つの用語に対し、最大 2000 ページもの膨大な Web データから用語説明を自動抽出する。これは多くの文書を取り扱うことで色々な用語説明を獲得するためである
- 用語説明抽出のとき、抽出する文の長さを柔軟に調整する
- 事前にコーパスを用意せず、動的に単語の文書頻度に基づき用語説明とテーマとの類似性を判定することで、多義の中からテーマにあった用語説明を獲得する

本システムを用いることにより、ポータルサイトの用語集を自動作成するための用語説明獲得が可能となった。

今後の課題として、より正確な用語説明を獲得するため、3.3 で獲得した用語説明パターンを用いスコア付けを調整する方法が考えられる。また今回スコア付けに用いた TF-IDF と RDF よりも良いスコア付けがあるかどうかを検討する。

謝辞

本研究を進めるにあたり、白井清昭 助教授ならびに島津明 教授には、数多くの御教示を頂きました。また、本研究に関して、多大な助言をしていただいた山田寛康 助手に心から感謝致します。そして、島津・白井研究室の皆様方には、研究に関する貴重な支援をして頂きましたことを心より感謝致します。