修 士 論 文

Webからの関連用語の自動獲得

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

星 正人

2004年3月

修 士 論 文

Webからの関連用語の自動獲得

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

210083 星 正人

提出年月: 2004 年 2 月

概 要

Web 上で利用可能な膨大な文書の中から必要な情報を探し出す手段として、ポータルサイトの利用が考えられる。ポータルサイトとは、あるテーマに関する基礎的な情報や、そのテーマに関連する用語集、リンク集などのコンテンツからなるサイトのことである。しかし、自分が知りたいと思っているテーマに関するポータルサイトが存在するとは限らない。したがって、テーマを入力するだけでそのテーマに関するポータルサイトを自動的に作成することができれば便利である。本研究ではポータルサイトのコンテンツのうち、用語集に注目する。さらに用語集の見出し語を抽出することに焦点を当て、用語集の見出しとなるようなポータルサイトのテーマに対する関連用語を自動的に獲得することを目的とする。

関連用語を Web 上から自動的に獲得するという処理は、大きく分けると「関連文書収集」、「用語候補抽出」、「関連度計算」の３段階の処理からなる。「関連文書収集」は、ポータルサイトのテーマに関連する文書を集めることであり、検索エンジンを用いてテーマを含む文書を収集する。「用語候補抽出」は、関連文書の中で特に良く現れると思われる用語を選び出す処理である。最後の「関連度計算」は、選び出された用語候補の中からテーマと関連性が深いと思われるものを出力する処理である。この処理は、検索エンジンを用い、テーマと用語候補の Web 上での共起確率を求め、関連度とする。具体的には、「用語候補抽出」によって出力された用語が単体で現れる文書数と、テーマと共に現れる文書数の比をとることで、その用語がどれだけテーマに特有の用語であるかということを測る。

３段階の処理のうち、２番目の「用語候補抽出」処理は、関連研究である専門用語抽出などで研究されている分野であり、本研究でもそれらを参考にしている。具体的には、名詞句を特定し、それらのスコアを計算することで用語候補にふさわしい名詞句を選ぶ処理である。そして、この段階の処理の違いにより最終的な出力が大きく変わるため、本研究ではこの段階の処理に関して３種類の手法を実装し、それぞれから得られた結果を比較し、考察した。３つの手法とは、「出現頻度に基づく手法」、「造語能力に基づく手法」、「組み合わせによる手法」である。

「出現頻度に基づく手法」は、名詞句の出現回数を数え、同じ要素数からなる名詞句の出現回数の平均で割ることによって、その名詞句が他の名詞句と比べてどれだけ多く現れているかということを求める。

「造語能力に基づく手法」は、専門用語抽出分野でよい成績を出しているとされる評価基準を用いた手法である。その評価基準とは、複合名詞の構成要素となりやすい単語ほど「造語能力」が高いとするものである。そして「造語能力」の高い単語によって構成されている用語が、その文書にとって重要であるという考えに基づき、スコアを計算する。

３つめの手法、「組み合わせによる手法」は、「出現頻度に基づく手法」と「造語能力による手法」の出力を組み合わせるものである。この段階の処理だけをみると非常に単純なものだが、「造語能力に基づく手法」、「出現頻度に基づく手法」それぞれの出力には異な

る特徴があり、さらに最終的な処理である「関連度計算」では前段階までのスコアに依存せずに関連度を求めるため、うまくいけば両手法のよい結果のみを最終的な出力とすることができる。

実験結果から、「用語候補抽出」段階の手法の違いにより、最終的な出力にもかなりの違いがみられた。単純に正解数だけを比較すると「出現頻度に基づく手法」が最も多く、「造語能力による手法」が最も少なかった。しかし、各手法が候補とする用語には特徴があり、正解数だけでは手法の優劣はつけられないことがわかった。具体的には、「出現頻度に基づく手法」は人名や製品の型番などを多く取り出す傾向があり、「造語能力に基づく手法」はテーマを部分文字列として含むものを関連用語候補として挙げやすいという傾向がみられた。後者は、本研究における関連度計算の手法にも一因がある。「造語能力に基づく手法」が候補としやすいテーマを部分文字列とする用語は、関連度の定義から、次の「関連度計算」の段階で決まって高い評価値をもつ。したがって、テーマを部分文字列とする用語が関連用語としてふさわしくないような分野に対しては、「造語能力による手法」の最終的な出力は不正解となる用語が多い。また、「関連度計算」段階の処理は共通なので、テーマを部分文字列とする用語に関する問題は「組み合わせによる手法」にも影響が現れている。しかし、この問題さえ克服できれば、「造語能力による手法」の正解率も向上し、「組み合わせによる手法」が２つの手法の異なる性質を持つ用語候補の両方を抽出できるようになると思われる。

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

インターネットの普及などにともない、膨大な量の情報が Web 上で利用できるようになった。しかし一方で、その膨大な情報の中から目的とする情報を探し出すことが困難な場合もある。Web から情報を得る際には検索エンジンがよく用いられるが、ヒット件数が多すぎる場合、または逆に全くヒットしない場合などに、キーワードを修正して必要な情報を得るためにはある種のコツのようなものが必要とされる。また、目的とする情報を含むページをみつけることができたとしても、その中に知らない用語などが含まれていれば、それについてさらに調べることが必要となる場合もある。

Web 上から必要な情報を効率よく入手するための方法としてポータルサイトの利用が考えられる。ここでいうポータルサイトとは、ある分野に関する基礎的な情報やさらに詳しい情報源などを提供するサイトのことである。しかし、自分が求めるテーマに関するポータルサイトがあるとは限らないし、あったとしてもそれを検索エンジン等を用いて探すのでは、手間がかかったり見落としてしまう可能性もある。そこで、自分が必要とするテーマに関するポータルサイトを自動的に生成できれば、必要な情報を手にいれるための手助けとなると考えられる。

ポータルサイトを自動生成するということは、「利用者が知りたいと思っているテーマを入力するだけで、そのテーマに関するポータルサイトが表示される」ということである。生成すべきポータルサイトのコンテンツとしては、テーマに関連する用語集、リンク集、FAQ などが考えられ、それぞれを自動的に作成する。これらのうち、ここでは用語集の自動生成に焦点をあてる。用語集の自動生成は大きく分けて以下の２つの手続きからなる。

1. テーマに関連する用語（見出し語）の獲得
2. 用語に対する説明文の獲得

本研究では、上記の１を行う。つまり、ポータルサイトのテーマを入力として受け取り、それに関連する用語を Web 上から自動的に獲得し、出力することを目的とする。

1.2　本論文の構成

２章では、関連研究について簡単に説明する。３章では、本研究の考え方やシステムについて、関連研究と比較しながら説明する。４章では実験方法と実験結果を示し、考察を行う。５章では本研究のまとめと提案手法の改善のための今後の課題などを述べる。

第2章　関連研究

2.1　専門用語抽出

ある分野の文書の中から重要と思われる用語を求める研究として、重要語抽出や専門用語抽出と呼ばれる研究がある。簡単に言えば、与えられた文書からその文書を特徴づける用語、あるいはその文書特有の用語などを重要語、専門用語として探し出す研究である。具体的な方法としては、文書内での出現頻度等の統計的情報を利用する手法が多い。これらは、情報検索（Infomation Retrieval）の研究分野であり、旧学術情報センター（NSCSIS）（現国立情報学研究所）[1]による NTCIR(NACSIS Test Collection for Information Retrieval systems) のタスクとしてコンテストが行われたこともある。そのテストコレクションは公開されているため、異なる手法の精度の比較なども可能である。

これらの研究の代表的なものとして中川ら [1][2][3] の研究がある。中川らの手法は、NTCIR テストコレクションにおいて良い成績を出していることが示されている。中川らは、出現頻度ではなく連接頻度に注目している。連接頻度とは、ある語に連接する語の異なり数の多さのことであり、ある語の直前または直後に現れる語の種類の多さのことである。すなわち、その語（単名詞）がいかに多くの複合名詞に含まれやすいか、言い換えればいかに複合名詞を構成しやすいかということを計る。これにより、各単名詞にスコアを与え、さらに複合名詞においては構成要素である単名詞のスコアの相乗平均をとることで複合名詞全体のスコアを求める。この結果から、単名詞と複合名詞を同様のスコアで比較し、そのスコアが高いものを専門用語として抽出している。

2.2　Web を利用した用語抽出

自然言語処理の分野において、Web 上の情報をいかに有効に利用するかという研究も盛んに行われている。Web 検索エンジンもそれらの研究の一部ということもできるが、さらに発展させて利便性を高めるための研究が行われている。例えば藤井ら [4][5][6] は、World Wide Web を辞典のように利用するための研究として、Web テキストに出現する新語の検出や、HTML タグ情報などをもとにした説明文の抽出を行っている。また、清田ら [7] は Web 文書を自動的に集め、HTML タグ情報や頻度情報に基づき重要文を求め、自動要約やインデックス作成を行っている。

[1] http://www.nii.ac.jp/

Web を対象とする研究のなかにも、専門用語抽出に似た研究もある。重要語抽出や専門用語抽出が与えられた文書を抽出の対象としているのに対し、Web を対象とする研究では処理対象を動的に変化させられるなどの特徴がある。Web を対象とする用語抽出研究の中でも、佐藤ら [8] の研究は本研究と関連性が非常に高い。佐藤らの研究とは細かなところで違いはあるものの、「ある用語を受取り、それに関連する用語群を返す」という全体像が同じであり、さらに内部処理も似た部分が多い。佐藤らの研究の具体的な方法については第 3 章において本研究との比較として詳しく述べるが、ここでは全体的な流れや考え方について示す。

佐藤らは、「専門用語」というものを以下のようにとらえている。

1. 特定の分野で広く、または、それなりに使われている。
2. 一般語ではない。
3. 定義や説明が存在する。
4. 関連する専門用語（関連用語）が存在する。

そして「用語は分野を指す」という考えに基づき、「関連用語」というものを、「与えられた用語（が示す分野）の専門用語」としてとらえている。

佐藤らの提案方法は、以下の 3 ステップからなる。

1. コーパス作成
2. 重要語抽出
3. フィルタリング

第 1 のステップでは、「与えられた用語が示す分野」としてのコーパスを Web から獲得する。簡単に言えば与えられた用語に関連する文書を集めるということになる。第 2 のステップでは、中川らの方法に基づき、重要語を抽出する。第 3 のステップは、Web を対象とする用語収集の最大の特徴ともいえる、「関連文書以外の文書」の情報をもとに出現頻度情報の比較を行うことで、関連用語としての妥当性を測っている。この第 3 のステップに関しては、ほぼ同じことを本研究でも行っているので、第 3 章で詳しく述べる。

第3章　関連用語自動獲得システム

3.1　関連用語の自動獲得

提案手法を説明する前に、本研究の基本的な考え方を述べる。本研究の目的は、ポータルサイトのテーマとして与えられた用語（キーワード）に関連性の高い用語を動的に獲得し、獲得された用語を「関連用語」としてポータルサイトの用語集の見出し語として出力することである。この「関連用語」という考え方については、「専門用語」よりは少し自由度の高いものとして考えている。

そもそも「専門用語」という概念の定義も明確ではないが、佐藤らの先行研究などから、「ある分野特有の用語」という印象を受ける。佐藤らの研究ではそれを明確に示すものとして「一般語ではない」ことを条件とし、内部処理において一般語と判断されたものは候補から除外している。しかし、本研究の目的はポータルサイトの用語集を作るための見出し語の獲得であり、与えられたテーマに関連性が深いものであれば、たとえそれが一般語であっても用語集に含めるべきだと考える。よって本研究では「専門用語」ではなく「関連用語」という表現を用いる。

しかし、佐藤らの研究もまた「関連用語」という表現を用いている。佐藤らの「関連用語」という考え方を簡単にまとめれば、「与えられた用語（が示す分野）に包含される分野の専門用語」ということになる。しかし、本研究では包含関係という考えによって関連用語を定義しない。ポータルサイトのテーマとなる語とよく共起する用語は関連性が高いとみなし、関連性さえ高ければ、それは用語集に含めるべきという考えに基づく。

また、関連研究としての「専門用語抽出」の説明でも触れたが、「専門用語抽出」というのは一般に「ある分野の文書を入力とし、その文書内での重要な語を出力するタスク」を意味する。それに対し、本研究は入力は用語（ポータルサイトのテーマ）であり、その用語と関連のある用語の集合を出力とする。したがって、用語抽出の対象とする文書集合が動的に変化するという点などで自由度が高い。よって本研究全体を「用語抽出」と表現するのもふさわしくないと考えられる。

このような理由から、本研究は「専門用語抽出」ではなく「関連用語の自動獲得」と表現する。また、関連研究である佐藤らのいう「関連用語」と本研究での「関連用語」では上記のように考え方に違いがある。しかし、内部処理としては「専門用語抽出」の手法を取り入れているし、佐藤らの研究とも全体的な考え方や処理などは類似点も多い。次節以降、先行研究などとの類似点、相違点なども含め、具体的な手法について説明する。

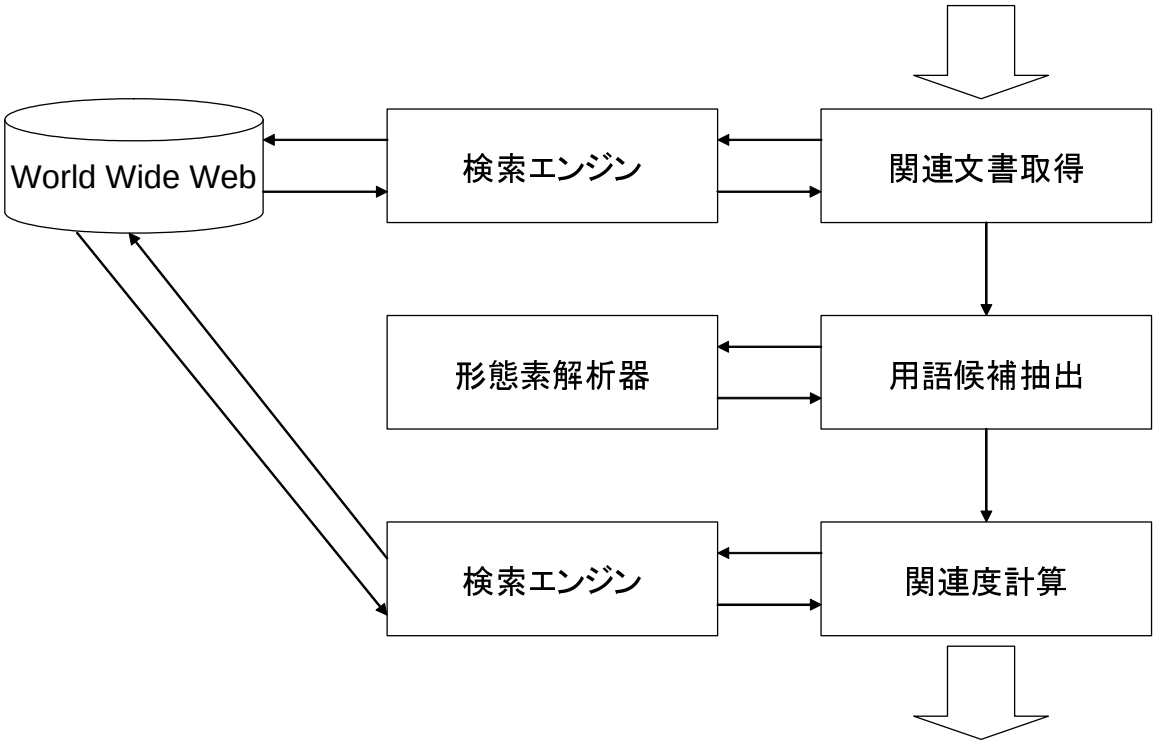

図 3.1: システム概要図

3.2 システム概要

本研究で作成したシステムは、入力としてポータルサイトのテーマ（キーワード）を受取り、それに関連性が深いと考えられる用語群を出力する。このシステムは大きく分けると以下のような構成要素からなる。

- 関連文書取得
- 用語候補抽出
- 検索エンジンを用いた関連度計算

処理の流れは図 3.1 のようになる。各ステップ間で受け渡されるデータは、「関連文書取得」から「用語候補抽出」へがテーマに関連する文書集合、「用語候補抽出」から「関連度計算」へが順位付けされた用語候補群となる。

以下、それぞれについて詳しく説明する。

3.3 関連文書取得

関連研究として挙げている重要語抽出などの研究は、処理対象が文書（群）である。それに対し、本研究はキーワードを入力として受け取り、まずはそのキーワードに関連する

文書群を自動的に収集することから始める。そしてこの段階の処理では、収集した文書群を次の段階の処理で扱いやすい形に整え、出力する。Web 上から関連文書を集めることで、コーパスが用意されていないような分野の用語集を作ることもでき、さらに新語や未知語などを含む文章を対象とすることができる。

本研究での具体的な処理は、以下の 3 段階からなる。

- 関連ページ URL 取得
- 関連ページ取得
- 文書整形

それぞれについて説明する。

関連ページ URL 取得

システム全体の入力として受け取ったキーワードを、検索エンジン goo[1]に渡し、出力ページを取得する。一回の処理では上位５０件までの結果しか得られないので、繰り返し処理により上位５００件分までの結果を取得する。そしてパターンマッチにより、参照先ページ（キーワードを含むページ）の URL を取得する。

佐藤らは、この段階の処理として goo の他に infoseek[2]を用いている。さらに goo に渡すキーワードを「(用語）とは」「(用語）という」「(用語）は」「(用語)」の４種類にしている。これは、用語に関する説明文を得るためであるが、本研究では説明文の獲得は対象外としたため、これらの処理は行っていない。

関連ページ取得

「関連ページ URL 取得」によって得られた URL のページを取得する。ただし、Web 上には HTML 以外の文書も存在し、検索エンジンはそれらも含めて検索結果とする。pdf 形式をはじめとする HTML 以外の文書は今後の処理で用いることができないので、除外する。具体的な処理としては、拡張子が「.htm」「. html」で終わっている URL と、「/」で終わっている URL のみを取得している。

文書整形

ここまでで、最大５００件の HTML ファイルが取得されている。まず下処理としてこれらの HTML ファイルの文字コードの統一を行う。

次に、これらのファイルをどのような形で出力するかというについて検討を行った。今

[1]http://www.goo.ne.jp/
[2]http://www.infoseek.co.jp

後の処理において HTML タグ情報を用いるのであれば HTML 形式のままにする必要がある。ある単語が用語としてふさわしいかを詳細な HTML タグのパターンによって判定するという手法も考えたが、最終的に本研究では「ある単語が HTML タグに囲まれていたかどうか」という情報のみを用いることにした。よって HTML タグを改行に置き換え、改行の連続は一つの改行に置き換えたプレーンテキスト形式にした。

また、５００件の文書（ファイル）を個別のものとして扱うか、一つの文書集合として扱うかについても検討した。IDF などの「出現する文書数」を用いる処理を行うのであれば、文書は個別のままにするか、単一の文書にするならば文書の境界を明らかにしなければならない。しかし、この情報は本研究では利用しないことにしたので、収集された最大５００件の文書を一つの文書にまとめた。

よって、「関連文書取得」段階から「用語候補抽出」段階へと渡されるデータは、キーワードに関連する文書群を１ファイルに結合したプレーンテキストファイルになる。

3.4 用語候補抽出

関連文書から、その文書内で重要と思われる用語を選びだす段階である。前段階の処理である「関連文書取得」で得られたテキストから関連用語の候補を選び、次の処理である「検索エンジンを用いた関連度計算」に選出した候補を渡す処理である。具体的には、後述する３つの手法それぞれによって選ばれた候補上位各１００語を関連度計算の段階に渡す。

この「用語候補抽出」の段階の処理は、文書集合からの「専門用語抽出」や「重要語抽出」そのものであり、本研究でもそれらの手法を参考にしている。しかし、本研究はこの段階で完結しない。次の段階の処理「検索エンジンを用いた関連度計算」は、Web テキストを対象とする研究が特定の文書集合からの専門用語抽出と最も異なる点であり、最大の特徴とも言える。専門用語抽出では、与えられた文書のみで再現率と精度を向上させるような処理をしなければならない。しかし Web テキストを対象とする本研究では、精度を向上させるのは次の関連度計算の役割となる。つまり、この用語抽出の段階では再現率を向上させることに重点を置く。

用語候補抽出の処理は、さらに以下の２つに分けられる。

- 名詞句抽出
- 名詞句に対するスコア付け

それぞれの処理について、以下で詳しく述べる。

3.4.1 名詞句抽出

用語抽出段階の処理をするにあたっては、まず「用語」をどうとらえるかという問題がある。この問題は難しく、様々な議論が行われているが、本研究では単純に「名詞句」に

限定する。「ある分野特有の表現」という意味では、動詞などの名詞以外の品詞を含む場合も考えられる。しかし、それらが「用語」として適切であると考えられるのは非常に稀なケースである。また、処理効率の観点も考慮し、それらは対象としないことにした。

本研究では、最終的なシステムの実装に先立ち、いくつかの用語候補抽出法を検討し、予備実験なども行った。それらについて簡単に説明し、本研究で最終的に「用語」の候補とした「名詞句」について説明する。

用語候補の検討

3.5 節で述べるように、候補となっている用語が「関連用語」としてふさわしいかを示す尺度「関連度」を Web 上での共起確率をもとに求める本研究では、どのような文字列を候補としても「関連度計算」を経ることである程度妥当な結果が得られるはずである。つまり名詞、動詞、ある程度の長さを持つ文、さらには記号の連続であっても、それが「その分野で特有の表現である」（他の分野では現れない）かどうかを判断することができる。

そこで最も単純な方法として考えられるのは、文字単位での n-gram すべてを候補とし、関連度計算を行うという方法だ。これによって得られる結果は既存の辞書情報などにまったく依存しない上、たった一度でも現れた表現に関してもスコア付けが行われるため、もっとも理想的な結果が得られると考えられる。しかし、たった１００文字の文書であっても最大で５０５０回もの関連度計算を行う必要があり、処理効率の観点から現実的ではない。

現実的に考えられる手法として、文字単位の n-gram に対して、文字数の最大値を決め、それらの出現頻度を数えて頻度の高いもののみ関連度計算を行うという方法を検討した。この方法も既存の辞書情報に依存しないという特徴を持つ。しかし、やはり n-gram すべてを候補とするのは無駄が多い。計算量を減らすために文字数の最大値を小さくすれば、用語として適切であるかも知れないものを候補にすることができなくなってしまう。そして、文字の頻度で用語候補の優先順位付けを行うため、頻度の高い助詞などの機能語が優先されやすいという問題があり、この手法も採用しなかった。

また、文字数を制限せず、かつ辞書情報等に依存しない方法として、候補の長さを一文字ずつのばしたときの出現頻度を比較するという方法を検討した。この方法の基本的な考え方は以下の通りである。例えば「自然言語処理」に関連する文書内での文字列の出現頻度を考える場合、「自」と「自然」と「自然言」の出現回数は「自然言語」の出現回数とあまり変わらないと考えられる。その場合、用語としては「自然」などは不適切であり、最も長い文字列である「自然言語」が適切であろう。よって、この段階で「自然言語」を候補に加える。また、「自然言語」に続く文字は何種類か考えられる。「の」や「は」などの助詞や、「処」などが続くと思われる。「自然言語」を候補に加えた段階で処理をやめず、さらに続く文字をみると、「自然言語処」のあとにはほぼ間違いなく「理」が来るので、「自然言語処」と「自然言語処理」との頻度はほぼ等しくなる。そのため、「自然言語処理」を候補に加える。一方、「自然言語の」や「自然言語は」などのあとにはさまざ

まな文字が続くと考えられるので、「自然言語処理の」や「自然言語処理は」に比べて、それらに一文字続いた文字列の頻度ははるかに小さくなる。そのため、「自然言語の」や「自然言語は」などは候補に加えない。このような処理を繰り返すことで用語として適切なものを選ぶことができるのではないかと考えた。この手法により用語候補を抽出する簡単な実験を行ってみたが、決まった言い回しなどの長い文字列や名詞句の途中などで区切られたものが候補に挙がったり、一種類の名詞句の部分文字列が何種類も候補に挙がるなどの問題があり、適切な用語候補を抽出することができなかった。

一方、未知語を認識する研究 [9] も行われており、ツールとして実装されているものもある。Web テキストの特徴である新語や造語を取り出すということに特化するのであれば、このようなツールを導入することも有望である。これは今後の課題として、本研究では採用しなかった。

このように、「用語」の候補をどのようにして選ぶかということに関して、いくつかの手法を検討した結果と、現実に「用語」とされているものの傾向などから判断し、「用語」の候補は「名詞句」に限定することにした。さらに重要な語であるかを判断するために「名詞句」にスコア付けを行い、スコアの高いもののみを関連度計算の段階に渡す。また、「名詞句」であることを判断するために形態素解析器を用いた。

本研究での「名詞句」

本研究で「用語」の候補とした「名詞句」に関しても、「用語」の定義と同様に、それをどうとらえるかという問題が生じる。ここでも問題を単純化し、茶筌 [10] による形態素解析結果から「単名詞」または「複合名詞」と判断されるものを「名詞句」とする。

ここでいう「単名詞」には「未知語」とされるものも含める。そして「複合名詞」は「名詞または未知語の連続」および「名詞または未知語で始まり、名詞、未知語、助詞「の」が続き、名詞または未知語で終わる句」であるものとする。後者は、「(名詞) の (名詞)」などのような表現も用語の候補に含めようという考えに基づく。

具体的には、関連文書内で上記の条件を満たす「名詞句」が現れたら、その部分文字列は候補とせず、その「名詞句」全体のみを候補に加えた。

用語候補の絞り込み

本研究における「名詞句」の定義は以上の通りである。しかし、これらすべてを候補とすると、まだ不要なものが多く含まれる。そこで、関連文書におけるすべての「名詞句」を用語候補として抽出するのではなく、以下の 2 つの条件のいずれかを満たすときに限り、用語候補として抽出する。

1. 名詞句が単独で現れる
   例　　日本語␣英語

例　　、統語解析、意味解析、文脈解析、
2. 名詞句の前後に何もない
例　⟨dt⟩形態素解析⟨dd⟩

1 に関しては、句読点やスペース、記号などで囲まれている名詞句を候補とすることを意味する。これは、関連用語はこれらの形式でよく出現するという観察に基づく。

2 に関しては、HTML タグで囲まれた名詞句を候補とする。これは、HTML 文書を対象とする処理の特徴といえる。一般的なプレーンテキスト文書を対象にする場合、1 の条件を満たすものだけを候補すれば、本来用語候補とすべき多くの名詞句を検出できないと思われる。しかし HTML 文書では、用語はタイトルタグや見出しタグなどで囲まれるだけでなく、さまざまな装飾タグによって囲まれていることが多い。下線や太文字など、強調して表現されている部分もタグに囲まれることになるし、用語としてリンクが張ってある文字列などもタグに囲まれている。つまり、Web ページの作成者が強調したり、説明などを加えている用語がこの条件に一致するということである。本研究では、HTML ファイルをプレーンテキストに変換する処理で HTML タグは改行に置き換えているので、名詞句が単独で現れる行がこれにあたる。

文書中に現れる名詞すべてを候補に加えると計算量が増大するので、本研究では上記の条件を満たす名詞句のみを候補に加えた。ただし、3.4.2 項で述べるスコアの計算においては、関連文書中のすべての名詞句の頻度情報を用いた。

3.4.2　名詞句に対するスコア付け

上記の方法により、用語の候補となる名詞句を抽出した。続いてその名詞句にスコア付けを行う。このスコアが高かったものが関連文書内で重要な用語であったということである。スコア付けに関しては三つの手法を実験した。三つの手法はそれぞれが独立したシステムとして、スコア上位１００件の用語を次の処理である関連度計算に渡す。

それぞれの手法の詳細を以下に述べる。

1. 出現頻度に基づくスコア付け
   重要語抽出などの研究でも基本となる、出現頻度に基づく方法である。対象とする文書内で多く出現する用語ほど重要であろうという考えに基づく。情報検索分野の研究の TF を求める処理であり、IDF にあたるものは考慮しない。これは、次の段階の処理「関連度計算」において IDF に相当する評価値によってスコア付けを行うため、この段階の処理では考慮する必要がないためである。
   このスコア付けの具体的な方法は、対象とする文書内での「名詞句」の出現回数を数える。実験の初期の段階では単純に出現回数で順位付けを行っていたが、その方法では短い名詞句ほど順位が高くなり、長い名詞句が候補に挙がりにくかった。そこで、出現回数を同じ n-gram（同じ長さの名詞句）の出現回数の平均で割るように修正し

た。これにより、長い名詞句であっても、相対的に頻度の高いものは高いスコアを与えた。
用語候補の出現回数を w(n) とすると、その単語のスコア score(w(n)) は式 (3.1) のようになる。n は用語候補に含まれる形態素数である。

$$score(w(n)) = \frac{w(n)}{\frac{sum_{i=1}^{j} w_i(n)}{j}} \tag{3.1}$$

2. 造語能力に基づくスコア付け
この手法は、関連研究である佐藤らの手法 [8] に基づいたものである。また、佐藤らは中川らの専門用語抽出の研究 [1] をもとにしている。
この手法の基本的な考え方は、２章でも触れたように、中川らの「連接頻度の高いものほど専門用語性が高い」という考えに基づく。この連接頻度のことを佐藤らは「造語能力」と呼んでいる。この「造語能力」を求めるためには、まず名詞句を決めなければならないが、佐藤らの研究では具体的な方法が示されていない。中川らは本研究と同じように「A の B」などの表現も対象としているが、佐藤らは実験結果を見る限りこれらを対象とせず、名詞の連続のみを名詞句としてとらえていると考えられる。しかし、具体的な説明がないため、「名詞句」の定義は本研究と同一とする。そしてスコアの計算方法は、中川らの手法に基づく。
候補語のリストを L、l 単語からなる候補を t とすると、造語能力のスコア $Imp(t, L)$ は、以下のようになる。$ws(w, L)$ は単語の造語能力で、このスコアが大きいほど複合語を作りやすいと考える。

$$Pre(w, L) = \text{“}L \text{ において、} w \text{ の直前に現れる単語の異なり数}\text{”} \tag{3.2}$$

$$Post(w, L) = \text{“}L \text{ において、} w \text{ の直後に現れる単語の異なり数}\text{”} \tag{3.3}$$

$$ws(w, L) = \sqrt{(Pre(w, L) + 1)(Post(w, L) + 1)} \tag{3.4}$$

$$Imp(t, L) = \sqrt[l]{\prod_{i=1}^{l} ws(w_i, L)} \tag{3.5}$$

3. 両手法の組み合わせ
４章で示すように、上記二つの手法が出力する用語群にはかなり異なる特徴がみられ

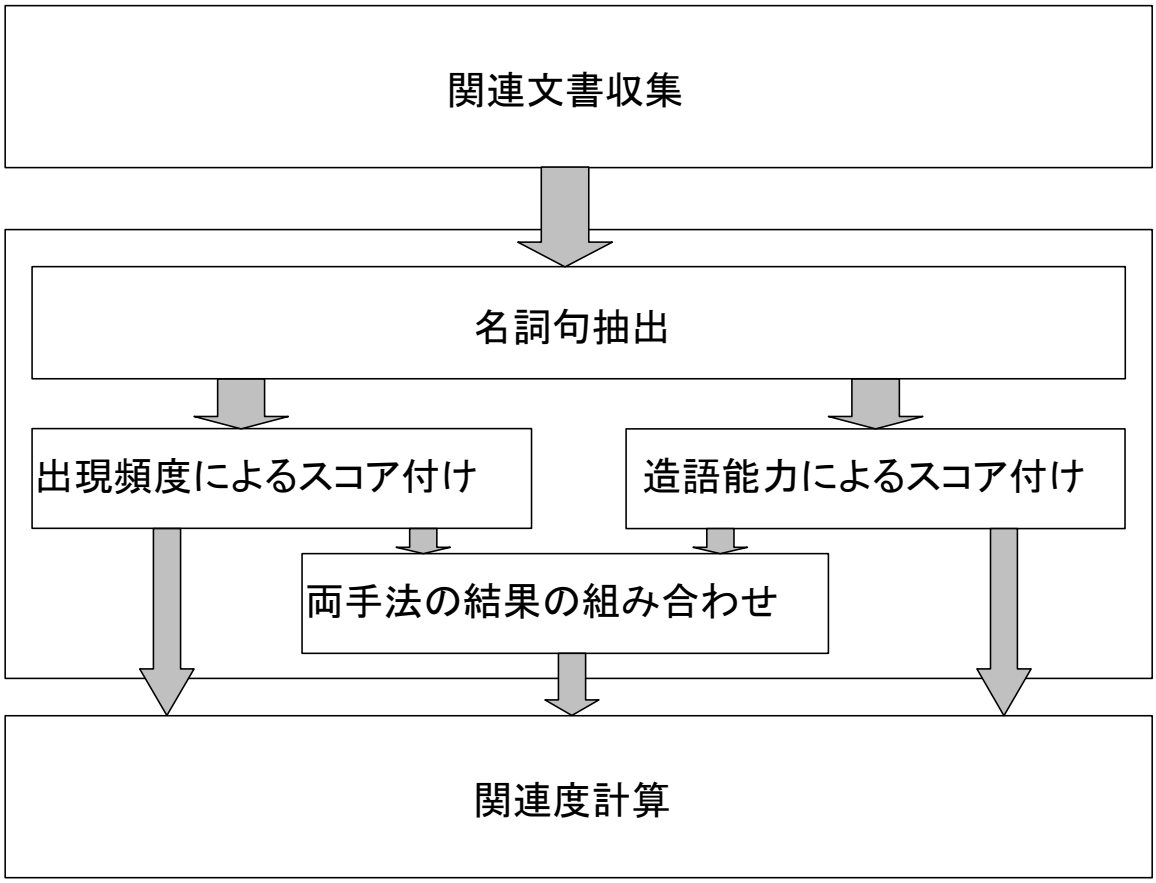

図 3.2: システム詳細図

た。これらはどちらが優れているとは一概には決めがたかったので、二つの手法による結果を組み合わせるシステムも実装した。これは単純に両手法の結果を上位から交互にとっていき、重複がないように上位１００件を選ぶ。
次の関連度計算の段階を含めて考えると、二つの手法を併用することは大きな意味がある。例えば、片方の手法がすべて関連用語としてふさわしくないような用語を候補として出力したとしても、関連度計算の段階でそれらは除外されるので、システム全体としては適切な出力が行われる。また、二つの手法には特徴があり、それぞれ異なる性質の用語を抽出するため、単一の手法を用いる場合よりも多様な用語を抽出できる。これについても詳しくは４．３節の考察で述べる。

本研究のシステムの関連用語抽出処理の詳細を図 3.2 に示す。まず「関連度計算」から渡されるのは最大５００件の HTML ファイルから作成されたプレーンテキストファイルである。そして「名詞句抽出」処理によって選ばれた名詞句に対し、「出現頻度」の手法と「造語能力」の手法でスコア付けを行う。「組合わせによる手法」は「出現頻度」と「造語能力」の結果を組み合わせる。「出現頻度」、「造語能力」、「両手法の組み合わせ」それぞれの名詞句上位１００件を、次の処理「関連度計算」に渡す。

3.5 検索エンジンを用いた関連度計算

この段階の処理は、Web 検索エンジンのヒット件数の比をとることで用語の候補がキーワードといかに共起するかということ、つまりどれだけ関連性が深いかということを求める。具体的には、検索エンジン goo を用いて前段階である用語抽出の結果１００語に対し、式 (3.6) の関連度を求める。t は用語、x はポータルサイトのテーマである。H(t) は用語 t を検索クエリとするときのヒット件数であり、H(t ∧ x) は t と x のアンド検索のヒット件数を表す。

$$\text{関連度} = \frac{H(t \wedge x)}{H(t)} \tag{3.6}$$

この値は、その用語が現れる文書のうち、キーワードと共に現れる文書数が占める割合である。つまり、キーワードが示す分野に特有の用語ほどこの値は大きくなり（１に近づき）、逆にその分野以外の文書にも多く現れる用語ほど小さくなる（０に近づく）。

キーワードと用語の関連度を、それぞれを含む文書の包含関係に基づいて説明する。t と x を含む文書集合の様々な包含関係を図 3.3-3.5 に図示する。キーワードを x、用語を t とする円はそれぞれが含まれる Web 文書を表し、円が重なっている部分が x と t が共起している文書、つまり t と x のアンド検索でヒットする文書を表す。

まず、用語間にあまり関連がないと思われる場合、つまり関連度が０に近い場合を示しているのが図 3.3 である。用語間に全く関連性がないと思われる、つまり円が重ならない場合も考えられるが、本研究ではそのような用語が候補になることはないので省略する。

次に、用語間の関連性が高いと思われる場合が図 3.4 である。これは、用語 t が出現するのはほとんどの場合キーワード x を含む文書内であるということであり、t は分野 x に特有の用語であると考えられる。さらに、t の円が x の円に完全に包含されている場合、t は x を含む文書以外では現れないということであり、関連度は最大値である１となる。

また、関連度計算を x と t が共起する文書数だけで求めることができないのは、図 3.5 のような場合を考慮する必要があるためである。このような場合、用語 t は一般的な語であり、どんな文書にもよく現れると思われる。このような場合、x と共起するからといって関連性が高いとはいえない。よって、図 3.5 のような用語のスコアを低くするために、式（3.6）の関連度はアンド検索のヒット件数を用語 t のヒット件数で割っている。

本研究では図 3.4 のような関係になる用語 t がキーワード x にとって重要であると考え、上記の関連度が大きいものを関連用語として出力する。

この段階の処理に関して、関連度の考え方は佐藤ら研究と本研究で共通であるが、処理内容は異なる。本研究では上記の式で求めた関連度によって候補の順位付けを行うが、佐藤らの研究ではこれらを用いてフィルタリングを行い、適切ではないと判断した候補を除外している。具体的には、佐藤らは上記の関連度が閾値を下回るもの、検索エンジンのヒット件数が一定以上に多いものと少ないものを除外している。これは、ヒット件数が多すぎるものは一般語であると判断し、少なすぎるものは分野内で一般的ではないものとであろうという考えに基づく。つまり、上記の関連度が低ければキーワードに対する関連用

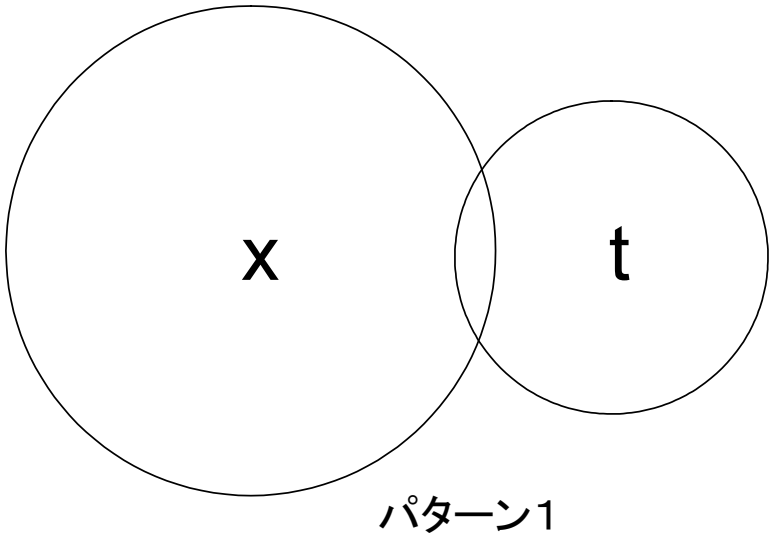

図 3.3: 用語の包含関係（パターン１）

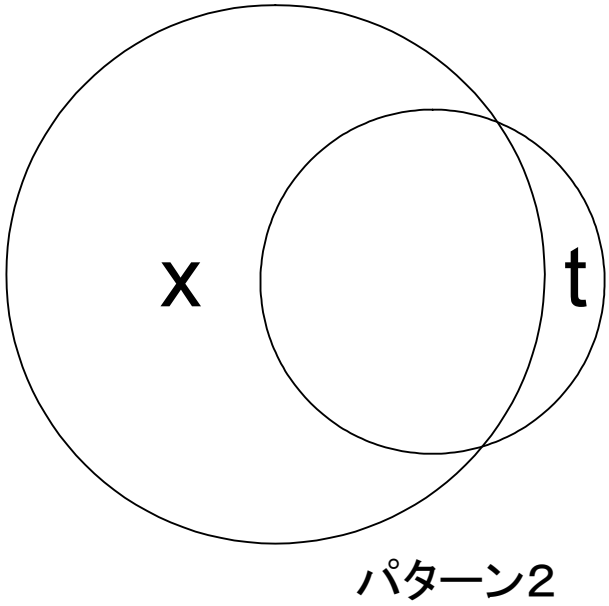

図 3.4: 用語の包含関係（パターン２）

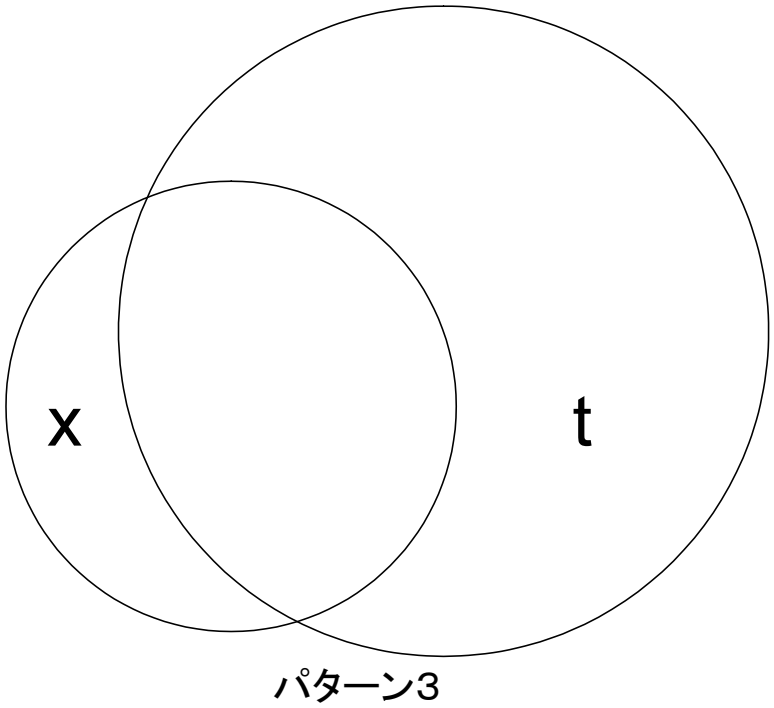

図 3.5: 用語の包含関係（パターン３）

語としてふさわしくないという考え方は共通だが、実際のスコア自体は 3.4.2 項で述べた中川らの「造語能力」によって決めている。

一方、本研究では３．１節で述べたように用語が一般語であるかどうかは考慮しないので、用語をクエリとした検索結果のヒット数によって用語を選別しない。また、この段階で求めた関連度によって候補を順位付けし、用語候補抽出処理までのスコアには依存しない。そして関連度の高いもの上位２０件を出力とする。用語集の見出し語ということもあり、出力する用語の数を動的に変化させるのではなく一定とした。

関連度は、理論上は最大値が 1 になるはずである。しかし、原因はよくわかっていないが、goo が示す H(t) の値が H(t ∧ x) よりも大きくなり、関連度が 1 を越える値になることがある。これらの値は、大抵 1.5 以下である。また、これとは別に、文字化けと思われる文字などに対し、関連度が 100 を越えるような異常な値になることもある。以上の考察から、goo が示すヒット件数のある程度の誤差は容認しつつ、文字化けしたときなど、用語候補として明らかに異常な値を除外するために、関連度が 2.0 以上のものは関連用語候補として出力しないことにした。

第4章　評価実験

4.1　実験の概要

３章では、用語候補のスコア付けの方法として、「出現頻度」、「造語能力」、「組み合わせ」の３つを提案した。実験は、これらをそれぞれ別のシステムとみなし、結果を比較した。具体的には、著者が選んだ２０テーマをそれぞれのシステムに渡し、それぞれのシステムによるスコアの高い上位２０語を関連用語として出力する。また、これらに対して、ポータルサイトの用語集の見出し語としてふさわしいかを判定し、手法の評価を行った。

4.2　実験結果

表 4.1 に、実験を行ったテーマ、対象とした HTML ファイル数（関連文書数）、候補となった名詞句の数を示す。

実験結果を表 4.2 に示す。表 4.2 は、与えられたテーマと、それに対する各システムの出力２０語のうち、正解とみなせる用語の数である。

表 4.2 の平均正解率によれば、「出現頻度」による手法が一番結果がよい。しかし、テーマ別にみれば、それぞれの手法の正解数にかなりばらつきがあることが読み取れる。テーマ「プログラミング言語」に関しては、組み合わせによる結果が他の結果よりよく、一方で「クラシック音楽」に対しては組合わせによる手法が最低の結果を出している。このように、どの手法が最も優れているかという結論を導き出すことはできない。また、テーマごとの正解数についてもある傾向が見受けられた。

まず、関連度計算の有効性について 4.2.1 項で述べ、手法の違いに関する考察は 4.2.2 項で、テーマの違いに関する考察は 4.2.3 項で述べる。

4.2.1　関連度計算の有効性

表 4.3 に、出現頻度による手法が出力した用語候補１００件と、関連度計算を経た最終的な出力を示す。用語は出現頻度によるスコアが高い順に並んでおり、「採用」となっているものが関連度計算によって上位２０件となった（システム全体の出力となった）用語である。

この結果から、「研究」や「pp」など、他の分野でも広く使われている表現は、出現頻度

表 4.1: 実験対象

	関連文書数	候補数
自然言語処理	402	2354
プログラミング言語	394	2440
Perl	398	2532
java	424	2634
windows	430	3048
マイクロソフト	329	2383
インターネット	483	2108
ハイブリッドレコーダ	398	8049
プラズマテレビ	362	2844
地上デジタル放送	409	1574
クラシック音楽	431	2743
MP3	388	2741
携帯電話	427	2806
ラーメン	436	1975
北陸	413	1410
スバル	424	2980
インプレッサ	416	1561
日産	420	2351
シルビア	437	1746
トレノ	439	3067

表 4.2: 実験結果

	出現頻度	造語能力	組み合わせ
自然言語処理	13	11	15
プログラミング言語	15	13	11
Perl	14	6	10
java	17	4	4
windows	8	5	7
マイクロソフト	8	7	6
インターネット	11	10	9
ハイブリッドレコーダ	10	2	9
プラズマテレビ	15	12	15
地上デジタル放送	11	7	9
クラシック音楽	14	2	2
MP3	14	9	13
携帯電話	11	14	10
ラーメン	14	10	9
北陸	10	8	10
スバル	11	2	6
インプレッサ	17	12	16
日産	12	6	13
シルビア	17	14	14
トレノ	13	6	13
平均正解率	12.75	8.00	10.05

が高くても関連用語としては不適切であると判定できていることがわかる。同様に、「コーパス」などのように、関連文書内での出現頻度がさほど高くない用語も、関連用語として出力されている。このように、ヒット件数に基づく関連度は、出現頻度だけでは捉えることのできない用語候補とテーマとの関連性を反映したスコアとなっている。したがって、本研究における関連度計算処理は有効であるといえる。

4.2.2 手法による傾向

システム全体の出力に関して、正解数で比較すると出現頻度による手法が全体的に良い成績を出している。しかし、それぞれの手法が出力する用語には異なる傾向がみられた。そして本研究における関連度計算処理との相性の問題なども明らかになった。以下、それらについて詳しく述べる。

出現頻度に基づく手法

この手法は、造語能力に基づく手法と比較すると様々な種類の用語を候補として挙げるという特徴がある。特に、人名を候補として挙げるのはこの手法だけだった。

しかし、人名などは関連用語としてふさわしい場合と、そうでない場合があると思われる。今回の実験においては「クラシック音楽」（表 4.4 参照）に対しては有名な作曲者や指揮者の名前は関連用語として正解とした。しかし「自然言語処理」に対する研究者の名前などは不正解とした。また、自動車や電機製品などの型番のような記号と数字の組み合わせによる用語（「AE １０１」、「AE ８６」など）も、この手法が最も多く候補として挙げていた。

造語能力に基づく手法

この手法では、造語能力が高い単名詞を含む用語ほど候補に上がりやすい。よって、造語能力が高いいくつかの単語を共通の部分文字列とする用語が複数候補として挙がることが多い。さらに、処理対象となる文書集合はすべてポータルサイトのテーマを含む文書からなるため、ポータルサイトのテーマ自体を部分文字列とする候補が非常に多いという特徴がある。

造語能力の高い単語からなる用語が候補となるのは妥当なことに思える。しかし同じ部分文字列を持つ用語ばかり出力されるのは、ポータルサイトにおける用語集の自動生成という観点からは望ましくない。手法ごとの差が顕著にみられたテーマ「クラシック音楽」の結果を表 4.4 に示す。

この例では、造語能力による手法の出力はすべて「クラシック音楽」を含む用語になっている。一方、出現頻度による手法で正解となっている用語は、人名や単名詞などが多い。人名は名詞句として含まれることが少なく、造語能力は低くなる。また、「クラシッ

表 4.3: 関連度計算

候補（1-50）	採用	候補（51-100）	採用
自然言語	採用	説明	
研究		開発	
ｐｐ		人工知能学会	
言語処理学会	採用	音声	
話者		黒橋禎夫	採用
言語		言葉	
情報処理学会		講義	
自然言語処理技術	採用	任福継	
機械翻訳	採用	テキスト	
情報		データ	
人間		言語理解	
必要		論文	
計算機		分野	
人工知能		荳	採用
電子情報通信学会		英語	
構文解析		現在	
表現		手法	
研究室		作成	
Ｖｏｌ		関係	
利用		研究分野	
松本裕治	採用	自然言語処理の研究	採用
形態素解析	採用	助教授	
情報処理学会自然言語処理研究会	採用	右膺肛	
意味		抽出	
対象		研究テーマ	
システム		研究会	
情報検索		言語表現	
情報処理学会研究報告		解析	
日本語		文章	
コンピュータ		ページ	
理解		青江順一	採用
情報処理学会論文誌		画像処理	
意味解析		学習	
技術		次	
自然言語処理研究会	採用	単語	
自然言語処理システム	採用	客体的表現	
言語処理		コミュニケーション研究会	
処理		コーパス	採用
知識		研究者	
教授		ｏｆ	
認識		日本認知科学会	
文		翻訳	
言語学		インターネット	
検索		長尾	
応用		ＮＬ	
プログラム		日本	
音声認識		自然言語処理入門	採用
機械翻訳システム	採用	人工知能学会誌	採用
蟷合函	採用	自然言語理解	採用
人		音声言語処理	採用

ク」や「音楽」以外の単名詞はそれ自体のスコアが高い場合もあるが、「クラシック」と「音楽」という高い造語能力が与えられた単語を部分文字列とする複合語の方が高いスコアを持つ。

さらに、Web を利用する本研究特有の原因もある。本来、造語能力によるスコア付けは専門用語抽出分野でよい成績を出すことが示されている。しかし、本研究では全体的には他の手法に劣る。これは、最終的な関連度計算を検索エンジンのヒット件数のみで求めていることによる。ここでもう一度関連度計算の式を示す。

$$\text{関連度} = \frac{H(t \wedge x)}{H(t)} \tag{4.1}$$

ここで、x はポータルサイトのテーマを、t は用語を表している。H(t) は用語 t をクエリとしたときの検索エンジンのヒット数、H(t ∧ x) は t と x のアンド検索のヒット件数を表している。つまり、用語 t がテーマ x を含む文字列である場合、この値は理論上常に 1 になるはずである。例えば、「クラシック音楽」and「クラシック音楽教室」の結果は、すべて「クラシック音楽」という用語を含むページであるはずなので、そのヒット件数を「クラシック音楽」のヒット件数で割れば 1 になるはずである。３章でも触れたように、ヒット件数には理論値と比べて多少のずれはあるが、そのことは問題ではない。問題となるのは、この値による関連度計算は、テーマを含む用語に対して非常に高く、そしてほぼ同一の得点を与えてしまうということにある。このことによって、用語候補１００件を選定する段階で、その中にテーマを部分文字列とする用語が２０件以上含まれていると、そのシステムはほぼ間違いなくテーマを部分文字列とする用語群を最終的な出力としてしまう。

テーマ「クラシック音楽」に対する出力は、特殊な例でもある。「クラシック音楽」を部分文字列とする用語の中に関連用語としてふさわしくないものが多い分野だったし、関連用語としてふさわしいものは人名など、造語能力では専門用語としての良さを測ることが難しい分野だった。テーマ「プログラミング言語」に対する造語能力による手法の出力は、「プログラミング言語」そのものを含むものばかりではなく、「関数型言語」や「手続き型言語」などの妥当なものを出力している。

組み合わせによる手法

組合わせによる手法に関して、候補１００件を選ぶ段階については出現頻度による手法と造語能力による手法の結果の組合わせなので、特筆すべき点はない。しかし、関連度計算を経た出力に関しては特徴がある。

まず、組合わせによる手法を試そうと思ったきっかけは、候補を選ぶ段階で出現頻度、造語能力による手法のそれぞれが候補とした用語すべてを関連度計算の段階に渡すことで、関連度が高いもののみを得ることができると考えたからだ。しかし、表 4.2 から読み取れるように、組合わせによる手法が最も正解数が多かったのは、同数を含めて５テーマに過ぎなかった。さらには、同じく同数を含む場合、６テーマにおいて正解数が最も少なくなった。組合わせによる手法の動機である「出現頻度と造語能力による手法の長所をと

表 4.4: テーマ「クラシック音楽」に対する出力

出現頻度	正否	造語能力	正否	組み合わせ	正否
♪		クラシック音楽教室		クラシック音楽ＣＤ	
クラシック音楽情報センター	正	クラシック音楽の世界		♪クラシック音楽	
作曲家別		関西クラシック音楽情報		クラシック音楽情報誌	
クラシック	正	♪クラシック音楽		クラシック音楽入門	
歌劇	正	クラシック音楽ＣＤ		クラシック音楽のＣＤ	
演奏家	正	クラシック音楽のページ		クラシック音楽の録音	
ヴァイオリン協奏曲	正	クラシック音楽の風景		クラシック音楽情報センター	正
マーラー	正	クラシック音楽の録音		クラシック音楽のページ	
ピアノ協奏曲	正	クラシック音楽館		クラシック音楽館	
２楽章		クラシック音楽関連		クラシック音楽関連	
ベートーヴェン	正	クラシック音楽作品名辞典		クラシック音楽情報	
１楽章		クラシック音楽入門		クラシック音楽教室	
ブラームス	正	クラシック音楽データベース		クラシック音楽データベース	
３楽章		クラシック音楽情報センター	正	クラシック音楽の世界	
弦楽四重奏曲	正	クラシック音楽情報誌		ｂｙ	
４楽章		クラシック音楽のＣＤ		クラシックのコンサート	
チャイコフスキー	正	日本クラシック音楽事業協会	正	クラシック	正
モーツァルト	正	クラシック音楽情報		クラシックの名曲	
シューベルト	正	クラシック音楽の部屋		クラシックＣＤ	
交響曲	正	クラシック総合情報		クラシックコンサート	

もに用いる」という観点からすれば、少し物足りない結果となった。

では、どうしてこのような結果になってしまったのだろうか。造語能力による手法の考察で触れた問題、すなわちテーマを部分文字列とする用語の問題が原因の１つと考えられる。実際、表 4.4 のように、「クラシック音楽」に対する出力はほとんどが造語能力による出力と同じでテーマ自体を部分文字列として含んでいる。これらの用語は検索エンジンのヒット件数をもとに評価した場合、高いスコアが与えられる傾向にあり、テーマを部分文字列とする用語を出力するシステムの影響をより強く受けている。そのために、２つの手法の長所を取り入れることができなかったと考えられる。

4.2.3　テーマによる傾向

手法の違いによる傾向の他に、テーマの違いによる傾向も考えられる。表 4.2 に示されているように、「インプレッサ」、「シルビア」（いずれも自動車の車種名）などの固有名詞においてはどの手法もよい結果となっている。しかし、「windows」、「マイクロソフト」などは、固有名詞であっても正解率は低い。これらは、出現する文書数が多く、「関連文書」としての傾向が弱いためではないかと考えられる。表 4.5 に、各テーマの検索エンジンでのヒット件数と３手法の平均正解率を示す。表 4.5 から、ヒット件数が多いほど正解率は下がるというある程度の相関が読み取れる。

また、テーマに関して、テーマ自体が多義である場合の問題もみられた。実験の例では「トレノ」というテーマは、自動車の車種名と想定していた。しかし出力の中に自動車と無関係な用語が現れていたため、調べたところ、あるゲームの町の名前として「トレノ」というものがあり、それに関する用語だった。それらを除くと、「トレノ」に関しても他の車種名の例と同様によい正解率となっていた。

4.3　考察

実験結果の分析により、提案手法の改善案がいくつか考えられた。本節ではそれらについて述べる。

4.3.1　テーマを部分文字列とする用語に関して

実験結果の傾向から、テーマを部分文字列とする用語に高いスコアを与える造語能力による手法と組合わせによる手法に不具合があることがわかる。よって、それらに関する対応策などを考えた。

まずはじめに思いついたのは、テーマを含む用語を除外するという単純な手法である。しかし、テーマ「ラーメン」に関しては、正解とした用語ほぼすべてが「ラーメン」を含むものであり、分野によっては不適切であると思われ、断念した。

表 4.5: ヒット件数と正解率

	ヒット件数	平均正解率（%）
自然言語処理	10600	65
プログラミング言語	46900	65
Perl	269000	50
java	761000	42
windows	1620000	33
マイクロソフト	211000	35
インターネット	1880000	33
ハイブリッドレコーダ	933	35
プラズマテレビ	30500	70
地上デジタル放送	21800	45
クラシック音楽	85500	30
MP3	812000	60
携帯電話	1630000	42
ラーメン	631000	58
北陸	434000	47
スバル	71800	32
インプレッサ	24700	75
日産	160000	52
シルビア	24000	75
トレノ	14200	53

次に、「ラーメン」に関する結果から考えられたのは、テーマの右側に他の用語が連接しているもののみを除外するという案だった。これは、「とんこつラーメン」などを残し、「ラーメン屋さん」などを除外するというものだ。「ラーメン」に関するデータを見ると、これでかなりの改善がみられた。ところが、テーマの右に名詞が現れる複合名詞の中にも、関連用語としてふさわしいものがある。例えば、テーマ「インプレッサ」に対する「インプレッサ WRX」や、テーマ「マイクロソフト」に対する「マイクロソフト認定技術者試験」など、適切な用語を除外することになる。名詞がテーマの右側に連接していても関連用語に含めるべきか、そうでないかを自動的に判別するのも非常に困難と思われる。

また、特定のパターンを除外することで不適切な候補を減らすということではなく、評価基準を変えることも考えた。テーマを部分文字列とする用語は関連度が 1 になることはわかっているので、それらのみを他の基準で評価するという方法である。例えば、候補である用語（テーマを含む）からテーマの部分文字列を除いた用語によって関連度を計算する。これを式で表すと式 (4.2) のようになる。

$$\text{関連度 2} = \frac{H((t-x) \wedge x)}{H(t-x)} \tag{4.2}$$

ここで、(t-x) は、用語 t からテーマ x を除いた文字列を表す。しかしこの手法では、「塩ラーメン」の場合、分母は「塩」のヒット件数になるため関連度は低くなってしまう。つまり、連接している用語が一般的であればあるほど関連度は低く見積もられる。また、元の関連度の式と比べて、式の定義自体も変わるため、元の関連度の式で評価する他の候補（テーマを含まない用語）と同列に比較できないという問題もある。また、テーマの前後両方に他の語が連接している用語にも対応できない。

これらのように、決定的な改善案は考えられなかった。しかし、この点を改善することができれば、造語能力による手法の精度が向上するだけでなく、組合わせによる手法は両手法のよい候補のみを取り出すという本来の目標が達成できるはずである。いずれにせよ、テーマを含む用語に対しては何らかの手段によって改善を行うことが強く要求される。

4.3.2 関連文書に関して

実験結果から、検索エンジンによるヒット件数が多いテーマほど正解率は下がるという傾向が見られた。これは、どこにでも現れる用語に関しては、集めた文書の「専門文書性」が低いことによると思われる。つまり、文書内にテーマが現れるからといって、その文書全体がテーマに関して述べているものとは限らないということであり、その傾向はどこにでも現れる用語ほど強いと考えられる。例えば、「インターネット」のようにさまざまな文書に現れる用語の場合、「インターネット」という用語を含んでいるからといってその文書が「インターネット」自体について述べている文書であるとは限らない。

これに関して、本研究では「関連文書収集」段階で検索結果上位５００件を無条件に取り出していることに問題がある。検索エンジンの内部処理によって、テーマ（検索キー

ワード）に関連性が高いものから順に並んでいるはずではあるが、この順位付けはやはり不十分だと思われる。よって、「関連文書収集」段階を改善することで、精度の向上は望める。

これに関して具体的な実験は行っていないが、例えば収集した各文書内でのテーマの出現回数を数え、フィルタリングを行ったり重み付けを行うという手法が考えられる。具体的には、テーマが一度しか現れない文書は関連文書から除外したり、テーマの出現回数が多い文書内で共起する用語に対して高いスコアを与えるような重み付けを行う。また、HTML タグ情報を利用し、タイトルタグにテーマが含まれているページに出現する用語を優先するなどの方法が考えられる。

4.3.3 関連用語候補に関して

本研究では、「用語候補抽出」処理から「関連度計算」に渡す用語の候補数を１００件としている。しかし、表 4.3 からわかるように、用語候補抽出処理での順位が低い用語の中にも関連用語としてふさわしい候補も含まれている。同様に、１００位以降にもふさわしい用語が含まれている可能性もある。よって、「用語候補抽出」処理から「関連度計算」に渡す用語候補の数を増やすことで再現率の向上が期待できる。

4.3.4 多義であるテーマに関して

先ほどテーマ「トレノ」の例を示したが、テーマが多義であるということは十分起こり得ることである。テーマが多義である場合の対処法として、文書分類のような処理を行うことが考えられる。関連文書を個別に分析し、文書ごとに用語の傾向が異なれば、それらは異なる語義に対応した別の分野の文書であると判断できる。しかし、このような処理には計算コストがかかるので、多義であるかわからない（多義でないかも知れない）テーマすべてに対し行うのは現実的ではないと考える。

しかし、ポータルサイトのテーマはユーザが与えるものであり、テーマを表す語が多義であっても、ユーザにとっては語義は一意に決まっている。したがって、システム全体の入力としてテーマ以外にも語義を一意に決めるヒントとなるような用語を入力してもらう方法などを考えた方が現実的であろう。

第5章　おわりに

本研究では、ポータルサイトの自動生成を目的とし、用語集の見出し語となるような関連用語をWeb上から自動的に獲得するためのシステムを３種類実装し、比較実験を行った。

実験結果では、テーマが検索時のヒット件数が少ない固有名詞である場合を中心に、適切な用語群を得ることができた。また実験結果から、３種類のシステムのうち最善のシステムを決めることはできなかったが、それぞれの傾向などからさらなる改善の方向性を見いだすことができた。

考察でも述べたように、今後は以下の点の改善が望まれる。

- 関連文書取得処理の改善
- キーワードを部分文字列とする用語に対する処理
- 評価基準等

これらを改善することで、各手法の精度を向上させることができると思われる。特にキーワードを部分文字列とする用語に対し、適切な処理を行うことができるようになれば、組合わせによる手法が常に良い成績を出すと予想され、この手法のみで十分な出力が得られると期待できる。

謝辞

本研究を進めるにあたり、熱心に御指導してくださいました白井清昭助教授に心から感謝致します。本研究および関連分野に関して適切な御助言をくださいました島津明教授ならびに山田寛康助手に深く感謝致します。そして、研究室をはじめとする多くの方々の御援助によって本研究を行なうことができましたことを厚く御礼申し上げます。

関連図書

[1] Hiroshi Nakagawa. Automatic Term Recognition based on Statistics of Compound Nouns Terminology vol.6 No.2 (2000).

[2] 大畑博一, 中川裕志. 連接異なり語数による専門用語抽出. 情報処理学会研究報告「自然言語処理」 136-16 pp.119-126

[3] 湯本紘彰, 森辰則, 中川裕志. 出現頻度と連接頻度に基づく専門用語抽出 情報処理学会研究報告「自然言語処理」 145-17 pp.111-118

[4] 藤井敦, 伊藤克亘, 秋葉友良. 辞典的 Web 検索サイトの構築 情報処理学会 第９回年次大会 発表論文集 pp.129-132

[5] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化 電子情報通信学会論文誌 vol.J85-D-II No.2 pp.300-307

[6] Atsushi Fujii, Tetsuya Ishikawa Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from semi-Structured Text 情報処理学会 第９回年次大会 発表論文集 pp.129-132

[7] 清田陽司, 黒橋禎夫. WWW テキストの自動要約と KWIC インデックスの作成 情報処理学会研究報告「自然言語処理」 137-5 pp.31-38

[8] 佐藤理史, 佐々木靖弘. ウェブを利用した関連用語の自動収集. 情報処理学会研究報告「自然言語処理」 153-8 pp.57-64

[9] 浅原正幸, 松本裕治. 形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現個所同定. 情報処理学会研究報告「自然言語処理」 154-8 pp.47-54

[10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書