情 報 学 基 礎 48－3
（1997. 11. 13）

インターネット情報探索に適した複合語検出

神林 隆, 清水 奨, 佐藤 進也, 風間 一洋
NTT 光ネットワークシステム研究所
東京都武蔵野市緑町 3-9-11

サーチエンジンを用いて、インターネットという巨大な情報空間を探索する時に、目的とする情報リソースを特定しやすい語が検索関連語として提示されると、探索効率が上昇する。本稿では、情報探索の支援となる検索関連語として、複数の名詞から構成される複合語に着目し、情報リソースから複合語を抽出するための手法として、適切な文書セットを作成し、その文書セット中の出現頻度に応じて複合語を検出する手法を提案する。また、評価実験を行ない、文書セットによる複合語の分類、および、合理的な文書セットの１つの解としてディレクトリ単位が有効であることを示す。

Compound Term Extraction for World-Wide Information Discovery

Takashi KAMBAYASHI, Susumu SHIMIZU, Shin-ya SATO, Kazuhiro KAZAMA
NTT Optical Network Systems Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo

Search engines are generally used when exploring a very large information space, the Internet. If search engines present users with some adequate related terms, they can easily specify the search for information resources. In this paper, as related terms which aid world-wide information discovery, we focus on compound terms that comprise a sequence of nouns. Moreover, we propose a compound term extraction method. This method organizes an appropriate document set and extracts compound terms according to their frequencies in the document set. Results from the our experimentation indicate that the compound terms are divided into four types and that the documents in a single file-system directory are a reasonable choice for a document set.

1 はじめに

インターネット上には非常に多種多様な情報リソースが大量に存在する。その中から必要な情報を見つけ出すために、キーワードを入力すると目的とする情報リソースの概要やURLを提示するサーチエンジンがよく利用される。サーチエンジンの優劣の判断材料として検索可能な情報リソース数が使用されるが、実際には検索対象となる数が増えるにつれて、検索される情報リソースの数も増加するので、膨大な検索結果の中から本当に必要とする情報を探し出すことは困難になる。

そこで、検索条件の変更や、検索語の追加や変更などを行なうことで、検索結果をさらに絞り込む必要がある。しかし、再検索のための検索語の選択は難しく、検索しようとする対象分野についての十分な知識と経験がないと、検索結果を絞り込めなかったり、逆に検索結果から目的とする情報リソースが除外されたり、何も検索されなくなることもある。

検索語の選択を支援するために、検索されたURLと一緒に検索関連語を提示するサーチエンジンも存在する。たとえば、我々が研究中のIngridにおいては、図1のように、検索された情報リソースに含まれる重要なキーワードを検索関連語としてユーザに提示し、ユーザはこの検索関連語をDrag & Dropするだけで検索条件が変更できる[1]。また、原田らはODIN[2]で、他のユーザが同一情報リソースを閲覧する時に用いた検索語を関連語として提示した[3]。

このようなサーチエンジンでは、提示された検索関連語を再検索時に追加して元の検索語とのAND検索を行なうことで、効率的にWeb情報空間の絞り込みが行なうことができる。また、検索関連語の中から新たに検索語を選び出すことで、検索対象とする情報集合を少しづつ変更することが可能になり、Web情報空間を渡り歩くことが可能になる。

Web情報空間中の目的とする情報リソース集合を特定したり変更したりする行為を情報探索と呼ぶが、情報探索の支援には、どのリソースにも含まれるような一般的な単語よりも、情報リソースの内容の違いを明確に反映するより特殊な単語を検索関連語として提示する方が好ましい。

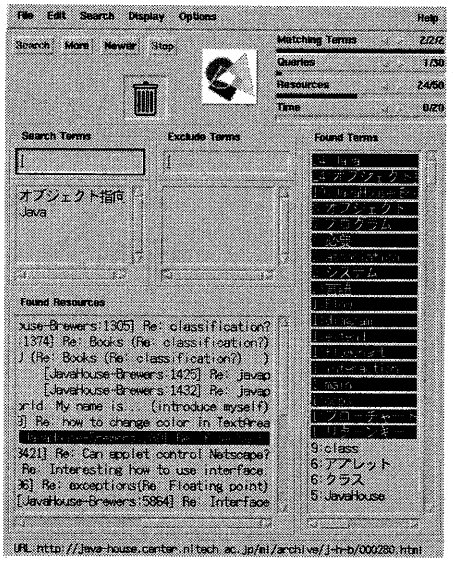

図 1: 関連語を提示する検索クライアント

本稿では、このような検索関連語として特に複合語に着目し、情報リソースから複合語を抽出する手法と、その評価実験と解析結果について報告する。第2節では、情報リソースからキーワードを抽出するための従来手法について述べる。第3節では、複合語の定義と有効性、および、複合語の文書中の出現頻度に注目した複合語検出手法について説明する。第4節では、本複合語検出手法の実験とその評価について論じる。

2 キーワード抽出手法

2.1 文章の分割

サーチエンジンのインデックスを作成したり、情報探索支援のために使用する検索関連語を抽出するためには、さまざまなフォーマットの情報リソースからテキストデータを取り出し、さらに適当な長さの語に分割する必要がある。

たとえば、日本語の場合には次のような方法が

存在する。

1. 文字種によって分割する。
2. n グラム統計を用いて分割する。
3. 形態素解析プログラムを用いて分割する。

下畑らは、テキストを文字種によって分割して語を抽出し、切り出した語が複合語である場合には、さらに単位語に分解した [4]。原らは、直前の文字種と後続の文字種の接続可能性を基にわかち書きすることで語の分割を最小限に押え複合語として抽出し、獲得した語の対について、文書のある範囲内での共起度を求め、キーワード抽出における重要度として使用した [5]。文字種による分割は、単語辞書を用いないために計算コストが少なくて済むが、文字種が切り替わる場所で単語分割を行なうので、複数の文字種を含む複合語 (たとえば、「ゴミ捨て」や「問い合わせ先」など) に対応することができない。

中渡瀬や森らは、n グラム統計を用い、字面処理でキーワードを獲得した [6][7]。n グラム統計も単語辞書を必要としないが、キーワード獲得に要する計算コストが膨大になるだけでなく、検索インデックスに登録すべきデータが元のテキストデータより大きくなるなどの問題がある。

インデックスのサイズを押さえながら、異なる文字種から構成される語もキーワードとして抽出するためには、形態素解析が利用されることが多く、我々も形態素解析方式を採用している。

2.2 形態素解析による分割の問題点

形態素解析は、単語辞書と連接規則に基づき、文章を辞書登録された単語に分割する手法である。そのために、単語辞書の品質に結果が大きく影響される。たとえば、文章中に単語辞書に登録されていない語が出現すると、うまく解析できずに誤分割される。このような誤分割は、処理する対象に合わせて単語辞書を保守することで回避できる。

しかし、インターネット上の情報リソースを対象にした場合には、膨大な処理結果の中から誤分割を発見したり、次々に生まれる新造語や固有名詞に合わせて辞書を保守しなければならない。この作業を実際に人手で行なうのは不可能に近い。

3 複合語検出

3.1 複合語の定義

本稿では、「複数の名詞から構成される単語」を特に複合語と呼ぶことにする。ただし、ここで述べる名詞とは、形態素解析の結果名詞として解釈された語のことを示し、必ずしも人間が名詞として認識できる単位とは限らない。

たとえば、juman version 2.0[8] に付属の辞書をそのまま用いると「イントラネット」という語が「イン」「トラ」「ネット」という 3 つの単語に分割される。このように細かく分割された単語を個別に見ても、その単語の示す意味は正しく判別できない。

3.2 情報探索における複合語の有効性

情報探索において、ユーザに複合語を提示するのは、非常に有効な手段である。

前述のサーチエンジン ODIN の検索ログを解析することによって、複合語が次のような目的に使用されていることが判明した。

1. 検索対象範囲の絞り込み／拡大
   - 複合語の生成／消失
     例: さだ ⟷ さだまさし
   - 複合語の延長／短縮
     例: オブジェクト指向 ⟷ オブジェクト指向設計
   - 複合語の合成／分解
     例: さだ AND まさし ⟷ さだまさし
2. 検索対象範囲の変更
   - 複合語の語尾の変更
     例: オブジェクト指向設計 ⟷ オブジェクト指向言語
   - 複合語の語頭の変更
     例: 荒井由実 ⟷ 松任谷由実

さらに、日本語の「オブジェクト指向」が、英語では“Object Oriented”となることからわかるように、検索キーワードにおいては、日本語の複合語は英語のフレーズに相当する。また、新造語や部署名などの固有名詞は、複合語であることが多く、「イントラネット」や「金融システム営業本部」などは、これらの代表例である。

また、「オブジェクト」という語は一般的であると共に、まったく別の意味で使われることがある。1997年10月中旬の時点でgoo[9]で検索したところ、「オブジェクト」は35745件も検索された。このような語を使っても、情報をほとんど絞り込むことができず、情報探索の支援に使うにはあまり有効ではない。しかし、「オブジェクトファイル」なら920件、「オブジェクト指向」なら12026件になるように、複合語はある語に対する下位概念を示し、情報空間中の情報リソース集合を特定する場合には、非常に有効であると言える。

3.3 既存の複合語抽出手法

名詞が複数個並んでいるものをすべて複合語にしても、必ずしも適切なキーワードとしては抽出できない。

小川らは、正規表現を用いたキーワードパターンに一致するキーワード候補を集め、それらを構成する単語の構文的・意味的性質より、不要な候補を削除した[10]。この方法では、重要な複合語候補が削除されないように、事前にキーワード抽出規則を十分に検討する必要がある。しかし、非常に多種多様な情報が頻繁に更新／追加されるインターネットの情報リソースの場合、規則に対する十分な評価が難しく、また新造語への追従性が劣ると思われる。

中川らは、形態素解析を行なった後の単名詞について、前後の連結数を求め、それらの積により複合語の重要度を決定した[11]。この手法では、連接数の長い複合名詞ほど高い重要度を持つことになる。しかし、情報探索においては、ほとんど使われない特殊な長い複合名詞が数多く抽出されるより、使用頻度を考慮して複合名詞が抽出される方が望ましい。

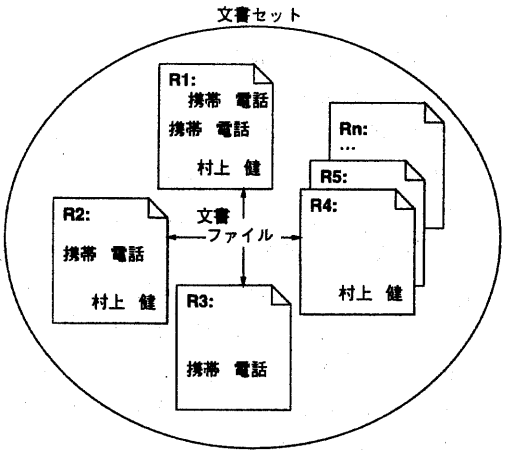

図 2: 複数の文書ファイルからなる文書セット

3.4 1文書単位の複合語検出

文章中であるフレーズ(連続した名詞)が複数回繰り返し出現している時、そのフレーズは文章中で重要な語として用いられていることが多い。

この性質を利用して、形態素解析の結果得られた単語のうち、出現順序が連続した名詞がある閾値の回数だけ出現した場合に、その連続した名詞の組を複合語として検出する手法について、我々はすでに報告した[12]。その時は、形態素解析プログラムとしてjuman version 2.0を使い、出現回数の閾値は2、検出対象を1文書(1ファイル内の文章)単位とした複合語(成句)検出を行なった。

しかし、1文書単位で処理しているので、文書サイズが小さくなると、文章内のフレーズの繰り返しが減少し閾値を越えることがないので、複合語が検出できなくなる。また、文章の作者の所属部署や名前なども、その文書を特定する重要な語であるにも関わらず、1文書内では1回しか出現しないので、キーワードして抽出されないことが多かった。

図2に示す例では、文書ファイルR1で「携帯」と「電話」が2回出現するので、「携帯電話」と

いう複合語が抽出される。しかし、それ以外のファイルでは、1文書内での出現頻度は1であるので、「携帯」と「電話」と別々の単語として抽出される。R1, R2, R4の作者の名前「村上健」という固有名詞も、1文書単位の複合語検出では抽出不可能である。

3.5 HTMLファイルの特性

インターネットで公開されている情報リソースの大部分はHTMLファイルである。またCunhaらによると、Unix上の通常の文書ファイルのサイズが1Kバイトから4Kバイトの範囲に主に分布するのに対して、HTMLファイルは256バイトから512バイトの範囲に主に分布する[13]。

このように、HTMLのようなハイパーテキストでは、たとえば出版の世界における本に相当する文章であっても、単一のファイルにならずに、章や節，段落などの構造に応じて複数のファイルに分割され、それらがハイパーリンクで結合される。この結果、1文書単位の複合語検出がうまく行なわれないことも多い。

3.6 文書セット単位の複合語検出

前節で述べた問題点を解決するには、複合語検出の対象を1文書単位ではなく、より大きい単位にすればよいことが直観的にわかる。しかし、インターネット上のすべての情報リソースに対して複合語検出を行なうのは、計算コストの点で現実的ではない。

そこで、できるだけ内容的に類似した文書をまとめて文書セットとして作成し、それに対して複合語検出を行なう手法を提案する。これによって、計算コストを押えつつ、複合語をうまく検出することが可能になる。

たとえば、図2の例において、{R1, R2, ... , Rn}をある文書セットだとし、検出閾値を2とすると、R2, R3では1回しか出現しない「携帯電話」も、文書セット内での出現回数は4であるので複合語となる。同様に、「村上健」も文書セット内での出現回数が3であるので、出現するすべての文書R1, R3, R4で検出されるようになる。

3.7 文書セットの選択方法

今回提案する手法では、検出される複合語の質、および、検出に必要な計算コストは、文書セットの選択方法に大きく依存する。

文書セットの選択方法としては、

1. HTMLファイルのハイパーリンクをたどり文書セットを作成する。
2. Webサーバが持つすべてのファイルで文書セットを作成する。
3. ディレクトリ単位に文書セットを作成する。
4. 1ファイルのみで文書セットを作成する。

などが考えられる。

HTMLファイルではさまざまな種類のハイパーリンクに対して単一のタグしか使わないので、一旦文書をHTMLでハイパーテキスト化し、そのHTML文書が他の情報との間に相互にリンクを張ると、元の文書を機械的に抽出するのは非常に困難になる。元の文書と他の情報との関連性が乏しい場合も多く、単純にハイパーリンクをたどる文書セットの選び方は適切とは言えない。

Webサーバ全体で文書セットを作成する方法は、Webサーバが巨大である場合に複合語検出に要する計算コストが膨大になり現実的ではない。1ファイルのみでは、文書セットでは小さ過ぎて複合語検出がうまくいかないというは前述した。

そこで、適切な大きさの文書セットを作成する1つの解として、本稿では同一ディレクトリに含まれるファイルを単位とすることを考え、それを検証する実験を行なった。

4 実験と評価

4.1 実験の概要

処理対象のデータとしては、一般に公開されている企業のWebサーバから、日本語で記述されたHTMLファイルを、トップページから幅優先

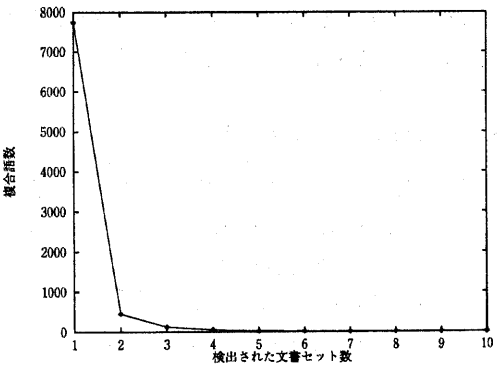

図 3: 文書セットと複合語の相関関係

で探索して 100 ファイル収集し、それを 1 つの文書セットとして、それを計 10 個作成した。また、日本語形態素解析プログラムには、juman version 2.0 を使用した。この文法辞書や形態素辞書には何も手を加えていない。この 10 個の文書セットに対して、日本語形態素解析を行なった後に、ファイルごとに複合語候補を検出して、評価に使用した。

4.2 文書セットと複合語の相関関係

図 3に、文書セットと複合語の相関関係を示す。x 軸はある複合語の検出文書セット数を、y 軸には複合語数を示す。検出文書セット数が多いほど、一般的な複合語と考えられる。

複合語は全部で 8419 個検出されたが、このうち 92% はある特定の文書セット内でしか検出されなかった。検出される複合語は辞書の未登録語であり、それらは特殊語や新造語であることが多いので、特殊語の使用や新造語の発生は Web 情報空間内で局所性を持つことが推測された。

4.3 複合語の検出頻度の変化

文書セットの違いによる複合語の検出頻度の変化を評価するために、各文書セットに対して、各複合語が検出したファイル数を求めて、この検出ファイル数を元に順位付けを行なった。この順位をランクと呼ぶ。

表 1: 複合語のタイプ

		頻繁	
		○	×
広い	×	I (図 4)	II (図 5)
	○	IV (図 7)	III (図 6)

これらの値をもとに、複合語が

- 広く使われているかどうか
- 頻繁に使われているかどうか

という 2 つのパラメータを用い、表 1のようにタイプ I から IV まで、4 つのタイプに分類した。また、これらのタイプに属する複合語のうち特徴的なものを、x 軸に文書セットのランクを取り、y 軸にはその複合語の検出ファイル数の対数を取り、図 4から図 7に示す。

タイプ I は、使用範囲は広くないが、その限定した範囲の中では非常に頻繁に使われる語である。企業名や部署名などの使用範囲の限られた固有名詞が多い。このタイプは、本複合語検出手法で検出するのに適している。

タイプ IV は、使用範囲も広く、また、その使用範囲ごとでも頻繁に使われる語である。このような語は非常に一般的であると言えるので、複合語検出という手法を使わなくても検出できるように、形態素解析の単語辞書に登録した方が効果がある。

タイプ III は、広く使われてはいるが、その範囲内ではあまり頻繁に使われていない語であり、タイプ IV に比べると特殊な語である。

タイプ II は、使用範囲も狭く使用頻度も低い語であり、語の特性は判断しにくい。新造語はこのタイプに属することが多く、時間的経過とともに、タイプ I かタイプ III に移行していくと思われる。

4.4 文書セットの選択とタイプの変化

次に、文書セットの選択方法を変更した場合に、検出される複合語のタイプが変化すること

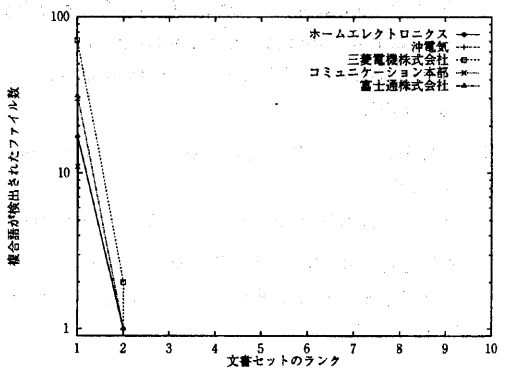

図 4: タイプIに属する複合語

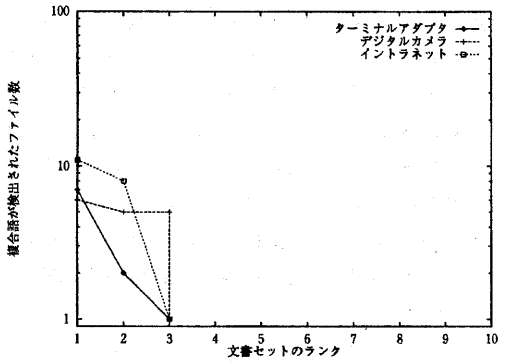

図 5: タイプIIに属する複合語

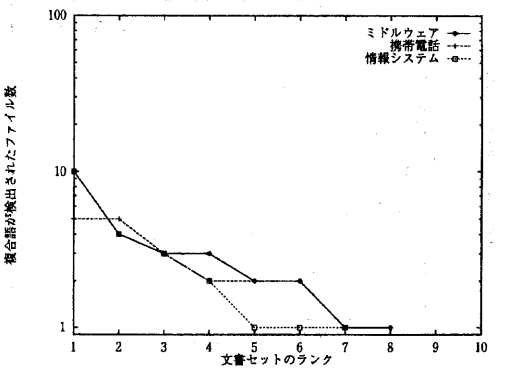

図 6: タイプIIIに属する複合語

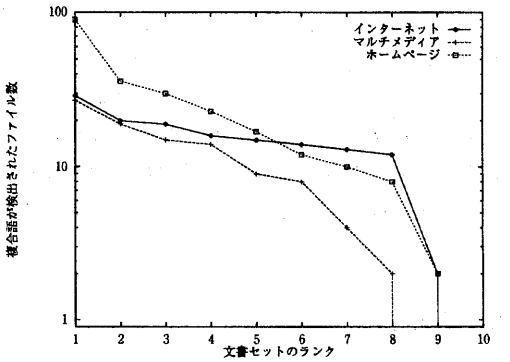

図 7: タイプIVに属する複合語

を調べた。具体的には、ある文書セットごとに、100個の文書を先頭から10個ずつに分け新しい文書セットを作成し複合語検出を行ない、タイプの変化した複合語があるかどうかを調べた。

タイプが変化した複合語の例として、元の文書セットでは、タイプIに属した複合語「富士通株式会社」が文書セットを変更した時の様子を図8に示す。新たな文書セットにおいては、この複合語はタイプIIIに属するものとなる。

このように、文書セットの取り方によって、複合のタイプが変化するものもある。複合語の検出には、文書セットを適切に選択することが重要である。

4.5 ディレクトリを単位とする文書セット

最後に、複合語検出の文書セットとして、同一ディレクトリに含まれるファイルを単位とする方法について検証を行なった。

各複合語について、元の文書セットで検出されたファイルのパス名を求め、それが同一ディレクトリに含まれるファイルから検出されているかどうかを調べた。その結果、元の文書で検出された総複合語数に対する同一ディレクトリで検出された複合語数の比率は、10文書セットの平均で、62.76%であった。

また、1ディレクトリ上位のファイルとの共起により検出された複合語についても同様に調べたところ、6.2%しかなかった。同一ディレクトリの場合と比較して、1/10しか検出されない。

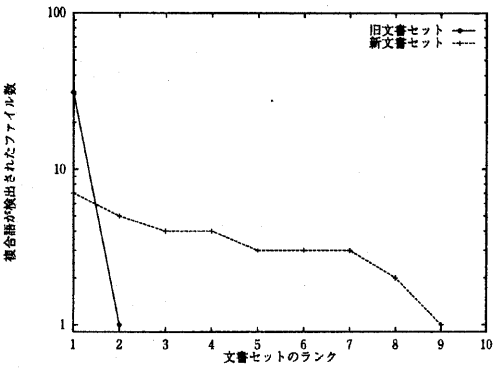

図 8: タイプが変化する複合語

これらの値は、「実際に文書セットを作成する時に、文書または文書の一部を同一ディレクトリに保存することが多い」という経験則をある程度裏付けるものであり、ディレクトリ単位に文書セットを作成することは、最適解とは言えないが、ある程度合理的な解の１つであると判断できる。

5 おわりに

本稿では、情報探索の支援となる検索関連語として複合語に注目し、情報リソースから複合語を抽出するための手法として、適切な文書セットを作成し、その文書セット中の出現頻度に応じて複合語を検出する手法を提案した。また、評価実験を行ない、文書セットによる複合語の分類、および、合理的な文書セットの１つの解としてディレクトリ単位の文書セットがあることを示した。

本稿は形態素解析の誤分割により生じた複数の単名詞を複合語として抽出することに着目したが、文書から抽出されるのは複合語だけでない。正しく形態素解析された単名詞も重要語として抽出される。今後は、このような単名詞も含めた重要語全体を見て、複合語検出のための適切な文書セットを作成する方法に関して検討を進めたい。

参考文献

[1] Paul Francis, Takashi Kambayashi, Shin-ya Sato, Susumu Shimizu: "Ingrid: A Self-Configuring Information Navigation Infrastructure", Proceedings 4th WWW Conference, Boston, pp.519–537, 1995.

[2] "ODIN – Open Documentary Information Navigator", <http://kichijiro.c.u-tokyo.ac.jp/odin/>.

[3] 原田昌紀, 清水奨: "WWW 検索システムにおける不特定多数の操作履歴の活用", 情報処理学会研究報告 97-DPS-81, 1997.

[4] 下畑光夫, 杉尾俊之: "文字種切り出しと複合語分解によるキーワード抽出", 情報処理学会研究報告 97-NL-120, pp.83–88, 1997.

[5] 原正巳, 中島浩之, 木谷強: "単語共起と語の部分一致を利用したキーワード抽出法の検討", 情報処理学会研究報告 95-NL-106, pp.1–6, 1995.

[6] 中渡瀬秀一, 大本晴夫: "統計的手法によるテキストからの重要語抽出メカニズム", 情報処理学会研究報告 95-FI-39, pp.41–48, 1995.

[7] 森信介, 長尾真: "n グラム統計によるコーパスからの未知語抽出", 情報処理学会研究報告 95-NL-108, pp.7–12, 1995.

[8] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真: "日本語形態素解析システム JUMAN 使用説明書 version 2.0", 1994.

[9] "goo", <http://www.goo.ne.jp/>.

[10] 小川泰嗣, 望月雅子, 別所礼子: "複合語キーワードの自動抽出法", 情報処理学会研究報告 93-NL-97, pp.103–110, 1993.

[11] 中川裕志, 森辰則, 松崎知美: "日本語マニュアル文における名詞間の連接情報を用いたハイパーテキスト化のための索引語の抽出", 情報処理学会研究報告 96-NL-116, pp.65–72, 1996.

[12] 神林隆, 清水奨, 佐藤進也, Paul Francis: "インターネット情報探索に適したキーワード抽出", 情報処理学会研究報告 97-NL-118, pp.79–84, 1997.

[13] Carlos R. Cunha, Azer Bestavros, Mark E. Crovella: "Characteristics of WWW Client-based Traces", <http://cs-www.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html>.