自然言語処理 118－13
（1997. 3. 22）

インターネット情報探索に適したキーワード抽出

神林 隆, 清水 奨, 佐藤 進也, Paul Francis
NTT ソフトウェア研究所
東京都武蔵野市緑町 3-9-11

インターネット上の情報を探索する場合、サーチエンジンは便利な道具である。しかし、インターネットは巨大で非均質な情報空間であるため、従来のデータベース技術をそのまま応用しただけでは、ユーザにとって使いやすいシステムになるとは限らない。優れた情報探索を実現するためには、検索語として適切な語がキーワードとして選ばれなければならない。本稿では、我々が分散型検索システム Ingrid で試みている成句検出を用いたキーワード抽出手法について論じる。Ingrid では、キーワードを切り出す際に成句検出処理を行なう。これにより、検索語として使われやすい特殊語や新造語、固有名詞をキーワードとして抽出できる。最後に、実験を行ない、その有効性を示す。

Keyword Extraction for World-Wide Information Discovery

Takashi KAMBAYASHI, Susumu SHIMIZU, Shin-ya SATO, Paul Francis
NTT Software Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo

Search Engines are convenient tools for exploring the internet. However, since the internet is a very large and heterogeneous information space, usual database technologies cannot be directly applied to the internet. Quality of information discovery depends on extraction of important keywords. In this paper, we propose a keyword extraction method using phrase detection which we tried in Ingrid — a project to implement a high quality information discovery in the internet. Ingrid detects phrases when it isolates keywords. Some of deteced phrases are special keywords, new created keywords or proper nouns which are often used as search terms. Finally we indicate its effectiveness from experimental results.

1 はじめに

インターネットの急速な普及とWWWの出現によって、電子化されたアクセス可能な情報がインターネット上に氾濫するようになった。しかし、どこにアクセスすれば本当に必要な情報が得られるのか知ることは難しい。このための手段としてAltaVista[2]やInfoSeek[3]のようなサーチエンジンが利用されることが多いが、これらはインデックスの作成に、従来のデータベース技術をそのまま応用したものであるため、インターネットの巨大性や非均質性に完全に対応できていない。

本稿では、インターネット上の優れた情報探索の実現のために、我々が研究中の分散型検索システムIngrid[1]で試みているキーワード抽出手法について論じる。第2節では、インターネット上のサーチエンジンが抱える問題を指摘し、第3節では、Ingridにおける問題の解決策について述べる。第4節では、解決策を実現するため、成句検出を用いたキーワード抽出手法について論じ、第5節では、成句検出の評価実験について報告する。

2 サーチエンジンの抱える問題

サーチエンジンは、ユーザが検索語を入力すると、検索語を含むURL(情報リソースの所在)を検索し、提示するプログラムである。ロボットと呼ばれるプログラムを巡回させWWWサーバの情報リソースを収集し、検索用のインデックスを事前に作成するのが一般的である。インターネット上で必要な情報リソースを発見するための手段として、サーチエンジンは必要不可欠であるが、次のような本質的な問題が存在する。

2.1 巨大性にともなう問題

サーチエンジンの優劣を決定する判断基準の一つは検索可能な情報リソース数であるが、検索対象数に応じてヒット数も増加し、必要とする情報を探すのが困難になる。たとえば、AltaVistaで「NTT」という単語で検索した場合、ヒット数は10万以上となり、内容どころか一覧をすべて見ることもできない。

この問題を解決するためには、検索語の変更を含む検索条件の見直しによる情報の絞り込みが必要になる。しかし、再検索のための検索語の選択は難しく、情報を絞り込めなかったり、逆に全くヒットしなくなることも多い。うまく絞り込むためには、検索対象の分野についての十分な知識と経験が必要になる。

2.2 非均質性(多様性)にともなう問題

インターネット上で扱われる情報は一般的に、HTMLファイルやテキストファイルのようなファイル形式にのみ従っていればよく、中身の書き方に関しての制約はない。公開された情報リソースは情報提供者のスタイルが反映され、言語や文字コードの違い、口語調／文語調の違い、使用する用語の違いなど、さまざまな違いが存在する。

従来のリレーショナルデータベース(RDB)やYahoo[4]などのディレクトリサービスのように、収集された大量の情報リソースに対して、データ項目ごとの分割や人手による関連キーワードの付与は多大の労力を伴うため、サーチエンジンでは一般的に全文検索を行なうことが多い。このため、文書の題名や著者名などのメタな属性を用いて情報リソースを限定することが難しいだけでなく、無関係な情報リソースまで検索されることが多い。また、情報提供者とサーチエンジン管理者が完全に分離するロボット方式では、WWWサーバごとの違いを吸収するのも難しい。

3 Ingridのアプローチ

前節で指摘した問題点を考慮して、我々は分散した情報リソースを管理するためのインフラストラクチャとしてIngridを設計した。情報リソースは従来のサーチエンジンのように集中管理されずに、インターネット上に分散した複数のサーバによって管理される。情報を公開して検索可能にするには、情報リソースを解析して、特徴を表すキーワードの組、情報リソースの所在を示すURL

などの情報を含むリソースプロファイル(RP)を作成し、これをサーバに登録する。情報提供者はサーバを立ち上げてRPを登録するだけで情報を提供でき、情報提供者が検索される情報リソースを管理できる。

サーバは、登録されたRP間にキーワードの同一性に基づくリンクを生成し、自律的にトポロジーを構築する。Ingridでは、検索クライアントがこのトポロジーのリンクをたどることで情報探索を行なう。

3.1 RPの作成

検索対象となるRPは、次の手順で作成する。

1) 元データ抽出と文書整形 情報リソースに対して、ファイルの解凍などのファイル操作、文書タイプの認識、エンコーディングの検出、パラグラフ単位の記述言語の認識を行ない。RP作成の元になるデータを抽出する。この時点で、文書整形や余計な部分の削除も行なう。たとえば、メイルやニュースの引用部分では、本来参照として表現すべきものをテキストの中に混在して表現することが多い。これが次の手順に悪影響を与えることがあり、特別に対処する必要がある。

2) 単語分割 複数の言語で書かれた情報リソースの混在や、複数の言語が混在使用されている情報リソースに対応するため、元データ抽出過程で得られた言語情報に基づき、各言語ごとの処理モジュールを使って単語の分割を行なう。現在、英語と日本語の処理モジュールが実装済みである。日本語の場合、単語を区切る目印が陽に存在しないので形態素解析プログラムを用いる必要があり、現在はjuman version 2.0[5]を使用している。分割後に、品詞情報と不要語のリストに基づいて、不要な語を除去する。

3) 単語重み付け 分割した単語の重要度を判定するために、$tf \cdot idf$ 法[6]と文書構造から得られる情報を用いて重み付けを行ない、キーワードとして抽出する。

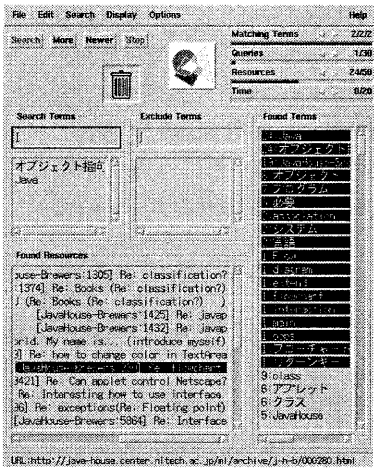

図 1: Motif版 Ingrid Navigator

3.2 関連語の提示

Ingridの検索クライアントは、検索条件の見直しをサポートする機能を持っている。図 1に、Motif版Ingrid Navigatorの画面を示す。ヒットした情報リソースだけでなく、RPに登録されているキーワード中で特に重要な語が、関連語として一緒に表示される(画面右側)。

ユーザは情報リソースを閲覧するだけでなく、関連語を使い、検索語の追加や変更を簡単な操作で行なうことができる。また、関連キーワードを見るだけでなく、ある関連語を含む情報リソースの閲覧、さらに、関連語に基づく情報リソースの類似性判定も行なえる。

4 成句検出

適切な関連語が提示されれば、情報探索に要する時間全体を短縮できる。そのためには、キーワードとして適切な単語に分割する必要がある。

4.1 形態素解析の問題点

形態素解析では、使用する辞書の品質が結果に大きく影響する。文章中に未登録語が出現すると、たとえば「福岡市」が「福」と「岡市」に誤分割されることがある。これを防ぐためには、辞

書を十分にメンテナンスしなければならないが、インターネットで使われる特殊語や新造語、さらに、分割されたら意味を持たない人名や企業名などの固有名詞を、事前にすべて網羅して辞書に登録するのは不可能である。しかし、このような語ほど検索語として使われることが多い。

4.2 成句検出手法

ある情報リソースで、ある単語の組が複数回繰り返されることは、その組で表現される語が情報提供者にとってかなり重要な語であると推測できる。このような繰り返しは文章中だけでなく、題名や章名、図や表のキャプションなどの文書構造に関連する部分や、HTMLのリンク名などの文書間の関係を示す部分でも発生する。

Ingridでは、形態素解析の後処理として、この性質を利用した成句検出を導入した。これは、形態素解析により分割された形態素の出現順序に着目して、出現順序が連続している形態素の組が同一情報リソース内である閾値以上出現した場合に、その組を成句として検出し、以後、それを一つの単語とみなす処理である。

現在は、名詞が2回以上出現する場合を対象としている。辞書未登録語を暫定的にサ変名詞として扱っており、これも成句検出の対象となる。カタカナの組については、それらが1つのまとまった語になることが多く確認されているので、閾値を1回にしている。図 2に示す例では、「形態」「素」「解析」という組が2回連続して出現しているので「形態素解析」として扱う。

成句検出により、不十分な辞書を使用した形態素解析において、情報提供者の意図を損なう形に分割された単語を復元できる。また、辞書に登録されていない特殊語や新造語、固有名詞も一つの単語として抽出できる。

4.3 成句検出と関連語表示

関連語表示機能を持つサーチエンジンとしてはIngrid以外に、Rcaau検索エンジンMondou[7]や

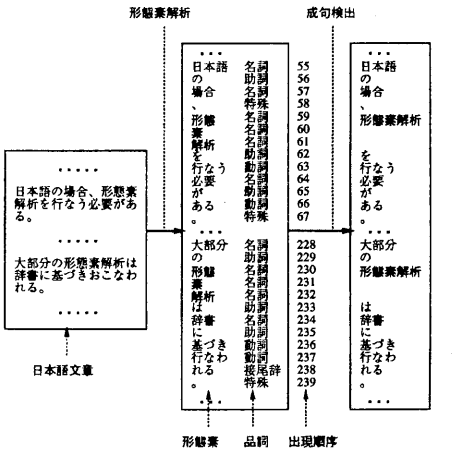

図 2: 成句検出処理

ODIN[9]がある。

Mondouは、抽出したキーワード群に対してデータマイニング手法を用いて関連語を抽出する[8]。これも形態素解析を行なうため、結果は辞書の品質の影響を大きく受け、検索関連語として不適切な語も提示されることが多い。

これに対して、ODINとIngridは2つの異なるアプローチを行なっている。ODINでは、ユーザが入力した検索語と閲覧した情報リソースの履歴から相関関係を解析し、同一情報リソースを閲覧するにいたった検索語を関連語として提示する[10]。この方法はユーザが重要だと思うキーワードを抽出する方法である。辞書が不要であり、ユーザが実際に入力した検索語を使用するために適切な語が提示されやすい。しかし、閲覧回数の少ない情報リソースに対してはよい結果が得られず、新しい情報や特殊な情報には向いていない。

Ingridの成句検出は、情報提供者が重要だと思うキーワードを抽出する方法である。辞書が不要であり、かつ、検出された成句は $tf \cdot idf$ 法による重みが高くなる特徴を持つ。

4.4 関連研究

成句検出のような複合語を扱った研究は、数多く報告されている。[11]では、形態素解析の結果

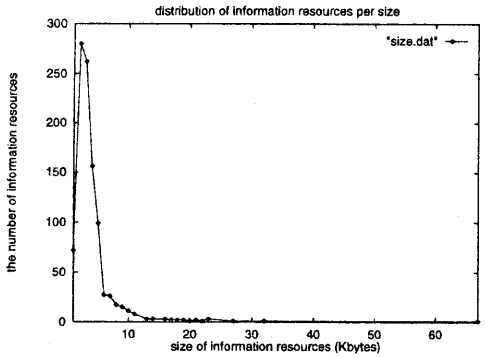

図 3: サイズごとの情報リソースの分布

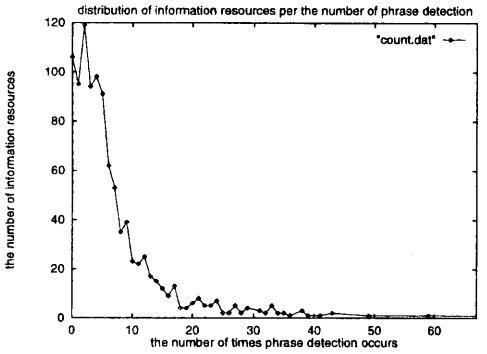

図 4: 成句検出発生回数ごとの情報リソースの分布

得られた形態素列から複合語キーワード抽出アルゴリズムを用いて複合語を抽出する。[12] では、直前と直後の字種の連結性に基づき複合語を切り出し、ある文章の範囲内での語の共起度を計算し複合語の重要度を決定する。[13] では、形態素解析を行なった後、複合語の候補として連続している単名詞の組をすべて拾い出し、単名詞の前後連結数を求めそれらの積により複合語の重要度を決定する。いずれの研究も、形態素の出現順序と回数を調べるだけで成句を検出する Ingrid の手法に比べると処理が複雑であり、[11] を除き、重要度決定のためのデータも大量に必要とする。

5 評価実験

成句検出の有効性を評価するために、情報リソースのサイズとの関係、および、成句の検索語としての適合性に関する評価実験を行なった。

5.1 実験構成

実験に使用する情報リソースは、Ingrid の主たる検索対象である HTML ファイルに限定した。18 の企業／団体のホームページからたどれる日本語の情報リソースを計 1008 個収集し、これに対して RP を作成し、成句検出を行なった。実験には juman version 2.0 をそのまま使用した。

5.2 情報リソースのサイズとの関係

HTML ファイルは、本の章や節、段落に相当する構造に応じて、複数のファイルに分割されることが多く、1 つの情報リソース当たりのサイズも小さくなる。図 3に、サイズごとの情報リソースの分布を示す。86.2% の情報リソースはサイズが 5K バイト未満であり、平均サイズも約 3.5K バイトであった。

このような情報リソースに対して、成句検出がどれだけ発生したのかを図 4に示す。1 情報リソースあたり平均 6.8 回もの成句検出が発生しており、89.5% の情報リソースで最低 1 回の成句検出が起きている。これは、HTML ファイルのように細分化された情報リソースに対しても、成句検出が十分動作することを意味する。

また、成句検出が 1 回も起きなかった情報リソースのサイズを調べたところ、その 78.3% が 2K バイト未満であった。このような場合に、どのように成句検出を行なうかは、今後の課題である。

5.3 成句の検索語としての適合性

実験により検出された成句数は、全部で 4028 個であった。これらを解析して、表 1のように分類し、検出に失敗したと思われるものに関しては、その症状によりさらに細く分類した。成句検出の効果は、単語の誤分割の復元と辞書未登録の固有名詞などの抽出であることから、少なくとも 4 割の成句検出は成功したと言える。どのタイプにも分類されなかったものについては、今回は明

表 1: 検出された成句のタイプと割合

分類タイプ		比率
誤分割が避けられたもの		15.9%
固有名詞と思われるもの		24.6%
失敗	意味をなさないもの	1.6%
	品詞設定ミスによるもの	12.1%
	検索語として長過ぎるもの	14.4%
上記以外		31.4%

確な評価基準を設定することができず、成功したとも失敗したとも言えない。以下に、失敗した原因と解決方法を示す。

意味をなさないもの　RP の元データ抽出後の文書整形の段階で、引用部分を完全に処理しきれなかったために発生する。たとえば「ク・コンピュ」が検出された。この原因は形態素解析の前段階処理にあり、成句検出では対応できない。

品詞設定ミスによるもの　形態素解析が接尾辞や副詞を名詞と誤認識してしまうために発生する。たとえば「サラダ油少々」が検出された。これを解決するには、他の品詞が名詞と誤認識されないように辞書を充実させる必要があるが、そのコストは名詞の場合よりかなり小さい。現在、そのようなコンセプトの辞書を作成中であり、評価を行なう準備を進めている。

検索語として長過ぎるもの　形態素の組を最長連結して成句を作ることに起因する。今回は 4 語以上連結したものをこのタイプに分類した。たとえば「オブジェクト指向プログラミングツール」が検出されたが、これはあまりにも特化された語であり、トポロジーの構築や関連語表示に用いるには不適切である。むしろ、「オブジェクト指向」や「プログラミングツール」のように分割されていた方がキーワードとして適切である。長過ぎる成句に関しては、キーワードとして適切な長さに切ることが必要となる。

6　まとめ

本稿では、情報リソースからキーワードを抽出する際に、適度な長さの単語を切り出すことがインターネットでの優れた情報探索の実現に重要であり、それを達成するための技術としての成句検出手法とその効果について説明した。今回の実験では、9 割の情報リソースに対して成句検出の効果が確認でき、成句のうち少なくとも 4 割はキーワードとして適切であり、成句検出が単語分割に有効であると考えられる。

成句検出により単語は適切に分割できるが、キーワードとして抽出されるか否かは、重み付け処理に依存する。情報リソースのサイズが小さく単語の頻度がすべて 1 である時など、頻度に基づく $tf \cdot idf$ 法では重要な語をうまく選び出せない。このような場合でも重要な語を抽出する方法については未検討であり、今後の研究課題である。

参考文献

[1] Paul Francis, 神林隆 他: "Ingrid: A Self-Configuring Information Navigation Infrastructure", Proceedings 4th WWW Conference, Boston, pp.519–537, 1995.

[2] "AltaVista Search: Main Page", <http://altavista.digital.com/>.

[3] "Infoseek", <http://www.infoseek.com/>.

[4] "Yahoo!", <http://www.yahoo.com/>.

[5] 松本裕治, 長尾真 他: "日本語形態素解析システム JUMAN 使用説明書 version 2.0", 1994.

[6] Gerard Salton: "Automatic Text Processing", Addison-Wesley, 1988.

[7] "Home Page of rccau -mondou-", <http://www.www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/>.

[8] 河野浩之, 長谷川利治: "WWW データ資源検索におけるデータマイニング手法", 情報処理学会研究報告 96-DBS-108, pp.33–40, 1996.

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

[10] 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.

[11] 林淑隆, 獅々堀正幹, 伊与田敦, 津田和彦, 青江順一: "複合語キーワードの効率的抽出法", 情報処理学会研究報告 94-NL-104, pp.63–70, 1994.

[12] 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.

[13] 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.