自然言語処理 129－4
（1999. 1. 20）

goo/InfoBeeが目指す自然言語処理

稲垣 博人　大久保 雅且　杉崎 正之　田中 一男
NTTヒューマンインタフェース研究所

インターネットの普及により個人が自由に情報を発信できる世界となった．このような情報流通社会において，goo/InfoBeeは水先案内人として絶えず革新を続けている．米国のインターネットの普及に伴い日本でもインターネットが普及し始めたころ，米国では既にインターネットを検索する種々のサイトが立ち上がっていた．NTTは，このようなインターネットサーチエンジンの必要性を予見し，97年3月gooサービスを立ち上げた．このインターネットサーチエンジンは，米国Inktomi社のエンジン部分を利用し，InfoBee自然言語処理技術を用い，日本語向け・日本人向けにしたインターネットサーチエンジンである．インターネットから情報を集める機能としてクローラがあり，さらに検索に必要なインデックスを作成するインデクサ，そしてそのインデクサが作成したインデックスをもとに検索を行うテキストリトリーバからなる．これらの機能は，自然言語処理が高度に融合したソフトウェアコンポーネントである．もちろん，インターネットサーチエンジンも単にインターネットの検索ができるだけでなく，種々のサービス，たとえば，フリーメイル，ディレクトリ情報サービス，各種データ提供サービスなど種々のサービスが必須機能となってきている．これらのサービスを統合的に提供するために，goo/InfoBeeは，種々の自然言語処理技術を用いた種々のサービスを実現してゆく．本稿では，goo/InfoBeeの基本的機能である，サーチエンジンにおける自然言語処理技術について述べ，さらに将来的に目指す自然言語処理について概観してゆく．

Natural language processing for the search engine goo/InfoBee

Hirohito Inagaki,　Masaaki Ohkubo,　Masayuki Sugizaki, and　Kazuo Tanaka
NTT Human Interface Laboratories

Not only mass-media but also personal media can easily be published in the Internet world, so many people began to think this is a huge network and we need navigator to traverse the Internet world. First, the States began to create several Internet search engines to navigate Internet. Corresponding such trends, NTT launched "goo" Internet search engine utilizing Inktomi's search engine technologies and InfoBee natural language processing technologies. Inktomi had been a small venture company that integrated the HotBot Internet search engine, but their scalable search engine technologies and Internet technology were better matching to InfoBee Japanese natural language processing technology. There are three parts in the Internet search engine. One is the crawler that gathers HTML texts from the Internet world. Second is the indexer that makes index from gathered HTML texts. Third is the text retriever that yields answer from a user query using index. The crawler, indexer, and text retriever utilize Japanese natural language processing. Of course, not only Internet search service but also several services such as mail services, information directory services, and information providing services are important as a portal service. In this paper, we review the basic characteristics of goo/InfoBee search engine and several services that utilize Japanese natural language processing as a portal service.

1 はじめに

近年のインターネットの普及により，多くの人が検索エンジンをインターネットの水先案内人として用いるようになってきた．特に，インターネットの場合，個人が発信する情報の比率が高く，かつ，整備されていない荒野のごとき情報発信の場である．このような情報発信の場では，自らが必要とする情報を探すのは難しく，水先案内人が必要とされる．この水先案内人である，goo/InfoBeeでは，インターネットを旅する人の水先案内人として種々のサービスを提供し，これらの旅人のポータル（入り口）となるべく，日々革新を続けているのである．その基本をなすのが，インターネットサーチエンジンである．このインターネットサーチエンジンでのキーとなる要素技術としては，インターネットに点在する情報を収集する機能(クローラ)および，それらの収集した情報を特徴付けする機能(インデクサ)，さらに，ユーザからの検索要求に対して適切な情報を提示する機能(テキストリトリーバ)の3機能がある．これらの機能がうまくかみ合うことにより，インターネット上にある大量の情報の中から，ユーザに必要な情報を抽出し提示することが可能となるのである．もちろん，同様なことを人間が行なうことも可能である．たとえば，Yahoo!などはクロール，インデックスの機能をサーファーと呼ばれるインターネット専門集団により実現している．おもしろい情報，話題の情報などを人手でクロールし，それらの情報を整理し，ディレクトリ形式にして提示している．goo/Infobeeは，これらの作業をすべて計算機上で行ない，自然言語処理技術を活用して，人間ができないような大量の情報を自然言語処理に基づき的確に処理していくのである．もちろん，水先案内人として，検索機能だけでなく，情報をディレクトリ形式に提示する情報ディレクトリ表示機能や，情報情報要約機能，情報トレンド提供機能などの各種の自然言語処理を応用した機能・サービスが実現または検討されている．本稿では，goo/InfoBeeにおける基本機能であるサーチエンジン機能および，自然言語処理技術を利用した，情報ディレクトリ表示機能，情報速覧・要約機能，情報トレンド提供機能などに着目し，goo/InfoBeeが求める自然言語処理を概観してくゆく．

2 goo/InfoBeeの検索機能

検索機能は，インターネットサーチエンジンが提供する最も基本的な機能である．gooにおいても，検索機能は基本的機能である．gooのサーチエンジン機能は，米国Inktomi社[1)]の技術を活用したサーチエンジンである．Inktomi社は，自社ではサーチエンジンサービスを行なわず，システムインテグレータとして，米国最大のHotBotや，オセアニアのサーチエンジンであるAnzwers，南米のサーチエンジンであるradarUOLなどを手掛け，かつワールドワイドパートナシップにより，これら世界のサーチエンジンと提携している．元々，Inktomi社は，UCバークレイの大学の教授であるDr. Brewerが中心となって築いた会社である．UCバークレイでは，自然言語処理の研究ではなく，比較的安いワークステーションを高速のネットワークで多数接続することにより超並列計算機の機能を実現するNOW(Networks Of Workstations)の研究を行なっている．これは，現在流行の計算機クラスタのコンセプトを実現したものである．各ワークステーションを接続する個別ネットワークとしては，Myrinetのネットワークを利用している．計算機クラスタでは，このネットワークを介してクラスタを構成する各ノードと通信を行ない，計算に必要な情報をやり取りする．NOW応用研究の一研究成果として検索エンジンがある．複数台からなるインターネットのクローラによりインターネット上の大量のHTML文書を収集，インデックス化し，それを複数台からなるNOWテクノロジを活用した検索サーバがユーザからの大量な検索要求を瞬時に検索してゆく．例えば，ユーザからの検索要求があると，NOWによりネットワークで複数の検索サーバに対して検索要求を伝播し，各検索サーバが検索要求をそれぞれ独立に計算することにより，単独の計算機で計算するより短時間で結果を出力することが可能になるのである．もちろん，インターネットは，日々膨張しているので，インターネットの情報量の増加に呼応して，検索能力がスケーラブルに拡張できる必要がある．このようなスケーラブルな機能の拡張は，計算機クラスタが最も得意とする点である．ノードを増設することにより計算能力を増加させることができる．

このような，計算機のスケーラビリティや計算能力を活用した一成果として，検索機能が実現されているのである．なぜゆえ自然言語処理を専門としない会社からでもこのようなプロダクトが出せるかといえば，それは，英語の特徴による点が大であると考えられる．つまり，英語はスペースで単語が切り分けられているため，自然言語処理を用いずとも，単語を認識できる．名詞であれば単数形，複数形の違いを考慮し，動詞であれば，活用

形を考慮する程度で良い．ある意味では，英語の場合，自然言語処理などのように複雑で，重い処理を行なわなくても，ある程度の処理が可能ということである．さらに，検索というタスクを単純化し，検索対象の文書をキーワードの集合と考え，キーワードの集合をインデックス化し，ユーザの検索クエリに対して，最も一致するキーワード群を持つ文書をスコアに応じて提示するタスクと捉えればよい．

このような英語の特徴を生かした検索エンジンではあるが，日本語の場合は，英語と同様な方法ではローカライズできない．そのため，gooは，NOWを使ったスケーラブル計算機技術とInfoBee自然言語技術を融合させ，日本語の日本人のためのインターネットサーチエンジンとしてローカライズされた検索エンジンである．もちろん，日本人のためのサーチエンジンであるから，日本語を中心にインターネットをクロールし，日本語文書のインデックスを作成する．さらに，ユーザからの検索要求を形態素解析し，適切な検索要求に変換するなどの処理を実現しなければならない．このように，gooでは，自然言語処理と密接に結び付いた検索エンジンである．以降で，インターネットサーチエンジンを構成する3要素技術について詳細に述べる．

2.1 goo/InfoBeeの要素技術

goo/InfoBeeなどのインターネットサーチエンジンでのキーとなる要素技術は，以下の要素である．

- クローラ技術: 情報を収集する技術
- インデクサ技術: 収集した情報を特徴付けする技術
- テキストリトリーバ技術: ユーザからの検索要求に対して適切な情報を提示する技術

これらの要素技術を組み合わせることにより，インターネットサーチエンジンが構成される．図1に要素技術がどのように組合わされているかを示す．ユーザはインターネットサーチエンジンのフロントエンドであるWWWサーバと会話する．検索処理においては，WWWサーバは検索エンジンであるテキストリトリーバによりユーザ検索要求に対して最も適切な文書情報を提供する．

2.1.1 クローラ技術

クローラ(別名スパイダー，ロボット，ワンダラーとも呼ばれている)は，検索を行なう対象文書を収集する機能である．イントラネットでは，通常，これらの検索対象となる文書は，ファイルサーバ等に蓄積されている場合が多いため，LAN経由で容易に対象文書に対してアクセスできる．その場合，対象とする文書のディレクトリやマシン名を指定することにより検索対象文書を特定することができる．一方，インターネットでは，外部のリソースにアクセスする方法は，通常HTTPやFTPである．これらのプロトコルを通して，外部のリソースをアクセスすることになる．FTPなどでは，ユーザ名を指定したり，Anonymusで特定のディレクトリ配下の外部リソースをアクセスすることが可能になる．HTTPでは，特定のディレクトリ配下の外部リソースにアクセスするだけでなく，外部リソースがHTML文書であれば，そのHTML文書の中からハイパーリンクで他の外部リソースを直接参照することも可能となるのである．このような外部リソースがハイパーリンクで接続されているような網の目のネットワーク状態を辿りリソースを収集するため，クローラとかスパイダーとかの名称でインターネットの文書情報収集機能は呼ばれるのである．クローラは通常シードと呼ばれる，クロールの基点となるリストとして持つ．そして，クロールして得たHTML文書のリンク先をさらにシードとして取り込み，インターネット上の多くのHTML文書を収集する．

このようなハイパーリンクを辿るクロールで一番問題となるのが，クロールされたくないページの処理と，同じハイパーリンクや巡回リストを辿ってしまった場合の処理である．

クロールされたくないページをいかにクロールしないようにするかだけでなく，逆に，ぜひ自分のページを検索されるようにクロールしてほしいという場合もある．クロールしてほしくないという場合には，robots.txtといわれるクロールされたくないHTML文書集合を記述する規約に基づき，クロールしないように設定することを推奨している．つまり，あるサイトのマシンをクロールする際に，robots.txtがあるかないかを確認して，robots.txtがある場合には，その規約に応じて，クロールすべきHTML文書を特定し，クロールを行なうのである．図2にrobots.txtの例を示す．このファイルをWebサーバのドキュメントルートに置く．図の例では，すべてのクローラ(User-agent)に対してクロールを行なわないように設定する場合の記述例である．

もちろん，robots.txtを記述できなかったり，記述が適切でなかったり，設定が困難であったりする場合があ

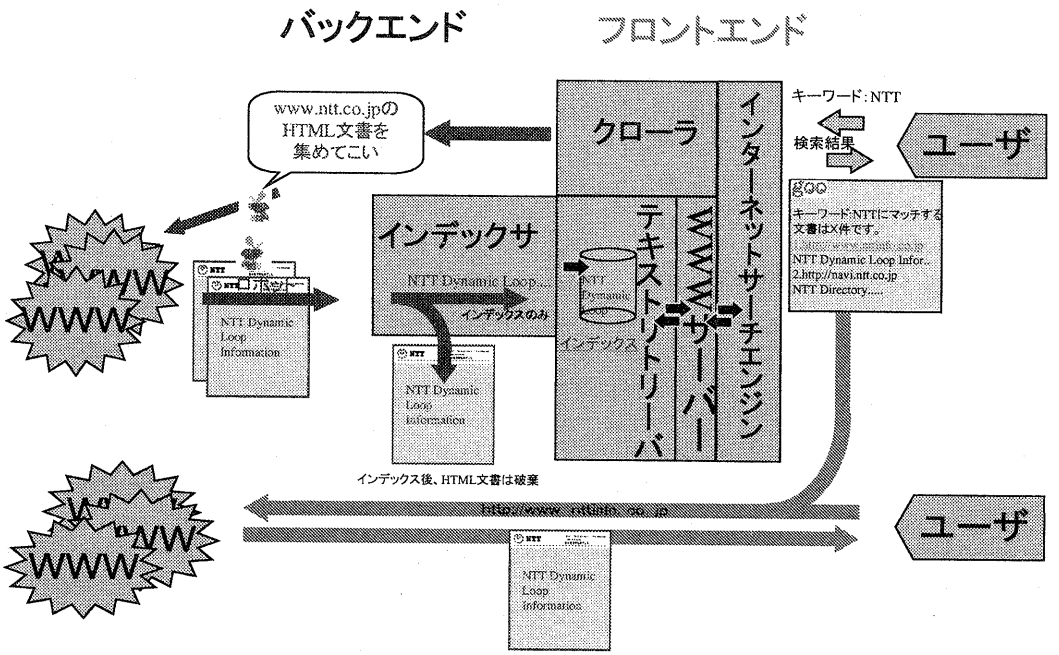

図 1: インターネットサーチエンジンの仕組み

```
User-agent: *
Disallow:/
```

図 2: robots.txt の例

る．特に日本のサイトでは，robots.txt を記述していないサイトが米国に比べて多い．そのような場合には，クローラ側でネガリスト，つまり，クロール側で，クロールしてはいけないリストを作成し，そのネガリストに応じてクロールを行なう必要がある．もちろん，逆にクロールしてほしいサイトについては，同様のクロールリストを作成し，クローラ側でそのクロールリストに応じてサイトの情報を収集する必要がある．

さらに，goo のように複数台でクロールを行なう場合，同じハイパーリンクをたどらないよう複数台で連携をとってクロールすると共に，巡回リストにはまらないように，URL の深さをチェックするなどの機能が必要となる．また，クロール時には，サイトに対して，過度の負荷をかけないように，帯域幅に応じてクロールしている．

インターネット上の情報は絶えず更新されているため，goo ではフレッシュなクロールデータを目指して，日々新しいインターネットの情報を収集することを行なっている．それぞれのサイトの HP の更新頻度や更新情報はクロール側では把握できない．そのため，クロール側として，できるだけフレッシュなデータを収集し，フレッシュなデータで検索ができるようにしなければならない．そのため，頻繁に更新される情報については，頻繁にクロールし，それほど頻繁にクロールされない情報については，ある程度間隔をおいてクロールする必要がある．goo では，前者の機能をマイクロクローラの機能で実現し，後者については，通常のクローラで行なっている．マイクロクローラによって，たとえば，新聞社や株式市場など，日々や時間単位に情報が更新されるサイトの情報を絶えずフレッシュな状態にできる．このようなフレッシュでなければ価値がない情報に対して，マイクロクローラは有効である．

HTTP プロトコルを利用して HTML 文書を収集する上では，日本語特有の問題は発生しない．つまり，HTML 文書が英語であろうが，日本語であろうが同様な処理がなされる．しかし，goo では，日本人のための

日本語のポータルサイトを目指しているため，URLが日本語を用いたHTML文書を持つかどうかを確認しながらクロールを実施している．このような処理により日本国内のjpドメインだけでなく，海外の日本語サイトも含め1700万URLを一挙にクロールしているのである．

2.1.2 インデックサー技術

クロールした情報を次に，高速に検索するために前操作を行なう必要がある．単に検索するだけであれば，クロールしたHTML文書群をgrepなどを使って文字列検索してもよいわけであるが，1700万URLを高速に検索しようとすると問題は非常に難しくなる．検索を高速に行なう方法としては，種々の方法が提案されている．例えば，文字や単語のn-gram法や，テキスト情報をwaveletなどの別種の変換を行ない特徴付けによる検索を行なう方法などがある．gooでは，日本語を基にしたキーワードインデックスファイルを作成し，それに基づいて検索するタイプである．キーワードインデックスでは，英語の場合，スペースで単語が分離するため，活用形や複数形のルールをプログラムしておけば，キーワードを分割することができる．さらに，冠詞など文書中には頻出するがキーワードとしてあまり意味をなさない[2]ような語は除きインデックスが行なわれる．日本語の場合には，単語がスペースで分離されているわけではない．そのため，適切な単語にキーワードを分離しなければならない．単語を切り出すには，字面で切り出す手法もあるが，goo/InfoBeeでは，高精度かつ，高速な単語解析が可能である形態素解析法を利用して，キーワードを抽出している．そのため，適切なインデックスを短時間に作成できる．形態素解析の解析精度としては，98%の精度で単語を解析できる．この形態素解析された単語の中で，インデックスとして不要な語を除き，インデックスを作成する．さらに，インターネットでは，複数の文字コードが存在するため，該当文書の文字コードが日本語かどうかを判断し，かつ，日本語のどの文字コード(SJIS，EUC，JISなど)で記述されたかを適切に判断し，形態素解析を行なわなければならない．

2.1.3 テキストリトリーバ技術

テキストリトリーバ技術は，インデックスに基づいて，ユーザからの検索要求に対して，適合した文書を提示する技術である．goo/InfoBeeでは，ユーザからの検索要求から適合文書の出力まで1秒以内でレスポンスすることを設計指針としている．図1に示すように，インターネットサーチエンジンでは，複数の技術からなるコンポーネントを組み合わせているため，ユーザからの検索要求を1秒以内とするため，検索結果を出力するまでに通るコンポーネントのすべてを最適化しなければならない．例えば，WWWサーバ自体のレスポンスも高速化しなければならない．CGIを使うなら，Perlなどの起動に時間のかかるCGI言語を用いずC言語インタフェースを用いたり，CGIでなく，WWWサーバ特有の高速アクセスの手法，例えば，windowsNTのインターネットインフォメーションサービス(IIS)のISAPIエクステンションなどを利用しなければならない．

テキストリトリーバは，これらのCGI等から直接呼び出される．CGI側は，検索要求を適切なメッセージに変換して，テキストリトリーバに検索要求を出す．例えば，適合文書を出力する件数や，検索要求の中に記述されているAND，ORなどの論理演算を指定する．さらに，インターネットサーチエンジンであるgooでは，追加機能として，以下のような機能がある．

- 検索キーワードの追加
- 日付などの指定
- 検索先のURL（ロケーション）の指定
- イメージ，音声等のデータタイプの指定

これらは，検索対象であるHTML文書を，HTML文書が作成された日付，URL，データタイプなどから特徴付けているのである．もちろん，イントラネット等のように専門化したシステムでは，インターネットよりもさらに多種の特徴付けが必要となる．例えば，議事録では，何の会議の何回目で，いつ行なわれ，参加者は誰であるのかなどの対象とする文書により特徴付けられる項目を個々に設定できなければならない．

テキストリトリーバでは，検索要求に含まれているキーワードと検索対象文書のキーワードとの一致をみて適合文書を決定するが，適合した文書をただ単に出力するだけであると，インターネットの場合，多いものでは何万件というURLが適合してしまう場合がある．そのため，適合した文書を順序付けし提示する必要がある．gooでは，キーワードの頻度や，ドキュメントの長さ，URLの深さ，タイトルにキーワードが含まれているかなどの情報を基に適合文書のスコアを算出し，スコアの高い文書から順に出力する．もちろん，インデックスファイルが複数存在し，複数のテキストリトリーバで情

報を検索するような比較的大規模な分散検索サーバの場合，分散環境での検索を考慮したスコアリングが必要となる[3]．

また，検索結果を出力する場合も，単にタイトルやURLを出力するだけでなく，該当文書の内容を表す要旨を提示し，検索結果の中からユーザの要求する文書をすばやく見つけられるようにしている．これは，単にHTML文書の先頭から100語出力したのではなく，InfoBeeの速覧技術[4]を応用し，最も適切と考えられる文をHTML文書から抽出して要旨として表示しているのである．文書に記述されている，話題を表す”～について”や”まず始めに”などの話題の手掛かり句をもとに重要文を抽出し，その重要文の中で最も適切な文を100語の要旨として抽出している．

さらに，単純に検索するだけでなく，類似文書を検索する機能など，キーワード検索機能を支援する必要がある．類似検索機能とは，ある検索要求に対して出力した適合文書をもとにその適合文書に類似した文書をさらに検索する機能である．これは，例えば，検索要求が最初は，漠然としていたものが，適合文書をみることにより，さらに明確化された場合や，適合しているかどうかは判断できるが適合している文書を表す適切なキーワードが思い浮かばない場合などには非常に有効な手法である．

3 goo/InfoBeeの自然言語処理応用技術

インターネットサーチサイトは，ボリュームの拡大，つまり，検索対象のドキュメントの量，日々のアクセス数だけでなく，サービス・機能面での拡大が行なわれている．例えば，gooでは，基本的検索機能だけでなく，情報ディレクトリ表示機能，Web経由のメイル機能，有料・無料のコンテンツ提供機能，グリーティングカード機能など，サイトを入口とするための多種のサービスが提供されている．

情報ディレクトリ表示機能は，98年5月にリニューアルした際に追加された機能である．イントラネットの場合，ある程度文書等が体系化されているため，検索結果を体系化しやすいが，インターネットのように，体系化されていないリソースの場合は，検索結果を単純にスコアに基づいて表示するだけでなく，検索結果をある種の体系化された状態で見ることにより，検索結果の把握が効率化される．

さらに自然言語を応用した機能として，次の機能が検討されている．適合文書を見易くしたり，書かれている内容を短時間に把握するための情報速覧・要約機能や検索ログに基づく情報トレンド提供機能などである．以下で，これらの自然言語処理を活用した種々の機能について述べる．

3.1 情報ディレクトリ表示機能

情報ディレクトリ表示機能は，情報を分類する自然言語処理技術を利用した機能である．情報を分類する場合，あらかじめ情報を分類するための箱が決定されている場合と，情報を分類するための箱は用意されておらず，ある文書集合が与えられ，その文書集合に対して適切な分類体系を作る場合の2種類の方法がある．あらかじめ分類体系が決定されている場合には，それぞれの分類に応じて文書集合を割り当てていくのである．我々は，この割り当て処理において，それぞれの分類に適した文書例を学習させ，各分類を特徴付けるキーワードを抽出し，既学習済みのキーワードと分類対象のキーワードとの意味的距離算出し，最も意味距離が近いものに分類する手法[5]を利用している．この場合，多くのキーワードとの意味距離のキーワードを計算するため，意味距離の計算や分類体系を特徴付けるキーワードの指定が分類精度に影響する．また，計算速度を高速化するための種々の絞り込みや分類の推定処理などが必要となる．分類体系を用意して，対象文書を分類体系に割り当てる方法では，新しい分類が必要になったり，1つの分類体系に属す文書集合が多くなり過ぎてさらに細分化が必要になった場合などの分類体系の動的変化に対しては弱い．その場合には，ある文書集合が与えられた時，最も適切な分類体系を作り直す必要がある．全く分類体系が与えられていない場合，1から分類体系を構築していかなければならないため，分類体系として適正なものを構築することは難しい．そのため，初期はある程度の分類体系を用意し，対象文書の増加や変化に応じて適切な分類体系に変化させるなどの両方の手法を融合させた手法が必要になると考えられる．

3.2 情報速覧機能

大量の検索された情報を短時間で理解・把握するために種々の方法が検討されている．提示されている情報の中から重要な情報をピックアップして可視的に提示する文書可視化提示による支援技術を我々は研究してきた[6),7]．その中で，文書の構造として話題構造を提案し，文書中の話題をコントラストをつけて表示する情報

速覧技術を提案している．これらの情報を構造化して提示する支援技術は，特に，サーバ側だけでなく，クライアント側での使用を考慮しなければならない．そのため，形態素解析レベルの表層解析を中心に自然言語処理を行なうことによりドメインに依存しない処理を実現した．

この話題とは，文書中の章・節・段落など，あるブロックの意味内容の中で中心となっている事柄を意味する．つまり，竹下ら[7]の定義する「文章中で記述されている内容の内，記述されている同じ内容をその文章中の言葉で表した語」を話題とした．つまり，文書内のブロックの意味内容を明示する語又は名詞句相当語句が，個々のブロックにおける話題となる．この話題を文章中で構造化したものが話題構造である．話題構造としては，明示的に話題が提示される大局話題と，局所的に話題が変化する局所話題の２つの話題を考慮した．大局話題は入れ子になりうるが，局所話題は大局話題のある一部分で局所的に発生する．それぞれの話題は，話題を導入する手がかり句と話題となりうる名詞句に後置される話題マーカを利用して決定される．例えば，大局話題なら，“まず”，“第一に”，“最初に”，“このため”，局所話題なら，“例えば”，“１つに”など，話題を導入する語句を持つ文で，かつ，“～について”，“～が”などの話題マーカを後置する名詞句が話題として抽出される．このようにして抽出された話題を順次文書に対して提示することにより，文書が話題によりブロック化でき，文書の把握が容易にすることが可能となる．

3.3 情報要約機能

大量の検索された情報を短時間で理解・把握するための手法として，情報要約技術による手法も検討されている．情報要約手法としては，3.2で述べた話題構造に基づく要約手法と，深層解析として，事象解析[6]を用い，事象という概念で文書を構造化することにより要約を生成する手法の２手法[8]が研究されている．

話題構造に基づく要約では，要約において重要な文とは，要約対象である文に記述された話題であるとの定義のもとに，話題を抽出し，話題を構成する文から要約を生成する手法である．3.2で述べた手法により話題が抽出され，その話題に応じて単文又は文を重要度付けし，要約として重要な文を決定する．この場合，要約対象の文をほとんど変更せず，そのまま結合して要約として出力する方法を用いている．文としては，連用中止形で接続された複文を単文化する程度で，なるべく入力の文をそのまま出現順に出力する要約生成を用いている．この場合，形態素解析または話題解析という表層的解析だけで処理されているので，ドメインに依存せず，非常に短い時間で処理できる．

一方，事象解析に基づく要約手法では，文書中で明確に事象を述べている内容を構造化し，誰が，何に対して，いつ，どこで，どうしたなどの情報を構造化する．この事象構造に基づき，要約を生成する．この深層解析では，文書中の事象に基づく深層構造を解析するとともに，事象に関係する項目を構造化し，重要度付けして情報を取捨選別することも行なっており，事象に関する情報という意味でのフィルタリングも行なえる．この事象解析により，より高度な要約を生成することが可能となる．要約文を生成する手法として，出現順に重要であるとする単純な手法やマスメディアなどにより多くのメディアで重要であると判断された事象を重要であるとする事象の重要度決定方法[9]を用いることにより，より高度な重要度付けが可能となる．例えば，出力するリソースに合わせて，出力文字数を変更したり，出力レイアウトを適切な形にすることもでき，単にパソコンに検索結果を要約として出力するだけでなく，携帯電話やページャなどあらゆるリソースに対して適切な要約コンテンツを提供することができるようになる．

3.4 情報トレンド提供機能

情報トレンド提供機能とは，検索のログ情報を基にトレンドなどの利用情報を提供する機能[10]である．検索サイトでは，検索に使用されたキーワードの累積情報などの統計的な情報を持ってはいるが，検索に使用されたキーワードの時間分析に基づくトレンド情報の提供は実現されていなかった．インターネットのサーチエンジンのログから，検索に使用されたキーワードと検索された時間情報を取得し，さらに，ブラウザのクッキー情報を利用して，検索したユーザ数を推定することにより，多くのユーザが考えている検索におけるトレンドを抽出する手法である．例えば，花見の時期であれば，花見に関連する“桜”，“花”，“花見”，“開化”，などのキーワードが瀕出し，現在のトレンドは花見であることが明確にされる．これらのトレンド情報を基に，サイトは，花見に関連するグッズ，開化情報などタイムリーな，情報流通・物品販売が可能となっていくのである．もちろん，これらの語が関連するということをシソーラス情報を利用して関連付けすることも可能である．例えば，“桜”と“花”はシソーラスでは上位・下位関係として関

係付けられるが，“花見”と“開化”などのように，関連が抽象的な場合，関連付けは難しい．そこで，同一情報へのHTTP要求で使用された検索キーワードであるのか，各検索キーワードの使用頻度の時系列の相関があるのかなどの情報を基に，検索キーワードをグループ化し，そのグループ語の利用頻度が高いものをトレンドとして抽出した．この情報トレンド提供機能により，検索されたログ情報から現在話題になっている事柄，皆が興味を持って検索されている内容，言葉などの利用情報の提供サービスが実現できる．

4 まとめ

インターネットの水先案内人として，インターネットサーチエンジンgooは日本最大級のドキュメント数とアクセス数を持つ．このインターネットサーチエンジンは，米国のInktomi社のNOW検索エンジン機能とInfoBee自然言語処理機能とが融合した，日本人のための日本語のサーチエンジンである．サーチエンジンはクローラ機能，インデクサ機能，テキストリトリーバ機能の3基本機能より構成され，これらは密接に連携している．

もちろん，gooでは，検索機能だけでなく，種々のサービス・機能が実現されてきた．その中で，自然言語処理を応用した機能としては，情報ディレクトリ表示機能，情報速覧・要約機能，情報トレンド提供機能が実現または検討されている．情報ディレクトリ機能は，検索された文書集合の提示に有効な手法である．情報の速覧・要約機能は，検索情報の短時間での理解・把握を支援する機能である．情報トレンド提供機能は，検索されたログ情報からトレンド，皆が興味を持っている事柄などを提供する機能である．これらの自然言語処理を応用した種々のサービス・機能を実現することによりgoo/InfoBeeはさらにユーザにとってポータルとなっていくと考えられる．

参考文献

1) http://www.inktomi.com.

2) H. P. Luhn. The automatic creation of literature abstract. *IBM Journal, Vol2*, 1958.

3) 森大二朗, 大森信行, 田中一男. 分散型大規模文書検索システムに関する一検討. 情報処理学会デジタルドキュメント研究会,DD15-2, 1998.

4) 稲垣博人, 早川和宏, 田中一男. 話題構造および文意味内容に基づく文書可視提示方式の提案. 情報処理学会第57回全国大会,4R-10, 1998.

5) 杉崎正之, 森大二郎, 大久保雅且, 田中一男. 疑似カテゴリ生成によるテキスト自動分類の高速化について. 情報処理学会第57回全国大会,2V-3, YEAR=1998.

6) Hirohito Inagaki and Tohru Nakagawa. An abstraction method using a semantic engine based on language information structure. *Coling-92*, 1992.

7) 竹下敦, 井上孝史, 田中一男. テキストの概要把握支援のための話題構造抽出. 情報処理, 1996.

8) 稲垣博人, 早川和宏, 田中一男. 情報流通向けテキストコンテンツ要約手法について. 情報処理学会デジタルドキュメント研究会,DD15-3, 1998.

9) 稲垣博人, 早川和宏, 田中一男. 類似意味内容の統合による伝達型電子化文書要約方式の提案. 情報処理学会第57回全国大会,4R-11, 1998.

10) 大久保雅且, 杉崎正之, 井上孝史, 田中一男. Www検索ログに基づく情報ニーズの抽出. 情報処理学会論文誌,vol.39,no.7,pp.2250-2258, 1998.