データベースシステム 109－46
（１９９６．７．２６）

意味解析機能を備えたWWW検索システム

鳥居 肖史
沖電気工業(株)　研究開発本部　関西総合研究所

我々は、WWW の検索システムを開発している。本システムの目的は、(1) 検索者が手軽に検索できることと、(2) システム運用者が辞書を低いコストで作成し更新できることである。前者の目的のために、(a) 日常的に用いる意味が曖昧なキーワードを使うための意味解析機能、(b) 不必要な情報をフィルタするための情報フィルタ機能、(c) 検索者が検索式を修正する手間を省くための検索式修正機能を備える。また、後者の目的のために、(d) システムが辞書をコーパスから作成するための辞書作成機能を備える。本稿では、本システムの各機能とシステムの構成、動作について述べる。

Retrieval system based on WWW with semantic analysis

Shouji Torii
Kansai Lab. , Research & Development Group , Oki Electric Industry Co., Ltd.

We are developing the retrieval system based on WWW. Aims of our system is to save not only user's labor but also system administrator's labor. The following are function we equipped this system for user. (a) Semantic analysis to be able to input ambiguous keywords. (b) Informational filter to reduce information like noise. (c) query modify function. Also , the function we equipped for system administrator is automatic dictionaries generation that utilizes text corpora.

1 はじめに

近年、インターネットでの World Wide Web (WWW) の利用が爆発的に増加し、商用やボランティアで WWW の検索サービスが立ち上がっている。インターネットに分散する WWW のデータベースを検索するため、ベクトル空間法 [1] の WAIS や WebCrawler[2] などがよく用いられる。しかし、そのようなシステムには、検索者サイドとシステムを運用する運用者サイドに以下のような問題点がある。

2 問題点の整理

2.1 検索者サイドの問題点

2.1.1 曖昧なキーワードの処理

日常会話では意味が曖昧な言葉を頻繁に用いるので、検索においても意味が曖昧なキーワードが使えるなら、日常会話に近い状況で手軽に検索できると考えられる。特に、検索の目的自体が不明確な状況では、検索者が特殊な専門用語を入力する必要があるよりも、意味が曖昧なキーワードが使えるほうが利便性が高い。

ここで、図1のように検索式が

“アメリカ and コンピュータ and 会社”

であるとする。“会社”には、“メーカ”や“銀行”や“デパート”などがある。しかし、サーチャーなら、「“コンピュータ”業界では“アメリカ”の“メーカ”が有力であること」が広く知られていることなどから、

“アメリカのコンピュータ・メーカ”

についての情報を意図していると考えるであろう。つまり、曖昧語である“会社”の意味を“メーカ”とする。

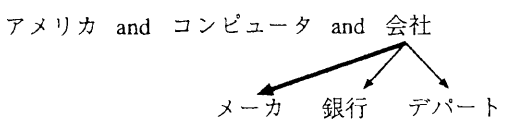


図 1: 曖昧なキーワードの処理

通常、システムは検索洩れを防ぐために、曖昧語をとり得る全ての意味に置換して検索する。その結果、検索結果には、検索者が意図しない文書（ノイズ）が多数含まれることになり、検索者は大量の検索結果の中から自分が意図した文書を探す必要が生じる。従って、検索者が手軽に検索できるためには、日常的に用いる意味が曖昧なキーワードを、検索者が意図したキーワードに変換する機能（**意味解析機能**と呼ぶ）が必要である。

2.1.2 ノイズの出力

システムは検索式により定まる文書集合を出力する。そのため、以下の例のように、検索式が意味が明確なキーワードのみからなる場合であっても、ノイズも出力する。

図2のように検索式が

“コンピュータ or ワープロ”

であり、データベースには文書1と2があるとする。文書の内容はキーワードが象徴するように、文書1はコンピュータでの文書処理、文書2は高速演算である。ここで、サーチャーなら、「“コンピュータ”と“ワープロ”に共通する代表的なことが文書処理であること」から、

“コンピュータやワープロでの文書処理”

についての情報を意図していると考えるであろう。その結果、文書1だけを検索結果とする。

文書	キーワード
1	コンピュータ DTP
2	コンピュータ マルチプロセッサ

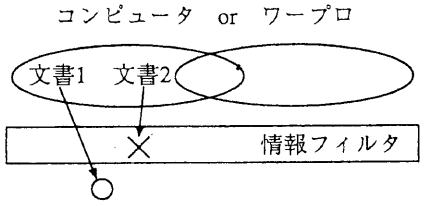


図 2: ノイズのフィルタリング

しかし、システムは、検索式の集合が文書1と文書2を含むので、ノイズである文書2も出力する。このように、システムが出力するのは検索式が定める文書集合なので、検索結果にはノイズが含まれることになり、検索者は検索結果の中から自分が意図した文書を探す必要が生じる。従って、検索者が手軽に検索できるためには、検索式が定める文書集合からノイズを除去する機能（**情報フィルタ機能**と呼ぶ）が必要である。

2.1.3 検索式の修正

検索者が目的とする文書を得るまでに、検索者は検索結果を参考にして検索式を修正して検索する。

例えば、図3のように検索式

“アメリカ and コンピュータ and 会社”

で検索した後、

“日本 and コンピュータ and 会社”
“コンピュータ and 会社”

などに修正する。

従って、検索者が次に入力しそうな検索式を推測し、検索式の修正を代行する機能（**検索式修正機能**と呼ぶ）があれば、検索者はより手軽に検索できるようになる。

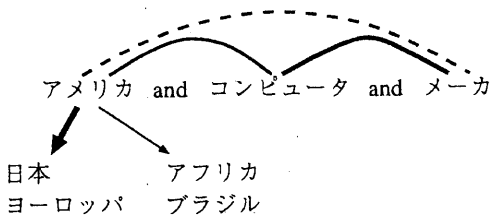


図 3: 検索式の修正

2.2 運用者サイドの問題点

2.2.1 辞書の作成と更新

我々のシステムは WWW の検索システムである。WWW では、科学技術用語や時事解説で現れる言葉などの多数の新しい言葉が現れる。また、検索者サイドからは、新しい情報を検索したい要望が高い。そのため、新しい言葉が使われるようになるたびに、その新しい言葉を用いて検索できるようにする必要がある。

従って、システム運用者が辞書を低いコストで作成し更新できるようにするためには、システムが辞書を作成し、新しい言葉が現れるたびにシステムが辞書を更新する機能（**辞書作成機能**と呼ぶ）が必要である。

3 適用技術

3.1 検索者サイドの適用技術

3.1.1 意味解析機能

一般に、言葉の意味は背景にある知識である文脈を利用することによって明確になる。例えば、検索式

“アメリカ and コンピュータ and 会社”

において、サーチャーが“会社”を“メーカ”に無意識に変換するメカニズムは、以下のように説明できる。“アメリカのコンピュータ”を話題にしている時には、“メーカ”を連想する頻度のほうが、“銀行”や“デパート”を連想する頻度よりも高い。従って、“会社”を“メーカ”と解釈すると考えられる。

このメカニズムを一般化して、意味解析機能のアルゴリズムを考案した。言葉には、意味が広い言葉と意味が狭い言葉がある。前者は文脈に依存して意味が大きく変化するが、後者はほとんど変化しない。“会社”は前者であり、“メーカ”と“銀行”、“デパート”は後者である。前者を**曖昧キーワード**、後者を**明確キーワード**と呼ぶ。また、あるキーワードから他のキーワードを連想する頻度を**連想価**と呼ぶ。アルゴリズムは次の二つの辞書を使用する。

- 意味辞書
  曖昧キーワードとその意味の候補である明確キーワードの対応を記憶する辞書
- 連想辞書
  明確キーワードの間の連想価を記憶する辞書

ここで、曖昧キーワードとその意味の候補である明確キーワードは意味的距離が近い。また、共起関係の強さ[3, 4]が高いほど連想価が高い。

まず、検索式の中の曖昧キーワードの意味の候補である明確キーワード（候補キーワードと呼ぶ）を意味辞書を参照して求める。次に、候補キーワードごとに、連想辞書を参照して、検索式の中の他の明確キーワードとの連想価を求める。検索式の中の他の明確キーワードとの連想価が高い候補キーワードほど、曖昧キーワードの意味として高い得点を与える。

3.1.2 情報フィルター機能

データベース中の文書のキーワードは、その文書の内容を象徴している。検索式

“コンピュータ　or　ワープロ”

からサーチャーが、内容が文書処理である文書1だけを取り出すメカニズムは、以下のように説明できる。検索者が意図している文書は、“コンピュータ”と“ワープロ”の両方に関連する何らかの情報である。この両方に関連する代表的な情報は文書処理である。従って、文書1だけを取り出す。

このメカニズムを一般化して、情報フィルター機能のアルゴリズムを考案した。互いに関連する情報を象徴するキーワードの間の連想価は高い。従って、検索式が定める集合が含む文書ごとに、連想辞書を参照して、検索式のキーワードとの連想価が高いキーワードが付いている文書ほど、高い優先度を与える。

3.1.3 検索式修正機能

検索式には、検索者が重視しているキーワードと軽視しているキーワードがある。検索者が検索式を修正する時には、軽視しているキーワードを他のキーワードに置換するか除去すると考えられる。検索式

“アメリカ　and　コンピュータ　and　会社”

において、検索者が重視しているキーワードは“コンピュータ”と“会社”であり、軽視しているキーワードは、“アメリカ”である。ここで、軽視しているキーワードと置換するキーワードは、次の基準で選んでいると考えられる。まず、(1)意味的距離が近いことである。更に、修正後の検索式は何らかの情報を意図するので、修正後の検索式のキーワードが互いに関連が高いことである。つまり、(2)修正後の検索式のキーワードは互いに連想価が高いことである。“アメリカ”と意味的距離が近いキーワードには、“日本”,“ヨーロッパ”,“アフリカ”,“ブラジル”がある。これらのうち、“コンピュータ”との連想価が高いのは、“日本”,“ヨーロッパ”である。従って、“アメリカ”を“日本”に置換するか除去するなどして、検索式を修正する。

このメカニズムを一般化して、検索式修正機能のアルゴリズムを考案した。検索者がキーワードを重視している度合（優先度）を計算するために、「検索式において他のキーワードとの連想価が高いキーワードほど優先度が高い」と仮定する。まず、連想辞書を参照して、優先度を計算する。優先度が低いキーワードから順に以下の処理を行なう。意味辞書を参照して、意味的距離が近いキーワードを探す。連想辞書を参照して、意味的距離が近いキーワードのうち、検索式の他のキーワードと連想価が高いキーワードを探す。優先度が低いキーワードを、以上の処理で求まるキーワードで置換する。あるいは、検索式から除去する。

3.2 運用者サイドの適用技術

3.2.1 辞書の作成

人手での辞書の作成はコストが莫大であるため、我々はコーパス中の共起データをもとに、単語間の共起関係の強さと意味的距離を自動獲得する方式を提案している[3, 4]。我々のシステムは、この方式で最新のコーパスを使うことで、システムが新しい言葉を含む意味辞書と連想辞書を作成、更新する。

4 意味検索システム

本システムはキーワードを意味レベルで操作するので、我々は本システムを**意味検索システム**と呼んでいる。意味検索システムの構成を図4に示す。検索者は、近年普及している Netscape

Navigator[1]などの WWW Browser を使用して、WWW の検索をする。意味検索エンジンは、上述の意味辞書と連想辞書の他に、文書ごとのキーワードを記憶するインデックス・ファイルを使用する。辞書作成サブシステムは、辞書作成機能を備え、Robot[5] が収集するコーパスを用いて、意味辞書と連想辞書を自動作成する。キーワード抽出サブシステムは、キーワード抽出機能を備えコーパスを用いて、インデックス・ファイルを自動作成する。WWW server[6]は、外部コマンド実行機能(Common Gateway Interface:CGI)により意味検索エンジンを実行する。

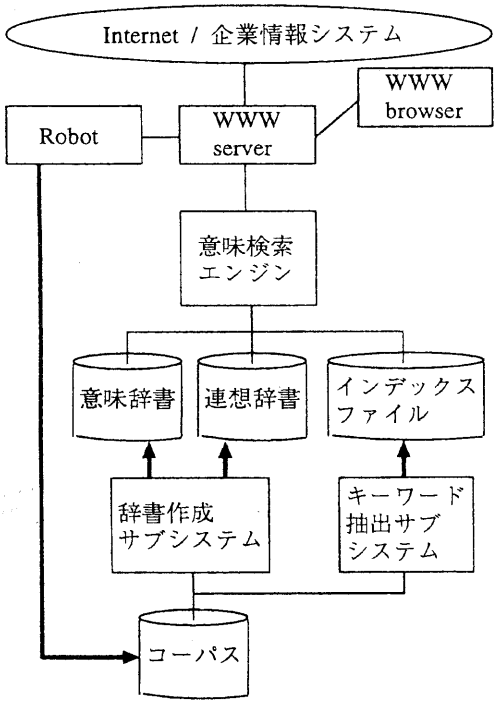


図 4: システム構成

意味検索エンジンの動作を図5に示す。動作のモードには、通常検索モードと修正検索モードがある。通常検索では、まず、検索者が入力した検索式中のキーワードに対し意味解析機能の処理を行なう。明確キーワードのみからなる検索式でインデックスファイルを検索する。その検索結果に対し情報フィルタ機能の処理を行ない、文書集合を出力する。修正検索モードは、通常検索モードとは、インデックスファイルの検索の直前に検索式修正機能の処理を行なう点が異なる。

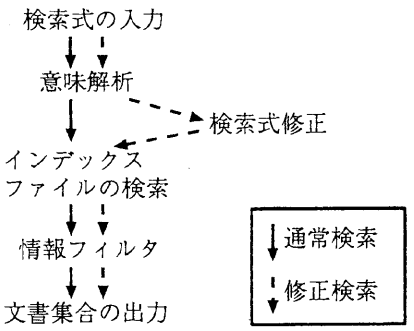


図 5: 意味検索エンジンの動作

意味解析機能、情報フィルタ機能、及び検索式修正機能の各処理では同一の連想辞書を用いる。また、意味解析機能と検索式修正機能の各処理では同一の意味辞書を用いる。従って、各処理がそれぞれ異なる辞書を用いる場合に比べて、システム全体で記憶容量を効率的に利用でき、辞書作成のコストも少ない。

現在までに、キーワード抽出サブシステムと Robot を除く部分を Windows NT3.51 [2]パソコン上に構築した。意味辞書と連想辞書は、コンピュータ関連のジャーナル(約15万文)のコーパスを用いて作成した。通常検索の処理の例を図6(a)に、修正検索の処理の例を図6(b)に示す。両方とも検索式は、

"アメリカ and コンピュータ and 会社"

である。

5 おわりに

本稿では、我々が開発している WWW の意味検索システムについて述べた。本システムは(a)意味解析機能、(b)情報フィルタ機能、(c)検索

[1]Netscape Navigator は Netscape Communications Corporation の登録商標です。

[2]Windows は、米国マイクロソフト社の登録商標です。

式修正機能、(d) 辞書作成機能を備えることを特徴とする。これらの機能を備えることにより、検索者が手軽に検索できることと、システム運用者が辞書を低いコストで作成し更新できることが可能になる。今後の課題は、キーワード抽出サブシステムの開発と、Robot が収集するコーパスからの辞書の作成である。

参考文献

[1] 細野公男他：情報検索論　認知的アプローチへの展望, 丸善 (1994)

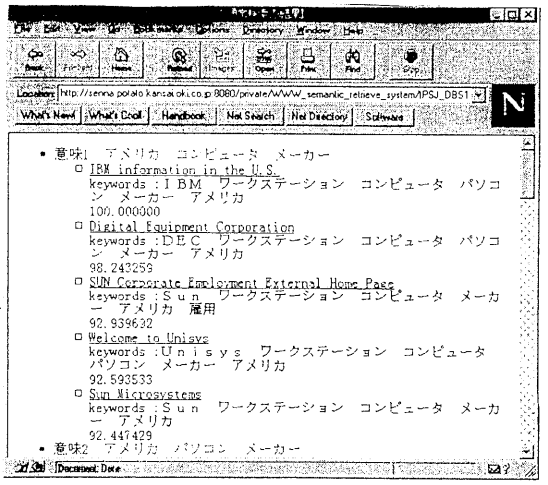
[2] Brian Pinkerton : Finding What People Want: Experiences with the WebCrawler , The Second International WWW Conference Fall '94:MOSAIC AND THE WEB

[3] 松平正樹, 山本由起雄, 坂本仁：共起データを用いた単語の意味ネットワークの作成, 情報処理学会第 42 回（平成 3 年前期）全国大会 7E-7

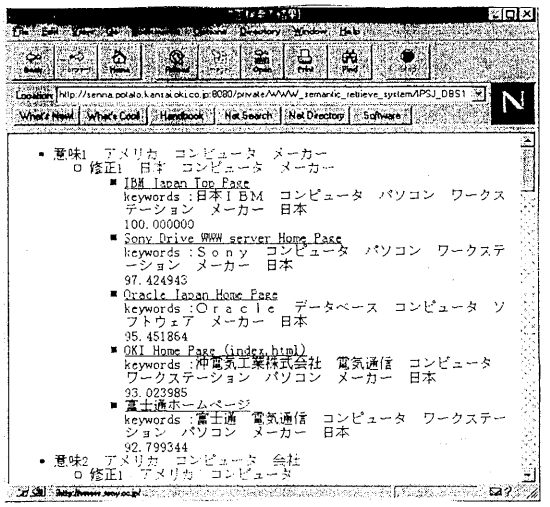
[4] 池野篤司：分野依存を考慮した単語間類似度の獲得と利用, 人工知能学会 AI シンポジウム'94

[5] List of Robots : http://*info.webcrawler.com/mak/projects/robots*/active.html

[6] W3C httpd : http://*www.w3.org/hypertext/WWW/Daemon/Status.html*



(a) 通常検索の例



(b) 修正検索の例

図 6: 意味検索システムの使用画面

[3]画面写真のビューアは Netscape Navigator を使用しています。
各社名、各製品名は各社の商号、商標または登録商標です。