◇学生卒業論文

自然言語処理によるWWW検索システムの研究

—インターフェイスの観点から—

Research for a Searching System on WWW Utilizing Natural Language Processing

—From the Viewpoint of Interface—

松本章代
Akiyo Matsumoto

現在、インターネットの普及からインターネット人口の爆発的な増加が起こっている。しかも、その中には、これまでコンピュータの経験がほとんどない、という人も決して少なくない。インターネットは情報の宝庫であるが、利用の仕方がわからなければ宝の持ち腐れでしかない。今後、益々増える新たなインターネット利用者のためには、誰もが情報を利用できるようなシステムが必要である。

そこで、まず従来の情報検索との対比から、使いやすいインターフェイスを持つ検索システムを構成するための、基本的アプローチを提案する。次に、そのアプローチに基づき制作したシステムの機能と構成を示す。最後に、ユーザーインターフェイスについて、特に自然言語処理を含む知識処理の観点から考察する。

目次

1 はじめに

私達日本人の生活の中に、抵抗なくコンピュータを取り込むためには、何が必要であろうか。新しいものを定着・普及させるには、それがどんなに便利なものでも、操作が短期間で習得できることが必要条件である。このことから、ユーザーインターフェイスがその鍵を握っていると考えられる。

近年、インターネットが随分ブームになっているが、今後もコンピュータネットワークは益々広がっていくだろう。それに従い、個人が得られる情報量は増加するが、膨大な情報から必要な情報を得ることは難しくなる。つまり、特殊な技能を持った人間だけではなく、誰もが情報を利用できるようなシステムが必要となると考えられる。

そこで本論文では、自然言語処理の技術を利用し、誰もが日本語のフレーズ（文）で情報を検索できるような検索システムを作成して、インターネットインターフェイスについて考えていこうと思う。

また、現在インターネット上で主流となっている検索サービスは、分野を特定しないすべての情報にリンクしているものだが、本論文では検索サービスの専門化・統合化も「使い易さ」「便利さ」に通じるところがあると考え、同時に提案する。

そこで、実際にシステムを構築するにあたり、現在既にインターネット上で提供されている就職情報サービスを対象に、統合された就職情報検索サービスを作成した。

2 日本語検索システムへのアプローチ

2.1 従来の日本語検索における注意点

日本語で検索する場合、一般に注意点と言われていることがらがいくつかある。

まず、日本語は英語などと違い、単語ごとにスペースで区切らない。そのため、どこまでが一つの単語なのかコンピュータが見分けることは難しい。つまり、「はし（橋）」で検索すると「はしる（走る）」までヒットしてしまう仕様のものが、日本語検索サービスにはある。

次に、英単語をカタカナ表示する場合には文書ごとの違いがある。例えば「Server」を「サーバ」と表記しているか「サーバー」と表記しているかで、キーワードの一致方法によっては別々の語句として扱われる。

また、半角と全角の問題もある。文書中に英単語や数字が含まれている場合に、どちらをどのように検索するかも問題となる。

このように、文章ではなく単語による検索でさえ、現状の検索システムでは検索者側が注意しなければならない点が多く、日本語検索システムの問題点になっているといえる[1]。

2.2 制作する日本語検索システムの方向性

2.2.1 日本語でも誰もが簡単に検索

「日本語で誰もが簡単に検索」を目指し、先に述べたような問題点は、検索側の注意点とするのではなく、やはりシステム側で解決する必要がある。

具体的な方法を以下にまとめる。

- 名詞の判別と検索式の自動生成

まず、日本語の単語を判別する方法であるが、JUMAN[1]という形態素解析ソフトをインデックス作成時と検索式生成時に利用することにより、実現した。JUMANは日本語の文章を単語に分解し、品詞を判別することができる。

さらに、現在の検索サービスはANDや

ORなどの論理演算子を日本語中に埋め込んで、複数のキーワードを組み合わせる方法が主流であるが、よりわかりやすい検索サービスとするため、日本語（自然言語）だけによる検索を提案したい。

・完全一致／部分一致の選択

検索のマッチング方法を、完全一致か部分一致かを選べるようにする。例えば、完全一致の場合、「ソフト」と「ソフトウェア」という単語は別の語と判断されるので「ソフト」と入力しても「ソフトウェア」を検索することはできない。「コンピュータ」と「コンピューター」なども同様である。それに対して部分一致の場合は、「ソフト」が「ソフトウェア」という単語の一部に相当するのでマッチしていると解釈するのである。これによって、検索者の意図に、よりそった検索ができるものと考えられる。

・カッコの認可

日本語の文章から、複数の論理演算子を使った検索式が生成される場合がある。それらを前方優先で解釈すると、不都合が生じる場合がある。

例えば「コンピュータのソフトまたはソフトウェア」と入力した場合、前方優先では「（コンピュータ and ソフト）or ソフトウェア」という解釈をしてしまうことになるのである。そこで、特にまとまりを指定したい箇所に半角カッコ'('')'を使うことで、よりユーザーの意図にそった検索式が作成されるようにする。つまり、この場合は「コンピュータの（ソフトまたはソフトウェア）」と入力すれば、正しく解釈されるようにした。

カッコは厳密には自然言語とは言えないが、この場合、カッコが果たす役割が大きいことと、まとまりを表すものとしてカッコは自然な概念に近いことから、カッコを認めることにした。

・半角文字と全角文字の同一視

データの情報源であるファイル中には、半角文字と全角文字が混在している。そのファイルからインデックスファイルを作成した場合、英数字等の半角・全角は別の文字として認識されるため検索者の期待通りの検索ができるとは言いがたい。

そこで、半角カナについては、EUCで扱えないという問題点があるため全角に変換し、英数字に関しては半角に統一してインデックスを作成する。また、当然入力フォームから入力された文字例にも同様の処理を行う。

・処理の段階化

検索精度と速度の向上のため、ユーザの要求に対して結果が出力されるまでの間に、一度確認のプロセスを含むようにする。そうして段階化することによって処理が2つに分かれることになり、待ち時間も二分され、ユーザが感じる「待たされる時間」は減ると考えている。また、検索結果に不満を持ち検索し直す場合、入力フォームまで一気に戻るのではなく、AND／OR／NOTの変更・完全一致／部分一致の変更だけなら中間の確認の画面に戻って検索し直すことも可能になる。逆に、生成された検索式が、ユーザにとって不本意だった場合も、長い時間待たされて一気に検索結果が表示されてしまうのでなく、まず検索式の確認ができる方が効率的である。

2.2.2 検索サービスの専門化・統合化

インターネットで提供されている情報を、効果的かつ有効に探し出すことを検索サービスの目的とするのであれば、キーワードによる検索だけでなく、さまざまな検索手段が考えられる。検索サービスはデータ量とその検索スピードを競っていた段階から、次の段階へと変わりつつあるように思える[1]。

本論文ではここで、検索サービスの専門化・統合化を提案する。各分野ごとに統合された検索サービスが一つずつあり、そこへ行けばそのものに関する情報が確実に手に入る。それがベストだと考える。

そこで、実際にシステムを構築するにあた

り、現在既にインターネット上で提供されている就職情報サービスを対象に、「統合された就職情報検索サービス」を作成することにする。現在、リクルート社など、就職情報誌を発行している各社の提供する就職情報がインターネット上に点在している状況なので、それらをまとめて検索できるサービスにしたいと考える。つまり今回は、情報源を得るため WWW 空間にロボットを走らせ就職に関する情報を集めるというのではなく、既存のデータベースを統合するという形をとった。

ちなみに、現在、就職情報データベースにおいてフリーワード検索ができるのは、私が調査した限りリクルート社が提供するデータベースのみである。

3　システムの仕様

◎作成目的

「自然言語処理による WWW 検索システムの研究」

◎利用目的

「統合された就職情報データベースの利用」

◎機能

「インターネットの Web ブラウザ上から日本語による検索」

ユーザが Web ブラウザ上から就職情報について検索したい文字例を自由に入力すると、自動的に各就職情報データベースにアクセクして検索が行われ、検索結果は URL として一覧表示される。

◎開発環境

- ハードウェア
  - ・SPARC Station 20
- OS
  - ・SUN OS 4.1.4

◎使用プログラミング言語

- C＋＋
- Perl

◎使用ツール

- newjuman 1.0（日本語形態素解析）
  JUMAN は日本語形態素解析のためのツールである。
  計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発されたシステムである。
- Web Whacker（Web 自動巡回）
  インデックスを作成するためにデータを取得する必要がある。
  大量のデータを Web ブラウザから取得するためには、自動巡回ソフトを使うことが不可欠である。

◎データ提供元（使用データベース）

- リクルート社「リクルートブック on the NET」
- 毎日コミュニケーションズ社「Career Space」
- 静岡新聞社「しごとのかんづめ」
- 他
  「統合化」を目指す以上、就職情報に関するすべてのデータベースを網羅したかったのだが、個人が、しかも卒論という色々な制約の中で構築するシステムとしては、３つ程度が限界であり、今回は断念した。

◎検索の機能

- 検索対象
  - ・全文
- キーワード一致方法
  - ・完全一致
  - ・部分一致
- 論理演算子
  - ・A and B　A かつ B
  - ・A or B　A または B
  - ・A not B　A だが B ではない
- その他の機能
  - ・カッコによるまとまりの指定
  - ・半角文字と全角文字の同一視

◎インデックスファイルの生成法

1.各データベースのデータ（HTML ファイル）を取得する。
（→自動巡回ツールを利用）
2.取得したファイルを連結する。（インデックスファイルは各データベース一つずつ）
3.半角文字を全角に変換するなど、HTML ファイルを perl スクリプトで整形する。
4.JUMAN により形態素処理する。
5.その出力からインデックスファイルを作る。
（→自作ツールを利用）

◎検索処理過程

1.情報処理のための入力をブラウザ上から受け付ける。（→日本語のデコード）
2.JUMAN によりこれを形態素処理する。
3.その結果から検索式を生成する。
4.その検索式から、インデックスファイルを検索する。
5.ブラウザ上に検索結果を一覧表示する。

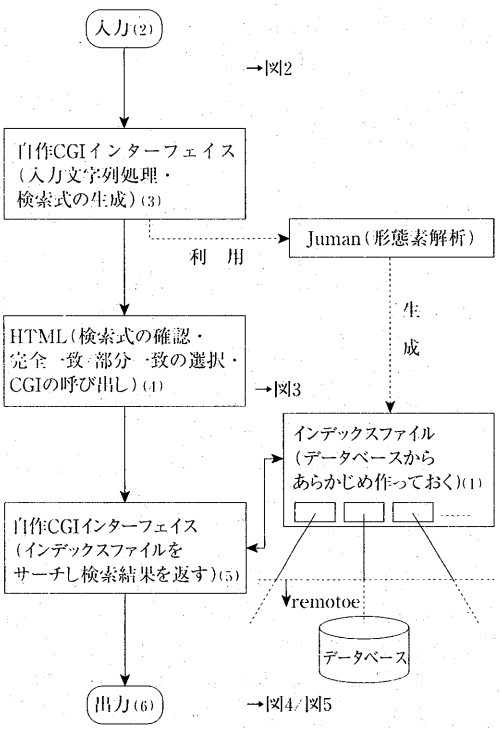

図１：システムの構成図

4　システムの構成

図１にシステム構成を示す。
図２～５に実行の様子を示す。
以下に図１の解説を示す。

(1)インデックスは対象データベースのデータファイルを元に、私の自作ツールと JUMAN を使い、あらかじめ作っておく。
(2)HTML から入力文字列を受け付ける。→図２
(3)CGI が入力文字列処理・検索式の生成をして、HTML として出力する。
(4)出力された HTML は、検索式の確認・完全一致／部分一致の選択・CGI の呼び出しを行う。→図３
(5)CGI によりインデックスファイルをサーチし検索結果を返す。
(6)ブラウザ上にリンク形式で出力する。→図４・図５

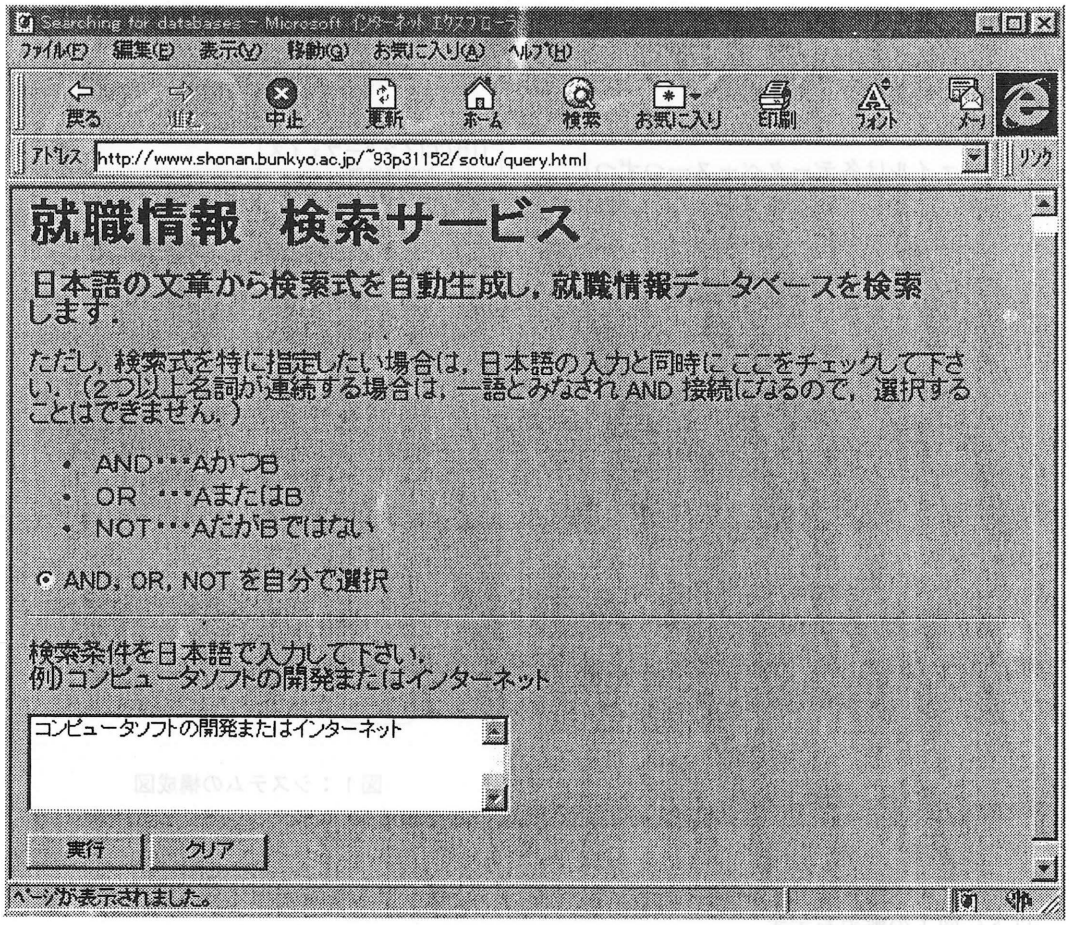

図 2: 検索サービス。検索条件の入力までしたところ。

= HTML =

```
<title>Searching for databases</title>
  (…中略…)
<input type="radio" name="USER" value="y">AND, OR, NOT を自分で選択
<hr>
検索条件を日本語で入力して下さい． <br>
例）コンピュータソフトの開発またはインターネット <p>
<textarea type=text name="INPUT" rows=5 cols=40 size=40,5>
</textarea>
<p>
<input type="submit" value="実行 ">
<input type="reset"  value="クリア ">
</form>
```

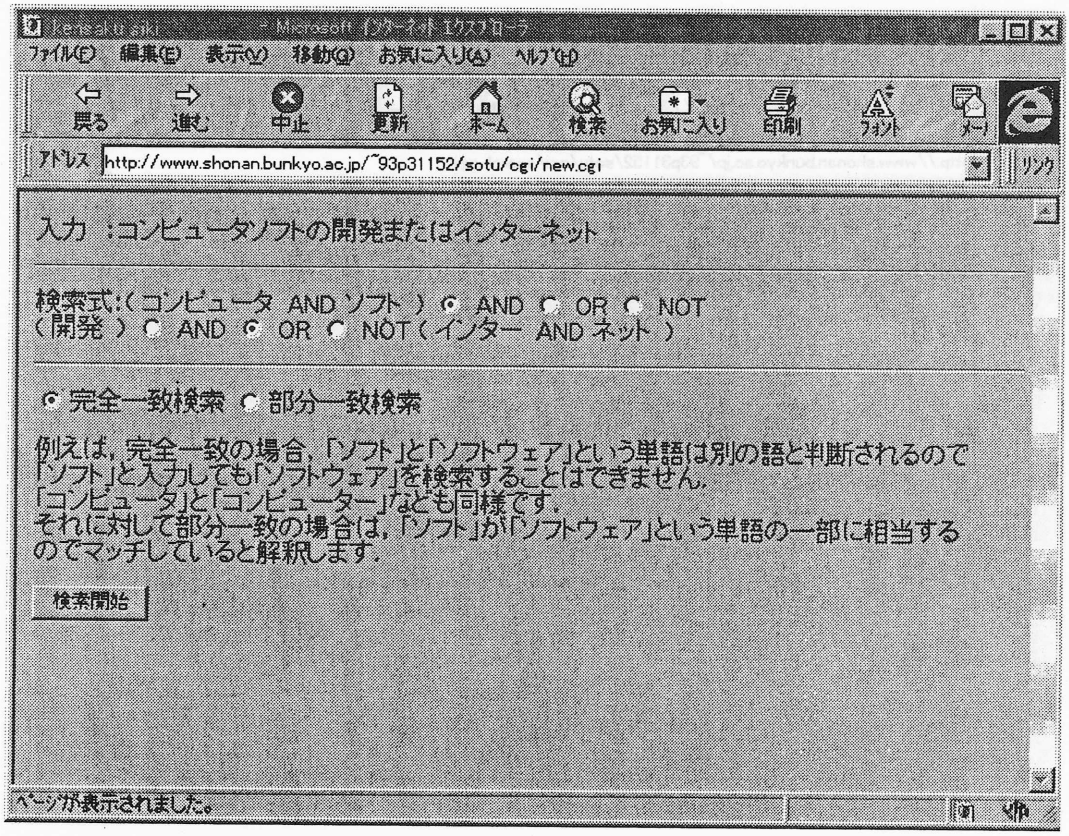

図 3: 図 2 で実行ボタンを押した結果。AND, OR, NOT を選択できる。

= HTML =

```
<title>kensaku siki</title>
<form method="POST" action="search.cgi">
入力 :コンピュータソフトの開発またはインターネット
<hr>
検索式 : <input name="ao0" type=hidden value="( コンピュータ ">( コンピュータ
<input name="ao1" type=hidden value=" AND ソフト "> AND ソフト
<input name="ao2" type=hidden value=" )"> )
<input name="ao2" type=radio value=" AND " CHECKED> AND
<input name="ao2" type=radio value=" OR "> OR
<input name="ao2" type=radio value=" NOT "> NOT
  (…中略…)
<input type=radio name="KAN" value="y" checked> 完全一致検索
<input type=radio name="KAN" value="n"> 部分一致検索 <p>
  (…中略…)
<input type="submit" value="検索開始 ">
</form>
```

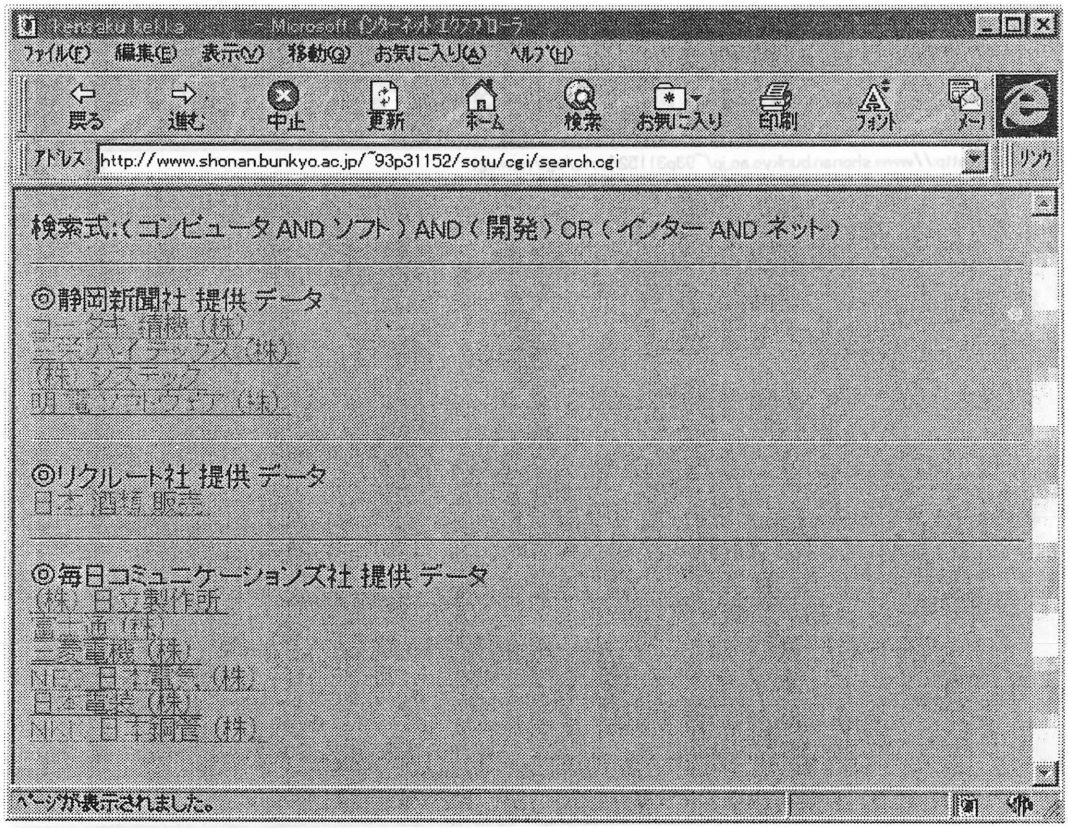

図 4: 図 3 で完全一致検索をチェックし，検索開始ボタンを押した結果。

= HTML =

```
<title>kensaku kekka</title>
検索式：( コンピュータ AND ソフト ) AND ( 開発 ) OR ( インター AND ネット )<br>
<hr> ◎静岡新聞社 提供 データ <br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/52.html> コー タキ 精機 (株) </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/68.html> 三栄 ハイ テックス (株) </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/70.html> (株) システック </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/350.html> 明 電 ソフトウェア (株) </A><br>

(…以下省略)
```

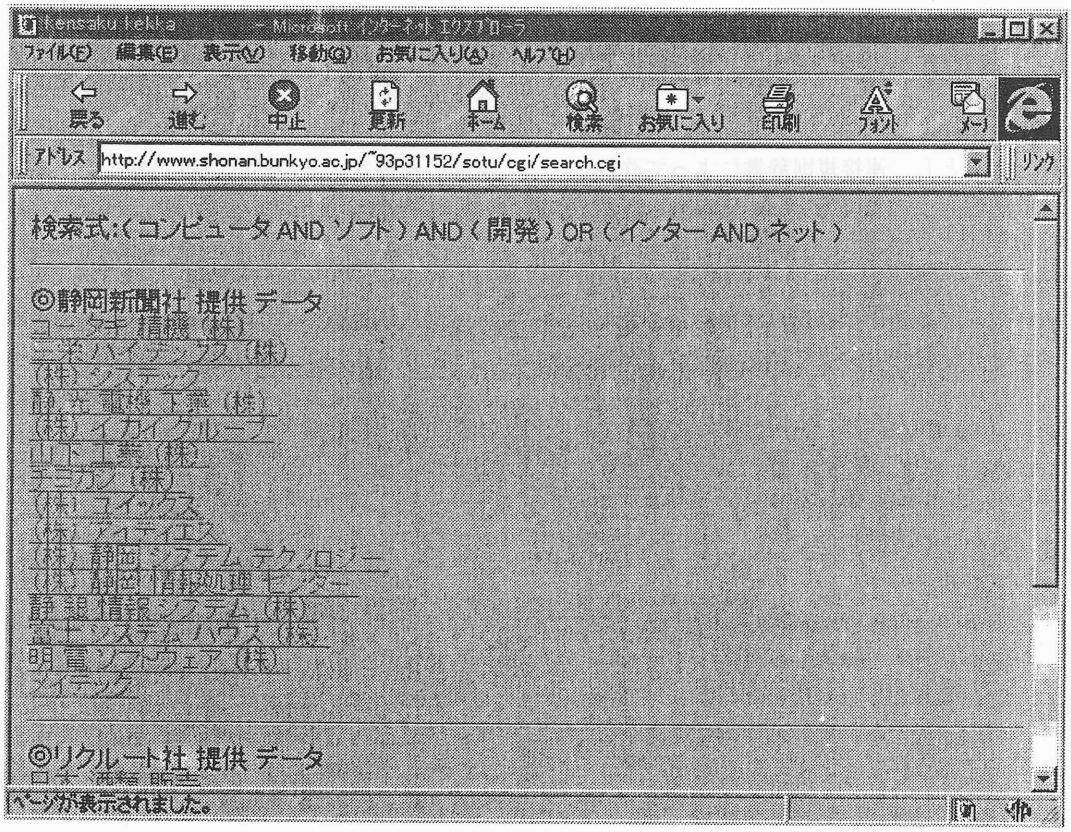

図 5: 図 3 で部分一致検索をチェックし，検索開始ボタンを押した結果。

= HTML =

```
<title>kensaku kekka</title>
検索式：( コンピュータ AND ソフト ) AND ( 開発 ) OR ( インター AND ネット )<br>
<hr> ◎静岡新聞社 提供 データ <br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/52.html> コー タキ 精機 (株) </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/68.html> 三栄 ハイ テックス (株) </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/70.html> (株) システック </A><br>
<A HREF=http://www.chabashira.co.jp/job-shizuoka/index/data/72.html> 静 光 電機 工業 (株) </A><br>

(…以下省略)
```

5 JUMANによる形態素解析の仕組み

JUMAN[3]は、EUCコードの日本語文字列を入力とし、連接規則辞書によって許容された形態素からなる束（lattice）状の構造を出力とする。ただし、結果の表示については、後方最長一致の解を一つだけ表示する。（曖昧性のあるものすべてを表示させることもオプションを付けることにより可能）

5.1 形態素解析アルゴリズム

JUMANの解析アルゴリズムは、入力としてあたえられた日本語の文字列に対する次の基本動作よりなる。改行をもって一つの入力文字列の終了とする。

- ある特定の位置からはじまるすべての可能な形態素を辞書引きによって得る。
- 辞書引きによって得られた個々の形態素に対して、その直前の位置に存在するすべての形態素との連接可能性のチェック、および、コストの計算を行なう。
  連接可能性のチエックによって連接不可能とわかった形態素間の接続は行なわれない。また、その位置で連接可能なもののうち最良（最小）のコストと比較して.jumanrcのコスト幅によって定義される数値以上のコスト差をもつ形態素の接続は行なわれない。

本形態素解析では次の二種類のコストが使用される。

形態素のコスト：個々の形態素に与えられているコストである。形態素のコストは、.jumanrcで定義された「品詞コスト」、「形態素コスト重み」、および、形態素辞書で定義された「見出し語のコストに対する重み」、の三つの数値の積として計算される。

連接コスト：二つの形態素の連接のコストである。連接規則辞書で定義されたコストと.jumanrcで定義された「連接コスト重み」の積として計算される。

形態素解析の一つの解析結果は形態素からなる列であり、その総コストは、それに含まれる形態素のコストの総計、および、各形態素間の連接コストの総計の和である。ただし、列の先頭と末尾には、それぞれ、「文頭」および「文末」と呼ばれるダミーの形態素があると仮定する。

一般に解析結果は、文頭および文末を両端点とする束状のグラフ構造になる。このグラフ内の文頭から文末へいたるそれぞれの経路が一つの可能な解析結果を表わすことになる。

連接可能性のチェックおよびコストの計算の詳細について説明する。最初の辞書引きは文字列の先頭で行なわれ、その直前には「文頭」が存在すると仮定される。また、「文頭」がもつコストは0である。

最初に辞書引きが行なわれるのは入力文字列の先頭（すなわち、「文頭」の直後）である。以降、辞書引きが行なわれるのは入力文字列中において文頭からの有効な連接をもつ形態素の直後の位置である。

形態素解析の最後の段階では、入力文字列の末尾に現れる有効な形態素と「文末」との連接可能性のチェックとコスト計算が行なわれ、解析結果としてのグラフ構造が得られる。

5.2 未定義語の取り扱い

未定義語に関しては、本システムでは、入力文字列中のあらゆる位置から未定義語が存在する可能性を考慮している。ひらがなおよび漢字以外の文字については、同種の文字の終りまで（カタカナ、アルファベット、数字等）をまとまった語として一つの未定義語と解釈する。ひらがなと漢字については、一文字ずつを未定義語の候補と考えている。

切り出された未定義語に対する連接関係を定義するためには、未定義語をどのような品詞と考えるかは、オプション定義ファイル（.jumanrc）内の「未定義語品詞」によって

自由に指定できる。ただし、未定義語品詞は、文法辞書で定義された品詞名（品詞再分類がある場合には品詞再分類名も指定する必要がある）のいずれか一つでなければならない。

未定義語に対するコストは同じくjumanrcの「品詞コスト」欄で定義できるので、これに高いコストを与えることにより、未定義語を含む解析結果の優先度を下げることができる。

5.3 オプションファイル（.jumanrc）の記述例

```
(文法ファイル
        "93p31152/sotu/newjuman/dic")
(辞書ファイル
        "93p31152/sotu/newjuman/dic/
        JUMANTREE")
(未定義語品詞（名詞サ変名詞）
(品詞コスト
    ((*)              10)
    ((未定義語)     5000)
    ((特殊*)         100)
    ((動詞)          100)
    ((形容詞)        100)
    ((判定詞)         10)
    ((助動詞)         10)
    ((名詞*)         100)
    ((指示詞*)       100)
    ((副詞*)         100)
    ((助詞*)          10)
    ((接続詞)        100)
    ((連体詞)        100)
    ((感動詞)        100)
    ((接頭辞*)        10)
    ((接尾辞*)        10)
(連接コスト重み100)
(形態素コスト重み１)
(コスト幅200)
    (HASHTABLE
    (HASHSIZE 16384)
    (HASHLIMIT 32768)
    (HASHLIMIT-S 32768)
    (HASHFULL .85)
)
```

5.4 JUMAN出力例

・元テキストファイル

```
%  cat  test.  txt
静岡県にある出版社
```

・オプションなし（整形して出力）

```
%  juman  <  test.  txt
静岡県  (しずおかけん)    静岡県  固有名詞
に      (に)              に      格助詞
ある    (ある)            ある    連体詞
出版    (しゅっぱん)      出版    サ変名詞
社      (しゃ)            社      普通名詞
```

・−cオプション（形態素情報をコードで出力）→このシステムではこれを使用

```
%  juman  −c  <  test.  txt
静岡県  しずおかけん  静岡県  ６３００
に  に  に  ９１００
ある  ある  ある  １１０００
出版  しゅっぱん  出版  ６２００
社  しゃ  社  ６１００
```

・−eオプション（完全な形態素情報を出力）

```
%  juman  −e  <  test.  txt
静岡県  しずおかけん  静岡県  名詞  6
固有名詞  3＊0＊0
に  に  に  助詞  9  格助詞１＊０＊０
ある  ある  ある  連体詞  11＊０＊０＊０
出版  しゅっぱん  出版  名詞  6  サ変名
詞２＊０＊０
社  しゃ  社  名詞  6  普通名詞1＊0＊0
```

・−mオプション（曖昧性のある部分だけその候補すべてを出力）

```
%  juman  −m  <  test.  txt
静岡県（しずおかけん）  静岡県  固有名詞
```

に	(に)	に	格助詞
に	(に)	に	名詞接続助詞
ある	(ある)	ある	連体詞
出版	(しゅっぱん)	出版	サ変名詞
社	(しゃ)	社	普通名詞
社	(やしろ)	社	普通名詞
社	(やしろ)	社	固有名詞

6 検索式の生成

6.1 JUMAN の出力からの検索式の生成過程

1. クエリー（入力文字列）を JUMAN によって単語列に分割する。
2. 名詞、未定義語[2]以外の品詞を持つ語を、ストップワードとして除去する。
3. ストップワード（群）の存在した場所を境界としてセグメントと呼ぶ単位へ分割する。
4. 各セグメントについて構成要素を AND 接続する。
5. さらに、以下の規則性に基づいて、AND/OR/NOT 接続し、検索式とする。
6. また、ユーザも希望によって AND/OR/NOT を選ぶことができ、それに応じても検索式が生成できる。[3]

・A なB	A and B
・AB	A and B
・A でB	A and B
・A のB	A and B
・A によるB	A and B
・A にあるB	A and B
・A かB	A or B
・A またはB	A or B
・A だがB ではない	A not B

etc...

例）ソフトを作りインターネットを扱う

1.→ソフト、を(ストップワード)、作り(ストップワード)、インター、ネット、を(ストップワード)、扱う(ストップワード)

2.→ソフト、インター、ネット

3. (ソフト)(セグメント)、(インター、ネット)(セグメント)

4.→(ソフト)、((インター) AND (ネット))

5.→(ソフト)AND((インター)AND(ネット))

6.2 検索式の生成例

例1：ソフトを作りインターネットを扱う

→(ソフト)AND((インター)AND(ネット))

例2：ソフトを作るまたはインターネットを扱う

→(ソフト)OR((インター)AND(ネット))

例3：ソフトを作るがインターネットは扱わない

→(ソフト)NOT((インター)AND(ネット))

例4：パソコン(ソフトまたはソフトウェア)

→(パソコン)AND((ソフト)OR(ソフトウェア))

7 考察

7.1 現在の検索サービスの問題点

ここでもう一度、改めて従来のシステムの問題点を整理しておく。

検索結果量が膨大…有名な検索サービスはリソースが大きく、検索時間がかかるだけでなく、場合によっては何百という検索結果が返される。この中には、意図しないものも多数含まれていることがめずらしくない。その中から必要な情報を探すことは大変である。

その一方、分野が特定された検索サービスは、規模が小さいものがあちこち点在している。

→専門化＆分野ごとの統合の必要性

日本語処理が不十分…日本語特有の処理の難しさが解決されずに、検索者側の注意点として残っている。

複数キーワード検索の不統一性…キーワードと論理演算子の組み合わせによる、複数キーワード検索は、論理演算子になじみのない人にわかりにくい。さらに使える論理演算子が検索サービスごとに異なる。

→コンピュータの知識のない人にもわかりやすいシステムを

情報処理を多少なりとも勉強した人や、プログラミング経験者にとっては当たり前の論理演算子。しかし、インターネット利用者はコンピュータの知識を持つ、という前提はもはや通用しない。英語さえ知らない小学生もインターネットを使う現在。インターネットを利用する上で最も基本となる情報検索が、難しかったり使いにくかったのではまずいであろう。

7.2 構築システムにおける自然言語処理

JUMANが単語や品詞の判別を誤った場合を除けば、具体的な数字では表せないが、ほぼ、検索者の意図にそった検索式が生成されると思われる。

自然言語処理による検索の実現には、とりあえず成功したと言える。

7.3 使い易いシステムとは

誰にとっても使い易いインターフェイスを考える場合、「簡単」で「便利」であることが要求される。つまり、簡単さと共に機能の充実も求められる。

2つのうち、どちらか一方だけを追求するのなら、簡単である。しかし、それでは使い易いシステムとは決してならない。その兼ね合いの難しさを痛感した。誰でも短期間で操作を修得でき、その後初心者でなくなったときも、ずっと使い続けることができるようなシステムでなければならない。

自然言語のみを入力とするのではなく、「AND」や「OR」を直接フォームに入力することも認めた方が、既に他の検索サービスに慣れた人にとっては便利かも知れない。そう考えたこともあった。しかし、それではシステムの機能が複雑になり、かえってわかりにくくなってしまうのではないだろうか。また自然言語で検索、というコンセプトから外れてしまう。そう考えて思いとどまった。

実際にシステムを構築することで、使い易いシステムとは操作の簡単さと機能の充実のバランスをとることだということを、改めて感じた。

7.4 今後の課題

「使い易さ」を機能面に中心に追求したわけだが、ユーザーインターフェイスにはGUI（グラフィカルユーザーインターフェイス）という概念がある。本研究ではGUIには一切触れてこなかった。

これに関しては、今後の課題としたい。

8 おわりに

知りたい事柄を入力して、インターネットという情報の海の中から探し出す。これを実現するには、日本語処理についての技術が不可欠である。従来から「かな漢字変換」や「校正機能」、また「シソーラス」といわれる同義語・類義語辞書の搭載などが求められているが、画期的な解決はされていない。

日本語の場合、文章中の単語の識別が難しい。また、表記も複雑である。漢字表記、ひらがな表記、カタカタ表記は混在しているし、拗促音の表記についても統一されていない。音引きについても「インターフェイス」「インタフェイス」「インターフェース」などバラバラである。こうした言語特性を理解して、本当に必要としている情報を探し出すということをコンピュータがもっと支援すべきである[6]。

それが、使い易く便利な日本語検索システムへの第一歩だと考える。

参考文献

[1]石川和也：WWW検索サービスを使いこなそう、インターネットマガジン96年9月号p212-219、インプレス

[2]清水奨：WWWサーバ上の検索システム構築、インターネット96年7月号

p130-139、CQ 出版社
[3]日本語形態素解析システム JUMAN 使用説明書 version 1.0
[4]菊井、鷲崎、林、砂場：インターネット情報ナビゲーションにおける多言語機能、情報処理学会“自然言語処理の応用に関するシンポジウム”、to appear(1995).
[5]林、菊井、鷲崎、砂場：WWW 情報空間における Resource Discovery と Navigation 支援、AI 研究会「メディアと情報処理」シンポジウム、(1995).
[6]中島由弘：情報の爆発に追いつかない日本語の情報検索技術、インターネットマガジン96年9月号 p294-295、インプレス

付録

A　著作権法について

A.1　サーチエンジンと著作権法

このシステムを作成するにあたって、一つ気になることがあった。

それは著作権の問題である。

しかし、「サーチエンジンが著作権法に触れるという話は聞いたことがないので、このシステムも大丈夫に違いない」と解釈し、著作権法についてはあまり考えてこなかった。

だが、本当にサーチエンジンは著作権法に触れないのか。そこで思いきって著作権法について調べ、サーチエンジンの、どの部分が著作権法に触れる恐れがあるかを考えた。

1つめは、Web 上に分散して存在する各データにリンクを張ることで、情報の提供を行うというところ。

2つめは、Web 上の各データを利用（コピー）し、それを元にインデックスデータベースを作るというところ。

問題はこの2点であろう。

A.2　リンクを張ることの是非

A.2.1　リンクを張ることの著作権上の意義

まず、リンクを張ることの著作権上の意義について考えてみる。

リンクを張ることとは、簡単に言えば、リンク先の URL をページ上に記述することであって、リンクの先のデータ自体をページ上に記述しているわけではない。

つまり、リンクを張ったからといって、リンク先のデータ自体をコピーしているわけではないし、リンク先にジャンプした利用者はリンク先のホームページを見ているだけなので、リンクを張ること自体が直接著作権侵害になるとは考えがたい[1]。

A.2.2　リンクを張られることによる不利益

次に、リンクを張ることがなぜ問題とされるかを考えてみる。

そもそも WWW は、リンクを張り巡らすことによって、世界中の情報を有機的に関連づけようという思想に根ざすものであって、リンクを張ること、リンクを張られることは、もともとインターネットの特性として折り込み済みであるように思われる。

しかし、「トップページからアクセスしてもらって次第に下の階層のページを見てもらいたいのに、下の階層のページにリンクを張られてしまうと、トップページをとばして、いきなり下の階層のページにアクセスされるようになって困る」というような意見も考えられる。

確かに、トップページの宣伝広告を見て欲しいからこそ、下の階層のページに魅力的なデータを置いてあるような場合には、そうした意見が出てくるのも理解できる。

しかし、そのような主観的な期待を、法的に保護すべき程度のものと言うべきかどうかは、かなり疑問がある。リンクを張るほうからすると、リンク先のホームページ開設者がそのような期待を持っているのかどうか判別するのは困難であるし、リンクが張られていなければ利用者が実際にトップページからアクセスすることが保証されているわけでもないからである[1]。

A.3 インデックス作成の是非

今度は、サーチエンジンの仕組みから、Web上のデータを利用してインデックスデータベースを作成することの是非を考える。

A.3.1 サーチエンジンのインデックス作成の仕組み

サーチエンジンのデータベースへのデータ登録方法には、手動で登録する方法と、URL取得ロボット（以下ロボットと略す）を使って自動的に登録する方法の、大きく分けて2つある。現在、大規模なサーチエンジンのほとんどが、ロボットを使い自動的に大量のURLを登録しているようである。

このロボットとは何か。

「ロボット」や「スパイダー」と呼ばれるフリーソフトウェアで、分散して存在する情報資源からURLやキーワードなどの情報を抽出し、インデックスデータベースを構築するためのツールである。

具体的には、あるHTMLが指定されると、まず、その文書に含まれているキーワードなどの情報を収集する。続いてさらにその文書に含まれているすべてのハイパーリンクを利用して別の文書を呼び出す。この文書に対してまた情報収集を行い、さらにリンクをたどるという作業を繰り返す。

このように文書同士がハイパーリンクしているというWWWの特性を利用して自動的に情報を集める仕組みになっている。

このように、情報空間を網羅的に探査し、各情報資源をインデクシングすることに対して、著作権の問題ははたして生じるのか。それが問題である。

A.3.2 インデックス作成と著作権

まず、データベースのインデックス作成において、著作権と関係がありそうな特徴を以下にまとめる。

(1) Web上のデータ（著作物）を一時的にコピーする（(2)の後、消去する）

(2) そのうちの一部(キーワード)を抜き出すなどの加工をし、インデックスを作成する

(3) それを複数の人数で利用する（私的利用とは言えない）ただし複製、公表、領布のいずれも行わない

(4) 著作者に経済的不利益は与えない

(5) 製作者に、サーチエンジンをインターネット上に公開することによる、直接的な経済的利益はない

まず、(1)で行われるコピーが問題ないかどうか、検討する。

WWW空間を巨大なデータベースとみなせば、著作権法の「データベースの著作権（著作権法12条の2）[2]」が適合される。

それによると、データベースは著作権法で保護される。しかし、すべてのデータベースが著作物として保護されるわけではなく、情報の選択的な体系や構成が独創性を有するものだけが保護される。これは、個々のデータが単なる事実の記述にすぎない場合、著作権が発生しない場合もある[1]、ということだが、Web上で普通に公開されているデータ群は、営利目的でないにしろ、著作物であると考えられる。また、ある程度まとまりをもった情報の集合体をデータベースから引き出し、再利用可能な状態で端末機に蓄積する行為（ダウンローディング）は、データベース著作物の一部複製として扱われ、データベース作成者の複製権が及ぶことになる[3]、ということである。

この場合はインデックスに加工してしまうので、「再利用可能な状態」とはならない。ということは、複製権は及ばない、と考えられる。

続いて、(2)・(3)で述べたインデックス作成とその私的利用を超えた利用について、問題がないかどうか、検討してみる。

通常、著作物を加工して利用する場合、著作者の許諾が必要になる。改作利用権により、著作物に手を加えて利用する権利は、著作者に専有されている[3]からである。

しかし、文化庁の「著作権法ハンドブック[4]によれば、「抄録の作成は著作権侵害にあたるか」という問いに対して、以下のように解を述べている。

「著作物を短くしたものとして抄録（アブストラクト）がある。抄録は著作物を短く短縮したもので、著作者の紹介、索引を目的とする。抄録は著作物を利用するものではあるが、利用した著作物にとって代わるものではなく、むしろそれを見た読者に著作物の存在を知らせ、それを読む興味や意欲を起こさせるものである。従って、こうした利用は著作権の侵害とはならないと考えられる。著作物の内容がほぼ感得できるものであるか、利用された著作物の存在を知らしめる程度のものであるか、によって判断すればよい。」（途中数カ所省略）

サーチエンジンのインデックスファイルは、その性質からも、索引という目的からも、非常に抄録に類似している。ここで述べられていることは、その判断基準からも、インデックスファイルにも当てはまると考えてよいように思われる。

A.4 結論

リンクを張ることでデータの提供を行うことも、Web上のデータからインデックスを作成することも、以上の検討と、(4)・(5)の性質から考えて、結論としては「問題なし」だと考えられるが、一応データ提供者に許可を取ることが望ましいのかもしれない。

こういった問題に関しては、今後ユーザ間の合意を形成していく必要があるだろう。

付録における参考文献

[1]宮下佳之他：ネットワーク時代の知的所有権入門、インターネットマガジン94年12月号～97年1月号、インプレス
[2]六法全書Ⅱ、平成8年度版、有斐閣
[3]半田正夫：著作権法概説［第七版］、1994、一粒社
[4]最新版著作権法ハンドブック、1991、文化庁

[1] インデックス作成にJUMANを利用するという方法は、参考文献[2]より学んだ
[2] p12参照
[3] 参考文献[5][6]からヒントをいただいた
（情報学部情報システム学科、松原康夫ゼミ卒）

〔担当教員から〕 松本章代さんは、当ゼミにおいて最も積極的かつ創造的な卒業生の一人です。この卒論は、自分のインターネットに関する興味とゼミで学んだ自然言語処理とを結び付け、広範囲の情報収集を行いながら構想を練り、ほとんど独力でシステムを構築したものです。自ら学び工夫する姿勢に富むとともに、他人の意見にも耳を傾ける謙虚さを併せ持つので、研究者ないしは教育者としての資質をもつと考えられます。彼女は社会に出るにあたって、専門学校の教員の道を選びました。当人が自分のやりたい道を選ぶように勇気付ける事ができたことは、教師としての本懐であります。

（情報学部教授・松原康夫）