社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＮＬ－158　(14)
2003／11／7

ページ内の意味的係り関係に基づく WWW ページ検索結果の絞り込みについて

松本章代† 小西達裕† 高木朗‡ 小山照夫¶ 三宅芳雄∫ 伊東幸宏†

† 静岡大学　〒432-8011 静岡県浜松市城北 3-5-1
‡ 株式会社 CSK 〒163-0227 東京都新宿区西新宿 2-6-1
¶ 国立情報学研究所　〒101-8430 東京都千代田区一ツ橋 2-1-2
∫ 中京大学　〒470-0393 愛知県豊田市貝津町床立 101
E-mail: †{gs2053,konishi,itoh}@cs.inf.shizuoka.ac.jp, ‡ Akira_1_Takagi@cii.csk.co.jp
¶ t_koyama@nii.ac.jp, ∫ ymiyake@sccs.chukyo-u.ac.jp

あらまし　本稿では，Web 検索エンジンの検索精度を更に向上させるため，Web ページ内に記述されている文章中における検索キーワード間の意味的係り受け関係を利用しようという試みについて述べる。キーワード間の意味的関係を表すような構造として，ここでは，センテンス，表，及び箇条書きに着目する。提案する方法は，まず，各々における検索キーワードの意味的関係を表すセンテンスや表の有無を調査し，関係があるとすれば，文中の位置やページ全体中の位置に対し加えて評価を行って候補ページをフィルタリングするというものである。また，本稿ではこの方法によって検索精度の向上することを実験的に検証する。尚，２つのキーワードを用いて検索する場合が多いことに着目し，本研究では当面 2 語による検索を研究対象とする。
キーワード　情報検索，サーチエンジン，自然言語処理，係り受け関係，構文解析，表解析，WWW

A Filtering Tool for WWW Search Engines based on Semantic Relation between Input Keywords

Akiyo Matsumoto† , Tatsuhiro Konishi† , Akira Takagi‡ , Teruo Koyama¶ ,
Yoshio Miyake∫ , Yukihiro Itoh†

† Shizuoka University　3-5-1 Johoku, Hamamatsu-shi, Shizuoka, 432-8011 Japan
‡ CSK Corporation　2-6-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo, 163-0227 Japan
¶ National Institute of Informatics　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
∫ Chukyo University　101 Tokodachi, Kaizu-cho, Toyota -shi, Aichi , 470-0393 Japan
E-mail: †{gs2053,konishi,itoh}@cs.inf.shizuoka.ac.jp, ‡ Akira_1_Takagi@cii.csk.co.jp
¶ t_koyama@nii.ac.jp, ∫ ymiyake@sccs.chukyo-u.ac.jp

Abstract　In this paper, we propose a filtering tool for WWW search engines. Because ordinary search engines select candidate pages based on appearance frequency of keywords, the candidate pages may contain every keyword. However, there are some cases that each keyword appears in individual context and they have no semantic relation. Such pages tend to be useless. We try to improve precision of a search engine by checking whether input keywords have a semantic relation in the candidate pages or not. Semantic relations between keywords are represented by using a structure, such as sentence, table, itemized list and so on. In this paper we deal with sentence and table as a structure to represent semantic relations. We propose a method to detect semantic relations between input keywords represented in the structure of sentence and table. We also show an experimental evaluation of our method.

Keyword　Information Retrieval, Search Engine, Natural Language Processing, Dependency Relation, Syntactic Analysis, Table Analysis, WWW

1．背景・目的

従来のWeb検索システムではページ内に指定したキーワードが存在することは保証してくれるが，出現キーワード間に意味的関係があることを保証するものではない。このことは，検索結果に不適当なものが含まれることの一つの原因となっていると考えられる。従って，キーワード間に適切な意味的関係を持つ可能性が高いと思われるページを優先的に扱うことが検索精度向上に有力と考えられる。

そこで我々は，Webページ内に記述されている文章中における検索キーワード間の意味的係り受け関係の強いページを優先的に出力するようなメタ検索エンジンの構築を目指す。

ページ中でキーワード間の意味的関係を表す構造には様々なものが考えられるが，今回は単一文内で，構文木として係り関係が認められるかどうか，表の中のキーワード配置で係り関係の存在が推定できるかどうかの2点について検討を行った。

また本研究では，2つのキーワードを用いて検索する場合が多いことに着目し，当面2語による検索を研究対象とする。

2．キーワード間の意味的係り関係

Webページの重要度に関わるファクタ[1]としては主に以下の項目があげられる。

- 検索語との適合度
- ページ更新日時（新鮮度）
- 参照履歴（人気度）
- リンク構造（引用度）
- ページタイプ

現在のWeb検索エンジンで主に採用されているのは，このうち適合度とリンク構造である。

適合度とは「そのページの主題に検索語が適合しているページほど重要」という考えを基とした指標であり，この推計に用いられている主な検索モデルには

- ベクトル空間モデル（vector space model）
- 確率型モデル（probabilistic model）

などがある[2]。各モデルとも，最終的に文書得点を算出する必要があるため，語の出現に関する何らかの統計量を利用せざるを得ない。一般に，各モデルにはtfとidfという2種類の統計量が用いられている。

tfとはterm frequencyの略で，ある文書に含まれているキーワードの頻度（出現回数）のこと，idfとはinverse document frequencyの略で全文書のうちで，あるキーワードが出現している文書の割合，の逆数のことである[3]。しかし，tf･idf法には問題点がある。新聞記事のようにある程度長い文書なら重要な単語が繰り返し出現するというtf法の仮定が有効であるが，検索対象の文書が繰り返しの少ない短文である場合，tf法の結果に差がでないことになる。これではidf法の単語の出現文書数だけで単語重要度を決めることになり，tf・idf法では高精度な文書検索は期待できない[4]。

我々は，このような統計的手法には限界があり，適合度の算出には検索キーワードやページ内の意味に踏み込んだ処理を行うべきだと考える。そこで本研究では，意味的係り関係に着目した。

単語間の係り受け構造を利用した研究としては立石ら[5]によるものがある。これは，クエリーを自然言語で受け付け，構文解析を行い，係る単語と係られる単語のペアの集合を作成して検索対象のテキスト集合から単語のペアを含む文を抽出し，係り受け関係の一致の度合いによって適合度を判定する手法である。しかし，この方法は，クエリー中で係りを持つ単語がともに含まれるセンテンスが文書中に存在しないと適用できない。このため，文章のみで構成される文書には適した手法ではあるが，Webページのような多様な構造を持つ文書に対応しているとは言い難い。

本研究では，検索キーワード間の意味的関係を抽出するために，HTML文書中でキーワード間の意味的関係を表しうる様々な構造に着目する。さらに，意味的関係を表す構造体内でのキーワードの位置や，ページ全体の中での構造体の位置に着目することで，その文書内における主題に対する検索キーワードの関連の強さを判定することを試みる。

3．センテンス構造による係り受け関係

3．1　方針

例文（図1）のように状態を表す語「人気」と実体を表す語「ノートパソコン」の場合，例文1のように「人気」が「ノートパソコン」に直接係る場合は意味的関係が強く，例文2のように「人気」が「ケース」に係るような場合は意味的関係が弱いと考えることができる。そこで本研究では，こうした関係を検出し，2語のキーワードが強い意味的関係を持つページを優先的に出力するようなメタ検索エンジンの構築を行う。また，語のタイプによって有効な係り受けのタイプも異なると考えられるため，キーワード対のタイプごとにそれらがとる係り受け関係の典型的なパタンを用意し，センテンス中のキーワードがこのパタンに一致する係りを持つ場合に強い意味的関係が存在すると判定する。

例文1）２００３年度の人気ノートパソコン

構文木
```
ノートパソコン【任意－名詞】
 |-人気【任意－名詞】
   |-年度[の]【値－単位名詞】
     |-２００３【任意－数詞】
```

例文2）人気ノートパソコンケースを販売しております

構文木
```
販売する ておる ます【任意－動詞】
 |-ケース[を]【対象格－準文名詞】
   |-ノートパソコン【任意－名詞】
   |-人気【任意－名詞】
```

図1：例文にCSKパーザを適用した結果

3．2　判定手法

まずは前処理として，対象HTML文書内から2つのキーワードを同時に含む文（以下，キーセンテンスという）を抽出する。その際，括弧を含む文については

整形を行い，長文については接続助詞で分割を行う。なぜなら，パーザの2語間の係り受け抽出の失敗事例を検証すると，「括弧が含まれる」場合と「長文が含まれる」場合に解析誤りを含む可能性が高く，従って構文解析を適用する前に，文を整形する必要があるからである。一方，同時にキーワードのタイプから適用する抽出すべき係り受けパタンを選択する。続いてキーセンテンスの構文解析を行い，2 語の係り受け関係を判定する。

現在は，キーワードパタンａとｉを「実体＋現象」，ｂとｇを「実体＋属性」，ｊを「実体＋値」，その他を「実体＋実体」と大まかに4つに分け，それぞれ以下の様な係り受けパタンを抽出するようにしている。

○「実体＋実体」
- 実体名詞＋連体助詞＋実体名詞
  例）大阪のホテル
- 実体名詞＋格助詞（場所位置格）＋ある（主格関係節属性）＋実体名詞
  例）大阪にあるホテル
- 実体名詞＋実体名詞
  例）大阪ホテル
- 実体名詞＋名詞一語＋実体名詞
  例）大阪の老舗ホテル

○「実体＋現象」
- 実体名詞＋連体助詞＋現象名詞
  例）Linux のインストール
- 実体名詞＋現象名詞
  例）Linux インストール
- 実体名詞＋格助詞＋現象名詞動詞形［する］
  例）Linux をインストールする

○「実体＋属性」
- 実体名詞＋連体助詞＋属性名詞
  例）為替のレート
- 実体名詞＋属性名詞
  例）為替レート

○「実体＋値」
- 値名詞＋連体助詞＋実体名詞
  例）中古のパソコン
- 値名詞＋実体名詞
  例）中古パソコン

なお，「実体＋実体」を除き，ルール適用にあたり語（キーワード）の順番を入れ替えることはできないものとする。

4．　表構造による係り受け関係

HTML 文書には，数多くの表が含まれている。表構造内には行・列方向間に関係があり，また，各行・列毎に違った意味を持っている。そして，その意味情報は各行・列の最上位の項目から判定可能である[6]。このように，表構造は，情報検索や情報抽出の分野に有益な情報を含んでいる。しかしながら従来の Web 検索エンジンは表内の関係を示すタグを取り除き，各項目を単語の羅列として索引化していた。そのため，表内に明示されている各項目間の関係を検索に反映させることができなかった[7]。

本研究では，検索キーワードが表中に出現するケースを分析し，適合ページである可能性が高いタイプを，以下の17通り（図2）に分類した。[1]

① **表見出し＋行見出し**
② **行見出し＋列見出し**
③ **2 語がともに表全体の見出しでセンテンスの係りを持つ**
④ **表見出し＋表の中**
⑤ **2 語が同一セル内でセンテンスの係りを持つ**
⑥ **親テーブルで子テーブルの見出し＋子テーブルの中**
⑦ **見出し用テーブル＋表テーブルの見出し**
⑧ **⑤のタイプかつ 2 語ともリンク**
⑨ **見出し用テーブル＋表テーブルの中かつリンク**
⑩ **表見出し＋表の中かつリンク**
⑪ **タイトル＋表見出し**
⑫ **タイトル＋行見出し**
⑬ **タイトル＋列見出し**
⑭ **タイトル＋見出し用テーブル**
⑮ **タイトル＋表の中かつリンク**
⑯ **タイトル＋表の中かつ同行のセルがリンク**
⑰ **タイトル＋表中に頻出**

すなわち，大きく分けると

①-⑦:2語がテーブル（表中に求める情報）
⑧-⑩:2語がテーブル（リンク先に求める情報）
⑪-⑭:1語がタイトル，1語がテーブル（表中に求める情報）
⑮-⑯:1語がタイトル，1語がテーブル（リンク先に求める情報）
⑰　:1語がタイトル，1語がテーブル中に頻出

ということになる。

この戦略をプログラムで実現するにあたり，テーブルには「表に見えるテーブル」と「レイアウトのために用いられているテーブル」があり，それを区別しないと正しい判定はできない，という問題点がある。

そこで，解決策として，「表に見えるテーブル」と「レイアウト用テーブル」の判別プログラムの作成を行い，テーブル判定の前処理として組み込んだ。

判定戦略として，
- 句点の存在の有無
- タグ以外のテキストの存在の有無
- 色の指定の有無
- 線の幅指定の有無
- セルの数
- リンクのあるセルの数
- 文字のあるセルの数

を独立変数として用い，目視で判定したデータを従属変数（教師データ）として C4.5[8]によって最適な決定木を作成し，それをもとにできるだけ「表」を取りこぼさないように改良を行った。

オープンテスト（対象：60 ページ，657 テーブル）の結果は，表 1 の通りである。目視とプログラムの判定が一致しなかったのは 5.8%+4.7%=10.5%で，およそ 9 割は正しく判定されている。

---

[1] ここでの「表見出し」とは 1 行 1 列目のセル，「列見出し」とは 1 行目のセル，「行見出し」とは 1 列目のセル，「タイトル」とはタイトルタグで囲まれた文字列のことを指す

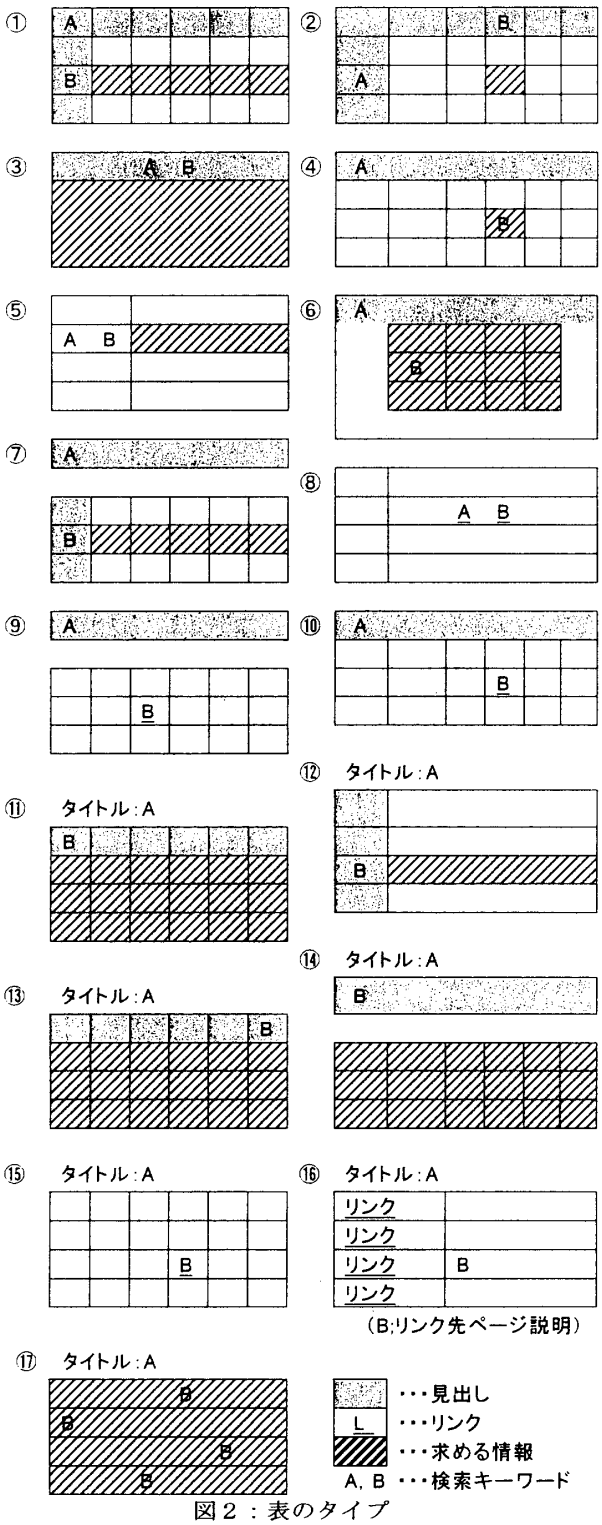

図２：表のタイプ

表１：表／レイアウト判定精度

		プログラム	
		表	レイアウト
目視	表	23.6%	5.8%
	中間	0.3%	0.8%
	レイアウト	4.7%	64.8%

５．システム構成

本システムは，既存の検索エンジンを利用したメタサーチエンジンである。現在のところ，実験にはGoogle[9]を用いている。尚，センテンスの係り受け判定には（株）CSK で開発された日本語パーザを利用しており，システム全体については Ruby で実装している。

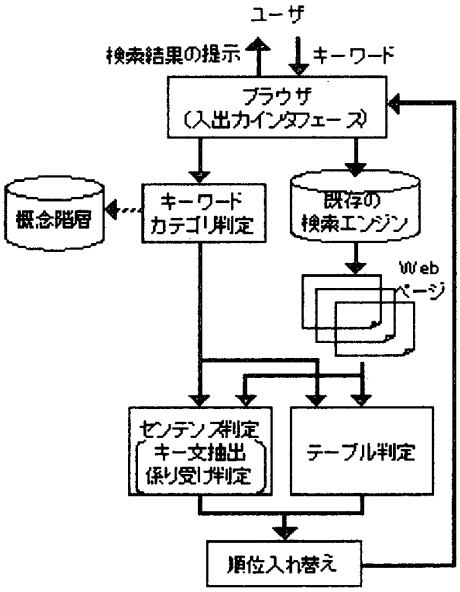

図３：システム全体図

６．実験

６．１　従来の情報検索と Web 検索の評価法

検索が成功したかどうかを評価するための伝統的な指標は，精度（precision）と再現率（recall）である。精度や再現率は，適合度による順位付けを行わない伝統的なブール型検索に適した評価手法である。しかし，Web ページの検索は①全文検索，②膨大なページ数，という２つの理由から，検索結果が数多く出力されてしまう[10]ため，適合度順に順位付け出力が行われる。従って，Web 検索を評価するには，いくつかの工夫を加える必要がある。適合度順出力の評価のために TREC(Text REtrieval Conference) [11]を中心とした検索実験で使用されている主要な指標には，精度（λ），再現率（λ），再現率 50%での精度，R 精度，平均精度などがある[2]。

また Web 検索において，ユーザは上位の結果しか参照しない傾向があるため，一般的に再現率（取りこぼしが少ないこと）より精度（不適合ページが少ないこと）が重視されている。

６．２　評価データの収集

我々のシステムを実験・評価するためには，テストコレクション，すなわち

・検索対象となる文書集合（データベース）
・検索質問の集合
・各検索質問に対して各文書が適合しているかどう

かの情報
が必要である。文書集合中に含まれる適合文書（正解文書）を洗い出す作業は容易ではない。そこで広大かつ多様なWeb空間においてどのようなデータ収集を行えば，我々に可能なレベル（コスト）で説得力のある評価データを揃えたい，という欲求を満たせるかを検討した。

2語キーワードで行われる検索の代表的パタンを網羅したテストセットを作成するため，2語キーワードでの検索の全体をどう捉えればよいかということを考え，検索意図でもって大分類を試みた。

Web上で検索エンジンを利用する際のユーザの目的とは，Web空間上で購入・予約・ダウンロードなど何らかのアクションをとることを欲してそれが可能なページを求める場合と，単に情報を得るために情報の記載があるページを求める場合とに大別できる。世の中は「もの」と「こと」から成り立っていると考えると，検索の場合も，「もの」もしくは「こと」に関する情報を得たいのであり，「もの」に関する検索は，ものの内包的性質の説明を求めるもの，もののインスタンスを求めるもの，ものの特定の属性値を求めるものに分けられる。「こと」にかかわる検索も同様である。

以上より，Web検索における意図の全体集合を次のように捉えることとする。

（ア）サービス提供
（イ）実体：概説
（ウ）実体：属性
（エ）実体：インスタンスリスト
（オ）現象：概説
（カ）現象：属性
（キ）現象：インスタンスリスト

それらの検索意図を持った際にどんな2語が指定されうるかを調べるため，「USJに行きたい」「ノートパソコンを購入したい」などと状況を設定して（ア）から（キ）の各タイプの検索意図を想定した検索課題を与え，その際どのようなキーワードで検索を行うかを尋ねるアンケートを実施した。そこから検索キーワードと想定する結果ページ（検索意図）の収集，分析を行った。尚，アンケートの回答者は，静岡大学情報学部または情報学研究科に在籍する学生約200名である。

アンケートで収集できた検索クエリー（有効件数：1947）のうち，1語で検索を行っているケースが21.5%（419件），2語が48.4%（942件），3語以上は30.1%（586件）であった。

さらに，検索キーワードについては，図4の11タイプに分類した。

この11タイプを元に，2語のペアごとに集計すると，上位12パタン（表2）で全体の75%以上を占めることがわかった。

そこで先述のアンケートで収集した約2000件の検索キーワードのうち，2語検索の上位12パタンの割合に従って代表50キーワード対を選び，Googleで検索を行って評価用のデータとして用いることにした。それぞれのキーワード対について検索結果上位100件を取得し，適合判定を人間（著者）が行った。パタンごとの検索課題数の内訳を表3に示す。

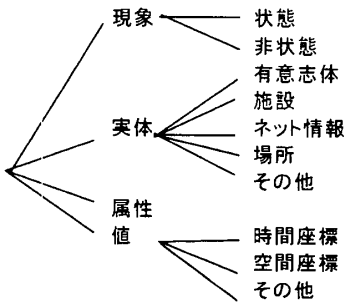

図4：検索キーワードのタイプ

表2：検索キーワードパタン

a	現象：非状態	実体：その他	153	16.5%
b	属性	実体：その他	94	10.1%
c	実体：その他	実体：その他	66	7.1%
d	実体：施設	実体：その他	60	6.5%
e	実体：場所	実体：施設	58	6.2%
f	実体：ネット情報	実体：その他	54	5.8%
g	実体：場所	属性	53	5.7%
h	実体：有意志体	実体：場所	45	4.8%
i	現象：非状態	実体：施設	38	4.1%
j	現象：状態	実体：その他	36	3.9%
k	実体：場所	実体：ネット情報	34	3.7%
l	実体：施設	実体：施設	34	3.7%
		上位12パタン計	725	78.0%

表3：実験用データセット検索課題数内訳

a	b	c	d	e	f	g	h	i	j	k	l	計
10	6	5	4	4	4	4	3	3	3	2	2	50

6．3　評価

実験利用データ5000ページのうちPDFなどのバイナリファイルを除いた4863ページにおける適合ページの割合は46.9%であったのに対し，係り受け関係を検出できなかったページを却下したときの精度は表4の通りである。表4中の「表＋セ判定」とは，表判定で①~⑰に該当しなかったページと同一セル内に2語があったページについてセンテンス判定を行った結果である。なお，表が存在するページの割合は(4863-901)/4863=81.5%，うち係り受けが検出できたページの割合は(1068+495)/(4863-901)=39.4%であり，同様にキーセンテンスが存在するページの割合は35.1%，うち係り受けが検出できたページの割合は57.9%である。

さらに，表4の「センテンス判定＋表判定」をキーワードタイプごとに集計（表5）すると，各パタンごとの精度／再現率にはかなりばらつきがあることがわかる。これはさらにパタンに即した戦略を立てる必要性を意味する。

続いて，係り受け判定の戦略によって重み付けを行いスコアを算出して並び替えを行った場合の，上位20位までの適合ページ数を比較する。

今回のスコアの算出は，センテンスによる係り受けが存在したら100点，表による係り受けが存在したら100点を与え，さらに（100-元々（Google）の順位）を加算する，という方法で行った。

50個の検索課題×20位＝1000件中，適合ページ数はGoogleが588件であるのに対し，本システムでは

684 件であった。すなわち 1 つの検索課題につき，上位 20 件に平均約 2 ページ，適合ページが増えたことになる。

表４：全ページを対象とした集計結果

		表分析判定	
		○	×
ユーザ判定	○	HIT	MISS
	×	FA	CR

（単位；ページ）

	HIT	FA	CR	MISS	対象外[2]	精度[3]
表判定	1068	495	1364	1035	901	68.3%
センテンス判定	715	273	294	425	3156	72.4%
表＋セ判定	1393	672	1910	888	—	67.5%

表５：キーワードタイプごとの集計結果

	精度	再現率
a	54.3%	45.4%
b	53.9%	49.9%
c	64.8%	68.5%
d	64.7%	51.1%
e	73.4%	69.5%
f	62.2%	42.4%
g	49.1%	61.4%
h	55.1%	77.3%
i	83.9%	46.4%
j	62.6%	73.6%
k	79.2%	52.1%
l	79.0%	78.0%

表６：Google と本システムとの比較

	Google	本システム
検索課題数	50	50
A:システムが適合と判定した文書数	1000	1000
B:データセット中の適合文書数	2281	2281
C:A に含まれる B の数	588	684
精度（C/A）	0.5880	0.6840
再現率（C/B）	0.2578	0.2999
F 尺度（精度と再現率の調和平均）	0.3584	0.4384
平均精度（MAP）	0.1968	0.2536
上位 5 件の文書集合における精度	0.6200	0.7600
上位 10 件の文書集合における精度	0.5780	0.7460
上位 15 件の文書集合における精度	0.5880	0.7107
上位 20 件の文書集合における精度	0.5880	0.6840
上位 30 件の文書集合における精度	0.3920	0.4560
上位 100 件の文書集合における精度	0.1176	0.1368
R 精度	0.2768	0.3088

さらに詳細に評価を行うため，NTCIR-Web[12]で用いられている評価尺度である，平均精度，R 精度，上位 5，10，15，20，30，100 件における精度，再現率－精度グラフ[2]を求め，Google との比較（表６，図４）を行った。これらは TREC [11]等で標準的に使用されているツールである trec_eval[13]を使用して算出した。２つの手法の性能差を調べる際に用いられる MAP(mean average precision)[2]という指標を用いて Google と本システムの性能差を比較すると，差は 0.0568 である。経験的に両者の MAP に 0.05 程度の差があれば有意差があるといわれているようであるが，より厳密に t-検定も実施したところ，危険率 1%で有意であった。

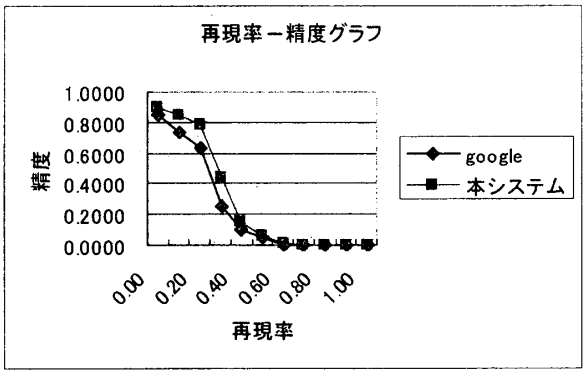

図４：再現率－精度グラフ

７．精度向上の工夫

同じタイプの係りを持っているとしても，その係りがページの中でどの程度の重要度であるのかを決める要因は，他にも存在していると考えられる。また，ページに重要な情報が掲載されているかを判定するためには，係りの強弱だけでなくページのタイプを特定することも必要なのでは無いだろうか。今後，さらに追加していきたいと考えている戦略を，以下に簡単に紹介する。

７．１　ページタイプ特定フィルタ

Web ページの特徴として，ジャンルの多様性が挙げられる。論文など学術的なものから，企業や商品の宣伝，個人の日記や掲示板などまでが区別なく混在している。これでは，検索キーワード間の意味的関係が強くても，検索意図にそぐわないページがあっても不思議ではない。例えば専門用語の意味が知りたいときに書籍紹介のページやシラバスが与えられても検索者は満足できないであろう。

そこで我々が作成したページタイプ特定フィルタは，掲示板や講習会案内といったページタイプを特定するコンテンツ解析フィルタである。検索意図に従って正解情報の抽出にも不正解情報の抽出にも利用できる。

現在，掲示板，メーリングリスト，授業案内・シラバス，講習会・セミナー案内，書籍紹介，商品紹介，代行サービス案内をほぼ特定できる。プログラム判定精度を表７に示す。

なお，ページタイプ特定フィルタは，検索時にユーザが意図に沿って選択し，利用できる形で提供したいと考えている。

---

2 キーセンテンス，テーブルがそれぞれ存在していないため CR・MISS に含めていないページ数

3 精度＝HIT/(HIT+FA)，また再現率は HIT/(HIT+MISS)で計算可能

表７：ページタイプ特定フィルタ精度一覧

ページタイプ	精度
掲示板ページ	80.9%
メーリングリストページ	100.0%
講習会・セミナーページ	72.4%
書籍紹介ページ	78.2%
商品紹介ページ	61.8%
シラバスページ	100.0%
代行サービスページ	69.8%

７．２　センテンスの時制の考慮

ニュース記事や日記は過去形の文となることが多い。そしてこのタイプの情報はあまり有用なページではないことが少なくない。そこで，キーワードが出現するセンテンスについて時制を調べ，誤った適合ページの検出を抑える手段とする。

７．３　キーワード間の係り受け関係の文中の位置

キーワードにマッチした語と語の間の係り受け関係において，同じタイプの係りだとしても，その文中の位置によって，重要度は異なると考えられる。

文中の接続助詞より文頭側や関係節の中に検索キーワード間の係り受け関係があるのと，構文木のヘッドに係り受け関係があるのとでは，文におけるキーワードの重要性は異なる。さらに，一文中に接続助詞が複数含まれる場合や接続助詞のタイプによる違いをも判定材料に含め，これらを利用して誤った適合ページの検出を抑える。

７．４　見出しとその支配範囲の特定

段落や表，箇条書きの見出しは，その支配下にある語（情報）に対し係り受け関係が存在すると考えられる。近傍の文字と比較し，サイズが大きい，色が異なる，フォントが太字になっているといったことを手がかりに見出しを特定し，また見出しの支配範囲も特定すれば，新たな係り受け関係を取得できるはずである。

７．５　ページ全体におけるキーワードの出現位置

ページ内において，検索キーワードが上部に出現する場合と末端部にしか存在しないような場合とでは，ページにおける語の重要度が異なることは明らかである。そこで，ページ全体中のキーセンテンスの位置を求め，誤った適合ページの検出を抑える手段として利用する。

表８：キー文が存在するページを対象とした集計[4]

	HIT	FA	CR	MISS	精度	再現率
表＋セ判定	916	358	209	224	71.9%	80.4%
表＋セ＋過去	912	344	223	228	72.6%	80.0%
表＋セ＋文中	900	318	249	240	73.9%	78.9%
表＋セ＋ペ中	904	334	233	236	73.0%	79.3%
全戦略	899	298	269	241	75.1%	78.9%

[4] 「文中」とは文中の係りの位置，「ペ中」とはページ中のキーセンテンスの位置，「全戦略」とは３つ全てを加味

８．おわりに

本研究では，２語キーワードによる検索において，センテンスや表の係り受け関係に着目することで検索精度を向上させることに効果があることを示した。

さらに，７章で紹介した工夫のうち，時制，文中の係りの位置，ページ中のキーセンテンスの位置については既に実装及び実験段階に入っており，現状では表８に示した程度の効果が認められている。今後もより一層戦略を改善することで，更に効果が現れることを期待している。

また現在は，表構造中のキーワード出現位置におけるキーワードパタンごとの戦略の検討に取り組んでいる。例えば「現象：状態＋実体：その他」の組み合わせの場合，状態を表す語は実体を直接修飾しなければならない可能性が高いため，①や②など，別々のセルにキーワードが存在しているケースはむしろ不適合ページである可能性が高い。

更に今後は，検索キーワードのタイプ等からインタラクティブに検索者の意図を推定できるシステムを目指し，検索意図を考慮してページ内に現れるべきキーワードの意味的関係を絞り込むなど，検索意図を利用して快適な検索を提供することに挑戦したいと考えている。

参考文献

[1] 福島俊一：Web サーチエンジンにおけるページ重要度計算，５８回情報処理学会全国大会公開パネル４

[2] 岸田和明，岩山真，江口浩二：検索実験の方法と実際:NTCIR ワークショップでの試み，Pre-meeting Lecture at the NTCIR-3 Workshop，2002

[3] 原田昌紀：サーチエンジン徹底活用術，オーム社開発局，1997

[4] http://venus.netlaboratory.com/salon/chiteki/jfs/02.html

[5] 立石健二，大庭直行，峯恒憲，雨宮直人：係り受け情報を利用した Web 上の日本語テキスト検索システム，情報処理学会研究報告．DD 13-7，pp. 47-54，1998

[6] 吉田稔，鳥澤健太郎，辻井潤一：表形式からの情報抽出手法，言語処理学会 第６回 年次大会，pp. 252-255，2000

[7] 岩口義広，鄭眠洙，獅々堀正幹，青江順一：WWW 空間上に存在する表構造の一索引化手法，情報処理学会情報学基礎研究会，FI 61-22，pp. 159-166，2001

[8] http://www.cse.unsw.edu.au/~quinlan/

[9] http://www.google.co.jp/

[10] 久野高志，安形輝，上田修一：情報検索システムとしてみたサーチエンジン，第 49 回日本図書館情報学会研究大会発表要綱，pp.47-50(2001-10-20)，2001

[11] http://trec.nist.gov/

[12] http://research.nii.ac.jp/ntcir/

[13] ftp://ftp.cs.cornell.edu/pub/smart/trec_eval.v3beta.shar