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

2006－DD－55 (2)
2006／5／26

表構造における意味的関係に基づくWWW検索精度の向上

松本 章代[†]，西口 直樹[†]，小西 達裕[†]，高木 朗[††]，
小山 照夫[‡]，三宅 芳雄[‡‡]，伊東 幸宏[†]

[†] 静岡大学 [††] CSKシステムズ [‡] 国立情報学研究所 [‡‡] 中京大学

Web検索エンジンに，ユーザが検索キーワードとして2つの語を入力した場合に，その2語が意味的関係を持って文書中に出現しているか否かを判定する，ランキング手法を提案する．キーワード間の意味的関係を表現する構造として，本研究では表構造を取り上げる．提案する手法に基づいて，既存検索エンジンの検索結果をフィルタリングするシステムを構築し，NTCIR-3WEBによるテストコレクションを用いて，本手法の精度を実験的に検証する．

Improvement in precision of WWW Search Engines based on Semantic Relation in Table Structure

Akiyo MATSUMOTO [†] Naoki NISHIGUCHI [†] Tatsuhiro KONISHI [†] Akira TAKAGI [††]
Teruo KOYAMA [‡] Yoshio MIYAKE [‡‡] Yukihiro ITOH [†]
[†] Shizuoka University [††] CSK Systems Corporation
[‡] National Institute of Informatics [‡‡] Chukyo University

In this paper, we propose a method to improve precision of WWW search engines. When Web search engines are used, the user inputs keywords. Web search engines accept a few keywords as a retrieval condition, and select candidate pages based on some statistics value such as TF/IDF and so on. We think that the keywords input by users have a semantic relation. We try to improve retrieval accuracy by checking whether two keywords have a semantic relation in the candidate pages or not. We deal with table as a structure to represent semantic relations between keywords. We construct a meta search engine which accept output of an ordinary search engine and select plausible ones by checking semantic relation. We use NTCIR-3WEB as a test collection, and we also show an experimental evaluation of our method.

1 はじめに

Web検索において，検索キーワードとして複数の語が入力される場合を考える．ただしWeb検索エンジンの大部分の利用者は，ごくわずかの検索キーワードしか利用しない傾向にある．風間らによれば，日本の検索エンジンODINの検索ログにおいて，そのほとんどが1〜2語（平均1.42語）であったという[1]．つまり複数の語が入力されるケースにおいては，特に2語で入力されることが多いと考えられる．その際，同義語・類義語を並べる場合もあるが，そうでない場合にはその2語が全く独立に選ばれるとは考えにくい．その2語は何らかの意味的関係を有すると考えられる．

複数キーワード間の意味関係を示す方法としては，文構造による記述が典型的であるが，実際にはそれ以外の構造でも記述は可能である．

例えば表構造の具体例を挙げると，「東南アジア情報」などのタイトルがついた表の行の見出しに「タイ」「ベトナム」などの国名が並び，列の見出しに「気候」「通貨」などとともに「宗教」という見出しが並んでいる場合，「シンガポールの宗教は…である」などの命題が表現されており，「シンガポール」と「宗教」の間に意味的な関係が構成されている．このように，文の構文解析では係り関係が

発見できなくとも，実際に目的とする検索意図に合致する意味関係にあることがありうる．

我々はこれまで，文構造および見出し構造を対象として検索精度の向上を検討してきた[2, 3]．今回は更なる精度向上を目指し，表構造に着目して意味的関係が認められるかどうかについて議論を行う．

表の中で二つのキーワードがどのような配置にある場合に，目的とする意味関係を記述している可能性が高いのかということを検討することが重要である．この結果に基づき，適切な意味関係にある可能性が高いテーブルを発見し，このようなテーブルを含む文書のランクを上げる検索手法を提案する．

2 表内の意味的な結びつき

HTML 文書には，数多くの表が含まれている．多くの場合，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[4]．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，岩口ら[5]も指摘しているように，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．そこで岩口らは，Web 空間上に存在する複雑な表構造を対象にし，表構造内の関係を保持したまま各項目を索引化する手法を提案している．但し，この手法は個々のテーブル毎にセルの位置情報を取得するものであり，ページ内に存在するすべてのテーブルを俯瞰的に把握できるものではない．一方，我々のシステムは，2 つのキーワードが異なるテーブル間にまたがって存在する場合，あるいはページのタイトルとテーブル内にそれぞれ存在している場合なども含め，キーワード間の意味的関係の強さを測ることができるようテーブルの構造を解析する．

そこで，表の構成要素を

- 表見出し … 表全体を支配する見出し
- 列見出し … 各列の見出し
- 行見出し … 各行の見出し
- 内容セル … 見出し以外

の 4 つと捉えることにする．これらの構成要素間には，役割に応じた意味的関係（修飾－被修飾関係）が存在していると考えることができる．

Table 1 表構造における係り受けパタン

(a)	表見出し+行（列）見出し
(b)	表見出し+内容セル（リンク）
(c)	表見出し+内容セル（(b) ではなく頻出）
(d)	表見出し+内容セル（(b),(c) ではなく同行セルがリンク）
(e)	表見出し+内容セル（(b),(c),(d) 以外）
(f)	行（列）見出し+同行（列）の内容セル（リンク）
(g)	行（列）見出し+同行（列）の内容セル（(f) 以外）
(h)	行見出し+列見出し
(i)	同一セル（見出し，内容セル，ページタイトル問わず）
(j)	ページタイトル+表見出し
(k)	ページタイトル+行（列）見出し
(l)	ページタイトル+内容セル（リンク）
(m)	ページタイトル+内容セル（(l) ではなく頻出）
(n)	ページタイトル+内容セル（(l),(m) ではなく同行セルがリンク）
(o)	ページタイトル+内容セル（(l),(m),(n) 以外）

なお，見出しが階層構造を持つ場合も在り得るが，それをプログラムで正しく把握することは困難であるため，現段階では見出しはフラットなものとみなして取り扱う．

HTML 文書内の表構造における 2 語が出現するパタンについて，修飾－被修飾関係を考慮し，何らかの意味があると思われるパタンを中心に整理を行った（Table 1）．

表見出しは，表全体の主題を端的に表したものであり，その意味で表の全ての情報と意味的関係を持つと考えられる．1 語が表見出し，もう 1 語が行（列）見出しならば，該当する行（列）の内容セルが検索要求を満たす（(a)）．

1 語が表見出し，もう 1 語が内容セルならば，内容セルに文章がつづられている場合には，表見出しとその文章には何らかの関係があるが，その文章におけるキーワードの持つ役割も適合／不適合に関わる重要なファクタである（(b)〜(e)）．

行見出し・列見出しは，一つの概念の複数の属性を列挙したり，一つのカテゴリに属する複数の概念を列挙したりするために用いられることが多い．

その範囲内では，該当する行または列の内容セルとの間に意味的関係を持つ（(f)，(g)）．見出しと内容セルが異なる列または行にある場合は，内容セルでは見出しが指定するものとは異なる概念，異なる属性についての記述がなされていると考えられるので意味的関係は薄いと考えてよい．列見出しと行見出しが両方用いられる場合は，双方が直交する属性を指定していたり，一方で概念を特定し他方で属性を特定するなど，両方の見出しが組み合わさって一つの条件を設定している（(h)）．

また，異なる内容セルは異なる概念や異なる属性についての陳述であるため，互いに独立した内容である．

尚，同一セルに2つのキーワードが存在した場合は，表見出し・行見出し・列見出し・内容セルに関わらず，文の係り受け判定に判断を任せるものとする（(i)）．

さらに，HTML文書の特徴として，リンクとページタイトルの存在を考慮すべきである．内容セルに存在するキーワードがリンクとなっている場合には，リンクとなっているところが内容セルの主たる内容であり，リンクの先にその詳細な情報が置かれている可能性も高いと考えられる．そこで，内容セルに関しては，キーワードがリンクになっている場合と，そうではない場合，それぞれ区別して分析を行うこととする．リンクになっていない場合について，キーワードがひとつの表内に頻出（暫定的に閾値を5回以上とする）するケースや，リンク集において，サイト名がリンクになっていてその横の説明文中にキーワードが存在しているケース（つまりキーワード自体がリンクではなくても，同行にリンクが存在しているケース）なども適合ページの可能性が高いと予想できることから分けて分析することとする．また，ほとんどのWebページにはページタイトルが付けられているが，そのスコープ内に表は含まれるものであり，表にとって無視できない存在である．ページ内の情報量に占める表の割合にもよるが，表のタイトルがページのタイトルになることさえある．そこで，ページタイトルと表との関係も合わせ，意味的な関係が密接である可能性が高いパタンを列挙した（(j)〜(o)）．

以上のパタンについて適合ページである可能性を検証するため，2語で検索した結果ページの中からページ中にテーブルタグが存在する500ページを対象として目視による分析調査を行った．すると(g)と(o)以外のパタンに関して，検索者が求める情報が，表の中またはリンク先に存在している可能性が非常に高い傾向にあることがわかった．(500ページ中，適合ページは236ページ(47.2%)であったのに対し，この13パタンのいずれかに該当するページは133ページあり，うち適合ページは118ページ(88.7%)であった)

そこで，(g)と(o)，さらに文の係り受け判定に依存する(i)を除いたパタンを，表の係り受け判定として採用することにする．

ただし，これを実際にプログラムで適用するためには「表」を正しく抽出できるということが前提となる．Wangら[6]が指摘するように，Webページに出現する多くのテーブルタグは，テーブルを表すために用いられるのではなく，レイアウトを制御することにも用いられる．彼らは，テーブルタグを用いて構成されたテーブルの中から「本物の表」を抽出するため，レイアウト構造（タグ）やセル内の字種などの特徴と，従来のテキスト分類の手法を組み合わせ，高い精度（精度：97.5%，再現率：94.3%）で判別を実現している．

確かに，レイアウトのために用いられるテーブルは，同じ行や列のセル同士に意味的な深い関係を持たないどころか，全く別のテーマを扱うことさえある．従って，我々にとっても，テーブルタグを用いて書かれているテーブルを「意味的に表を表わすテーブル（タイプA)」と「レイアウトを制御することだけに用いられるテーブル（タイプB)」とに弁別することは，正しく係りを判定する上で必要不可欠である．

そこで我々も，独自のアルゴリズムでテーブルタグをタイプAとタイプBに分類し，タイプBについては表の係り受け判定処理の対象外とするための判別プログラムの開発を行い，表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

Table 2　表 (タイプ A) ／レイアウト (タイプ B) 判定精度

n=675(テーブル)

		プログラム	
		タイプ A	タイプ B
目視	タイプ A	23.6%	5.8%
	中間	0.3%	0.8%
	タイプ B	4.7%	64.8%

- 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（対象：120 ページ，1909 テーブル）を従属変数（教師データ）として C4.5 [1]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タイプ A 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，Table 2 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，Wang らの精度には及ばないもののおよそ 9 割は正しく判定されており，我々は実質上使用できるレベルに達していると考えている．

なお，表の見出しをプログラムで抽出するにあたり，

- 表見出し：CAPTION タグ，(タイプ A のテーブルの) 1 行 1 列目セル，見出し用テーブル（隣接したタイプ A の見出しとしての役割しか持たないタイプ B のテーブル），親テーブル（セルの中にタイプ A のテーブルを含むタイプ B のテーブル）
- 行見出し：(タイプ A のテーブルの) 2 行目以降の 1 列目
- 列見出し：(タイプ A のテーブルの) 2 列目以降の 1 行目
- 以上の条件に該当していても，その中に句点が含まれる場合は，見出しではないとみなす

という条件に基づき処理を行うものとする．

3　システム構成

本システム（Fig.1）は，既存の検索エンジンを利用したフィルタリングツールである．

まず，ユーザによって Web ブラウザから検索キーワードが入力されると，既存の検索エンジンにキーワードを渡し，検索結果の Web ページを取得する．

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

Web ページは表構造，文構造，見出し構造のそれぞれについて意味的関係の強さを測られる（それぞれのユニットを表判定，文判定，見出し判定と呼ぶ）．

表判定ユニット（Fig.2）では，まず，ページ中に存在するすべてのテーブルについて，タイプ A かタイプ B かを判定する．続いて検索キーワードが存在するテーブルに対し Table 1 のパタン（(g),(o) を除く）に該当するか否かを調べる．その際，2 節で述べた定義に従い，タイプ B の中でタイプ A のテーブルを内部に含むテーブルは親テーブルとみなし，タイプ B のテーブル（但しタグを除く文字が含まれるセルが 2 つ以内のもの）とタイプ A のテーブルが続けて現れる場合，前者を見出し用テーブルと判定する．(見出し用テーブルの判定アルゴリズムはヒューリスティックに基づく)

文判定ユニット [2] では，各キーワードをパーザの概念階層辞書を用いて実体，現象，属性，値という 4 つのカテゴリのいずれかに分ける．そのキーワードカテゴリを用いて係り受けパタンの候補を挙げる．ついで，キーワードを含む文がそのうちのどれかのパタンに該当するか否かを調べる．

見出し判定ユニット [3] では，2 つのキーワードがそれぞれ見出しの一部として存在し，かつその見出し間に親子関係があるか否かを調べる．

各判定ユニットで，パタンにマッチしたページをキーワード間の意味関係が強いとみなし，一定のスコアを与える．システムはスコアの合計によってページを並べ替える．つまりより多くの判定ユニットで適合と認められたページほど上位にランクされる．

こうして並び替えられた検索結果はユーザに提示される．なお，スコアリング戦略は今後詳細に議論する必要がある．

4　評価実験

4.1　実験手法と結果

我々のシステムを客観的に評価するにあたり，国内の標準的なテストコレクションである NTCIR-3WEB を用いる．NTCIR-3WEB は，(i) 検索対象となる文書セットとしておよそ 100GB の Web ページ（NW100G-01）と (ii) 情報ニーズを記述した検索課題，(iii) 各検索課題に関する適合判定の結果から構成されている．

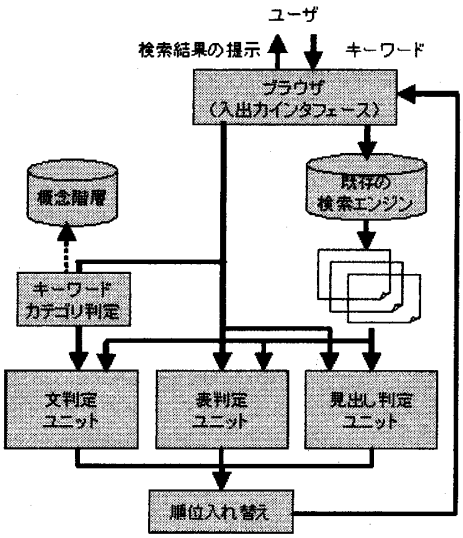

Fig. 1　システム全体図

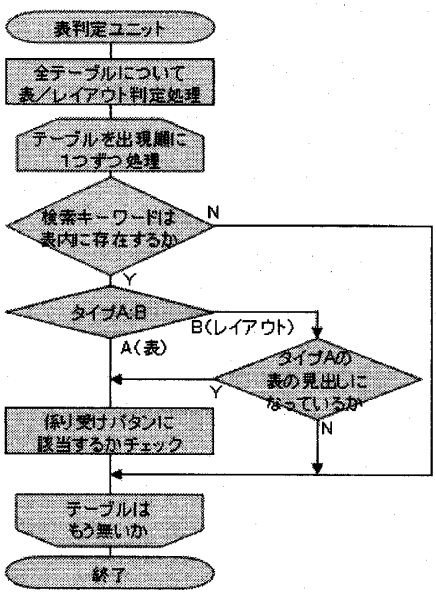

Fig. 2　表判定ユニット

Table 3　キーワードリスト

1	ロープワーク, 結び方
2	速読法, 効果
3	テーピング, 方法
4	変分法, 入門
5	亀, 寿命
6	オゾン層, オゾンホール, 人体
7	印象派, モネ, 美術館
8	イースター, 復活祭, キリスト
9	カプサイシン, とうがらし, 効能
10	アントシアニン, ブルーベリー, 視力
11	柴犬, 日本犬, 特徴
12	シフォンケーキ, 作り方, 菓子
13	京都, 寺, 神社
14	宮部みゆき, 書評, レビュー
15	スピーカー, 評価, 比較
16	資格試験, 情報処理, IT
17	憲法第九条, 解釈, 意見
18	加速器, 医療, 治療
19	アロマセラピー, アロマオイル, アロマキャンドル
20	湖, 水質, 透明度

NTCIR-3WEB では，検索エンジンに入力するキーワードについて，(a) 同義語・類義語を並べる場合，(b) 異なる観点の語を並べる場合，(c) 戦略 (a) と戦略 (b) の複合，の 3 つの戦略が区別され，検索課題ごとにいずれかのラベルが付与されている．そこで，47 の検索課題の中から，キーワードが 2 語で構成されているもの，及び 3 語で構成されておりかつ (c) に該当するもの，計 20 組を選んだ（Table 3）．

NW100G-01 から適合判定の結果が存在しているページを集め，namazu [2] で索引化を行い，(類似ページを除いた）検索結果上位 100 件を取得して，我々のシステムでランキングする．なお，各検索課題が想定する適合ページがすべて表構造に適しているとは限らないので，文構造・見出し構造による判定も加え，最終的な評価とする．

Table 4 に，namazu（TF/IDF）と我々の提案手法における表判定を含めない場合と含めた場合，それぞれのランキングの上位 20 位までの適合ページ数を示す．ここで，Table 4 における検索課題 14 と 19 は，適合ページが存在しないことが判明している．従って，14 と 19 を除いた 18 の検索課題について，namazu に比べ表判定を含めた提案手法は平

[2] http://www.namazu.org/

均約 1.22 ページ増えている．表判定を含まない場合と含む場合とを比較すると，検索課題 3,13,16,18 では適合ページ数の増加が見られるものの検索課題 6,7,12 で減少しており，トータルでは若干しか増えていない．表判定を加えたことで精度が下がった原因の一つは，4.2.3 節で述べる理由によるものである．もう一つは，片方の語が「作り方」など属性の場合の問題である．実体とその属性，というキーワード対の場合，ページ中の属性名が別の実体の属性であった，というのは不適合ページの典型例となる．表判定の場合，ページ中の属性の指す実体が何か，という厳密なチェックは行っていないため，誤って適合ページと判定してしまう危険性がある．この問題に関しては，キーワードを分類しタイプごとに有効な表判定のパタンを見出す，あるいは文判定と適切に組み合わせることで，対処できる可能性はある．

次に，各検索課題ごとの平均精度の平均，MAP(mean average precision) を求める．namazu の元々の順位では 0.433 であるのに対し，本手法では 0.498 であった．また，表判定を除き文判定及び見出し判定のみで算出した場合は 0.479 となった．なお，手法の性能比較を行う場合，MAP に 0.05 程度の差があれば有意差があると一般的に言われている．

Table 4　上位 20 件中適合ページ数

	namazu	表判定除く	表判定含む
1	18	18	18
2	5	5	5
3	3	2	3
4	0	0	0
5	3	4	4
6	11	10	9
7	2	9	7
8	16	19	19
9	1	1	1
10	19	17	17
11	0	4	4
12	6	10	9
13	8	10	11
14	0	0	0
15	0	0	0
16	0	2	4
17	14	13	13
18	4	4	6
19	0	0	0
20	5	7	7
計	115	135	137

4.2　考察

4.2.1　精度を下げる要因

本手法で不適合ページを拾い上げてしまうケースとは，以下のような事例である．

まず，キーワードが別の語を修飾し，検索意図から外れてしまったというケースである．キーワードが文のヘッドから遠く，文における重要度が低いといえる．

次に，検索意図にそぐわないページタイプがある．大きく以下の 3 つのタイプに分けられる．

- 個人的な内容 → ブログ，日記，体験談
- 過去の内容 → ニュース，キャンペーン・新製品などのお知らせ，過去のオークション・プレゼント，求人
- 見出しのみで中身（解説）がない → 書籍・授業・セミナー紹介

また，パタンにマッチした箇所が，ページの中での扱い（重要度）が小さく（属性値などの）知りたい詳細な情報が書かれていない，というケースがある．これを判断するためには，キーワードが存在している構造自身が，どれだけの情報量の支配範囲を持つかという情報が必要である．支配範囲がほとんど（もちろんリンクも）ない場合，少なくとも，そのキーワードの解説はページ内に存在しない可能性は高いと思われる．

4.2.2　取りこぼしの要因

本手法で適合ページを取りこぼす主な要因を調査したところ，以下に挙げるケースに該当する事例が多く見受けられた．

まず，キーワードの 1 語が見出しに含まれ，もう 1 語がその見出しの支配範囲下にある地の文に含まれる場合である．現在の見出し判定の戦略では，見出し間の親子関係を対象とするため，片方の語が地の文にしか含まれない場合は対象とならない．一方のキーワードの支配範囲下において任意の場所にもう一方のキーワードが出現さえしていればよい，としてしまうと多くの不適合ページを誤って適合とみなしてしまうことになるが，地の文でも

- リンク
- 頻出
- 1 行目
- カギ括弧による強調

などと条件を絞り込むことによって，精度を下げずに再現率を上げられる見込みはあると思われる．

2つ目は，片方のキーワードが「意見」「効能」など属性を表す語であった場合，その言葉を直接使わなくても，もう一方のキーワードの属性値について言及することは可能，というケースである．例えば，「速読法　効果」というキーワードで検索した文書において，「効果」というキーワードを用いずに速読法の効果について説明しており，同じ文書の別の箇所でたまたま速読法以外の「効果」について触れているようなケースがこれにあたる．

4.2.3　先行研究による実験結果との比較

今回は初めてテストコレクションとしてNTCIR-3WEBを用いて評価実験を行った．これまで，我々は自作したテストコレクションを用いて評価実験を行っており，Googleの検索結果をフィルタリングし上位20件中の適合ページ数などを比較している[2]．これによると，例えば今回と同じく20の検索課題の中に含まれている上位20件中の適合ページ数を比較した結果，Googleでは計260ページだった適合ページが文判定のみで296ページに増加している．

つまり，これまでの結果（＋36ページ）と今回の結果（＋22ページ）を比べると，今回は明らかに精度が低いと言わざるをえない．

これは，適合文書の判定基準の違いに起因するものと考えられる．例えば，「印象派　モネ　美術館」のNTCIR-3WEBによる判定基準は次の通りである．「適合文書は印象派に属する画家のみではなくその作品（題名だけでも良い）も紹介されていて，どこの美術館でそれらの作品が見られるかという情報を提供しているもの．」この「のみではなく」という部分の条件を満たすことが難しい．また，「京都　寺　神社」の場合は次のようになっている．「適合文書は，京都の寺や神社について場所以外にさらになにか説明している文章全般である．」これも「以外にさらになにか」の部分が困難である．すなわち，全体的に検索に用いるキーワードのみでは類推できないような細かい条件が存在し，これらのキーワードに関連する一般的なリンク集などが不適合ページと判定される傾向にあり，このことが本システムによる適合の判定を難しくしている．

5　おわりに

本稿では，検索キーワード間の意味的関係を利用して適合ページをフィルタリングする手法について議論を行った．

意味的関係を表す構造には，文，表，見出しなど，様々な構造がありうるが，本稿では，表構造に焦点を当て，その構造の中に現れた意味的関係を確認する手法を提案した．さらにそれを評価するために，既存の検索エンジンのフィルタリングツールを構築し，NTCIR-3WEBの評価用データセットを用いて実験を行い，システムの評価を行った．

評価においては，我々のシステムが元にした検索エンジンの精度を，MAPにおいて0.065向上させ，有意差があることを示した．

今後は4.2節で示した本戦略で不十分である点を見直し，更なる精度及び再現率の向上を目指したいと考えている．

参考文献

1) 風間 一洋, 原田 昌紀: Webサーチエンジン技術の高度化, 人工知能学会誌, Vol.16, No.4, pp.503-508 (2001).

2) 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏: 文構造における検索キーワード間の修飾ー被修飾関係に基づくWWW検索精度の向上, 電子情報通信学会技術研究報告, NLC2005-114〜125, pp. 7–12 (2006).

3) 西口直樹, 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏: 見出しの階層関係を利用したWWW検索精度の改善, 電子情報通信学会技術研究報告, NLC2005-114〜125, pp. 1–6 (2006).

4) 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).

5) 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).

6) Y. Wang and J. Hu.: Detecting tables in HTML documents. In LNCS, volume 2423, pp. 249–260. Springer-Verlag, (2002).