ロボット型サーチエンジン用ランキング手法の改善

佐々木 亮† 児玉 英一郎† 宮崎 正俊†

抄録 ロボット型サーチエンジンでの Web ページ検索のさいに行うランキングにおける問題点としてトピックドリフト問題が挙げられる．そしてトピックドリフト問題を表層的トピックドリフト問題と深層的トピックドリフト問題の 2 つに独自に分類し，それぞれについて説明する．このうち，表層的トピックドリフト問題の解決に当たり名詞孤立度を提案する．また，名詞孤立度を用いたランキング手法で，TF・IDF 法をベースとして用いた NIF・IDF 法について説明し，その効果を証明するために行った評価実験について述べる．評価実験において NIF・IDF 法は表層的トピックドリフト問題の軽減に有効であることが証明できた．

An Improvement of the Ranking Technique for Robot Type Search Engines

Ryo Sasaki† Eiichiro Kodama† Masatoshi Miyazaki†

Abstract A topic drift problem is in the problem of the ranking of a robot type search engine. The topic drift problem was classified original with a surface-thing and what depths-thing. Among these, the degree of noun isolation was proposed as the solution technique of a surface-topic drift problem, and NIF・IDF using it was proposed. In the evaluation experiment, it was proved that the NIF・IDF method has an effect in a surface-topic drift problem.

1. はじめに

近年，インターネットの普及によってその Web の利用者数が年々増加している．また，それに伴いインターネット上で閲覧できる Web ページの増加も著しく，現在では 30 億以上 [1] にもなっている．実際，2003 年 1 月 15 日現在で Google[1] には 3,083,324,652 もの Web ページが登録されている．このような状況のもと，ユーザは自分の必要とする Web ページを発見することが非常に困難となっている．このため，ユーザは，通常，サーチエンジンを使用し Web ページの検索を行っている．この様子を**図 1** に示す．

サーチエンジンには大きく分けてディレクトリ型サーチエンジン（Yahoo!など）とロボット型サーチエンジン（Google など）の 2 種類 [2] があるが，ディレクトリ型サーチエンジンは全て人手で管理されているため人手不足などの理由によって管理を続けていくことが限界と言われており，また，ロボット型サーチエンジンでは検索結果の提示においてランキングを行ってはいるが，不要な Web ページ（ノイズ）が混じることによって検索の精度が落ちてしまうなど問題も指摘されている．**表 1** にこの 2 種類のサーチエンジンの特徴を示す．

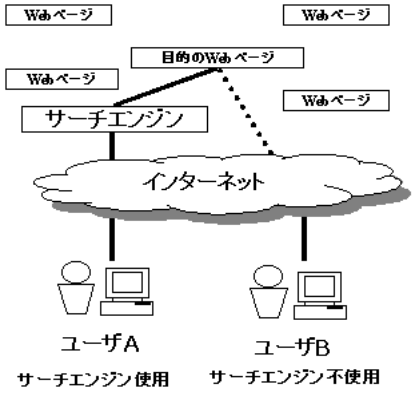

図 1: Web ページとサーチエンジン

表 1: 現在のサーチエンジンの特徴

	ロボット型サーチエンジン	ディレクトリ型サーチエンジン
Web ページの収集方法	ソフトウェアによる収集	人手による収集とユーザからの推薦
Web ページの分類	無し	人手によりカテゴリごとに分類
Web ページの検索方法	ソフトウェアによる検索	ディレクトリの階層をたどって検索
利点	登録されている Web ページの量が多い	質の良い Web ページが多い
問題点	検索結果の提示において不要な Web ページが混在してしまう	所有している Web ページ量が少ない

本研究では，今後も Web ページが増大 [3] していくことが予想されるため，ディレクトリ型サーチエンジン的アプローチによって，サーチエンジンを改良していく方法は難しいと考えた．そこで，前述のロボット型サーチエンジンの現状の問題点を解決する方向

†岩手県立大学ソフトウェア情報学部
Software and Infomation Science,
Iwate Prefectural University

で，より良いサーチエンジンの構築を目指す．このため，名詞孤立度という本研究独特の概念を用い，ランキングを行った際のノイズの軽減を目標とし，ランキング手法の改良によりランキングの精度を向上させることを目的とした．

本研究では，この目的を達成するにあたり，以下のことを行う．

1. 現状のロボット型サーチエンジンの分析
2. 現状のランキング手法の分析
3. 現状のロボット型サーチエンジン及びランキング手法の問題点の明確化
4. 上述の問題を解決するためのランキング手法の提案
5. 本論文で提案するランキング手法の評価実験

2. 現在のロボット型サーチエンジンとそのランキング手法

2.1. ロボット型サーチエンジン

ロボット型サーチエンジンではWebページ上に設けられた検索のためのユーザインタフェースにて，ユーザが任意の検索語を入力することによってWebページを検索する．検索語を受けたロボット型サーチエンジンはランキングを行った結果をもとにユーザへ検索結果を提示する．

ロボット型サーチエンジンは**図 2** のように構成される．

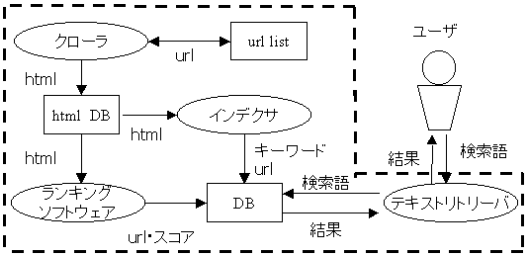

図 2: ロボット型サーチエンジンの構成とデータの流れ

また，ロボット型サーチエンジンを構成するそれぞれのソフトウェアの機能を以下に示す．

- クローラ

  Web 上に存在する Web ページを自動で巡回し収集するソフトウェア．

- インデクサ

  クローラで収集した Web ページに単語の切りだしなどの特徴づけを行うソフトウェア．

- ランキングソフトウェア

  インデクサにより切り出された単語ごとにスコアを計算するソフトウェア．

- テキストリトリーバ

  ユーザが検索語を与えることにより，その検索語に対応する Web ページを提示するソフトウェア．

2.2. 一般的なランキング手法

現在稼動しているロボット型サーチエンジンで用いられているランキング手法の 1 部について説明する．ランキング手法は，検索結果の上位に有用な Web ページを持ってくることで，ユーザが上位から見ていった場合に早く目的の Web ページに到達することを意図して導入されたものであり，現在のところ PageRank や TF・IDF 法など様々な手法が知られている．このうち PageRank 法と TF・IDF 法について順に説明する．

2.2.1. PageRank

PageRank 法 [4][5][6] は Google で用いられているランキング手法であり，Page 氏と Brin 氏によって提案された．この手法は多くの良質な Web ページからリンクされている Web ページは，良質な情報源であるとみなしてランキングを行う手法である．つまり，非リンク数が多い Web ページは PageRank が高くなり，また，非リンク数が多いページからリンクがあればその Web ページの PageRank が高いという相互関係で成り立っている．**図 3** に，この PageRank の概要 [7] を示す．

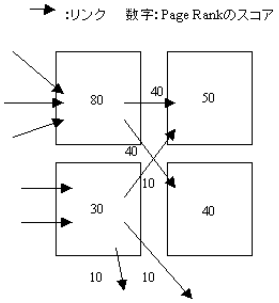

図 3: PageRank

2.2.2. TF・IDF 法

TF・IDF 法 (Term Frequency・Inverse Document Frequency)[8][9] は InfoSeek，goo などで用いられているランキング手法である [10]．TF 法は単語の重要度の評価の手法であり，同一文書で繰り返し出現する単語を重要視する．また，IDF 法は全文書中に，ある単語が含まれる文書が少ないほどその単語は文書の絞込みに役立つという考えに基づいた手法である．この TF 法と IDF 法を組み合わせたものが TF・IDF 法である．以下に検索語を t，検索対象の文書を d とした場合の計算方法を示す．

TF法　$TF(d,t) = \frac{\text{文書}d\text{中の検索語}t\text{の出現回数}}{\text{文書}d\text{中の全語数}}$

IDF法　$IDF(t) = \log(\text{全文書数} / \text{検索語}t\text{が含まれる文書数})$

TF・IDF法　$TF \cdot IDF(d,t) = TF(d,t) \times IDF(t)$

図 4: TF・IDF 法の定義

2.2.3. トピックドリフト問題

前述のように，ランキング手法はもともと検索結果の上位に有用な Web ページを持ってくることで，ユーザが上位から見ていった場合に早く目的の Web ページに到達することを意図して導入されたものであるが，ユーザが与える検索語によっては，検索結果の上位に検索意図とは関係の無い結果が混じってしまうことがある．これはトピックドリフト問題と呼ばれている．

現在，このトピックドリフト問題解決のために様々な試み [9][11][12][13] が行われているが，現状完全に解決された状態には至っていない．実際，現在運用されている商用ロボット型サーチエンジンで調査を行った結果，様々なトピックドリフト問題の事例が観測された．これを以下に独自の視点により分類して示す．

2.2.4. 表層的トピックドリフト問題

ユーザが与えた検索語が短い場合，Web ページの長い単語の 1 部としてヒットし，検索意図とは異なる結果が検索結果の上位としてランキングされる．

この表層的トピックドリフト問題の例としては，動物のリスを検索しようとしたときに，クリスマス，キリストなどが上位にくることがあげられる．

2.2.5. 深層的トピックドリフト問題

この深層的トピックドリフト問題は，表層的トピックドリフト問題と違い，ユーザの検索意図を特定できない限りは解決できない．これは一つの検索語でも複数の意味をもつ場合があること，単一の意味でも関連項目が複数あることによるものである．以下，この深層的トピックドリフト問題を分類して示す．

2.2.5.1. 多義性による場合　ユーザが与えた検索語に同音異義語にある場合，検索意図とは異なる意味でヒットした Web ページが上位にランキングされる．

例えば，テーブルで検索したときに家具のテーブルと表のテーブルが混在することがあげられる．

2.2.5.2. 漠然性による場合　ユーザが与えた検索語が漠然としていて検索意図を忠実に再現しきれていない場合，検索意図とは異なる Web ページが検索結果の上位としてランキングされる．

例えば，スピーカーの形状を知りたくてスピーカーと検索すると，価格や作り方などが混在することがあげられる．

2.3. 研究目的

本研究では，前述のトピックドリフト問題のうち表層的トピックドリフト問題をターゲットとし，TF・IDF 法をベースにこれを改善するための NIF(Noun Isolation Frequency)・IDF 法を提案する．また，実験により，本提案手法の有効性の評価を行う．

3. 名詞孤立度を用いたランキング手法の提案

本研究の対象となる表層的トピックドリフト問題を軽減するための独自の NIF・IDF 法の提案を行う．また、NIF・IDF 法で用いる名詞孤立度を求める方法について説明する．

3.1. 名詞孤立度

現在 TF・IDF 法では単純に html 文書から切り出された単語ごとにスコアを計算している．そこで，単語の前後に対象の単語とは別の品詞の単語を見つけ名詞孤立度を求める．名詞孤立度とは Web ページから切り出された単語が長い単語の 1 部であるかどうかを示す値である．例えば単語が「リス」であった場合，単語が含まれる Web ページでその単語の前後のつながりを見たときに「〜のリスは〜」のようになっている状態が孤立している状態であり,「〜からクリスマスが〜」のような場合は単語の一部に含まれているが，単語の前後に「ク」,「マス」とついているので，これは孤立していない．

以上を踏まえ，名詞孤立度 I の計算方法を以下のように提案する．

1. html 文章のタグを取り除き 1 次元化し，検索語の先頭の文字と最後尾の文字の位置をそれぞれカウントする．
2. 検索語の前にある別の品詞の最後尾の文字の位置と，検索語の後ろにある別の品詞の先頭の文字を同様にカウントする．
3. ここでカウントした文字の位置を先頭からそれぞれ a，b，c，d とする．
4. a-b 間の距離と c-d 間の距離を加算する．
5. 4 を相加平均する．

この名詞孤立度 I の計算例を，検索語「ブリ」，検索意図「魚」で検索したときに得られた html 文書を用いて示す．

- 検索語「ブリ」，検索意図「魚」で検索したときに得られた html 文書の 1 つを **図 5** に示す．
- **図 5** の html 文書のタグを取り除き，**図 6** に示すように 1 次元化する．
- このとき，a，b，c，d は **図 7** のようになる．

⟨html⟩ ⟨title⟩Britney Spears Home⟨/title⟩
#中略
⟨body⟩
⟨b⟩Gift from BRITNEY ⟨br⟩
〜ブリトニーからギフトが届きました ⟨/b⟩⟨/font⟩⟨br⟩
#以下略

図 5: 取得した html 文書

Britney Spears Home Gift from BRITNEY 〜ブリトニーからギフトが届きました

図 6: タグを取り除き 1 次元化した文書

Britney Spears Home Gift from BRITNEY 〜 (a) ブ (b) リ (c) トニー か (d) らギフトが届きました

図 7: a，b，c，d の位置

- 以上から，a，b，c，d のそれぞれの値は a=39，b=40，c=41，d=45 となるので名詞孤立度 I は以下のように求められる．

$$I = (|39 - 40| + |40 - 45|)/2 = 3 \quad (1)$$

3.2. NIF・IDF 法

NIF・IDF 法は第 2 章でも述べたように TF・IDF 法を改良した手法である．前述の名詞孤立度 I を用い，TF 法で求めた TF 値を除算し，NIF 値を求める．

以下にこの NIF の定義を示す．

3.2.1. NIF の定義

$$NIF = \frac{TF}{I} \quad (2)$$

例えば，ある Web ページにおいて事前に求められている TF 値が 50 で名詞孤立度が 10 のような場合はつぎのようになる．

$$NIF = \frac{50}{10} = 5 \quad (3)$$

さらに，TF・IDF 法と同様に IDF 値と NIF 値を乗算し NIF・IDF 値を求める．

以下にこの NIF・IDF の定義を示す．

3.2.2. NIF・IDF の定義

$$NIF \cdot IDF = NIF \times IDF \quad (4)$$

この場合，IDF 値は TF・IDF 値のものと同様になる．NIF 値が 5，IDF 値が 2.5 の場合，最終的にスコアはつぎのようになる．

$$NIF \cdot IDF = 5 \times 2.5 = 12.5 \quad (5)$$

ここで求めた NIF・IDF を新たなスコアとし，ランキングを行うことによって表層的トピックドリフト問題を改善する．

4. 評価

NIF・IDF 法を用いた評価実験について報告する．また，実験結果をもとにした考察についても述べる．

4.1. 評価実験の方法

本評価実験では NIF・IDF 法のもととなった TF・IDF 法と本 NIF・IDF 法とのランキング結果を比較する．このため，商用のロボット型検索エンジンで，ランキングに TF・IDF 法を使用している InfoSeek を用いた．

検索を行った際，検索結果上位 50 件について NIF・IDF 法を用いてスコアを再計算し，スコアの高い順にソートして新たな順位を与える．つぎに，TF・IDF 法と NIF・IDF 法の両方について上位 20 件での正解の Web ページ数とその割合を求め比較する．この評価実験は 5 つの検索語について行った．

実験の際に使用した検索の条件と検索語の選定基準はつぎの通りである．

- 日本語表記（片仮名，平仮名）である．
- 表層的トピックドリフト問題が起こっている．
- 単語 1 語で検索を行う．(検索に用いられる平均語数は 1.4 語であることが報告されている [2]．)

実験に使用した検索語とそれぞれの表層的トピックドリフト問題を**表 2** に示す．

4.2. 実験環境

本評価実験を行うために実験環境を構築した．**図 8** は実験に用いた環境とデータの流れである．

つぎに実験環境を構成する各ソフトウェアについて順に説明する．

- html 収集ソフトウェア（html_getter.java）

表 2: 実験に用いた検索語

検索語	検索意図	表層的トピックドリフト問題
ブリ	魚	「ブリトニー」など
ビス	ねじ	「サービス」など
サイダー	飲料	「インサイダー」など
フロ	風呂	「フローズンアイス」など
ふろ	風呂	「ふろしき」など

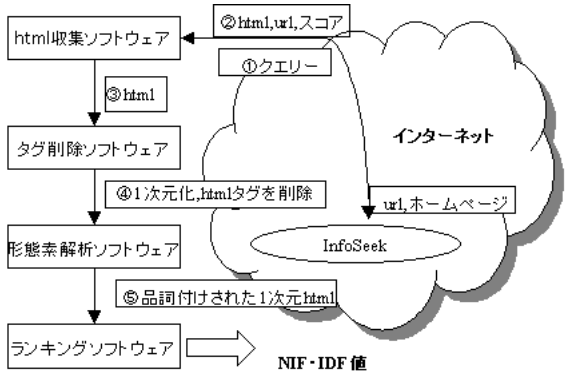

図 8: 実験環境の構成と動作

ロボット型サーチエンジンで TF・IDF 法を用いている InfoSeek に検索語のクエリー (質問) を渡し，検索結果の順位とそれに対応する html 文書，スコアを保存する．本ソフトウェアは JAVA を使用して作成した．

- タグ削除ソフトウェア（tag_del.pl）

NIF 値を計算する際に不要となる html 文書中の html タグの削除と html 文書の 1 次元化を行う．本ソフトウェアは perl を使用して作成した．

- 形態素解析ソフトウェア

既存の形態素解析ソフトである茶筅 [14] を用いた．この際，茶筅のオプションで出力するファイルには単語ごとの品詞のみを付与した．

- ランキングソフトウェア（calc.pl）

もととなる収集された html 文書から NIF 値を用い，スコアを再計算する．また，求められたスコアの高い順に結果をソートし表示する．本ソフトウェアは perl を使用して作成した．

4.3. 実験結果

前述の実験環境と検索語を用いて評価実験を行った結果について説明する．**表 3** に実験結果をまとめた．

表 3 において，可能な最高正解数とは表層的トピックドリフト問題が完全に解決された状態で得られる結果となる．各検索語においての詳細な結果は次のようになった．また，**図 9** は**表 3** の TF・IDF 法での正解率と NIF・IDF 法での正解率の変化をグラフ化

表 3: 実験結果

検索語	TF・IDF 法の正解数	NIF・IDF 法の正解数	可能な最高正解数
1:ブリ	12/20(60%)	16/20(80%)	20/20(100%)
2:ビス	0/20(0%)	3/20(15%)	3/20(15%)
3:サイダー	11/20(55%)	14/20(70%)	20/20(100%)
4:フロ	3/20(15%)	4/20(20%)	11/20(55%)
5:ふろ	6/20(30%)	13/20(65%)	20/20(100%)

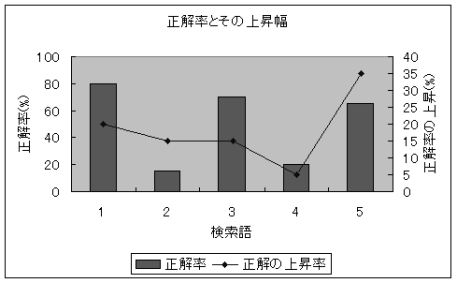

図 9: 正解率の変化

したものである．**図 9** において左側のグラフが TF・IDF 法での正解率，右側のグラフが NIF・IDF 方での正解率，折れ線がそれぞれの正解率の上昇を示している．また，検索語 1〜5 はそれぞれ**表 3** の 1〜5 に対応している．

4.3.1. 検索語「ブリ」，検索意図「魚」の場合

検索語「ブリ」，検索意図「魚」で検索した結果，上位 50 件中の正解の Web ページは 33 件であり，これは全体の 66%であった．このうちトピックドリフト問題を起こしていた単語は「ブリトニー」であった．

上位 20 件についての正解の Web ページ数は TF・IDF 法では 12 件であり，全正解数の 36.7%であった．NIF・IDF 法では上位 20 件中 16 件が正解の Web ページであり，全正解数の 48.5%となった．**表 3** から見ると正解率が 20%も上昇している．

4.3.2. 検索語「ビス」，検索意図「ねじ」の場合

検索語「ビス」，検索意図「ねじ」で検索した結果，上位 50 件中の正解の Web ページは 3 件であり，全体の 6%であった．このうちトピックドリフト問題を起こしていた単語は「サービス」,「ビステオン」,「ビスフェノール」,「テルビス」であった．

上位 20 件についての正解の Web ページ数は TF・IDF 法では 0 件であり，全正解数の 0%であった．NIF・IDF 法では上位 20 件中 3 件が正解の Web ページであり，全正解数の 100%となった．**表 3** から見ると正解率が 15%も上昇している．

4.3.3. 検索語「サイダー」，検索意図「飲料」の場合

検索語「サイダー」，検索意図「飲料」で検索した結果，上位 50 件中の正解の Web ページは 28 件であ

り，全体の 56%であった．このうちトピックドリフト問題を起こしていた単語は「インサイダー」,「アウトサイダー」,「サイダーハウス」であった．

上位 20 件についての正解の Web ページ数は TF・IDF 法で 11 件で全正解数の 39.3%であった．NIF・IDF 法では上位 20 件中 14 件が正解の Web ページであり，全正解数の 50%となった．**表 3** から見ると正解率が 15%も上昇していることがわかる．

4.3.4. 検索語「フロ」，検索意図「風呂」の場合

検索語「フロ」，検索意図「風呂」で検索した結果，上位 50 件中の正解の Web ページは 11 件であり，全体の 22%であった．このうちトピックドリフト問題を起こしていた単語は「フローター」,「フローズン」,「フロータント」,「フローラ」であった．

上位 20 件についての正解の Web ページ数は TF・IDF 法では 3 件で全正解数の 27.3%であった．NIF・IDF 法では上位 20 件中 4 件であり，正解の Web ページで全正解数の 36.7%となった．**表 3** から見ると正解率が 5%上昇していることがわかる．

4.3.5. 検索語「ふろ」，検索意図「風呂」の場合

検索語「ふろ」，検索意図「風呂」で検索した結果，上位 50 件中の正解の Web ページは 21 件であり，全体の 42%あった．このうちトピックドリフト問題を起こしていた単語は「ふろしき」,「しふろ」,「ふろむ」,「えむふろ」,「あふろ」,「ふろんてぃあ」であった．

上位 20 件についての正解の Web ページ数は TF・IDF 法では 6 件であり，全正解数の 28.6%であった．NIF・IDF 法では上位 20 件中 13 件が正解の Web ページであり，全正解数の 61.9%となった．**表 3** から見ると正解率が 35%も上昇していることがわかる．

4.4. 考察

実験結果から NIF・IDF 法での上位 20 中の正解の Web ページ数は増加していることがわかる．また，検索語「ビス」においては表層的トピックドリフト問題を完全に解決した状態となった．以上のことから NIF・IDF 法は表層的トピックドリフト問題の軽減に効果があると証明された．

5. おわりに

ロボット型サーチエンジンでのランキングおける問題点であるトピックドリフト問題を独自に表層的トピックドリフト問題と深層的トピックドリフト問題の 2 つに分類した．このうち，表層的トピックドリフト問題の解決，軽減を研究目的とし，その解決手法案として名詞孤立度を提案し，名詞孤立度を用いたランキング手法で，TF・IDF 法をベースとした NIF・IDF 法を提案した．また，提案した NIF・IDF 法の表層的トピックドリフト問題に対する効果を証明するために評価実験を行った結果，本手法が表層的トピックドリフト問題軽減に有効であることが証明された．

今後の課題として，本ランキングアルゴリズムの精度の向上と，深層的トピックドリフト問題に対するランキングアルゴリズムの考察を行いたい．

謝辞 本研究を行う機会を与えていただき，ご指導賜りました岩手県立大学ソフトウェア情報学部 宮崎正俊教授に心から感謝致します．また，本研究にあたり，日頃から熱心にご指導してくださいました岩手県立大学ソフトウェア情報学部 児玉英一郎助手に深く感謝致します．さらに，本研究を進めるために協力してくださった宮崎研究室の皆様，火曜日に行われていた院生ゼミの参加者の皆様，その他多くの方々に感謝致します．

参考文献

[1] URL http://www.google.co.jp
[2] 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, pp.503-508 (2001).
[3] URL http://www.kake.info.waseda.ac.jp/ matsuda/last/node6.html
[4] 山名早人，近藤秀和:サーチエンジン Google，情報処理，Vol.42, pp.775-780 (2001).
[5] URL http://www.rankwrite.com
[6] URL http://www.supportforums.org/pagerank
[7] URL http://www.kusastro.kyoto-u.ac.jp/ baba/wais/pagerank.html
[8] URL http://www.internetconference.org/ic99/papers/slide-s05p01.pdf
[9] 風間一洋，原田昌紀，佐藤進也:ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法，情報処理学会研究報告，2000-FI-59, pp.17-20 (2000).
[10] URL http://web.kyoto-inet.or.jp/org/kyo-yoh/kensaku/kensaku.html
[11] URL http://www.htmlhelp.com/ja/reference/html40/head/meta.html
[12] 大森貴博，笹塚清二，水谷正大:リンク情報を考慮した Web 検索システム，情報処理学会研究報告，Vol.99, pp.49-56 (1999)
[13] 黄林春, 林幸雄:リンク構造解析によるページの価値計算とネットワーク分析，電子情報通信学会技術研究報告，Vol.100, pp.13-18 (2000)
[14] URL http://chasen.aist-nara.ac.jp/index.html.ja