

ユーザ履歴を活用した検索システム

川前徳章, 青木輝勝, 安田浩

東京大学先端科学技術研究センター

〒153-8904 東京都目黒区駒場4-6-1 45号館 5F

TEL: 03-5452-5277 FAX: 03-5452-5278

あらまし キーワード入力による検索システムを用いて、ユーザが目的に合った文書を検索する時に、適切なキーワードを入力する事が求められるが、多くのユーザにとってこの作業は困難を伴う。本稿ではユーザの検索を支援するキーワードディレクトリの作成とその効果について述べる。キーワードディレクトリはディレクトリ形式で、ユーザが検索時に入力したキーワードに関連するキーワードを表示する。その作成にシステム側に残されたログを利用する。その結果、キーワードディレクトリにより、ユーザは容易に検索に適切なキーワードを変更、追加することが可能となり、検索の効率化が実現される。

キーワード 情報検索 情報共有 ユーザ支援 ユーザ履歴 検索効率 キーワードディレクトリ

The search system of utilizing users' search history

Noriaki Kawamae, Terumasa Aoki, Hiroshi Yasuda

Center for Advanced Research and Technology
The University of Tokyo

4-6-1 Komaba Meguroku, Tokyo 153-8904, JAPAN

TEL:+81(3)5452 5275/090 4813 2036

FAX:+81(3)5452 5278

Abstract

On using search system with inputting keyword, it needs users to input proper keyword for search documents. But it troubles many of them. In our paper, we describe how to make keyword directory that helps users' search activity and how much it effects. Keyword directory displays the keywords related keywords inputted by users in directory. To struck it, we uses logs that remain in the search system Therefore, using by keyword directory, users can select efficiently the proper keywords for search to change and add.

key words

Information Retrieval, Information Sharing, User Support, User History, Retrieval Efficiency, Keyword Directory

1 はじめに

www 情報空間が日々拡大を続け、それに比例して検索システムへのニーズも大きくなっている。しかし、情報量が増大するのに伴い、従来の検索システムではユーザの検索活動の効率化を支援することが難しくなっている。現状のシステムは、ユーザがキーワードを入力し、システムがあらかじめ www 空間内における www ページについてインデックスを作成し、この情報に基づいて検索を実行し、www ページの URL のリストを検索結果の一覧としてユーザに表示している。検索に適切とされるキーワードをユーザが入力するのは専門領域外の情報の検索や、ユーザがその検索のニーズを具体化していない場合には困難を伴う。

この解決策として従来、孤立している個々の検索活動を、検索システムに蓄積されたログを介することで、ユーザ間の検索要求、戦略の情報を共有することにより、個々の検索活動の支援を行なうことが考えられる。ログにはユーザを識別する手がかりとしての IP アドレス、検索に用いたキーワード、そしてそのキーワードで検索結果から辿った www ページの URL(Universal Resource Location)が含まれている。本稿ではこのログに着目し、ログを検索シーケンス、URL・キーワードリストというデータ単位に分離し、それを活用してキーワードディレクトリを作成する検索支援手法を提案する。本研究で実現したキーワードディレクトリは、ユーザが入力したキーワードを起点とした方向性のあるキーワードのディレクトリである。従ってディレクトリを辿る事によりユーザは再検索する際に、適切なキーワードへの変更、利用するキーワードの追加などを容易に行なえるようになる。

本稿では、第二章で従来のユーザのキーワードによる検索支援手法の関連研究、第三章で我々が提案する有向性のあるキーワードディレクトリについて解説し、第四章ではログを用いたキーワードディレクトリ作成のプロセスについて検討し、第五章でその結果と有効性について論じる。

2 キーワード提示によるユーザの検索支援

2.1 ユーザの検索行動

ユーザの検索行動は次のステップを踏む。

- (1)キーワードの入力
- (2)システムの返す検索結果の閲覧
- (3)キーワードの変更

この検索行動で、ユーザはキーワードの入力、検索結果の評価が求められるが、(1)に関しては、必要なキーワードが思いつかないなどの問題があり、(2)に関しては www ページの URL のリストなので、それらの中でニーズにあったページを発見するのに時間がかかるという問題がある。そこでこれらの問題を解決するのに従来を検索システムには次のユーザ支援が求められる[1]。

- (1)キーワード入力の支援
- (2)検索結果を構造化
- (3)検索結果のランキング

本稿ではこの中で(1)による検索の支援を目指す。

2.2 検索の効率化

www 空間における情報の増加に比例して、検索結果の数も増加している。従って適切な検索キーワードを入力しなければ、大量の検索結果のリストからユーザは必要とする情報を検索するのが困難になる。例えば、Fresheye[2]においてカレーと入力した場合、検索結果として表示される URL の数は 8324 件であり、内容どころかそのリストを全て眺めるのもユーザにとっては困難である。また、人によっては同じ検索要求を持ちながらカレーと入力する場合もある。すると 500 件となるが、現在のパターンマッチングの技術では表記のゆらぎに対応出来ず、カレーとは異なった検索結果が出てくる。その中にはオカリーナという、そのユーザには不必要な情報も検索結果として出てくることになる。

我々は検索の効率化を“検索活動における情報要求のステップ数を従来のシステムより少なくす

る”と定義し、www 空間の検索システムの評価として再現率と適合率に代わって”検索コスト“を定義する。検索コストはユーザが必要な文書の有無を判断するのに必要な文書の総閲覧数である。

2.3 キーワード提示による検索支援

ユーザの検索支援の既存研究に www ページから形態素解析によって切り出されたキーワードの共起集合から関連するキーワードを提示する手法として Mondou[3]があるが、日本語は形態素解析が困難で、辞書の品質によってユーザに役立つキーワードとして不適切な場合や計算コストの面から適用の限界がある。文書集合で特徴的に出現するキーワード群の関連性をグラフ上に示した[5]や、ユーザが利用したキーワードの共起（組み合わせ）に注目し、関連の強いキーワードをグラフ化する手法[6]があるが、無向グラフによりキーワード提示を行なっている。

キーワード提示は、関連するキーワードを無向グラフで提示するよりも、方向性を持たせてユーザに示すことでより高い効果が発揮できる。

3 ログから作成されるキーワードディレクトリ

3.1 ユーザの検索行動とキーワードディレクトリ

ユーザが検索を続けて行く上で、入力するキーワードを変更する理由は次のようになる。

- (1)入力したキーワードに対しての検索結果を見て、次に入力するキーワードを具体化する。
- (2)入力したキーワードに対しての検索結果が少ないので、次に入力するキーワードを抽象化する。
- (3)入力するキーワードを追加、削除する。
- (4)興味の対象が変化する。

そこでユーザのキーワード入力を支援するキーワード提示には次の機能が求められる。

- (1)(2)(4)⇒前後の関連の強いキーワードを提示
- (3)⇒追加すべきキーワードの提示

我々は、この機能を満たす、キーワードディレクトリという方向性を持つキーワード提示手法を

提案する。ユーザが“CGI”と入力した場合に提示されるキーワードディレクトリを図1に示す。

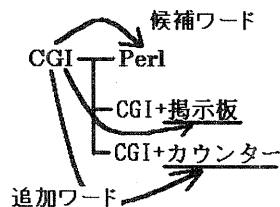


図1. キーワードディレクトリ

図1で“CGI”をルートワード、“perl”は候補ワード“掲示板”、“カウンター”は追加ワードと定義する。検索システムは検索結果として“CGI”を含む文書の URL リストと共に“CGI”を起点とするキーワードディレクトリを提示する。このディレクトリでユーザが検索するキーワードを“CGI”から“perl”へ変更すると、システムは直ちに“perl”を含む www ページの URL のリストと、“perl”をルートワードとするディレクトリを再帰的に生成する。本稿はこのディレクトリを作成する為に、ログの解析手法と、生成されたキーワードディレクトリについて検索効率の点で評価を行なった。

3.2 ログからのデータとキーワード・URL

サーバに蓄積されたユーザの検索ログを図2に示す。これを IP、KW、URL として分離する。IP はクライアント側の IP アドレス、KW はユーザが検索に用いたキーワード、URL はユーザがそのキーワードで出力された検索結果から辿った www ページの URL である。今回利用させていただいた商用サーチエンジンのログは一日分のもので 952666 ラインから構成されている。ログの利用により、ユーザの用いたキーワードを用いることになり、形態素解析によって抽出されたキーワードよりも使い易くなると期待出来る。

```
***.***.***.***(IP) - - [01/Sep/1999:00:00:05
+0900] "GET /?go=http://www.***.***.***.
(URL)"http://***.***.***?kw=***** (KW) "
```

図2: ログの形式

4 ログからの情報抽出とディレクトリ作成

ログからの情報抽出とディレクトリ作成の手順を図3に示す。

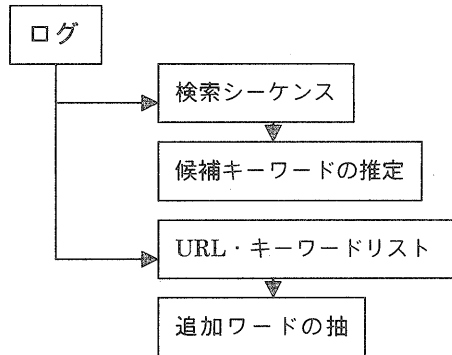


図3：ログ解析の流れ

4.1 検索シーケンスの作成

キーワードの変更パターンを抽出する目的でログから検索シーケンスリストを作成する。検索シーケンスはユーザをキーとし、そのユーザが用いたキーワードを時系列的に格納したハッシュであり、次のような構造を持つ。

$$kw\{u_i\}=\{kw_1, kw_2, \dots, kw_j\}$$

u_i = 識別子 i のシーケンス

kw_i = 時間 t に用いられたキーワード

ユーザ毎の識別は次の手順で行なった。

- 1) IP アドレス毎に分離
- 2) 同 IP アドレスでも時間間隔が 20 分以上ならば別シーケンスに分離

分離された検索シーケンスの総数は 159039 となった。

4.2 URL・キーワードリストの作成

各 URL のインデックスを形態素解析からユーザの用いたキーワードにする為に URL・キーワードリストを作成する。分類された URL 数は 123191

あった。ログにはキーワードとそのキーワードで辿った www ページの URL が記録されており、同一 URL を辿るのにユーザが用いたキーワードを、URL をキーとしたハッシュに、利用頻度、そのキーワードを含む www ページの総数と共に格納する。ハッシュは次のような構造を持つ。

$$from\{URL\}=\{kw_i, fre\{kw_i\}, hrefby\{kw_j\}, \dots\}$$

URL : 辿られた www ページの URL

kw_i : URL を辿るのに利用されたキーワード

$fre\{kw_j\}$: キーワード kw_j の利用頻度

$hrefby\{kw_j\}$: kw_j を含む URL の数 (検索結果数)

このリストにより各 URL は形態素解析によって分離されたインデックスから、ユーザの用いたキーワードとその利用頻度をインデックスとして持つ。

ログから検索シーケンス、URL・キーワードリストに分類され、検索シーケンスはユーザの興味・関心の特徴づけるキーワード群、URL・キーワードリストは URL のインデックス群 (各 URL の概念を文中のキーワードにより構成する語彙的連鎖 [4]) はユーザが利用したキーワードなので形態素解析によるものより使い易いものになっていると解釈出来る。

4.3 キーワードディレクトリの作成

4.3.1 候補キーワードの推定

候補キーワードの推定を行なう為に各検索シーケンス内のキーワードの時系列的变化に着目する。例として検索シーケンスにおいて次のようにキーワードが入っている検索シーケンスを示す。

$$kw\{u_i\}=\{\text{チーズ}, \text{パルメザンチーズ}\}$$

このシーケンスはユーザが検索キーワードをチーズからパルメザンチーズに変更したことを示している。候補キーワードを推定する為に、この検索シーケンスから次のようなハッシュを作成する。

from{チーズ}={パルメザンチーズ}

この from{チーズ}のハッシュにはチーズから変化したキーワードが格納され、今はパルメザンチーズが入っている。すべての検索シーケンスに同様の処理を行ない、次のようなハッシュを作成する。

from{チーズ}={kw_i,fre{kw_i},count{kw_i|チーズ},...}
 kw_i; 後続するキーワード
 fre{kw_i}; キーワード kw の使用頻度
 count{kw_i|チーズ}; チーズから kw_iへ
 変化した頻度

以後、ハッシュのキーkw_A (この場合はチーズ)を変化のスタートである事からスタートワード、そのハッシュ内のキーワードを後続ワード kw_i(i=1...n-1)と呼ぶ。スタートワードのまま検索を続行、あるいは中止する場合もあるので、キーワード kw_A の使用頻度 fre(kw_A)と、変換後のキーワードの頻度 count{kw_{Ai}|kw_A}には次の関係が成り立つ。

$$fre(kw_A) \geq \sum_{i=1}^{n-1} count(kw_i|kw_A)$$

そこで後続キーワードのリストに、その場合の後続キーワードとして kw_A を追加し、次の様に定義する。

$$count(kw_A|kw_A) = fre(kw_A) - \sum_{i=1}^{n-1} count(kw_i|kw_A)$$

kw_A は以降 kw_n として扱う。検索シーケンス内のキーワード変化に着目したハッシュを得ることで、各々のスタートキーに対して後続ワードに対する確率分布が得られる。その結果、kw_A をスタートキーとする場合のキーワードの変更の予測を行なうにあたって、ベイズの公式の適用が可能となる。本稿では適用に際し次のように確率分布を設定する。

$$P(kw_i | kw_A) = \frac{P(kw_A | kw_i)P(kw_i)}{\sum_{i=1}^n P(kw_A | kw_i)P(kw_i)} \quad (1)$$

$P(kw_i | kw_A)$ = ユーザがキーワード kw_A を利用した後に kw_i へ変更する確率

$P(kw_i)$ = キーワード kw_i を選択する確率

$P(kw_A | kw_i)$ = キーワード kw_i に変更する前にキーワード kw_A を利用したする確率

$$p(kw_i) = \frac{fre\{kw_i\}}{\sum_{i=1}^n fre(kw_i)} \quad p(kw_A|kw_i) = \frac{count(kw_i|kw_A)}{fre(kw_i)}$$

(1)はスタートワード kw_A に応じてその後続ワード kw_i の出現確率が影響を受け、その定量的関係が条件付き確率分布 $P(kw_i | kw_A)$ で定められる事を示している。kw_A のハッシュにあるキーワードの集合 kw_i はそれぞれ、出現確率 fre(kw_i)を持つ。例として最初に入力したキーワードを kw_A、その後にユーザが入力するキーワードを kw_i,kw_A (変化しない、あるいは検索の終了)とした場合を表 1 に示す。

表 1: キーワード遷移

kw _i	count(kw _i kw _A)	p(kw _i kw _A)	p(kw _i)	P(kw _i)/0.05
kw ₁	25	0.25	0.01	0.2
kw ₂	10	0.1	0.01	0.2
kw ₂	5	0.05	0.005	0.1
kw _A	60	0.60	0.025	0.5
Total	100	1.0	0.05	1.0

全体における変更後のキーワードの頻度 fre{kw_i} は、変換後のキーワード集合の頻度の和で正規化する。キーワード kw_A を知った場合の確率 (事後確率) $P(kw_i | kw_A)$ が、知る以前の確率 (事前確率) $P(kw_i)$ より大きければ、kw_i は kw_A の変更キーワードとして適当と判断する。表 1 の場合は kw_A というキーワードを知る事で、キーワードとして kw_i が事後確率が事前確率より大きくなっているため、kw_A の候補キーワードとして適当と判断し、kw_A をルートワードとしたディレクトリに候補ワードとして挿入する。

4.3.2 追加ワードの抽出

追加ワードを抽出する為に、URL・キーワード

リストの共起関係に着目する。共起とは異なる複数のキーワードが共に同一文書あるいは同一段落に現れることであり。この共起の相関を求めるのに丹羽[5]はキーワードの相関を、文書集合から選び出したキーワードの共起統計から算出し、グラフ上に表示した。その際にキーワード kw_1 から見

$$r(kw_1, kw_2) = \frac{|D(kw_1 \cap kw_2)|}{|D(kw_2)|}$$

たキーワード kw_2 の関連度を次の式で定義した。

$D(kw_1)$: kw_1 を含む文書の数

$D(kw_1 \cap kw_2)$: kw_1 と kw_2 を共に含む文書の数

佐藤[6]は

$$r(kw_1, kw_2) = \frac{|D(kw_1 \cap kw_2)|}{|D(kw_1 \cup kw_2)|}$$

$D(kw_1 \cap kw_2)$: kw_1 と kw_2 を共に含む文書の数

$D(kw_1 \cup kw_2)$: kw_1 または kw_2 を含む文書の数

この場合、キーワードの関連は kw_1 と kw_2 の関連は対照的になり、方向性のある関係が測定出来ない場合がある。例えば、図2の場合、 kw_1 に関して kw_2 の関連は強いが、 kw_2 に関して kw_1 はそうでもない判断出来る。丹羽[5]の関連度によれば $r(kw_1, kw_2) > r(kw_2, kw_1)$ である。

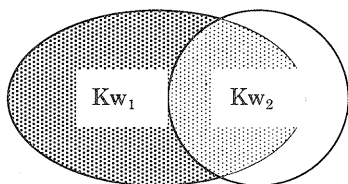


図4 : 共起の非対象

本研究では各 URL のインデックスの共起を求める。URL・キーワードリストは同一 www ページの URL を辿るのにユーザにより利用されたキーワードの集合であり、①ユーザが検索結果を評価する前に URL 内に共起するキーワードを知らせる事が出来る。②検索シーケンス内のキーワードの共

起と www ページ内の共起は異なるので、www ページ内にあるキーワードでなければディレクトリの有効性が発揮出来ないという理由で利用する。

$tf \cdot idf[7]$ はある文書 (D) に高い頻度で現れる一方、他の文書にはあまり現れないキーワード (t) の評価の指標である。tf (term frequency)、idf (inverse document frequency) は次の式で表される。

$$tf(D_i, t_i) = \frac{\text{キーワード } t_i \text{ の出現頻度}}{\text{文書 } D_i \text{ の形態素数}}$$

$$idf(t_i) = \log \frac{\text{データベースの全文書の数}}{\text{キーワード } t_i \text{ を含む文書集合の数}} + 1$$

上式の tf, idf の積 $tf \cdot idf$ がキーワード t_i の文書 D_i を識別する能力をあらわしている。

本稿は、ルートワード kw_A に対しての追加ワードを、 kw_A を含む URL 集合を識別する能力の高さにより kw_A と共起するキーワードの集合 $\{kw_{A1}, kw_{A2}, \dots, kw_{Ai}\}$ (URL・キーワードリスト) から抽出する為に tf, idf を変形した次の式を定義する。

$$tf(D_A, kw_{Ai}) = \frac{D_A \text{ において } kw_{Ai} \text{ を含む URL の数}}{kw_A \text{ を含む URL 集合 } D_A \text{ の URL の数}} \quad (2)$$

$$idf(kw_{Ai}) = \log \frac{\text{データベースの全 URL の数}}{kw_{Ai} \text{ を含む URL の数}} + 1 \quad (3)$$

(2), (3) を用いることで、キーワード kw_A と共起するキーワードに対して各々計算される (2), (3) の積はキーワード kw_A を含む URL の集合において共起するキーワードの頻度と判断でき、この値が高い順にスタートワード kw_A のディレクトリに追加ワードとして挿入する。

5 解析結果

5.1 作成されたキーワードディレクトリの表示

検索シーケンス、URL・キーワードリストからキーワードディレクトリが作成され、“釣り” というキーワードをルートワードにしたキーワードデ

ディレクトリを図 5 に示す。ルートワード以下のキーワードの上下関係が概念的に上下関係に対応している場合もあれば、この場合の様にキーワードの包含関係、つまりディレクトリの下位が上位のキーワードを含む場合があり、ユーザの検索活動が共有され、キーワード提示による検索支援に結びつくことが確認された。

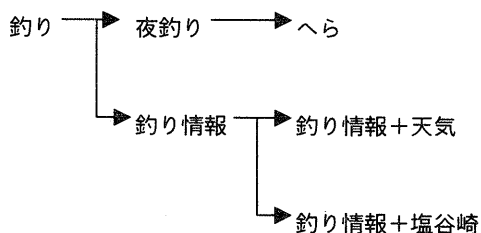


図 5：釣りをルートワードにしたディレクトリ

5.2 検索コスト

ユーザが、目的とする Web ページを発見する為に、出力結果の URL を閲覧する数を検索コストと定義する。入力したキーワード kw についてディレクトリが存在しない場合、検索コストを $cost\{kw\}$ として次のように定義する。

$$cost\{kw\} = fre\{kw\} \times hrefby\{kw\} \quad (4)$$

$fre\{kw\}$: kw の使用頻度

$hrefby\{kw\}$: kw を含む URL の数 (検索結果数)

キーワードディレクトリを利用した場合、ユーザの入力キーワードをルートワードとしたキーワードディレクトリが表示され、ユーザはディレクトリのキーワードを利用するか否かの選択が出来る。この時、検索コストはディレクトリに出現した全キーワードの検索コストの期待値とする。

例としてユーザが釣りと入力した場合、図 6 のような釣りをルートワードとするキーワードディレクトリが作成された。()内は (使用頻度, 検索結果数) である。ユーザは次の行動を選択出来る。

(1) “釣り” を含むキーワードの検索リストの閲覧

(2) “夜釣り” で再検索

(3) “釣り情報” で再検索

これらの行動を選択する確率はどれも確からしいものとする。キーワード “釣り” に関する検索コストは次の様に計算出来る。

$$cost\{釣り\} = \frac{1}{3} \times (62 \cdot 1280 + 21 \cdot 60 + 28 \cdot 537)$$

キーワードディレクトリが存在しない場合、ユーザの選択は(1)のみであり、その場合の検索コストは次のようになる。

$$cost\{釣り\} = 62 \cdot 1280$$

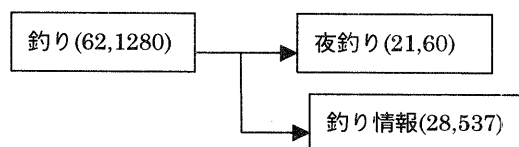


図 6：釣りをルートワードにしたディレクトリ

表 2：実験結果

①ルートワード	36667
②キーワードディレクトリがない場合の検索コスト	3524626
③キーワードディレクトリがある場合の検索コスト	113287
④ルートワード毎の検索コストの低下	93.0

表 2 は利用したデータから作成されたキーワードディレクトリに関して検索コストを計算したものである。キーワードディレクトリのルートワードの数が①であり、それぞれルートワードに関してキーワードディレクトリを用いず検索した場合の検索コストを (4) を用いて計算した合計が②であり、キーワードディレクトリを用いた合計が③である。④はキーワードディレクトリにより低下したコスト(② - ③)をルートワード数①で割ったもので、ルートワード一つあたりのキーワードディレクトリより低下したコストを示している。この結果、明らかにキーワードディレクトリがユーザの検索の効率化に寄与した事が確認出来た。

6 考察

6.1 キーワードディレクトリの効果

現在の検索システムにおいて、第一の問題が検索に必要とされるキーワードの想起である。とりわけユーザが新しい知識を得る為に検索を行なう場合、ユーザの知らない、あるいは思いつかないキーワードが対象となる www ページに含まれていることが多く、両者のキーワードの利用状況の相違が情報検索の効率化を妨げている。本研究で提案したキーワードディレクトリは、検索対象である www ページ内に含まれるキーワードを、他のユーザにより用いられたキーワードを候補あるいは追加ワードとしてキーワードを提示する検索支援手法を実現した。

6.2 ログの利用と検索効率

検索の効率化はキーワードディレクトリのみでは解決出来ない。今回はログを活用したが、そのログのうち、ユーザが入力したキーワードに対して検索結果から辿った www ページの URL は、検索結果として表示した URL のリストに大きく依存する。このリストにおいて、現在はランキングという尺度によって次元に並べられており、今後はこのランキングの改善が課題となる。このランキングで上位に現れない URL はユーザに閲覧されにくく、URL・キーワードリストにも登録され難く、ランキング自身が内容に関するランキングではないので、その下でユーザが辿ったログを蓄積、活用しているとキーワードディレクトリの追加ワードの精度にも影響が出てくると考えられる。今後の課題として、URL のリストを次の点で改善する事が求められる。

- 1) 内容の類似性に基づいた検索結果の構造化
- 2) ランキングの改善

これによりキーワードディレクトリの精度と同じに検索結果での検索の効率化を促進すると考えられる。

7 まとめ

本研究では、検索シーケンス内のキーワードの変化と、URL にユーザによって用いられたキーワードをインデックスとして用いることにより、方向性を持つキーワードディレクトリが作成されることを示した。従来は個々に独立していた検索活動をログを介して共有化することにより、ユーザの検索活動が共有化され、キーワードディレクトリがユーザが入力したキーワードに対し、他のユーザによって発見されたキーワード、あるいはそのキーワードを含む URL 集合内で高い頻度で出現するキーワードを提示する事で、従来は個々のユーザが行な必要があった検索活動のステップ数を削減し、さらにキーワードを絞り込む事によりユーザが閲覧すべき検索結果の URL のリストの数を減らす事で検索の効率化を実現した。

参考文献

- [1]川前徳章, 青木輝勝, 安田浩: "メタ検索システム構築の為のデータの構造化に関する研究"; 人工知能基礎論 & 第 45 回知識ベースシステム研究会 SIG-FAI/KBS-9902, pp.37-42, 1999.
- [2]"Fresheye" <http://www.fresheye.com/>
- [3]河野浩之, 長谷川利治: "WWW データ資源検索におけるデータマイニング手法", 情報処理学会研究報告 96 - DBS - 108, pp. 33 - 40, 1996
- [4] J.Morris, J. and G.Hirst: Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics*, Vol.17, No.1, pp.21-48, 1991
- [5]丹羽芳樹: "動的な共起解析を用いた対話的文書検索支援", 自然言語処理 96 - NL - 115, pp.99 - 106, 1996
- [6]佐藤進也, 原田昌紀, 風間一洋: "サーチエンジンへの問い合わせの解析", 情報処理学会研究会報告, 2000-FI-57-18, pp. 135-142, 2000.
- [7]Salton, G. Allan and Buckley, C.: Term - Weighting Approached in Automatic Text Retrieval, *Information Processing and Management*, Vol.14, No.5, pp.513 - 523, 1998