2P—03

サーチエンジンの検索結果の構造化

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

1 はじめに

www 空間上の情報が増加するのに伴い、この空間の中からユーザが必要な情報を検索するサーチエンジンへのニーズは高まっている。しかし、現在のサーチエンジンは検索効率の面で問題を抱えている。問題の一つに、ユーザに対し検索結果の確認に負担を強いていることがある。その原因は検索結果の URL のリストが次の状況にあるためと考えられる。。

①ページの内容は実際に確認するまで分からない
②サイトの構造とは無関係に出力される
③ユーザにとって必要なページがランキングの上位に出て来ない
④数が多すぎる

2 検索結果の構造化

本研究では検索結果の URL のリストを構造化することで、ユーザの検索結果の確認を支援し、検索における効率化を目指す。既存の研究では、検索結果の構造化に関してはカテゴリ分類、クラスタリング、サイト毎に分類する手法がある。カテゴリ分類[1]は、検索結果の URL がそれの属するカテゴリに分類されて表示されることになるが、あらかじめ分類基準を与えておく必要がある。クラスタリングには、LSI(Latent Semantic Indexing)[2]がある。LSI は文書を構成するタームをベクトルとした多次元空間の次元を減少し、いくつかのタームを合成した新たなベクトルの空間に置き換える手法である。タームの数が多い時には、システムにかかる負荷、形成されるクラスタ数に関して問題がある。サイト毎では、同じサーバ上にある web ページをディレクトリ構造とリンク関係を反映した[3]がある。検索結果の数やサイト自身が巨大であれば、構造化された結果が有効に活用されない問題がある。本研究ではサーチエンジンに残されたユーザの検索記録であるログを活用して、検索結果の構造化を実現する手法を提案する。

3 ログを活用した検索結果の構造化

本研究では検索結果を次の二段階の構造化を行なうことで検索の効率化を実現する。

①カテゴリに分類
②同カテゴリ内で類似したページをクラスタリング

3.1 ログの解析

カテゴリ、クラスタリングに使うタームはサーチエンジンに残されたログから抽出する。ログは次のような形式を持つ。

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

このログから分かることは、あるユーザ(IP)がいつ、どんなキーワード（KW）を用いて、どのページ(URL)にジャンプしたかということである。ログから各 URL 毎にユーザにより利用されたキーワードを集め、次のリストを作成する。

3.2 カテゴリタームの抽出

分類の基準となるカテゴリのタームはユーザが用いたキーワードである。検索結果のカテゴリとしてのタームの選択基準を次のように定める。

①ターム間の相関が低いものとする。
②ユーザによく利用されるターム

各タームの相関は次のように測定する。apple というターム（メインターム）と共起するタームのリストが次のように得られた。

co{apple}＝{Mac,果物,フルーツ,本,店,peach…}

メインタームの検索結果を出力する時、カテゴリのタームはこのリストから抽出することになる。それぞれのタームをメインタームとして共起するタームのリス

トは次のようになる。

co{Mac}={G4,モニター,ワイヤレス,Cube,OS X,…}
co{果物}={野菜,季節,スイカ,甘い,…}
co{フルーツ}={デザート,健康,甘い,…}

これらタームの相関を、それぞれ共起するターム t_i の頻度(fre{t_i})で数値化し、それぞれタームの利用頻度を乗じたベクトルとして積率相関係数を用いて次のように測定する。

$$r_{xy} = \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_t (x_t - \bar{x})^2 \sum_t (y_t - \bar{y})^2}} \quad (1)$$

x, y：比較するメインタームのベクトル
t：メインターム x, y と共起するターム

共通するタームが多いと相関は高くなる。メインタームと共起する全てのターム間の相関を求め、次の式が最大となるタームをカテゴリタームとして抽出する。

$$r_{term} = \prod_{j=1}^{n} \prod_{i=1}^{n} \left(1 - r_{x_i y_j}\right) \quad (2)$$

term：メインタームと共起するタームの部分集合
n：部分集合内のタームの要素数

3.2 カテゴリ内におけるクラスタリング

クラスタリングは各カテゴリ内の各 URL のタームを URL 集合を識別する能力の高さ [5]の値を用いて行う。次に同一カテゴリ内で各 URL に対して求められたタームの値を表1に示す。

表1 各 URL 内のタームの統計値

	Kw_1	Kw_2	Kw_3	Kw_4	・
URL1	0.02	0	0.007	0.004	・
URL2	0.03	0.01	0.006	0.001	・
URL3	0	0.005	0.003	0.003	・
:	:	:	:	:	:

この表に対し次の操作を行う。
①値の合計を求め、各セル内の値をそれで割る
②この表の行をMDL情報量基準を用いてプールする。

この操作によりプールされた URL の集合が各カテゴリ内でクラスタリングされた URL となる。

4 提案手法の評価

検索結果として得られたクラスタの評価をした。

表2 クラスタ内の相関・ターム数

評価基準	積率相関係数平均	平均ターム数
LSI	0.5243	4.1
提案手法	0.6214	3.2

この表から提案手法が LSI よりクラスタ内のターム数が少なく、URL 間の相関が高いのでクラスタ内の内容の類似性が高いことが分かる。カテゴリの区分のタームとして、ユーザが用いたキーワードを利用し、ユーザの利用頻度、ターム間の相関を反映したことが分類の指標としては適切であると考えられる。

5 まとめ

検索結果の URL のリストが内容に基づきカテゴリ別に分類されたことで、従来のサーチエンジンを用いた場合の内容を確認する為にページを閲覧すべき手間が省け、従来の問題点の①を解決し検索の効率化に寄与できた。今後は②～④の解決を目指す。

参考文献

[1] 岩崎正秀,中川こころ,高田善朗,関浩之:"可変なカテゴリ構造を用いた文書検索支援手法の実験的評価", 情処研報, 2000-DBS-120-1, pp.1-8,2000.

[2] Dumais,S. T.,Furnas,G. W.,Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In Proceedings of CHI'88: ACM, 281-285.

[3] 原田昌紀,佐藤進也,風間一洋 :"WWW ページ間の階層構造の推定と検索システムへの応用", 情処研報, 99-DBS-118-14, pp.105-112,1999.

[4] 川前徳章,青木輝勝,安田浩:"ユーザ履歴を活用した検索システム", 情処研報, 2000-DBS-122-15, pp.113-120,2000.