サイト評価情報を用いたWWW検索表示方法

3T－7

灰原 清太郎　柏井優希　川越恭二
立命館大学理工学部

1 はじめに

利用者のWWWの検索効率を高めるために、インターネット上ではさまざまな検索サービスが運営されている。その1つの方法として、ウェブサイトについてのサイト名・URL・紹介文などのデータをカテゴリ構造として、人間の手によって分類するディレクトリサービスがある。

しかし、増大するWWWに応じて、ディレクトリサービスの登録サイト数も増加している。そのため、1つのカテゴリに多くのサイトが含まれ、利用者の操作コストも増えてしまう。そこで、サイトごとの評価情報を算出して、その評価情報を用いて、並び換えて表示する方法を提案する。

2 現在のディレクトリサービスの問題点と解決策

WWWの利用者が利用している検索サービスは、ディレクトリサービスと検索エンジンの2種類に分類できる。

ディレクトリサービスは、検索エンジンと比較すると、提供する側の人手に頼ることが多い。より効率的な検索を利用者に提供するためには、手間がかかることになる。そのような現状の中で、

①分類された1つのカテゴリ内でもサイト数が多く、操作コストがかかる。

②管理が不十分で、不要なサイトが含まれる。

など、利用者にとって不便な点もある。

現在のディレクトリサービスでは、カテゴリ内の登録サイトは、辞書順や登録順で表示されている。

本研究では、登録サイトに対して、サイト評価情報を設定し、その評価順でソートすることにより、利用者の操作コストを軽減する方法を考える。これにより、上の①・②で述べたような不便さは抑えることができる。

また、ソートの条件を、利用者が与えやすくするような、検索表示方法を提案する。これにより、利用者は、複雑なインターフェイスではなく、簡単な1つのパラメータを変化することで、条件を操作できるようにする。

3 WWW検索表示方法の概要

3.1 サイト評価情報

登録サイトに対するサイト評価情報（サイトスコア）は、以下のような自動的に解析できる要素から算出されるものとした。

- アクセス度(サイトのアクセス数)－ページごとのアクセス数を、ページの階層を考慮して、算出する。
- 更新度(サイトの更新頻度と更新量)－最新更新日だけではなく、更新間隔から、更新頻度を算出する。更新されたページ数などから、更新量を算出する。
- 内容度(サイトのページ数とファイルサイズ)－サイト全体の総ページ数・総ファイルサイズ・平均ファイルサイズから算出する。
- リンク度（被リンク数、リンク関係）－サイトに対してのリンクされている数・関係から算出する。

重み付けをした要素の値の総和で、サイトスコアを算出する。このサイトスコアで登録サイトをソートし、表示する。

利用者は、この重み付けを自分の要求に合わせて変更でき、図1に示す簡単なインターフェイスを提案した。

「新鮮」側にパラメータを近づけると、更新度が強調されたサイトスコアになり、登録サイトは更新がよく行われてい

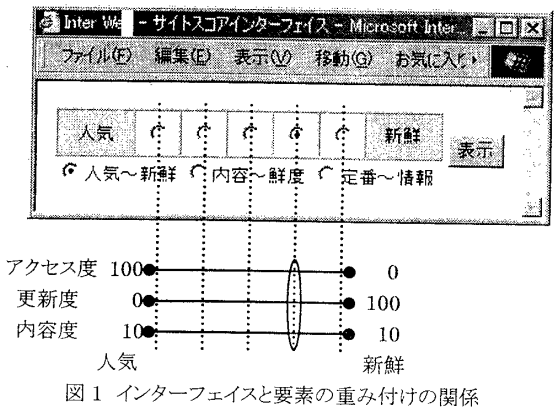

図1 インターフェイスと要素の重み付けの関係

Site Score based Ordering for WWW Search Result.
Seitaro HAIBARA, Yuki KASHIWAI, Kyoji KAWAGOE.
Ritsumeikan University.
1-1-1, Noji-Higashi, Kusatsu, Siga, 525-8577, Japan.

るサイト順に表示される。「人気」側に近づけると、アクセスが多いサイト順に表示される。

3.2 構成

全体の構成を図 2 に示す。収集ロボットは、登録サイト内のすべてのページを取得する。Proxy サーバログを解析し、各サイト・ページのアクセス数を抽出する。各カテゴリでのパラメータは、ユーザプロファイルデータベースで利用者ごとに管理され、クライアントブラウザの Cookie 機能で利用者を識別する。

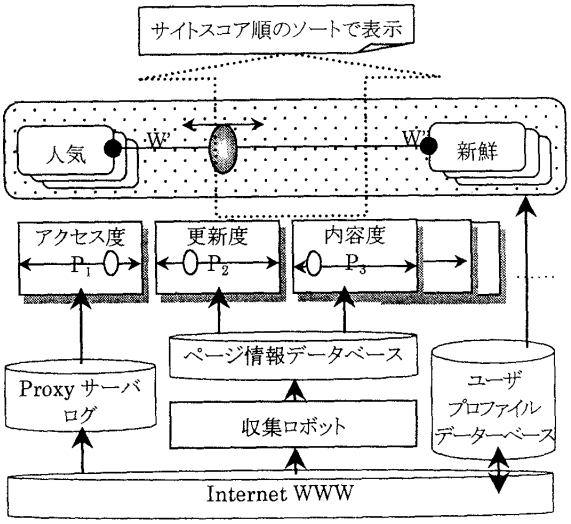

図 2 全体の構成

4 評価

この検索表示方法の効果を見積もるために、パラメータを変化させたときのサイトスコア順でのランク変動と、利用者の要求との適合とで、評価を行った。

22 個のサイトでのアクセス度・更新度・内容度の 3 つの要素を用いて、「人気～新鮮」間でパラメータを変更する。算出されたサイトスコア順(表1)での"ランクの変動数(変動したサイト数)と平均変動量(ランク増減量の絶対値での平均)"と、"ページビューによるランキングとのランク適合率"とを調査した(表 2)。

この結果から、検索表示はパラメータの変化に合わせて平均的にサイトのランクが変動し、利用者の要求に合わせられることができる検索表示であることが分かる。また、パラメータを「人気」側に近いほど、ページビューによるランキングとのランク適合率は高くなり、利用者の要求に近い検索表示であることが分かる。

表1 算出結果

	「人気」側	「新鮮」側
1	パーソナルキングダム(100)	SoftPlaza(85)
2	SoftPlaza(92)	Creative Farm(84)
3	CYBER PLAZA(90)	Career Up!(81)
4	インターネットファンクラブ(78)	いくじーず(81)
5	シネマスクランブル(64)	シネマスクランブル(79)
6	Creative Farm(63)	パーソナルキングダム(78)
7	旅 Web(61)	CYBER PLAZA(75)
8	Hanako Net(55)	旅 Web(66)

表 2 評価結果

	変動数	変動量	ランク適合率
人気	13	0.8	92%
⇕	14	1.0	89%
	15	1.0	82%
	14	1.3	73%
新鮮			61%

このような結果から、この検索表示は、1つのカテゴリでの登録サイト数が増えても、対応できると予想している。

また、このインターフェイスは、操作が容易で、利用者の操作コストを軽減できるため、多くの利用が期待できる。より詳細なインターフェイスを提供しても、操作が複雑になっては利用者には使ってもらえないと思われる。

5 おわりに

現在のディレクトリサービスの問題点を踏まえて、サイト評価情報(サイトスコア)に基づいて、ソートする検索表示方法を提案した。サービス提供側の人手に頼らないこのWWW 検索表示方法は、利用者の操作コストを軽減できることが分かった。

この WWW 検索表示方法は、利用者に実際のサービスとして提供して、その利用状況から、より利用者の要求にあった評価情報の算出などを、検討する必要がある。

参考文献

[1] NEC BIGLOBE CYBERPLAZA、
http://www.cplaza.ne.jp/
[2] 柏井優希、灰原清太郎、川越恭二:「時間制約下での WWW 検索スケジューリング方法」,情報処理学会第 58 回全国大会講演論文集,vol.3,3T-05(1999)