ディレクトリの利用履歴に基づくさまざまなドメインレベルでの嗜好特性について

6Z－5

幸 嘉平太、元田 敏浩、川崎 隆二
NTTソフトウエア研究所

1. はじめに

膨大な情報が発信されているインターネットに対し、求める情報のURLを検索するサービスへのニーズが高まっている。様々な手法の検索サービスが提供されているが、ディレクトリ型と非ディレクトリ型の2つに大分することができる。ディレクトリ型はURLの紹介情報をジャンル別に分類・階層化し、ディレクトリとして提供するものである。情報発信者からの申告によって情報の登録を行う。非ディレクトリ型はいわゆる収集ロボットにより情報を登録し、一括して全文検索を行うタイプである。既存の検索サービスは多種多様な利用者に対し、同一の検索インターフェースしか提供していないが、個々の利用者の嗜好特性を考慮し、利用者毎に個別のインターフェースを提供することで、より効率的な情報検索が可能になると考えられる。

以下、NTTが提供しているインターネットディレクトリサービスであるNTT DIRECTORY[i]の利用履歴を解析対象として、これら問題点の解決策であるパーソナルディレクトリの提案と、利用履歴から利用者の嗜好を抽出する技術について述べる。

2. パーソナルディレクトリの提案

ディレクトリサービスには、第1階層・第2階層・....・最下層を経由して紹介情報を得る階層部と、登録された紹介文に対する全文検索部の2つのインターフェースがある。階層部の構造は複雑であり、NTT DIRECTORYの場合、第1階層で約110、第2階層で約800、最下層は約1900のジャンルに分かれている。図1に階層構造の例を示す。目的のジャンルに到達するには、この階層構造を巧みに経由する必要があり、試行錯誤を要求する作業である。また、個々の利用者単位で見た場合、全く利用しないジャンルが提示されている、ジャンル分類も固定されているなど、利用者の利用特性を反映した表示・分類にはなっていない。利便性を考えた場合、利用頻度の高いジャンルは別途に提示すべきであり、既存の分類に当てはまらないジャンル利用特性がある場合、新規にジャンルを生成すべきである。このように、利用者個別の嗜好を把握し、固有のディレクトリを提示する、パーソナルディレクトリを提案する[ii]。パーソナルディレクトリを構築するためには、利用者の嗜好を把握する必要があるが、このためには「自己申告式」と「自動判断式」の2つが考えられる。われわれは、利用者の負担がほとんどない、自動判断式を研究中である。

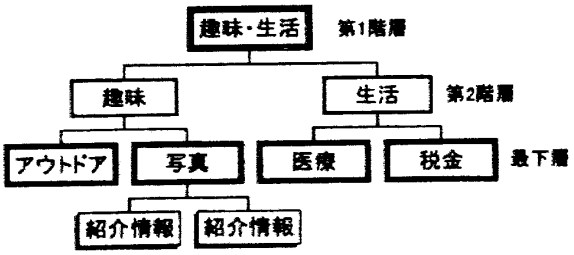

図1　階層構造の例

次節でサーバの利用履歴に基づいた嗜好の抽出法を考える。サーバから有意な嗜好特性を抽出することが可能であれば、その特性に基づいたパーソナルディレクトリを生成できる。

3. 利用履歴からの嗜好抽出

NTT DIRECTORYのアクセスログを分析対象とした。このログには、

IPアドレス:FQDN:ユーザID:アクセス日時:利用ジャンル

The property of users' desire on the log of directory service,
YUKI Kaheita, MOTODA Toshihiro, NAGAOKA Mitsuo,
NTT Software Laboratories, 3-9-11 Midori-cho, Musashino, Tokyo 180-0012, Japan

が記録されている。ここでユーザIDとは、個々の利用者を識別するために付与している HTTP Cookie-ID である。この Cookie-ID によって、利用者をブラウザ単位で識別することが可能となる。分析するドメインレベルによって特性が異なることが予想されるため、「ユーザ ID 単位」、「IP アドレス単位」、「ドメイン単位」の 3 単位で分析した。IP アドレス単位、ドメイン単位の分析によって、同じドメイン内に存在する利用者の集団的特性が得られる。なお、ここでのドメイン単位とは、セカンドレベル(例:ac.jp)までを 1 集団とした単位である。

以下、1997 年のある期間中(約 1 ヶ月)のログを対象に、利用ジャンルの偏りを反復度を算出することにより調査した。ここで、反復度を、

$$\frac{\sum(\text{各ジャンルの利用回数})}{\text{利用ジャンルの総数}}$$

と定義する。つまり調査期間が 1 ヶ月の場合、1 ヶ月・1 ジャンルあたりの平均利用回数である。例えば、3 種類のジャンルをそれぞれ 5 回、10 回、3 回利用した場合、反復度は(5+10+3)/3=6 となる。この反復度が大きいほど、特定ジャンルへの偏りが大きいと言える。以下、表 1 に 3 つのドメインレベルでの反復度を示す。ジャンル利用に偏りがなくとも、単純に利用回数の多い場合も高い反復度となるケースがあるため、利用したジャンル数の平均も併記した。反復度が高く、かつ利用ジャンル数の平均が低い場合が、最も偏っている状態だといえる。

		第 1 階層	第 2 階層	最下層
単位別反復度	ユーザ ID	4.50 (1.45)	3.89 (1.81)	3.61 (1.97)
	IP アドレス	5.03 (2.17)	4.35 (2.88)	3.45 (3.49)
	ドメイン	5.13 (23.8)	3.89 (32.1)	3.48 (29.1)

表 1 ドメインレベル別反復度 : 回／月・ジャンル
(括弧内は利用ジャンル数の平均)

表 1 での、IP アドレス、ドメイン単位の高い値は、主に利用回数の高さによるものと考えられる。平均利用ジャンル数の高さがそれを示している。しかし、ほぼ利用者個人を特定しているユーザ ID 単位でも、高い値が得られており、かつ平均利用ジャンル数も低い。このことは、ユーザ ID を単位としたパーソナルディレクトリの生成が可能であることを示している。利用者毎に、上位 3 ジャンルを通常のジャンル表示とは別枠に提示し、容易にアクセスできるようにした場合、その 3 ジャンルが全利用ジャンル中に占める割合を表 2 に示した。表 2 から、上位 3 ジャンルのみでも、ほぼ全利用ジャンルをカバーできることが分かる。

	利用ジャンルカバー率
第 1 階層	98.1%
第 2 階層	96.3%
最下層	95.3%

表 2　上位 3 ジャンルのカバー率

4. まとめ

ディレクトリサービスのログから利用者の嗜好を抽出できることが分かった。1900 近くに分類されている NTT DIRECTORY でも、実際に利用しているのはごく少数の偏ったジャンルであると言える。このような特性を用いて、利用者毎にパーソナルディレクトリが提供できると考えられる。今後は時系列の観点から、より的確な反復度を抽出できる期間や重み付けなどを探る予定である。また、全文検索部を利用した結果得られた URL 紹介情報にもジャンル属性が付随しており、これをさらなる情報源として活用することも可能である。反復度や平均利用ジャンル数だけでは、偏り特性を表すには十分ではないため、新たな表現を検討する必要もある。ディレクトリサービスは非常に利用回数の高いサイトであることから、パーソナルディレクトリを実際に運営するには、技術的にいくつかの困難があると考えられる。今後は全体設計を見据えた研究を進めていく。

---

[i] NTT DIRECTORY, http://navi.ntt.co.jp/

[ii] 幸 嘉平太他, 「ディレクトリの利用履歴に基づく個人への動的なジャンル適応方式の提案」, 信学技報 Vol. 97, No. 413