デジタル・ドキュメント　7－4
（1997．5．23）

WWW検索ログに基づくトレンド情報の抽出について

大久保 雅且　杉崎 正之　井上 孝史　田中 一男

NTTヒューマンインタフェース研究所

検索ログから多くの人に共通の情報ニーズを抽出できれば、情報収集計画の策定、検索インタフェースへの反映、特集の企画化など、効果的な検索サービスの提供へとつなげることができる。本稿では、(1)各ユーザの識別、(2)ユーザ毎の視点の違いを考慮した同一情報への要求に対する検索語のグルーピング、によって情報ニーズを抽出し、さらに、(3)情報ニーズ時系列の可視化、によって時間変化(トレンド)の把握が容易となる手法を提案する。また、現在インターネット上で提供しているNTT　DIRECTORY(WWWページの検索サービス)の検索ログに適用し、本手法の有効性を確認した。

Extracting the Trends of Information Demands by Analyzing a WWW Search Log

Masaaki Ohkubo, Masayuki Sugizaki, Takafumi Inoue, and Kazuo Tanaka

NTT Human Interface Laboratories

This paper proposes a method to detect information demands common for many people by analyzing an access log for a WWW (World Wide Web) search engine. The proposed method can effectively extract information demands by identifying each user, and by gathering various keywords used to retrieve equivalent information. The movement of the extracted demands is visualized so that we can easily understand and guess the trends of the demands. We applied the proposed method to the access log of NTT DIRECTORY, a directory service for WWW pages, and evaluated its effectiveness.

1. はじめに

情報のデジタル化技術の進展や、インターネットの爆発的な発展に伴い、ネットワーク上に蓄積される情報が激増している。このため、大量の情報の中から必要な情報を得る検索技術がますます重要となってきている。

ネットワーク上で検索サービスを提供する場合には、多くの情報の収集と検索の高速化・高精度化はもちろんであるが、さらに、必要とされている情報を的確に把握し、それを情報の収集や分類、あるいは検索インタフェースへ反映させることも重要と考えられる。

本稿では、検索サービスの充実を目的とした、検索ログに基づくトレンド情報の抽出について述べる。

2. 検索ログの利用と課題

2.1 検索ログの利用例

我々は現在、NTT DIRECTORY[注1]において、登録されたホームページの情報に対する全文検索サービスとして、InfoBee検索[田中,1996]を提供しており、多くの方に利用して頂いている。

NTT DIRECTORYでは、ジャンル別検索も同時に提供している。このため、InfoBee検索では、不特定多数による、分野を限定しない検索が主となる。図2.1にInfoBee検索における、検索語あたりの利用回数の分布を示す。1度しか使用されない検索語が全体の半分以上を占めており、検索語が非常に多岐に渡るというキーワード検索の特徴がよく表れている。

一方、何度も使用されている検索語もあり、多くの利用者が共通して欲しがっている情報の存在を示唆している。したがって、利用回数の多い検索語から多くの人に共通の情報ニーズを特定できれば、それを情報収集計画に反映させたり、簡単な操作でそれらの情報にアクセスできるようにメニュー化したり等、効果的な検索

[注1] http://navi.ntt.co.jp

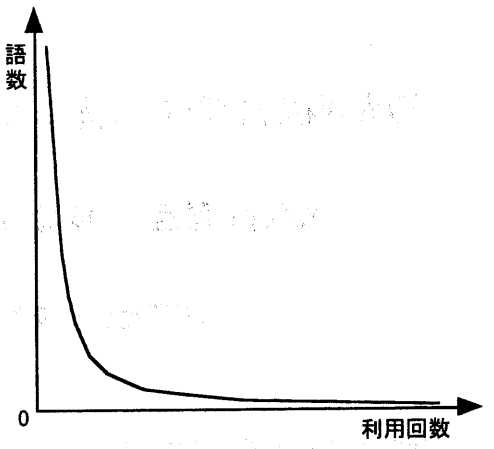


図2.1 検索語あたりの利用回数の分布

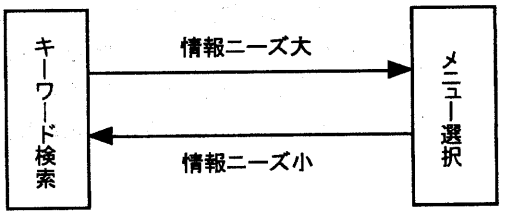


図2.2 情報ニーズと有効な検索インタフェース

サービスの提供へとつなげることができる（図2.2参照）。

2.2 検索ログ解析の問題点

検索ログは、「いつ、誰が、どのような検索を行ったか」の記録であり、前節で述べたように、そこから情報ニーズを抽出できれば、検索サービスへ反映させることができる。しかし、使用された検索語を単純に集計するだけでは以下のような問題点がある。

(1) 100回使用された検索語でも、1人の利用者が100回と、100人の利用者が1回では意味が異なる。

必要なことは多くの人に共通する情報ニーズの把握なので、それぞれの利用者を区別して集計しなければならない。

しかし、通常のWWWサーバのアクセスログには、クライアントのIPアドレスしか残らない。このため、例えば、DHCP[Droms,1993]等を用いてIPアドレスを動的に与えている場合や、ダイアルアップユーザによるアクセス、ファイアーウォール越しのアクセスなどの場合には、アクセスログに残されたIPアドレスと実際の利用者とは1対1では対応しない。

(2) 同じ情報を求める際でも、それぞれの利用者の持つ固有の視点から、異なる検索語を用いる。このため、「同一の情報を求めるために使用した検索語」をまとめて集計する必要がある。

例えば[杉崎他,1997]では、電子ニュースや新聞記事を対象として、各記事に出現する単語の種類と数に基づく分類手法によって同じ話題に関する記事をグループ化している。しかし、検索ログ解析では対象は個々の単語であることから、この手法は適用できない。

また、個々の商品名や省略形など時代を反映した新語が次々に使用されたり、検索の時点で利用者が「関連している」と見なした語を抽出する必要があることから、一般的な類語辞書では対応できない。例えば、「年賀状」と「当選番号」は、ある期間では「お年玉つき年賀ハガキの当選番号」を調べるために使用されることが多いため、同一の情報ニーズとしてまとめて集計した方がよいが、別の期間ではグループ化すべきではない。「サッカー」と「ワールドカップ」、「スキー」と「北海道」なども同様である。

(3) 利用者の情報ニーズは時と共に変化する。

例えば、10日間で合計100回使用された検索語でも、1日に100回と、1日10回ずつ10日間では「情報ニーズ」としての意味が異なる。利用者の情報ニーズを反映した検索サービスを提供するためには、検索語を集計して得られた情報ニーズが、今後どのように変化するかを予測し、それに対して最適なインタフェースを提供する必要がある。したがって、全体をまとめての集計や、ある時点での集計（いわゆるベストテン情報）ではなく、情報ニーズの時間的な変化（これを「トレンド」と呼ぶことにする）の把握が容易となるよう視覚化することが重要と考えられる。

3. トレンド情報の抽出法

3.1 各利用者の特定

各利用者を特定するために、HTTP-cookie [Kristol and Monrulli, 1997]を用いる。HTTP-cookieは、サーバが送信する小データで、クライアント（ブラウザ）側で保存され、次回同一サーバにリクエストする際にその小データも同時に送られる。この仕組みを利用して各利用者にIDを付与し、検索語と共にログに記録する。HTTP-cookieによるIDは、正確にはブラウザ単位での識別になるが、ブラウザと利用者とは1対1に対応していると仮定する。また、HTTP-cookieは、現在インターネット上での使用割合の多くを占めるブラウザで利用できるため、ある程度の区別が可能と考えられる。

3.2 関連語の抽出

同一の情報を求めるため使用された検索語が異なる理由は、

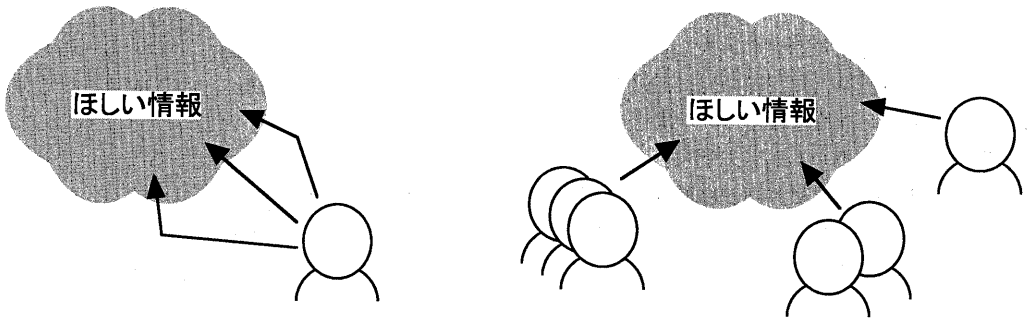
(A) 一人の利用者がいくつかの視点から複数の検索語を使用した

(B) 複数の利用者がそれぞれの視点から検索語を使用した

という2つに大別でき（図3.1参照）、これらを抽出できれば、ある特定の情報ニーズに対する検索語をグループ化できると考えられる。

まず(A)について考える。同一の利用者が複数回の検索を行う場合、

(a) ある情報を得るために様々な検索語を入力して試行錯誤しながらの検索



(A) 同一利用者による視点の違い　　(B) 複数利用者による視点の違い

図3.1 同一の情報を求めるために行う検索のイメージ図

(b) 以前の検索とは別の情報を求めるための新たな検索

の2種類がある。(a)は比較的短い時間間隔での連続する検索となり、(b)は前回の検索から比較的長い間隔をおいての検索となると考えられる。すなわち、同一の利用者によって使用された検索語は、その使用時間間隔が短ければ同じ情報を求めるために、長ければ別の情報を求めるために、それぞれ使用された可能性が高いと考えられ、検索語間の関連（これを個人内関連度と呼ぶ）を使用時間間隔の関数と見なすことができる。そこで、時区間Tにおいて、利用者iによる検索語x,yの使用時間間隔をd_i、その区間におけるxとyの個人内関連度を$f(d_i)$としたとき、

$$c_{xy} = \Sigma f(d_i)$$

を、Tにおけるx,yの時間関連度と定義する。

次に、(B)タイプの検索語集合について考える。ある一定の時期に、多数の利用者から同一の情報への要求があったとき、その検索に使用された語の使用頻度傾向は似ていると考えられる。すなわち、例えば1日単位で検索語の使用頻度を集計し、その時系列を比較すれば、ある時区間において同一の情報を求めるために使用されたかどうかの検出に利用できる（図3.2参照）。そこで、検索語の使用頻度の時系列の相関係数によって、同一情報に対する検索かどうかを判定する。時区間Tにおいて、検索語x,yの、時区間T/nごとの使用頻度を、それぞれ、x_i, y_i (i = 1, 2, ..., n)とするとき、相関係数r_{xy}は、

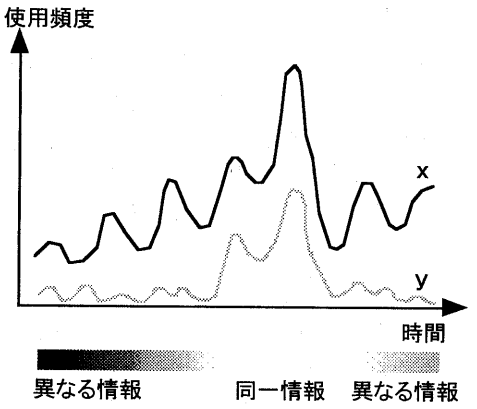


図3.2 検索語x,yの使用頻度の時系列と求める情報の関係のイメージ

$$r_{xy} = \frac{\Sigma(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2}}$$

によって求められる。ただし、$\bar{x}, \bar{y}$はそれぞれ、x_i, y_iの平均値である。

時区間Tにおいて、2つの検索語x,yが同一の情報を求めるために使用されたかどうかは、これら2つの値c_{xy}, r_{xy}を併用して判定する。同一情報に対する検索語の使用と判定された場合には、それらをグループ化し、使用頻度を合計す

ることによって、トレンド抽出に役立てる。

4. 実現法

以上の基本方針を実現し、InfoBee検索ログに関して解析を行った。検索語の集計時には、英数字の1byte化、英大文字の小文字化、1byteカナの2byte化等の正規化により、同一語に対する表記の揺れを吸収した。また、1日を単位として検索語を集計し、2週間分を1区間として各検索語間の時間関連度と相関係数を求めてグループ化する。

同一の利用者による検索要求の時間間隔と回数の関係を図4.1に示す。通常、検索は1度では終わらず、求める結果を得るために何度か繰り返して行うことが多いこと考えば、図4.1の曲線

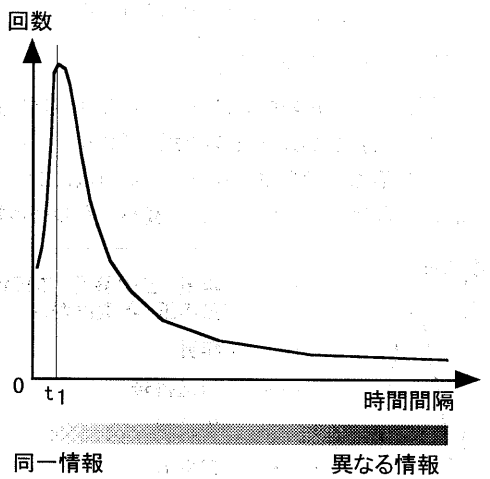


図3.2 検索の時間間隔の分布と求める情報の関係

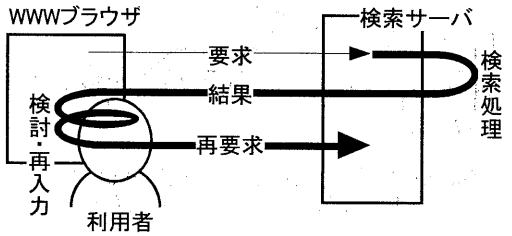


図4.2 時間間隔t_1前後の考え方

のピークを示すt_1は、検索語を入力してから結果を受けとるまでの時間（システムの処理時間＋通信時間）と、検索結果に基づいて次のアクションを決定して再リクエストをし、それが届くまでの時間の和と考えられる（図4.2参照）。したがって、t_1前後までは、(a)の検索を行っている可能性が高い。また、ある時間間隔を越えると(b)の検索を行っている可能性が高くなると考えられる。

これらのことから、個人内関連度を求める関数fを、利用者iの検索語x,yの使用時間差の最小値をd_iとしたとき、

$$f(d_i) = \begin{cases} a & (d_i = 0) \\ 1 & (0 < d_i < t_2) \\ (d_i - t_3)/(t_3 - t_2) & (t_2 \leqq d_i \leqq t_3) \\ 0 & (t_3 < d_i) \end{cases}$$

ただし、$0 < t_1 < t_2 < t_3$

と定義した。関数fでは、ANDやORなどと共に同時に使用された検索語(d_i=0)に関しては、特に関連度が高いと考えて、特別な値としている（図4.3）。

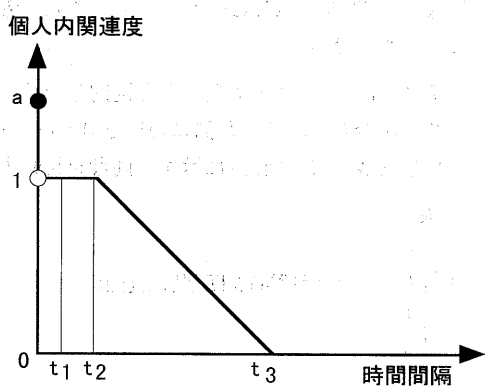


図4.3 検索語の使用時間間隔と関連度

また、2つの検索語x,yが同一の情報を求めるために使用されたかどうかを判定する関数gは、

$$g(c_{xy}, r_{xy}) = \begin{cases} 1 & (c_{xy} > c_l \text{ かつ } r_{xy} > r_l \text{ かつ } (c_l - c_h)r_{xy} + (r_l - r_h)c_{xy} + c_h r_h - c_l r_l > 0) \\ 0 & (\text{上記以外}) \end{cases}$$

とした（図4.4）。

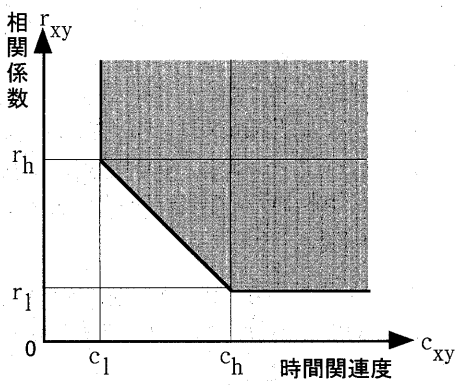


図4.4 関数gによる検索語のグループ化範囲

5. 評価および考察

5.1 利用者の特定

全アクセスの約90.8%はHTTP-cookieによるidを持つ利用者からのアクセスであった。残りの9.2%は、期間中に1回だけアクセスした利用者か、HTTP-cookieに対応していないブラウザの利用者か、HTTP-cookieの使用を拒否した利用者かのいずれかである。

HTTP-cookieによって利用者を区別した場合としない場合とで、2ヶ月間に使用された検索語の使用回数分布を図5.1に示す。両者を比較すると、1回しか使用されない検索語は全体の56%、1人しか使用しない検索語は全体の66%であった。また、1000回以上使用された検索語数と、1000人以上が使用した検索語数とでは、3倍以上の差があった。これらのことから、HTTP-cookieによる利用者の特定が有効であったことがわかる。

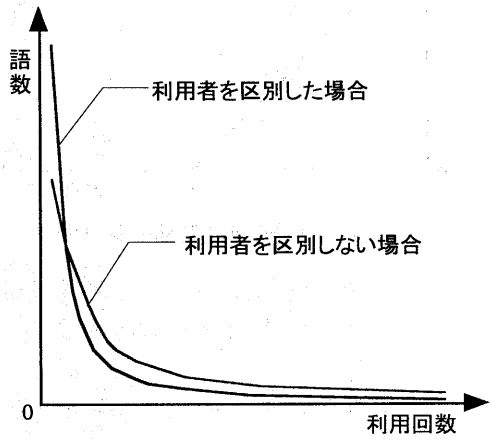


図5.1 検索語あたりの利用回数の分布

5.2 語のグループ化の効果

同一情報を求めるための検索語をグループ化することにより、

・個々の検索語の使用頻度は大きくないが、それらのグループ化によって情報ニーズを発見できる。

・通常時と比較して、ある特定の期間に、なぜ検索要求が増えたのかがわかる。

といった効果があった。

例えば、1月の中旬に使用された「年賀状」「お年玉」など数種類の検索語がグループ化されて顕著な山の発生になっている（図5.2[注2]）。これは、「お年玉付き年賀ハガキの当選番

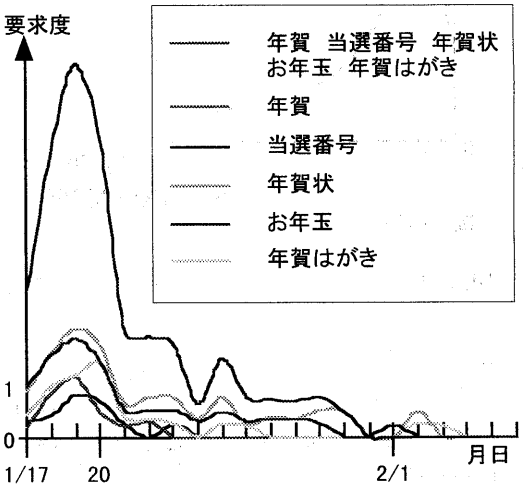


図5.2 検索語のグループ化の効果（その1）
（1月17日の「年賀状」の要求度を1とする）

注2 図5.2では、1月17日における「年賀状」という検索語の使用者数を要求度1とし、各検索語の使用者数を相対的な要求度として表している。以下、他のグラフでも同様である。

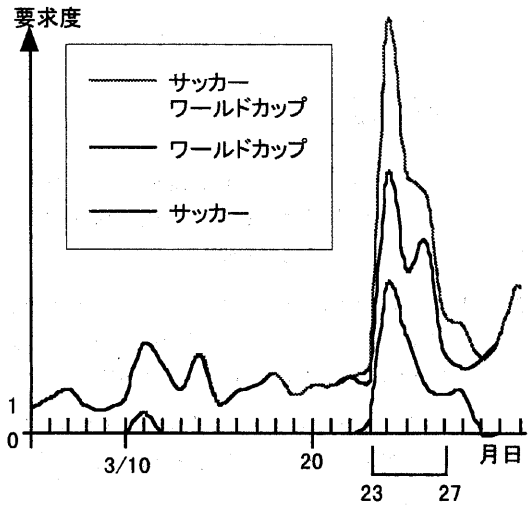


図5.3 検索語のグループ化の効果(その2)
(3月5日の「サッカー」の要求度を1とする)

号」に対する要求であったと予想される。個々の検索語の使用頻度は多くはないが、実際にはその情報に対する要求が多かったことが、検索語がグループ化されたことによって見えるようになった。

後者には、例えば、「サッカー」という検索語は、通常一定の頻度で使用されているが、ある時期に急激に増えている。この期間には、サッカーのワールドカップの予選が(現地時間で3月23日から27日まで)が開催されており、「サッカー」と「ワールドカップ」とがグループ化されたことで、これに関する情報への要求であったことがわかる(図5.3)。同様の例として、「宝くじ」と「グリーンジャンボ」なども あった。

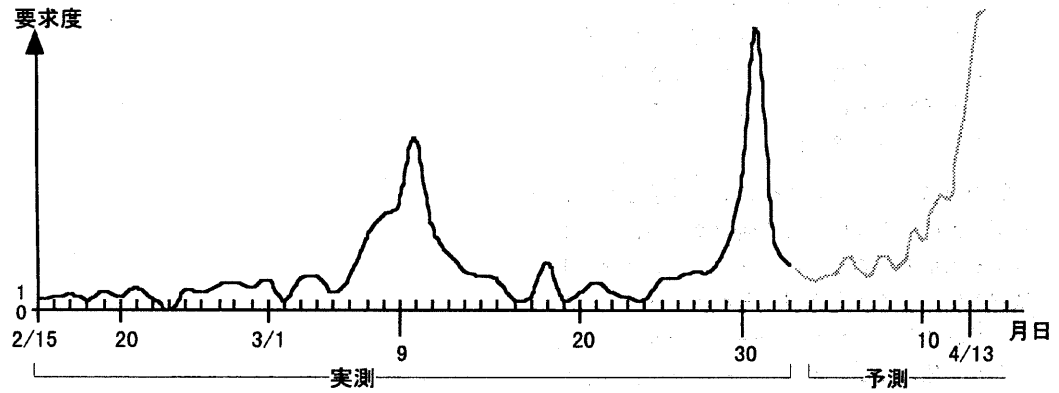


図5.4「F1」に関する情報要求の変化と予測 (2月15日の「F1」の要求度を1とする)

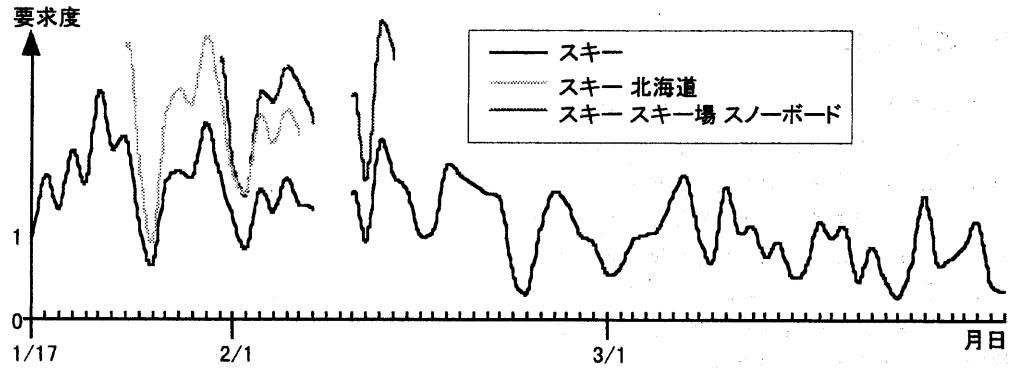


図5.5「スキー」に関する検索要求の変化 (1月17日の「スキー」の要求度を1とする)
(2月8, 9日はシステムメンテのために有効なデータが取得できていないので省略)

5.3 抽出したトレンド

検索語の使用頻度の時系列をグラフ化することによって、例えば、以下のような予測が可能になる。

- サッカーのワールドカップ予選の開催期間は、多くの要求があった（図5.3）。次は6月22日～28日に開催されるので、そのころにまた多くの要求が発生するだろう。
- 「F1」は、3/9にオーストラリア、3/30ブラジルで開催され、それぞれ検索要求が高くなっている。次は4/13アルゼンチンで開催されるので、そのころに要求が高くなるだろう（図5.4）。
- 「スキー」に関する検索は、2月中旬までは「北海道」や「スキー場」「スノーボード」などとグループ化されて多くの要求があったが、3月に入って減ってきている。この傾向が続けば5月頃にはかなり少なくなるだろう（図5.5）。

世の中には数多くのイベントが開催されているが、上記のように、この検索サーバにアクセスする利用者（の多く）が欲しているイベント情報が何であるかを把握できれば、「先回りして」情報を収集したり、企画化したりといったことが可能になる。また、シーズンに依存した情報に関する企画に関しても、タイムリーな開始や終了が可能になると考えられる。

6. おわりに

WWW検索ログに基づくトレンド情報の抽出について述べた。5章で示したように、本手法によって検出されたトレンド情報を用いれば、より使いやすい検索サービスの提供につながると考えられる。また、本稿で示した関連語のグループ化手法は、そのときの流行や検索利用者の視点を反映した関連語辞書の自動構築や、曖昧検索などへの応用も可能である。

今後は、曜日や時間帯によるユーザ層の違いを意識した解析や、トレンド情報の効果的な視覚化手法について検討していきたい。

参考文献

[田中, 1996] 田中, "InfoBee検索エンジンを用いたディレクトリ検索サービス", NTT技術ジャーナル, vol.8, No.8, 24-27 (1996).

[Kristol and Monrulli, 1997] D. Kristol and L. Montulli, "HTTP State Management Mechanism", RFC 2109 (Feb., 1997).

[杉崎他, 1997] 杉崎ほか, "情報分類を用いたトレンド・アウェアネスの支援", 情処研報 97-DD-6, 9-16 (1997).

[Droms, 1993] R. Droms, "Dynamic Host Configuration Protocol", RFC1541 (Oct., 1993).