WWW検索ログに基づく情報ニーズ傾向の把握

2AC－5

大久保雅且　杉崎正之　早川和宏　田中一男

NTT ヒューマンインタフェース研究所

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。これら大量の情報の中から必要な情報を得るための検索サービスは、不特定多数による様々な検索要求に応えられる反面、誰に対してもどのような要求に対しても画一的な検索インタフェースなので、使い勝手が良いとは言えない。

例えば、情報ニーズの大きなものをメニュー化して視覚化すれば、多くの人が容易な操作で情報にアクセスできるだけでなく、潜在的な情報需要の喚起にもつながると考えられる。本稿では、WWW検索ログに基づく情報ニーズ傾向の抽出法を提案する。また、実際の検索ログを解析し、その結果について考察する。

2. 検索ログ解析の問題点

我々は現在、NTT DIRECTORY[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[2]を提供しており、多くの方に利用して頂いている。情報検索は、利用者の情報要求の生の声であるから、そのログを解析すれば多くの人に共通する情報ニーズを抽出できる。しかし、同じ情報を求める際でも、それぞれの利用者固有の視点から、異なる検索語を用いている。このため、検索ログに出現する検索語の使用頻度を単純に集計しても情報ニーズとはならず、同一情報を要求した検索語をグループ化して集計する必要がある。

3. 検索語から情報ニーズへ

同一の情報を求めるために使用される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や立場に応じて様々な検索語を使用した

という2つに大別できる。これらを検出して各検索語間に関連の強さを定義できれば、同一の情報要求のために使用されたかどうかを判定できる。

まず(A)について考える。図1に、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示す。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

(b) 以前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。通常、1回の検索で求める情報を得ることは難しく、比較的短い時間間隔で何度も検索を繰り返す。図1の曲線のピークt_1はこの繰り返し周期を示しており、t_1前後までは(a)の検索をしていると考えられる。一方、(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。これらのことから、同一の利用者によって使用された検索語

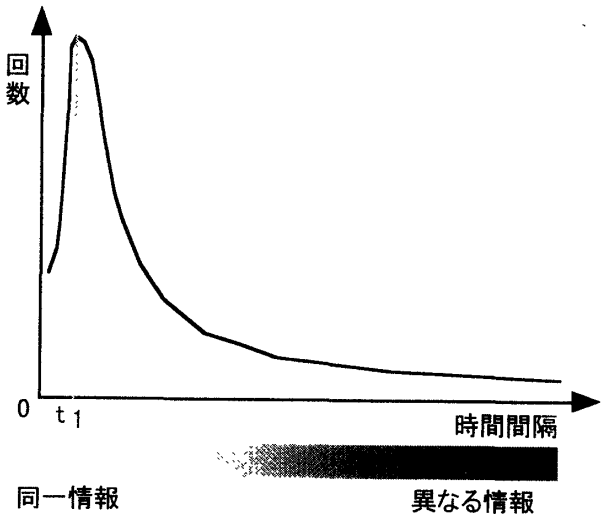


図1　検索の時間間隔の分布と求める情報の関係
(対象：1997年1月から3月までの検索)

Extracting Information Needs by WWW Search Log Analysis

Masaaki OHKUBO, Masayuki SUGIZAKI, Kazuhiro HAYAKAWA, and Kazuo TANAKA

NTT Human Interface Laboratories

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、検索語間の関連を使用時間間隔の関数と見なすことができる。これを間隔関連度と定義する。

次に、(B)タイプの検索語について考える。ある一定の時期に、多数の利用者が同一の情報を求めた場合、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。そこで、１日単位で検索語の使用頻度を集計し、その時系列の相関係数を時系列関連度と定義する。

２つの検索語が同一情報を求めたものかどうかは、上記２つの関連度によって判定する。ただし、例えば、「当選番号」が時期によって「お年玉つき年賀ハガキの当選番号」や「宝くじの当選番号」等、異なる情報要求のために使用されるように、それぞれの時期に応じてタイムリーな関連度を求めることが必要である。したがって、関連度の計算は区間を区切って行うこととする。

4. 実現および考察

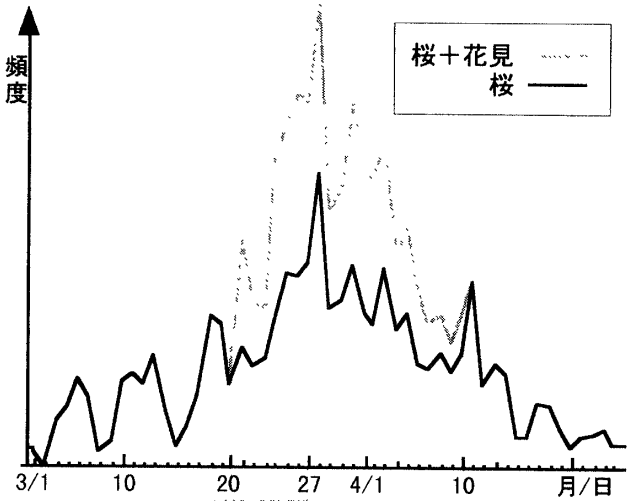
以上の基本方針に基づいてInfoBee検索ログの解析を行った。間隔関連度の計算では、まず各利用者ごとに、各検索語の使用時間間隔の最小値を求め、その値に応じて０から１までの値を付与し、その総和を間隔関連度とした。間隔関連度は１週間単位で、また時系列関連度は２週間単位で、それぞれ計算し、同一の情報を求めた検索語と判定された語を、その区間においてグループ化した。

図２に「桜」に対する解析結果を示す。図２において、例えば灰色の枠で囲んだ部分は、３月21日から27日までの間隔関連度の高い検索語上位３個を示している。また14日から27日までの区間の時系列関連度と、前記の間隔関連度の値に基づいて「桜」と「花見」はグループ化されている。

各区間での間隔関連度の高い語から、

- ３月中旬までは「桜前線」「開花」など、桜の咲き始める時期を求める要求が多かった。
- ３月下旬以降は、「京都」「高遠」「造幣局」「北海道」など、桜の名所に関しての情報要求が多く、その場所は時と共に北上していく。

ことがわかる。なお、東京における今年(1997年)の



花見	花見	京都	花見	花見	花見	京都	造幣局
気象庁	桜前線	花見	写真	京都	京都	花見	北海道
京都	写真	桜前線	開花	さくら	高遠	高遠	札幌

図2「桜」の使用頻度の推移と、間隔関連度の高い検索語　（■■■ はグループ化された検索語）

桜の開花宣言は３月21日に出されており、上記にほぼ一致している。

一方、３月下旬から４月上旬までは「桜」と「花見」は同一の要求としてグループ化されている。両者の合計を図２のグラフに灰色の線で示す。「桜」の頻度だけからは、それほど強い情報要求はなかったように見えるが、実際は、この期間に「桜」の「花見」に関する情報要求が非常に大きかったことが、グループ化によってわかる。

5. おわりに

検索ログに基づく情報ニーズ傾向の抽出法について述べた。各検索語間の関連度を求めることにより、利用者が本当に求めている情報の内容とその推移、および要求度合いの真の大きさの把握が可能となる。今後、結果の効果的な視覚化手法と、実際の検索サービスへの反映手法について検討していく。

参考文献

[1] http://navi.ntt.co.jp/

[2] 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).