1T-3

# Web 検索ログの検索時間間隔を用いた利用者の行動パターンの分析

柳　阿礼[†]　河村　春雄[†]　徳永　幸生[†]　杉崎　正之[‡]　池田　成宏[‡]

芝浦工業大学　工学部[†]

NTT レゾナント株式会社　技術マーケティング部[‡]

## 1. はじめに

インターネットの発達により，web を用いた情報発信が世界的規模で増え続けている．このような大規模な情報の中から自分の欲しい情報を見つけるための様々な検索システムが開発されている．

一般に，利用者は求める情報を探し出すため検索システムに検索式を入力し，試行錯誤しながら求める情報に近づいている．従って，この web 情報の検索ログには利用者の情報要求の生の声が潜んでいると考えられる．

そこで，web 検索サイトから未知の情報を検索する時の行動（検索ログ）を分析することにより，ある情報を得るために使用された検索式間の関連度を抽出する試みがなされている．更に，この情報の関連度を用い，検索式同士の背景に潜む構造や相互の関係から，検索の目的・情報取得の目的を探る議論がなされている．[1]

上記の議論，分析は 1997 年の検索ログを基になされてきた．図 2 は 1997 年の検索ログを分析して求めた検索の使用時間間隔の分布である．しかし，10 年前と現在とでは検索方法が異なる．例えば，かつては 1 ワード検索が主流であったのに対し，現在では 2 ワード以上の同時検索が主流となっていると考えられる．このような検索方法の変化に伴い，図 2 のグラフの形状も変化すると思われる．

そこで，本稿では，利用者が検索システムを利用した際の検索時間，検索式の内容が利用者ごとに記録されている，最近の膨大な検索ログデータから利用者の行動パターンを分析し，その特性を比較考察した．

## 2. 関連語の抽出

①人間の検索行動

通常，１回の検索で求める情報を得ることは難しい．図１のように，STEP1 - STEP2 間では，異なる検索式の入力や検索式の組み合わせを変えるなど，試行錯誤による連続した検索が行われるため，比較的短い時間間隔での頻繁な検索が繰り返されると考えられる．STEP2 - STEP3 間では，STEP2 の検索結果としてタイトルやコメントなどを含むため，ある程度内容を推測できる．よって，STEP3 からの後戻りは少ないと考えられ，比較的長い時間間隔での検索となる．STEP3 - END 間では，STEP3 において求める情報を得られた，あるいは得られないと判断した時点で一連の行動は終了するため，比較的長い時間間隔での検索となる．

これらより，STEP1-STEP2 間で入力された短い時間間隔の検索式群は，利用者にとって同一の要求を得るために使用された検索式群である可能性が高いと考えられる．

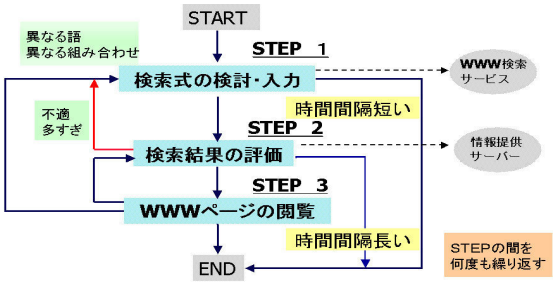


図１　WWW検索サービスの利用者の行動[1]

②検索回数の割合と検索の使用時間間隔の関係

1997 年の検索の使用時間間隔の分布は図 2 のようになる．最も検索回数が多い使用時間間隔を t1 とする．①に基づいた結果，t1 前後までは一連の検索行動である可能性が高いと考えられる．

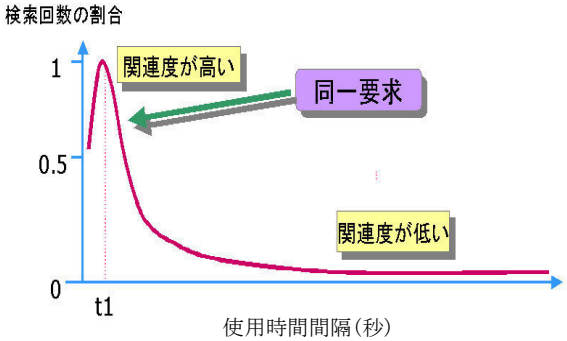


図２　検索の使用時間間隔の分布(1997 年)[1]

---

Extracting User's Action Pattern by Analyzing Interval of Time of a WWW Search Log

Are YANAGI[†]　Haruo KAWAMURA[†]

Yukio TOKUNAGA[†]　Masayuki SUGIZAKI[‡]

Naruhiro IKEDA[‡]

Shibaura Institute of Technology[†]

NTT Resonant Inc[‡]

### 3. 検索方法の変化

検索が行われた時間，検索式が順に並べられた膨大な検索ログデータから，隣接する検索が行われた時間の差を検索の使用時間間隔として抽出した．そして，使用時間間隔が 0 秒の時，つまり，同時検索の割合が時期に応じてどのように変化したかを調査したグラフが図 3 である．

約 1 年半前から現在に至るまでに同時検索は約 2 倍に増加していることが分かった．

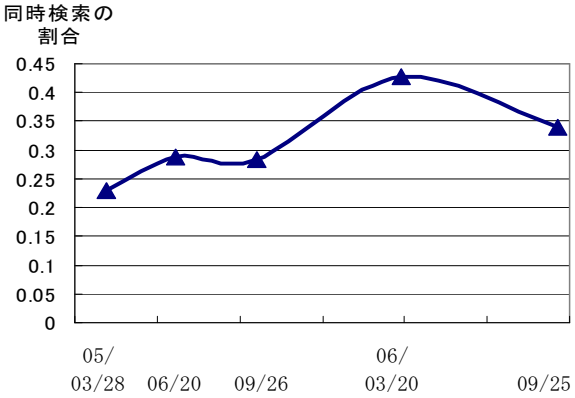


図 3　同時検索の割合の変化

### 4. 現在の検索の使用時間間隔の分布

3 と同様に，隣接する検索が行われた時間の差を検索の使用時間間隔として抽出し，使用時間間隔の分布を求めた．（図 4-1，図 4-2） 全 48 日分（2005/03/22～29，06/20～29，09/20～29，2006/03/20～29，09/20～29）の検索ログデータの各一日分に対し，それぞれグラフを作成した．

図 4-1 は 2005 年 6 月 20 日の検索ログデータから求めたグラフであり，1997 年の検索の使用時間間隔の分布（図 2）と同じ山形になっている．一方で，図 4-2 は 2006 年 9 月 28 日の検索ログデータから求めたグラフであり，山形になっておらず，図 2 とは異なる形状になっている．全検索ログデータのうち約 75%が図 4-1 の形状となり，それ以外は図 4-2 の形状になっている．

また，図 4-1，図 4-2 は共に使用時間間隔が 1 秒の時の検索回数の割合が最も高くなっている．このうち，1 秒後に再度同一の検索式を入力している場合が約 59%存在する．これは，利用者の検索を行う上での癖であったり，利用者が検索式を入力した後，検索結果として表示される web ページの URL の一覧を速く得たいと思うことから，ダブルクリックを行ってしまっているためと考えられる．通常，1 秒間で新たに検索式を入力することは難しいと考えられるが，実際に検索ログデータを分析した結果，1 秒間で一度入力した検索式に対して僅かに追加，削除を行い，訂正している場合が多いことが分かった．

更に，使用時間間隔が 2～5 秒の時に窪みができている．これは，一度入力した検索式を吟味してから再度検索式を入力している利用者が多く，（図 1　STEP1 - STEP2 間）一度検索式を入力してから，2～5 秒後までの間は再度検索式を入力する利用者が少ないためと考えられる．

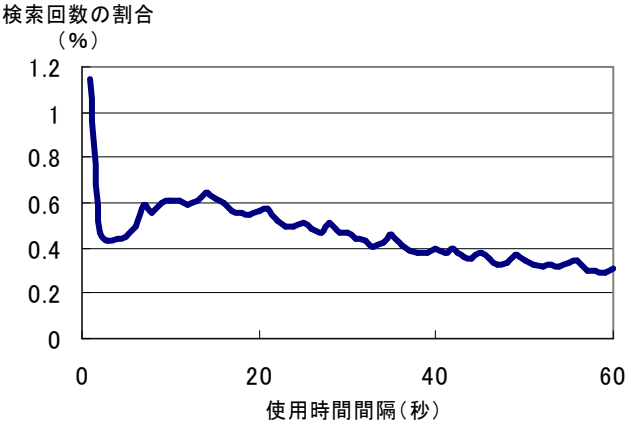


図 4-1　検索の使用時間間隔の分布(2005/06/20)

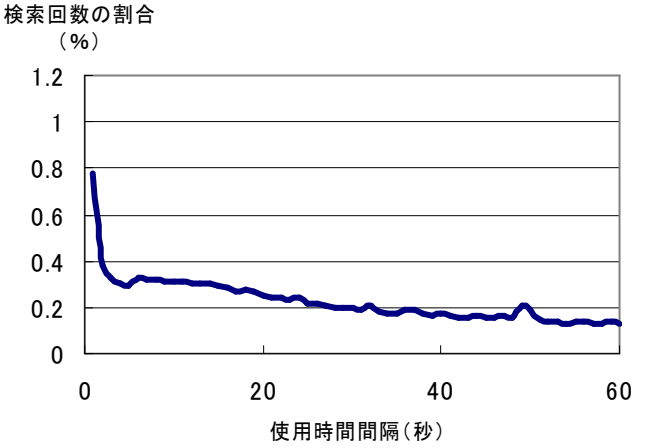


図 4-2　検索の使用時間間隔の分布(2006/09/28)

### 5. 今後の課題

4 で作成した様々なグラフは図 4-1 の形状，あるいは図 4-2 の形状の 2 種類に分類される．これらのグラフを基にモデル化を行う．モデル化を行う上では，グラフの曲線上の点における接線の傾きが変化する点を考慮する．

今後の展望として，モデルの持つ意味を考察し，関連度の精度を向上させる．さらに，関連度を可視化することにより，最初に入力した検索式が最終的にどのような検索式となって利用者が目的の情報に辿り着いたのかを調査する．そして，これらを基に詳細な情報ニーズを抽出することに結び付けたい．

### 参考文献

[1] 大久保雅且，井上孝史，杉崎正之，田中一男：www 検索ログに基づく情報ニーズの抽出，情報処理学会論文誌，Vol.39，No.7，1997.