D-006

Web 検索ログの検索時間間隔モデルに基づいた関連語の抽出

柳　阿礼[†]　河村　春雄[†]　徳永　幸生[†]　杉山　精[†]　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．研究の背景・目的

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に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語を入力し，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 情報の検索ログデータ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 検索システムから未知の情報を検索する時の行動（検索ログデータ）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語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，検索語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 年の検索ログデータを基になされてきた．図 2 は 1997 年の検索ログデータ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．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 ワード以上による AND 検索が主流となっていると考えられる．これまでは，検索語群が同じ概念を表す情報を得るために使用されているかの尺度として使用時間間隔のみを用いてきた．しかし，現在では AND 検索が多用されるようになった．従って，単に使用時間間隔ごとの使用頻度を集計した値[2]を関連度の計算に反映させたとしても，現在の人間の検索行動を考慮した関連度の重み付けにはならないと考えられる．

そこで，本稿では，検索の使用時間間隔の分析に加え，最近の膨大な検索ログデータ内に表れる，検索の前後において入力された検索語の変化に着目した．そして，より詳細に利用者の検索行動のパターンを分析し，その特性を比較考察した．さらに，これらを基に検索語間の使用時間間隔のみからでは定義することのできない検索語間の関連性を見い出した．

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，異なる検索語の入力や検索語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．一方，STEP2 おいて検索結果としてタイトルやコメントなどを含むため，閲覧しようとしている Web ページの内容をある程度推測できる．従って，STEP3 からの後戻りは少ないと考えられ，STEP3 を介した検索は比較的長い時間間隔での検索となる．また，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．従って，STEP1－END，STEP2－END，STEP3－END 間も比較的長い時間間隔での検索となる．

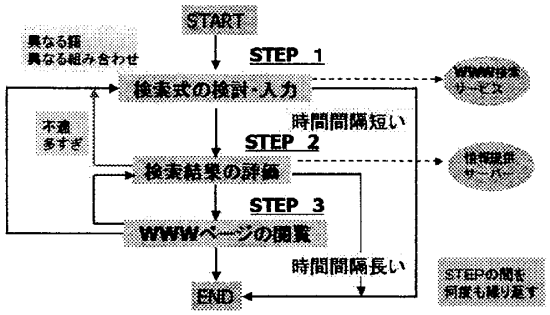

図 1　Web 検索システムの利用者の検索行動[1]

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の考えに基づいた結果，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．

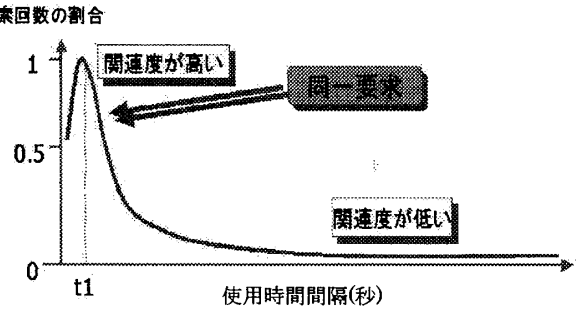

図 2　検索の使用時間間隔の分布(1997 年)[1]

3. AND 検索の割合

検索が行われた時間，検索語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，AND 検索の割合を調査したグラフが図 3 である．現在行われている検索の多くが AND 検索であることが分かった．これより，関連度を求める上では，

Extraction Related Words Based on Model of Interval of Time of a Web Search Log

AreYANAGI[†]　Haruo KAWAMURA[†]
Yukio TOKUNAGA[†]　Yasushi SUGIYAMA[†]
Masayuki SUGIZAKI[‡]　Naruhiro IKEDA[‡]
Shibaura Institute of Technology[†]
NTT Resonant Inc[‡]

AND 検索が行われなかった時に使用時間間隔から求められた関連度よりも，AND 検索が行われた時に求められた関連度の方が，検索語間の関連度に及ぼす影響が大きいと考えられる．単に使用時間間隔のみから関連度を求めるのでは，本来関連のある検索語群（使用時間間隔が 300 秒以上であったとしても明らかに関連があると思われる検索語群[1]）に対して高い関連度を付与できない可能性がある．逆に，関連の低い検索語群に対して誤って高い関連度を付与してしまう可能性もある．

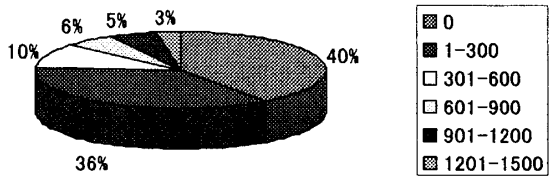

図 3　2006/09/28 の使用時間間隔ごとの検索回数の割合

4. 検索パターンの分類

3 で述べた問題点を解決するために，検索語間の使用時間間隔のみから関連度を定義するのではなく，検索パターンも踏まえて関連度を定義することを考える．検索ログデータ内の各検索において，前後の検索語の変化を分析し，以下の 5 種類（グループ）に分類した．検索ログデータ内（利用者約 40 万人，全検索回数約 120 万回）において，5 種類の検索パターンの存在する回数を求めたグラフが図 4 である．

グループ 1：A－A（AND 検索なし，前後で検索語が同一のパターン）

短い時間間隔の場合には検索語を入力後，結果として表示される Web ページの URL の一覧を早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と考えられる．長い時間間隔の場合には検索結果や Web ページを閲覧した後，“次へ”ボタンにより次のページへ切り替えていると考えられる．

グループ 2：A－B（AND 検索なし，前後で検索語が異なるパターン）

短い時間間隔の場合には検索語の入力ミスに気付き，再入力している場合が考えられる．長い時間間隔の場合には検索結果や Web ページを閲覧した後，検索語を変更した方が良いと判断し再入力を行っている，もしくは以前とは異なる情報を得るための新たな検索を行っていると考えられる．

グループ 3：AB－AB（AND 検索あり，前後で構成されている検索語が全て同一のパターン）

グループ 1 と同様な検索の特徴が表れる．

グループ 4：AB－AC（AND 検索あり，前後で検索語の一部分が同一のパターン）

検索語の一部にグループ 2 と同様な検索の特徴が表れる．

グループ 5：AB－CD（AND 検索あり，前後で構成されている検索語が全て異なるパターン）

短い時間間隔で新たな検索語を 2 語再入力することは難しいと考えられる．長い時間間隔の場合にはグループ 2 の長い時間間隔の場合と同様な検索の特徴が表れる．

1 のパターンに関しては，同一の検索語となるため，関連度は定義しない．3 のパターンに関しては，A と B の関連度は A と B を AND 検索した際の関連度として定義される．4 のパターンに関しては，B と C は共に A と AND 検索されている．従って，AB－AC 間の使用時間間隔が大きくても B と C の関連度は高くなると考えられる．このパターンの場合の関連度は，使用時間間隔に依存しない，一定の値として定義すべきであると考える．2，5 のパターンに関しては，関連がある場合とない場合とが考えられる．そこで，確実に関連があると考えられる 3，4 から利用者が同じ概念を表す情報を得るための検索にかかる平均検索時間を求める．この値を閾値とし，使用時間間隔が閾値以下であれば，関連度を使用時間間隔の単調減少関数として定義し，使用時間間隔が閾値以上であれば関連度を 0 と定義すべきであると考える．

例えば，A（t1）－A（t2）－AB（t3）－BC（t4）のように，順番に検索語が入力されたとする．ここで，A，B，C は検索語，t1，t2，t3，t4 はそれぞれの検索語が入力された時間を示している．同一の検索語 A が含まれる A（t1）－A（t2）－AB（t3）までを同じ概念を表す情報を得るための検索とし，この検索にかかった検索時間（t3-t1）を求める．検索ログデータからこの検索を抽出し，利用者 1 人当たりの平均検索時間を求め，全ての利用者の平均検索時間を求めた．平均検索時間は約 1571 秒であった．10 年前には平均検索時間を約 300 秒としており，使用時間間隔が 300 秒以降の検索語群の関連度を 0 と定義していた．[1] 現在の平均検索時間は 300 秒より短くなるだろうと予想していたが，結果より非常に長いことが分かった．これは，明確な目的を定めずに検索を行っている利用者が増えているためであると考えられる．

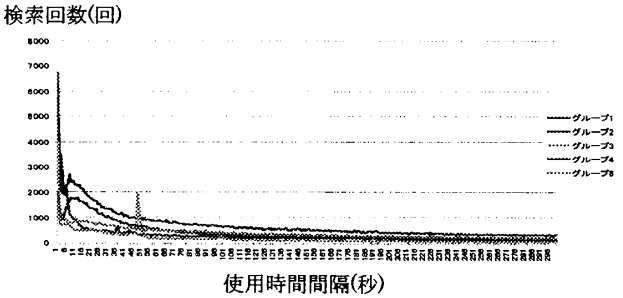

図 4　2006/09/28 の各検索パターンの存在する回数

5. 今後の課題

本稿で述べた関連度の定義を基にモデル化を行う．モデル化を行う上では，AND 検索時における関連度の値 ，5 種類の検索パターンの存在する割合からそれぞれのパターンの関連度の定義を全体のモデルにどう反映させるかを考慮する．

今後の展望として，新たなモデルを作成し，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語が最終的にどのような検索語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，これらを詳細な情報ニーズの抽出に結び付けたい．

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.

[2] 柳阿礼，河村春雄，徳永幸生，杉崎正之，池田成広：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，第 69 回情報処理学会全国大会，2007.