検索の使用時間間隔の分布を用いて抽出される関連語の評価

柳　阿礼[†]　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. 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．利用者は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）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]

そこで，本稿では，最近の膨大な検索ログデータ（利用者が検索システムを利用した際の検索時間，検索語の内容が利用者ごとに記録されている）から検索が行われた時間の差を使用時間間隔として抽出し，前後の検索語の内容を分析することで利用者の行動パターンを明らかにした．これらを基に新たな時間間隔関連度を定義し，特徴ベクトルを用いて単語間の距離から求めた関連度と比較，検証を行った．

2. 検索の使用時間間隔の分布の作成

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．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．また，STEP2 における検索結果にはタイトルやコメントなどが含まれるため，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る．すなわち，STEP1 - STEP2 間では，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，STEP3 を介した検索は比較的長い時間間隔での検索となる．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．

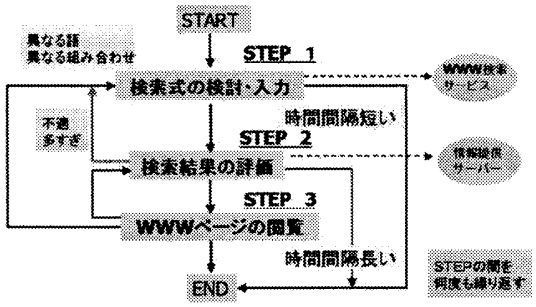

図 1　Web 検索システムの利用者の検索行動[1]

② 検索の使用時間間隔と検索回数の関係

2006 年 9 月 29 日の検索ログデータを分析した結果，検索の使用時間間隔の分布として図 2 を得た．ここで，最も検索回数が多い使用時間間隔を t1 とすると，①で述べたように，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．

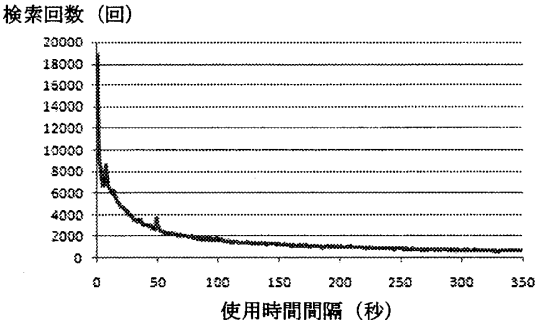

図 2　検索の使用時間間隔の分布
（2006 年 9 月 29 日）[1]

3. 時間間隔関連度の算出

検索の使用時間間隔の分布を基に，使用時間間隔から時間間隔関連度を求める assoc 関数を作成した．図 2 より t1=10 秒，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 t2=335 秒と設定した．

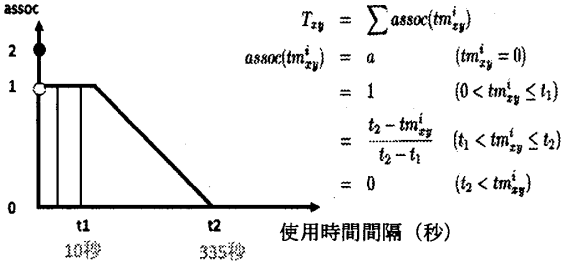

図 3　assoc 関数[1]

Evaluation of related words extracted by Analyzing Interval of Time of a WWW Search Log

Are YANAGI[†]　Yukio TOKUNAGA[†]　Kiyosi SUGIYAMA[†]
Masayuki SUGIZAKI[‡]　Naruhiro IKEDA[‡]
Shibaura Institute of Technology[†]
NTT Resonant Inc[‡]

4. 関連度の算出

時間間隔関連度を用い，2 種類の手法で関連語を抽出した．

① 時間間隔関連度

STEP1.

同一利用者によって使用された各検索語間の使用時間間隔 $tmini_i[x,y]$を求める．同じ検索語が複数回使用されている場合には，最小値を使用する．

STEP2.

STEP1 で求めた使用時間間隔を図 3 の assoc 関数に適用し，同一利用者における時間間隔関連度 $assoc(tmini_i[x,y])$を求める．

STEP3.

STEP1，2 を全ての利用者に対して行い，各検索語間の関連度の総和 $T_{xy}=\Sigma assoc(tmini_i[x,y])$を求める．この値が時間間隔関連度となる．

② 特徴ベクトルによる cos 関連度

求めた時間間隔関連度 T_{xy} の値を用いて，単語 x の特徴ベクトル W_x を $W_x=(T_{x1},\cdots,T_{xj},\cdots,T_{xn})$とし，特徴ベクトルを用いた単語間の距離 Dis_{xy} を三角関数$\cos\theta$で定義する．[2]

5. 関連語の抽出

2006 年 9 月 29 日の検索ログデータを対象に，検索語「ホテル」について，時間間隔関連度によって求めた検索語と関連度の高い上位 10 位までを図 4 に示す．同様に，特徴ベクトルによる cos 関連度によって求めた関連語の上位 10 位までの関連語を図 5 に示す．

6. 考察

図 4 の時間間隔関連度には「京都」，「東京」といった地名や「予約」，「格安」といった「ホテル」に関する要求を表す関連語が上位に表れる．一方，図 5 の特徴ベクトルによる cos 関連度には「宿泊」，「ビジネスホテル」といった「ホテル」と似た意味を持つ関連語が上位に表れる．また，特徴ベクトルによる cos 関連度には明らかに検索語と関連がないと思われる関連語も含まれている．

さらに，図 4 の時間間隔関連度で上位に表れている「京都」，「東京」は図 5 の特徴ベクトルによる cos 関連度では，それぞれ 302 位，122 位となっていた．地名は「ホテル」以外の様々な単語とも一緒に使用される可能性が高いため，特徴ベクトルによる cos 関連度では順位が下がったと考えられる．一方，「ホテル」と「観光」の時間間隔関連度の上位には共に地名が多く含まれている．地名以外の単語と一緒に使用されることが少ないため，特徴ベクトルによる cos 関連度では上位に表れていると考えられる．

関連語	時間間隔関連度
京都	71.36923077
東京	69.58769231
予約	46.97846154
大阪	36.72
横浜	36
札幌	26
格安	26
沖縄	24
新宿	22.24
名古屋	22

図 4 「ホテル」の関連語
（時間間隔関連度）

関連語	特徴ベクトルによる cos 関連度
観光	0.628086307
宿泊	0.596906217
ビジネスホテル	0.594811068
大丸	0.535510367
高速バス	0.483562318
市バス	0.474865303
ラーメン	0.456936468
グルメ	0.456767689
夜行バス	0.45555701
ランチ	0.449028909

図 5 「ホテル」の関連語
（特徴ベクトルによる cos 関連度）

7. まとめ

時間間隔関連度を用いると，検索語と一緒に（AND 検索）調べられた単語が上位に表れる．これらは検索語に対し追加候補となる検索語群であると考えられる．

一方，特徴ベクトルによる cos 関連度を用いると，調べられ方の似ている単語が上位に表れる．これらは検索語に対し置換候補となる検索語群であると考えられる．

本稿では，時間間隔関連度を用いた手法と特徴ベクトルによる cos 関連度を用いた手法とで抽出される 2 種類の関連語について，違いや傾向を明らかにした．今後は，どのような場合にどちらの関連度を用いるのが適切であるのかについて，様々なケースについて分析する．

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] 杉崎正之, 牧野俊朗 田中一男: www 検索ログを用いた次検索候補単語の提示方法の検討，情報処理学会論文誌.