Web検索ログに基づく複数の関連度を利用した情報ニーズ検索支援方法の提案

柳　阿礼[†]　徳永　幸生[†]　杉山　精[†]　杉崎　正之[‡]　望月　崇由[‡]
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，検索語間の関連度を抽出し，検索語同士の背景に潜む構造や相互の関係から，情報取得の目的を探る議論がなされている[1]．

本稿では，Web検索システムの利用者の検索行動を分析し，膨大な検索ログデータからその特性を考察する．これを基に，時間間隔関連度と特徴ベクトルによるcos関連度を定義し，各々における関連語を抽出する．さらに，これら2種類の関連語を相互に利用することを念頭に置いて，関連度可視化システムを構築する．そして，出力結果を洞察することにより，人間の検索行動における法則性や情報ニーズを抽出する．

2. 利用者の検索行動の分析

2.1 利用者の検索行動

図1に検索行動を整理した．通常，Webで検索を行う場合，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，STEP3に至ると，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検索行動は終了する．

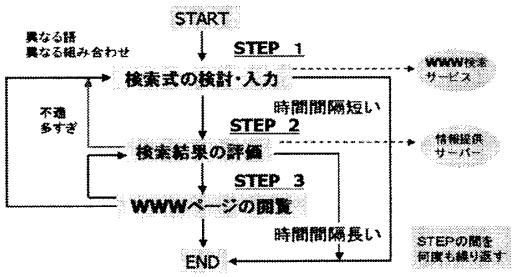

図1　Web検索システムの利用者の検索行動[1]

Proposal of information needs retrieval supporting method using two or more related levels based on a WWW Search Log
Are YANAGI[†]　Yukio TOKUNAGA[†]　Kiyoshi SUGIYAMA[†]
Masayuki SUGIZAKI[‡]　Takayoshi MOCHIZUKI[‡]
Shibaura Institute of Technology[†]
NTT Resonant Inc[‡]

2.2 検索の使用時間間隔と検索回数の関係

検索ログデータから検索の使用時間間隔の分布を求めた（図2）．ここで，0秒付近を除き，最も検索回数が多い使用時間間隔をt1とすると，2.1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．

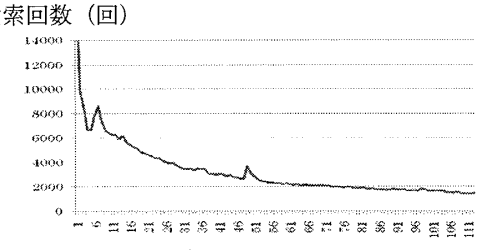

図2　検索の使用時間間隔の分布の例（2006.09.29）[2]

3. 関連度の算出

3.1 assoc関数

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関数を定義した（図3）．また，一連の検索行動の可能性が高いと考えられるt1=10秒，同じ情報を求めるための検索と別の情報を求めるための新たな検索の境界値と考えられるt2=52秒と設定した．

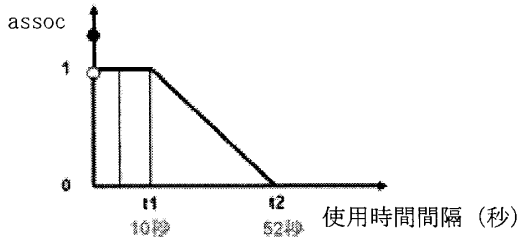

図3　assoc関数

3.2 時間間隔関連度

STEP1.

同一利用者によって使用された検索語xと検索語y間の使用時間間隔tminii[x,y]を求める．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検索語xと検索語yの時間間隔関連度assoc(tminii[x,y])を求める．

STEP3.

STEP1，2を検索語xと検索語yの全利用者に対し行い，時間間隔関連度の総和 Txy=Σassoc(tminii[x,y]) を求める．この値を検索語xと検索語yの時間間隔関連度と定義する．

$$
\begin{aligned}
T_{xy} &= \sum assoc(tm^i_{xy}) \\
assoc(tm^i_{xy}) &= a && (tm^i_{xy} = 0) \\
&= 1 && (0 < tm^i_{xy} \le t_1) \\
&= \frac{t_2 - tm^i_{xy}}{t_2 - t_1} && (t_1 < tm^i_{xy} \le t_2) \\
&= 0 && (t_2 < tm^i_{xy})
\end{aligned}
$$

図4　検索語xと検索語yの時間間隔関連度

3.3 特徴ベクトルによる cos 関連度

3.2 で求めた時間間隔関連度 Txy の値を用いて，検索語 x の特徴ベクトル Wx を Wx=(Tx1,…,Txj,…,Txn) とし，特徴ベクトルを用いた検索語間の距離 Disxy を三角関数 cos θ で求める．この値を特徴ベクトルによる cos 関連度と定義する[4]．

3.4 時間間隔関連度と特徴ベクトルによる cos 関連度の特性

時間間隔関連度を用いると，検索語と一緒に（AND 検索）調べられた検索語（関連語）が上位に表れる．これらは，入力された検索語に対し追加候補となる検索語群である．検索語と時間間隔関連度で上位に表れた関連語の関連度の変化を考察すると，情報ニーズを抽出できると考えられる．

一方，特徴ベクトルによる cos 関連度を用いると，ある検索語を共通の検索語として，調べられ方が似ている検索語（関連語）が上位に表れる．これらは，入力された検索語に対し置換候補となる検索語群である．検索語と特徴ベクトルによる cos 関連度で上位に表れた関連語はある共通の概念を持って使われていると考えられる．

4. 関連語可視化システムを用いた情報ニーズの抽出

4.1 関連度可視化システムの概要

中央に示された検索語と関連度の高い関連語を検索語の周囲から順に配置する関連度可視化システムを構築した．2006 年 9 月 29 日の検索ログデータを用いて，検索語「横浜」について，特徴ベクトルによる cos 関連度の場合の本システムの出力結果を図 5 に示す．同様に，時間間隔関連度を求め，表形式で掲載した（表 1 に「鹿児島」，表 2 に「東京」，表 3 に「京都」，表 4 に「横浜」，表 5 に「札幌」のそれぞれにおいて，時間間隔関連度の高い上位 10 位までの関連語を示す）．

4.2 情報ニーズの抽出

図 5 から「横浜」と特徴ベクトルによる cos 関連度の高い関連語に「鹿児島」，「東京」，「京都」，「鹿児島」といった地名が存在する．「横浜」，「東京」，「京都」，「鹿児島」は地名を共通の概念として調べられ方が似ている．

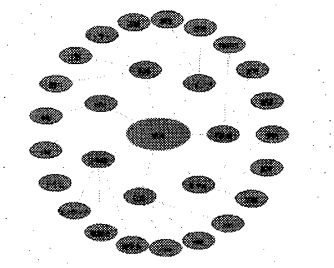

図 5　「横浜」の出力結果

次に，「鹿児島」，「東京」，「京都」，「横浜」，「札幌」のそれぞれについて時間間隔関連度を求め，比較，考察した．

まず，「ホテル」が「鹿児島」を除く 4 県全てにおいて 1 位である．このことから「ホテル」と地名で検索する利用者が多いことが分かる．一方，「鹿児島」でも「温泉」，「観光」，「旅行」といった検索語が見られることから，「ホテル」で検索するのではなく，まず，「温泉」，「観光」で検索する利用者が多いのではないかと考えられる．「京都」に関しては，「ホテル」や「旅館」で検索する利用者も多いが，9 月ということもあり，「紅葉」，「温泉」，「グルメ」で検索する利用者が多いと考えられる．

また，「京都」を除く 4 県全てに交通手段が含まれる．「鹿児島」では「自転車」，「東京」では「路線図」から「電車」と「高速バス」，「横浜」では「バス」，「札幌」では「バス」となっている．それぞれの地域で主流の，あるいは，便利な交通手段が選ばれていると考えられる．

さらに，デパートに着目してみると，「東京」では「大丸」，「京都」では「伊勢丹」，「横浜」では「高島屋」，「札幌」では「大丸」の人気が高いことが分かる．

最後に，「東京」では「大学」，「横浜」では「中華街」，「アウトレット」，「札幌」では「金券ショップ」，「ラーメン」など特定の地名だけに表れている関連語もある．

表 1　「鹿児島」

風俗
温泉
福岡
アミュプラザ
熊島
ja
観光
自転車
旅行
イオン

表 2　「東京」

ホテル
風俗
路線図
大丸
大阪
観光
地図
大学
公園
高速バス

表 3　「京都」

ホテル
紅葉
観光
伊勢丹
宿泊
大丸
温泉
旅館
グルメ
高島屋

表 4　「横浜」

ホテル
観光
バス
風俗
中華街
高島屋
アウトレット
不動産
レストラン
元町

表 5　「札幌」

ホテル
北海道
大丸
金券ショップ
ラーメン
東区
居酒屋
三越
豊平区
バス

5. まとめ

本稿では，時間間隔関連度と特徴ベクトルによる cos 関連度を定義し，各々における関連語を抽出した．また，関連度可視化システムを構築した．時間間隔関連度と特徴ベクトルによる cos 関連度の特性を相互に生かすことによって，人間の検索行動における法則性や情報ニーズを抽出するのに本システムは有用と考えられる．

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] 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，1T-3，(Mar 2007).

[3] 柳阿礼，徳永幸生，杉崎正之，池田成広：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，FIT2007 第 6 回情報科学技術フォーラム，D-006，(Sep 2007).

[4] 柳阿礼，徳永幸生，杉山精，杉崎正之，池田成広：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，第 70 回情報処理学会全国大会，2R-8，(Mar 2008).