1M-1

WWW 検索ログに基づく時系列に着目した検索語間の関連性の分析

小澤　泰輔† 　望月　崇由‡ 　松田　達樹‡ 　徳永　幸生† 　杉山　精*
芝浦工業大学大学院 工学研究科† 　NTT レゾナント株式会社‡ 　東京工芸大学*

1. はじめに

インターネット利用者は，WWW（World Wide Web）上に存在する大量の情報から自分の欲しい情報を得るために，WWW 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く．この WWW 検索システムに記録される検索ログデータには，検索者の情報要求の生の声，すなわち情報ニーズが潜んでいる．そこで検索語間の関連度を WWW 検索ログから抽出し，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研究がなされている．

ある特定の時期に検索数が上昇した検索語には，検索者の検索行動の意図が潜んでいると考えられる．本稿では，各検索語の検索人数を長期的な時系列変化として捉え，そこから検索語の特性を分析した．

2. 時間間隔に基づく関連度

大久保らは検索時の時間間隔に基づき assoc 関数（図 1）を定義し，検索語間の関連度（時間間隔関連度）を算出した[1]．これは検索者がある目的の情報を得るために，検索語同士のペアがどの程度一緒に検索されているかの指標となる．

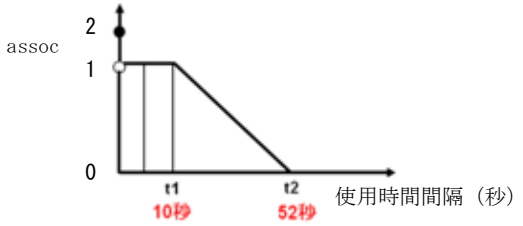

図 1. assoc 関数

柳らはある期間の WWW 検索ログに対して，時間間隔関連度とこれを要素とした特徴ベクトルによる cos 関連度を算出した．これを様々な仮説の下で相互に利用しながら試行錯誤し，検索者の情報ニーズを抽出する試みを報告している[2]．さらにここで得られた特定の検索語に特徴的な情報ニーズを効率的に抽出するために，筆者らは 2 種類の関連語を組み合わせて利用できる関連度可視化システムを構築した[3]．

時間間隔関連度は，1 日や 1 週間といったある一時的な期間の WWW 検索ログをまとめ，固定された期間の検索傾向を調べる場合に有効である．時系列に着目した検索傾向を調べるには，時系列ごとに検索語間の関連度を算出し，ログを総当たりで調べなければならない．そこで本稿では，時系列による検索語間の関連性を調べるために新たな関連度を定義し，分析を行う．

3. 時系列に着目した関連度の算出と分析手法

3.1 時系列に基づく関連度の算出

長期間にわたる WWW 検索ログデータを用い，検索語別の検索人数を日単位で区切り，検索語それぞれに対して日ごとの検索人数を要素としたベクトルを生成する．そして検索語のペアに対し，両ベクトル間の cos 類似度を時系列に基づく検索語間の関連度とする．この関連度が高い検索語同士は，似通った原因で検索人数の変動が起こっている可能性が高い．

3.2 検索人数の変動のカラースケール表示

ある検索語（注目語）に対し，他の検索語との関連度を算出する．注目語との関連度が高い順に検索語をソートした．縦軸を検索語，横軸を時系列とし，検索人数の値を要素としたマトリックスを形成した．これをカラースケールで表示することにより，注目語と，検索語と変動の似ている検索語の検索人数の変動を俯瞰的に調べることができる．

4. WWW 検索ログを用いた分析

4.1 日ごとに分割した時系列による分析

2009 年 11 月 1 日から 2010 年 3 月 31 日の 151 日間に WWW 検索サイトに記録された検索ログから，この期間の総検索人数の上位 30,000 語を対象とし，注目語を選定して時系列の変動による類似度を算出した．これを類似度の高い順にソートし，カラースケール表示した．なお，2010 年 1 月 8 日から 11 日の間はログデータ未取得のため，ベクトルの要素としては外している．その結果，検索語の変動の傾向として，おおよそ 3 つのパターンに分類できることがわかった．

① 曜日などの周期的な影響

図 2 の上部に現れる，Excel や書式といった検

An Analysis of Association between the Search Words focused on Time Series based on a WWW Search Log
† Taisuke OZAWA (m110032@shibaura-it.ac.jp)
‡ Takayoshi MOCHIZUKI (mochizuki@nttr.co.jp)
‡ Tatsuki MATSUDA (t.matsuda@nttr.co.jp)
† Yukio TOKUNAGA (tokunaga@shibaura-it.ac.jp)
* Kiyoshi SUGIYAMA
† Graduate School of Engineering, Shibaura Institute of Technology
‡ NTT Resonant Inc.
* Tokyo Polytechnic University

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

索語は平日に多く検索される．また DS や動画といった検索語は図 2 の下部に現れ，土曜や日曜といった休日に多く検索される．このように，曜日などの周期的なパターンを示す検索語の存在が明らかとなった．

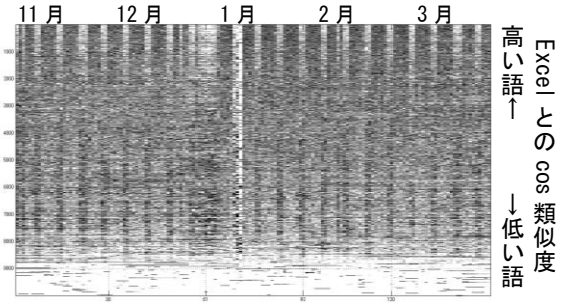

図 2. 「Excel」を注目語とした日ごとのカラースケール（色が濃ければ検索人数が多い）

② イベントなどの時期的な影響

時期が決まっているイベントと検索語が関係している場合，そのイベントの時期に合わせて検索人数が増減する．クリスマスの場合，アクセサリーや家電などの検索語と類似度が高い．

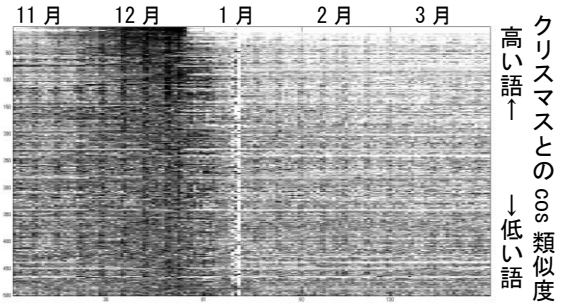

図 3. 「クリスマス」を注目語とした日ごとのカラースケール（類似度上位 500 語抜粋）

③ ニュースなどの突発的な影響

地震などのニュースや，WWW 検索サイトが一定の期間に特集した検索語などは，一時的に検索人数が急上昇し，その後急激に検索されなくなる．これらの検索語は検索人数の変動のしかたが極端であるため，他の検索語を注目語としたカラースケールでは下層に現れやすい．

4.2 曜日と時間帯による分析

長期間の時系列変化から検索語の特性を調べる場合，①の週の周期性による影響が強いことが分かった．また，テレビ番組を調べる場合など，時間帯に関しても検索人数の増減にかかわる影響が大きいと考えられる．そこで，7 個の曜日と 24 の時間帯を掛け合わせて 168 の要素によるベクトルを検索語ごとに作成し，同様にカラースケールを用いて分析した．

Excel を注目語とした場合，4.1 の分析結果では仕事やプログラミングに関連する検索語とともに平日に検索されていることが示されたが，曜日と時間による分析においても同様の傾向が表れた（図 4）．しかし，時間帯をみると，これらの検索語は平日の中でも 9～11 時台，13～19 時台に検索人数が増加している．すなわち昼休みと考えられる 12 時台を除いた仕事中の時間帯に多く検索されているといえる．

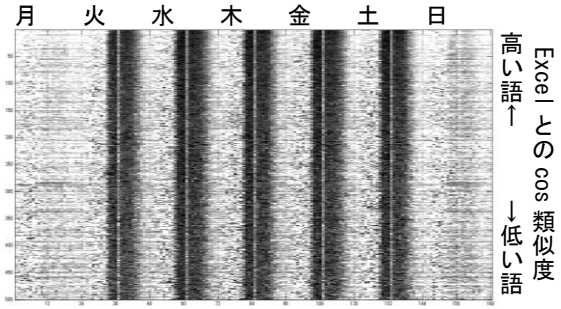

図 4. 「Excel」を注目語とした曜日と時間帯のカラースケール（類似度上位 500 語抜粋）

5. 時間間隔関連度を組み合わせた分析

時系列による類似度から抽出された検索語群に対し，検索人数の上昇のピークとなっている期間の時間間隔関連度による関連語を抽出することで，検索語ごとに特有な検索目的を詳細に比較できると考えられる．

図 3 を例に挙げると，「歩行者天国」はクリスマスとの類似度が上位であるため，この時期に歩行者天国に関する情報ニーズがあると考えられる．歩行者天国の時間間隔関連語を調べたところ，検索数が多いクリスマスの時期に「銀座」と関連度が高く，一緒に検索されている．また銀座はこの時期に，「アバクロ」や「松屋」など歩行者天国沿いの店と時間間隔関連度が高い．このためクリスマスの時期に服飾小物等のお店の情報を求める目的で，銀座の歩行者天国を検索する人が多いと推測できる．

6. まとめ

長期間にわたる WWW 検索ログに対し，時系列による検索語間の類似度を用いて同時期に検索人数が上昇した検索語群を抽出し，検索人数の変動の法則性を探った．さらにこれらの検索語群を比較し，時間間隔関連度を用いて分析を行った．その結果，時系列に関連した検索語ごとに特有な検索目的を推測できた．

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，徳永幸生，杉山精，杉崎正之，望月崇由：“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”，情報処理学会第 71 回全国大会，5P-3，2009. 3

[3] 小澤泰輔，杉崎正之，望月崇由，徳永幸生，杉山精：“Web 検索ログに基づく関連度可視化システムによる情報ニーズの抽出”，情報処理学会第 72 回全国大会，2Q-5，2010. 3

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.