

WWW 検索ログに基づく情報ニーズの抽出

大久保 雅且[†] 杉 崎 正 之[†]
井 上 孝 史[†] 田 中 一 男[†]

WWW (World Wide Web) 検索サービスで使用された検索語から情報ニーズを抽出できれば、より効果的な情報収集や情報提供が可能になる。しかし、同じ概念を表す情報を求める際でも、それぞれの利用者の持つ固有の視点から様々な検索語が用いられる。このため、使用された検索語を単純に集計するのではなく、その時期に「同義語として用いられた」検索語を判定して集約しなければ、情報ニーズの本当の強さや傾向を求めることはできない。本論文では、使用された検索語の間に関連度を定義し、その値によって関連する検索語をグループ化する手法を提案する。関連度は、(1) 検索要求の時間間隔から同一情報への要求かどうかを判定する方法と、(2) 各検索語の使用頻度の時系列を求めそれらの間の相関関係を用いる方法、の2つの観点から定義する。また、一定の期間ごとにこれらの関連度を求めることによって、「その期間における同義語」をタイムリーにグループ化する。さらに、本手法を実際のWWW検索サービスのアクセスログに適用し、各情報ニーズの真の強さや傾向を把握できるだけでなく、関連語の提示によって要求解釈の支援が可能となることを示した。

Extracting Information Demand by Analyzing a WWW Search Log

MASAAKI OHKUBO,[†] MASAYUKI SUGIZAKI,[†] TAKAFUMI INOUE[†]
and KAZUO TANAKA[†]

This paper proposes a method for detecting the information demands of a large group of people by analyzing the keywords used on a WWW (World Wide Web) search service. In general, a variety of keywords are used to retrieve information on the same topic. These keywords differ according to each user's viewpoint. Therefore, they, that is related words in a sense, must be gathered and summed up to extract information needs accurately. In this paper, relationships between two keywords are measured by time interval between them, and by the correlation coefficient of the number of uses per day. By calculating these relationships once in a certain period of time, for example one week, and by combining them effectively, keywords for the same topic can be grouped together. We applied the proposed method to the access log for an actual WWW search service, and found that it was useful for interpreting each request as well as for understanding the true strength and trends of information demands.

1. はじめに

情報のデジタル化技術の進展や、インターネットをはじめとするコンピュータネットワークの普及・発展にともない、ネットワーク上に蓄積される情報が急増している。特に近年は、World Wide Web (WWW) を用いた情報発信が激増を続けており、WWW ページの検索サービスも数多く提供されている^{1),2)}。このようなサービスでは、検索の高速化・高精度化はもちろんであるが、必要とされている情報を的確に把握して、

それを情報の収集や分類、あるいは検索インタフェースへ反映させることも重要である。多くの人に共通の情報要求(これを情報ニーズと呼ぶ)を抽出して、それをあらかじめメニューとして提示すれば、多くの人々が容易に情報アクセスできるだけでなく、潜在的な情報需要の喚起にもつながる。また、情報ニーズの大きさに基づく、いわゆるランキング情報は、それだけで有力なコンテンツとなりうる。

情報検索は、利用者の情報要求の生の声であるので、検索ログの解析によって情報ニーズの強さや傾向を抽出できる。しかし、同じ概念を表す情報を求める際でも、それぞれの利用者の持つ固有の視点から、異なる検索語が用いられる。たとえば、「桜の花見」に関す

[†] NTT ヒューマンインタフェース研究所
NTT Human Interface Laboratories

る情報の検索には、「桜」「花見」以外にも、「さくら」「桜前線」「開花」「満開」などの語、さらには、花見時期の天気予報を求める語や桜の名所などが検索語として用いられることが予想される。このため、使用された各検索語の単純な集計ではなく、「同じ概念を表す情報を求めるために使用された検索語」、すなわち関連の強い語を抽出し、それらをまとめて集計しなければ、情報ニーズの本当の強さを求めることはできない。本論文では、WWW 検索サービスで使用された検索語の間の関連度を求め、関連の強い語のグループ化と関連語の提示による、情報ニーズ抽出手法について述べる。

語と語の関係を記述したものとしてはシソーラスがある。しかし、不特定多数を対象とする、分野非限定の WWW 検索サービスでは、個々の商品名や省略形など、時代を反映した新語が次々に使用されるため、汎用のシソーラスでは対応できない。

検索対象となる文書集合から単語間の関係を求める方法として、統計的手法を用いてシソーラスを自動的に構築する方法³⁾や、HTML 文書中のタイトルやハイパーリンクなどから抽出された単語の共起に基づく方法⁴⁾、検索途中の文書集合中の特徴的な語群間の共起統計を求める方法⁵⁾などが提案されている。しかし、情報の提供側と要求側との情報量のバランスは一般には一致しないため、検索対象となる文書集合から求められた語の関係が情報ニーズを反映しているとは限らない。たとえば、「当選番号」は、ある期間では「お年玉つき年賀ハガキの当選番号」を調べるために使用されることが多いので、「年賀状」や「年賀ハガキ」などと強い関連を持つが、別の時期では「宝くじ」や「かもめー」などとの関連の方が強くなる。このように、検索に用いられる語の間の関連の強さはその時期に応じて変化しているが、検索対象となる文書集合の量や質がこれらをタイムリーに反映していることは希である。

文献 6) では、利用者が入力した検索語と閲覧した情報との相関から関連語辞書を構築する方法を提案しているが、検索結果数が少ないと関連語が抽出できないことから、新しい情報や特殊な情報には向いていないという問題点がある。

一方、大量のデータの中から規則や知識を発見する手法として、近年、データマイニングが注目されており、POS データの分析などに応用されている⁷⁾。データマイニングでは、対象データに頻出する組合せなどから、データ間に内在する関係を導く。しかし、日本語の WWW 検索サービスでは 90%以上が単一の検索

語入力との報告がある⁶⁾ことから、この手法を検索ログに直接適用することは困難である。

本論文では、検索ログに出現する検索語の使用頻度と、各検索語の使用された時間間隔とから語の関係を求める方法を提案する。本手法によれば、検索サーバに蓄積されている文書や背景知識に依存せず、情報ニーズを直接反映したタイムリーな語関係を求めることができる。また、関連の強い検索語をグループ化して集計することにより、情報ニーズの強さをより正確に求めることができる。以下、2 章では、検索語と情報ニーズとの関係を求めるための関連度の計算手法に関する基本的な考え方について論じる。3 章では、2 章で示した基本方針を実際の検索ログの解析に適用する場合の実現手法について述べる。4 章で実験結果を示すとともに考察を行う。

2. 関連語の抽出

本論文で対象とする WWW 検索サービスは、利用者から検索式（検索語の論理式）を受け取り、該当する WWW ページの URL の一覧を検索結果として返すものとする。

同じ概念を表す情報（以下、簡単に同じ話題と記す）を求めるために使用される検索語が異なる理由は、

- (1) 1 人の利用者がいくつかの視点から複数の検索語を使用した。
- (2) 複数の利用者がそれぞれの視点や立場に応じて異なる検索語を使用した。

という 2 つのタイプに大別できる。これらを検出するために各検索語間の関連度を計算し、関連の強い語どうしをグループ化する。

2.1 1 人の利用者の視点の違い

まず (1) のタイプについて考える。WWW 検索サービスを利用して自分のほしい情報（WWW ページ）を探す際の利用者の行動は、以下の手順となる（図 1 参照）。

STEP1 ほしい情報を得るための検索式を検討し入力する。

STEP2 検索サービスから受け取った URL 一覧を見て、自分の要求に合致する URL があるかどうかを評価し、該当する URL を選択する。

STEP3 情報提供サーバから得た WWW ページを閲覧する。

通常、求める情報が 1 回の検索で得られることは少ない。STEP2 で検索結果が不適だったり多すぎたりするときには、STEP1 に戻って、異なる検索語を入力したり、検索語の組合せを変えたりなどの試行錯

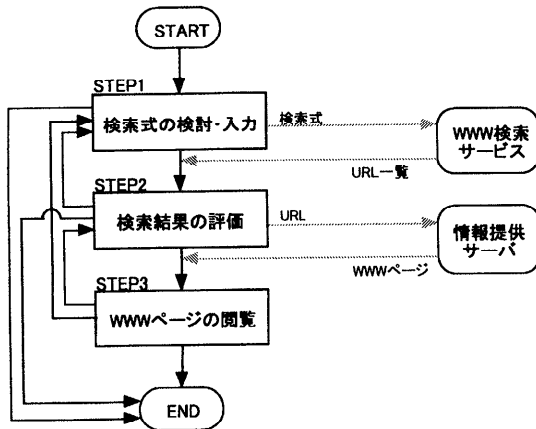


図1 WWW 検索サービスの利用者の行動

Fig. 1 Action model for WWW search service users.

誤による、連続した検索を行う。また、STEP3における閲覧の結果、目的の情報ではなかったと判断したときには、STEP2に戻って別のURLを選択したり、STEP1に戻って再度の検索を行う。このように、上記ステップのそれぞれの間を何度も繰り返しながら、求める情報を得るために行動する。この一連の検索行動における検索語は特定の情報を得るために使用されているので、これを検出すればよい。本論文では、同一利用者による検索の時間間隔に着目する。

多くのWWW検索サービスでは、STEP2の検索結果として、URL以外にタイトルやコメント、WWWページの一部などを含むため、その内容をある程度推測できる。このため、STEP3からの戻りは比較的少なく、STEP1とSTEP2との間での比較的短い時間間隔での頻繁な繰返しが多くなると予想される。一方、求める情報を得られたとき、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。検索終了後、別の情報が必要になって新たな検索を行うまでには、他の情報を閲覧したり、情報取得行動そのものを中断したりするため、この間の時間間隔は比較的長くなる。

これらのことから、STEP1とSTEP2との間の頻繁な繰返しに着目すると、同一の利用者によって使用された検索語は、その使用時間間隔が短ければ同じ情報を求めるために、長ければ別の情報を求めるために、それぞれ使用された可能性が高いと仮定できる。そこで検索語間の関連を使用時間間隔の単調減少関数と見なし、この関数を $assoc$ と呼ぶことにする。たとえば、利用者 i が時刻 s_0 に「桜」で検索し、 s_1 に「花見」、 s_2 に「開花」、 s_3 に「花見 and 開花」でそれぞれ検索したとする(図2参照)。「桜」「花見」「開花」の3つの検索語の使用された時間間隔は表1ようになる。

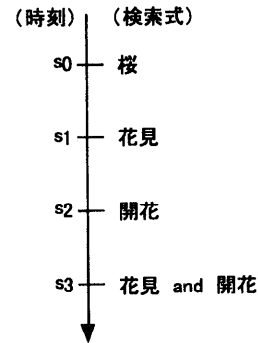
図2 利用者 i による検索のようす

Fig. 2 Retrieval time and keywords by a user.

表1 各検索語の使用時間間隔

Table 1 Time intervals in use between keywords.

	花見	開花
桜	$s_1 - s_0$ $s_3 - s_0$	$s_2 - s_0$ $s_3 - s_0$
花見		$s_2 - s_1$ $s_3 - s_2$ $s_3 - s_1$ 0

このように、同一利用者が同じ検索語を使った検索を何度も行う可能性があるため、2つの語の使用時間間隔は複数考えられるが、本論文では、このうちの最小値に着目する。利用者 i が使用した検索語 x, y の使用時間間隔の最小値を $tmin_i[x, y]$ と表せば、図2の場合、

$$tmin_i[\text{桜}, \text{花見}] = s_1 - s_0$$

$$tmin_i[\text{桜}, \text{開花}] = s_2 - s_0$$

$$tmin_i[\text{花見}, \text{開花}] = 0$$

となる。このとき、利用者全体での総和

$$T_{xy} = \sum assoc(tmin_i[x, y])$$

を x, y の間隔関連度と定義する。なお、関数 $assoc$ の具体的な与え方については、3章で考察する。

2.2 異なる利用者による視点の違い

間隔関連度は、同一利用者によって使用された複数の検索語の関連を求めているが、同じ話題を求める際に、様々な検索語を使用する利用者が少ない場合には、うまく関連度を計算できない。たとえば、「桜の花見」に関する情報を得るとき、「桜」と「花見」の両者を試してみる利用者が多くないと、これらが同じ情報を求めるために使用されたかどうかをうまく判定できない。そこで、これを補い、本章の初めに述べた(2)のタイプの検索語集合を求める手法について考察する。

本論文の目的は、情報ニーズの強さや傾向の把握で

ある。ニーズに一定の傾向（増減やその変化度合い）があれば、それを求めるための検索語も同一の傾向を示すはずである。たとえば、「桜」と「花見」は、ある一時期に特定の情報（「桜の花見」）を求めるために使用されると予想されるため、両者は同じような傾向で使用頻度が増減すると考えられる。そこで、1日単位で検索語の使用頻度を集計し、その時系列の相関係数によって、同一情報に対する検索かどうかを判定する。検索語 x, y の n 日間の使用頻度を、それぞれ、 X_j, Y_j ($j = 1, 2, \dots, n$) とするとき、相関係数 R_{xy} は、

$$R_{xy} = \frac{\sum(X_j - \bar{X}) \cdot (Y_j - \bar{Y})}{\sqrt{\sum(X_j - \bar{X})^2 \cdot \sum(Y_j - \bar{Y})^2}}$$

によって求められる⁸⁾ので、これを時系列関連度と定義する。ただし、 \bar{X}, \bar{Y} はそれぞれ、 X_j, Y_j の平均値である。

2.3 語のグループ化

関連の強い語は同じ話題を求めるために使用されたと考え、間隔関連度と時系列関連度に基づいて検索語をグループ化する。以下、検索ログに含まれる検索語の集合を $W = \{x_1, x_2, \dots, x_m\}$ 、 x_i を含むグループを $S[x_i]$ とし、初期値として、 $S[x_i] = \{x_i\}$ ($i = 1, 2, \dots, m$) とする。

x_i と x_j が同じグループに属するかどうかを判定する2値関数を $gcheck(x_i, x_j)$ とすると、2つの語の間隔関連度 T_{x_i, x_j} の大きな検索語対から順に、

$gcheck(x_i, x_j) = 1$ ならば $S[x_i]$ と $S[x_j]$ を併合という処理を行うことで検索語をグループ化することができる。間隔関連度が等しい場合には、時系列関連度の大きい対を先行して処理する。

次に、 $gcheck$ について考察する。各グループ $S[x_i]$ は、同じ話題を求めるために使用された語の集合なので、 $S[x_i]$ に含まれる語の間には、

$$\forall u, \forall v \in S[x_i] \text{ に対して,} \\ T_{uv} > T_0 \text{ または } R_{uv} > R_0 \quad (1)$$

が成り立つ必要がある。ここで、 T_0, R_0 は、2つの語に関連があると判定できる間隔関連度と時系列関連度のそれぞれの閾値である。ところで、 R_{uv} は、 u と v とが同じ話題の検索に用いられなくても大きな値をとりうる。たとえば、週末によく使われる語のように一定の周期で頻度が推移する語どうしや、たまたま同時期に行われたイベントに関連する語どうしなどは、それぞれ時系列関連度が大きくなる。そこで、2つのグループの併合には、少なくとも1対の語に関しては間隔関連度が大きいことを条件とする。すなわち、 $gcheck$ を、

$$gcheck(x_i, x_j) \\ = 1 \text{ iff } \exists u \in S[x_i], \exists v \in S[x_j] \text{ に対して,} \\ T_{uv} > T_0 \quad (2)$$

かつ

$$\forall u \in S[x_i], \forall v \in S[x_j] \text{ に対して,} \\ T_{uv} > T_0 \text{ または } R_{uv} > R_0 \\ = 0 \text{ otherwise}$$

と定義する。

関数 $gcheck$ によってグループ化を行う場合には、式(2)の条件により対象となる2つのグループを間隔関連度の大きな検索語対が必ず「橋渡し」する。したがって、ここで示したグループ化処理を行った場合には、2語以上から成る各グループに含まれる検索語について、式(1)の関係以外に、

$$\forall u \in S[x_i] \text{ に対して } \exists v \in S[x_i] \text{ が存在し,} \\ T_{uv} > T_0 \quad (3)$$

が成り立つ。なお、これらの処理は、各検索語の属するグループの検索(FIND)と、2つのグループの併合(UNION)という集合に対する基本操作で行え、効率的なデータ構造やアルゴリズムが提案されている⁹⁾。

3. 実験

我々は現在、NTT DIRECTORY[☆]において、登録されたホームページの情報に対する全文検索サービスとしてInfoBee検索¹⁰⁾を提供しており、多くの方に利用していただいている。本論文で示した手法について、実際のInfoBee検索ログ(対象期間:1997年2月28日~5月29日の13週間)を用いて評価実験を行った。なお、ここでの検索語とは、検索式中で空白で区切られた文字列を指す。したがって、「はなみ」と「お花見」と「花見」はすべて異なる語として扱われる。集計に際しては、英数字の1byte化、英大文字の小文字化、1byteカナの2byte化など、表記に関して簡単な正規化を行う。また、HTTP-cookie¹¹⁾によって利用者IDを発行し、それぞれの利用者を区別する^{☆☆}。以下では、検索語の使用頻度と使用人数とは同義で用いる。

まず、間隔関連度を求める際の関数 $assoc$ を決めるために、同一利用者による連続した検索要求の時間間隔 $tint$ と回数との関係を調べた(図3)。図3の曲線は、 $tint = t_1$ にピークを持つが、これは前章で示した検索行動における、STEP1とSTEP2の間の繰り返し周期を示していると考えられる。すなわち、 t_1

[☆] <http://navi.ntt.co.jp>

^{☆☆} HTTP-cookieによる識別は、正確にはブラウザ単位での識別である。

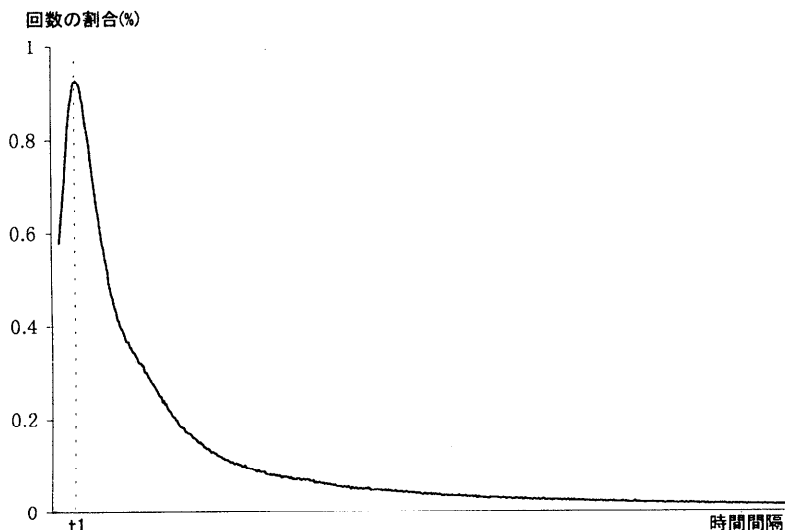


図3 検索の時間間隔の分布

Fig. 3 Distribution of time interval for consecutive request in a WWW search.

は、検索語がサーバに届いてから結果を受けとるまでの時間（システムの処理時間+通信時間）と、検索結果を評価して検索式を再検討・入力し、その検索リクエストが再度システムに届くまでの時間（人間の評価時間+通信時間）の和と見なせる。したがって、 t_1 前後までは一連の検索行動の可能性が高い。一方、ある時間間隔 t_3 を超えると、別の情報を求めるための新たな検索を行っていることが予想される。これらのことから、利用者 i の検索語 x, y の使用時間差の最小値を $tmin_i[x, y]$ としたとき、関数 $assoc$ を、

$$assoc(tmin_i[x, y]) = \begin{cases} a & (tmin_i[x, y] = 0) \\ 1 & (0 < tmin_i[x, y] \leq t_2) \\ (t_3 - tmin_i[x, y]) / (t_3 - t_2) & (t_2 < tmin_i[x, y] \leq t_3) \\ 0 & (t_3 < tmin_i[x, y]) \end{cases}$$

ただし、 $0 < t_1 < t_2 < t_3$

と定義する。関数 $assoc$ では、AND や OR などの検索条件とともに同時に使用された検索語 ($tmin_i[x, y] = 0$) に関しては、特に関連度が高いと考えて、特別な値としている（図4）。今回の実験では、 $a = 2$ 、 $t_2 = 60$ 秒、 $t_3 = 300$ 秒として間隔関連度を求めた。

検索語のグループ化の計算は1週間ごとに行う。グループ化は、過去1週間の検索ログに基づく間隔関連度と、過去2週間のログに基づく時系列関連度によって行う。また、情報ニーズの抽出という本論文の目的から、過去1週間で5人以上に使用された日数が3日間以上ある語のみを、その週の計算対象とする。 $gcheck$ で用いる R_0 は、危険率1%で独立でないかと判定できる相関係数の値とする。本実験では、

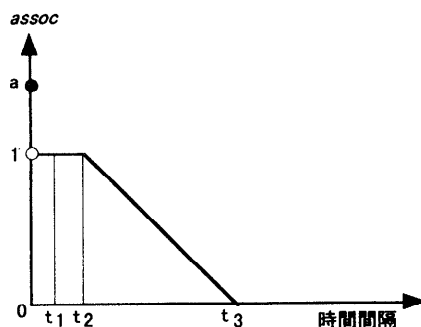


図4 時間間隔と間隔関連度の関係

Fig. 4 Time interval vs. $assoc$ value.

自由度 $= 14 - 2 = 12$ であることから、 $R_0 = 0.661$ となる⁸⁾。

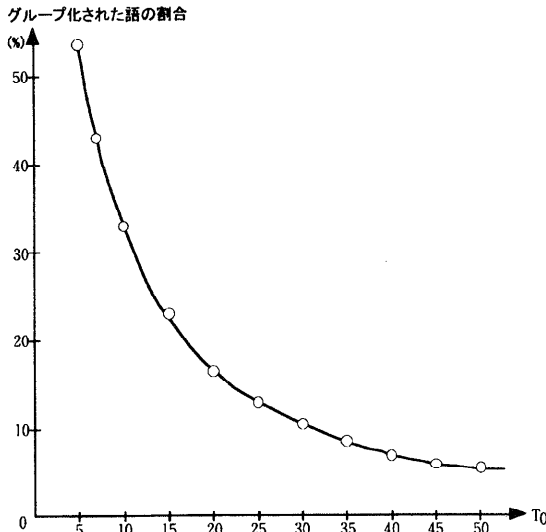
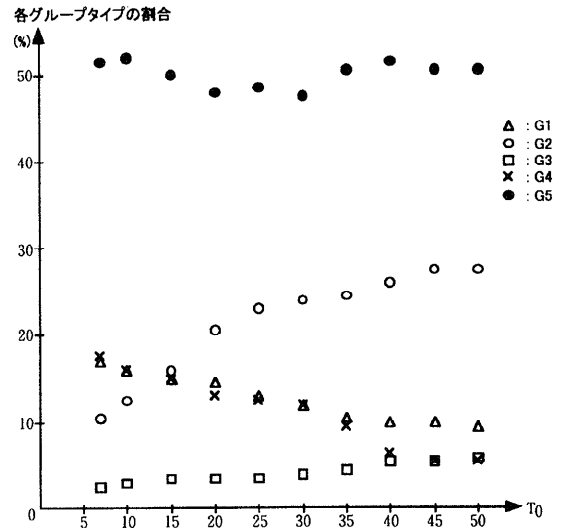
4. 結果および考察

4.1 結果と考察

各週において、検出された検索語は平均 94778 語で、そのうち計算対象となった語は平均 1640 語であった。また、時系列関連度が R_0 以上となる検索語対の数は平均 84731、 $T_0 = 20.0$ のとき間隔関連度が T_0 以上となる検索語対の数は平均 12921、グループ化された語数は平均 272 語、グループ数は平均 124 であった。

グループ化された語には、以下のような特徴がみられた。

(G1) 意味的に近い語や同義語。企業名や商品名などのアルファベット表記とカタカナ表記、正式名称と略称など。例：{オートバイ, バイク}, {jis,

図5 T_0 とグループ化された語の割合の関係Fig. 5 T_0 vs. grouped words ratio.図6 T_0 と各グループタイプの割合の関係Fig. 6 T_0 vs. group type ratio.

日本工業規格}, { ml, メーリングリスト }, など.

(G2) 複数の語がまとまって1つの概念を表すもの. 特に2語以上から成る英語の固有名詞(商品名やサービス名, 歌手名など)に多かった. 英単語は空白で区切って入力されるため, 別の語として集計されるが, 実際にはそれらがまとまって1つの意味を持つものがグループ化されたと考えられる. 例: { white, house }, { home, page }, など.

(G3) 特定の事件やイベント, 季節などに関連して, 一時的に同義で用いられたと考えられる語. 例: { 桜, 花見 }, { サッカー, ワールドカップ }, { 競馬, 天皇賞 }, など.

(G4) 新聞社どうしなど同業種の企業, より上位の概念によってまとめられる語. 例: { 旅館, ホテル, 民宿 }, { 姓名判断, 占い, 四柱推命 }, など.

(G5) 特定の情報を求める人が多かったと考えられるもの. 例: { 携帯電話, 無料, phs, プレゼント }, { パソコン, 中古, 販売, 価格, 通販 } など.

T_0 の値とグループ化された語の割合との関係を図5に, T_0 の値と G1~G5 の割合の関係を図6に示す. 図6では, 各グループ内の2語の組合せが, G1 から G5 のどのタイプに属するかを決定し, 重複を省いて集計した. たとえば, { オートバイ, バイク, 中古 } というグループに対しては,

G1: { オートバイ, バイク }

G5: { オートバイ, 中古 }

G5: { バイク, 中古 }

という3つの組合せとなり, G1, G5 として集計す

る. また, 対象となる2語を組合せたときに固有名詞となる場合にのみ G2 として集計した. したがって, { white, house } は G2, { 健康, 食品 } は G5 とした. 図6から明らかなように, 本手法では, 特に G5 のタイプのグループ化が多く行われる. この点は, 従来のシソーラスなど, 同義語を中心とした語のグループ化と最も異なる点と考えられる. また, T_0 を小さくすると, 同一利用者による使用が少ない語どうしのグループ化が増加するため, 特に, G1 と G4 のタイプが増えることが, 図5, 6より分かる.

また, T_0 の値により, ある語と同じグループに入れられる語の種類や数は異なる. たとえば, $T_0 = 50.0$ のときには, { 東京, ホテル } がグループ化され, $T_0 = 10.0$ では, { 東京, 東京都, 大阪, 横浜, 京都, 神戸, ホテル, ビジネスホテル, 宿泊, 旅館 } がグループ化された. 前者からは, 『「東京のホテル」に対する情報ニーズが大きい.』, 後者からは, 『「大都市の宿泊施設」に対する情報ニーズが大きい.』と, それぞれ読みとることができる. T_0 の値をどう設定すべきかは, どのような情報ニーズを把握したいのか, それぞれの立場に応じて決定すればよいと考えられる.

次に, 間隔関連度のみでグループ化を行った場合と, 時系列関連度と組み合わせてグループ化を行った場合とを比較した. 間隔関連度のみでグループ化を行う場合には, 関数 $gcheck$ は,

$$\begin{aligned}
 &gcheck(x_i, x_j) \\
 &= 1 \quad \text{iff } \forall u \in S[x_i], \forall v \in S[x_j] \text{ に対して,} \\
 &\quad T_{uv} > T_0 \\
 &= 0 \quad \text{otherwise}
 \end{aligned}$$

となる。このため、3語以上のグループを生成する際の条件が厳しくなり、2語から成るグループの割合が大きくなる(図7)。

なお、検索ログを予備調査したところ、たとえば競馬のように、大きなレース(桜花賞や皐月賞等)が短い周期で行われると、「競馬」と「桜花賞」、「競馬」と「皐月賞」がそれぞれの週では同義語的に用いられていると見受けられた。このため本実験では間隔関連度を求める期間を7日、時系列関連度を求める期間を14日とした。これらの期間の設定値の決め方については

2語から成るグループの割合

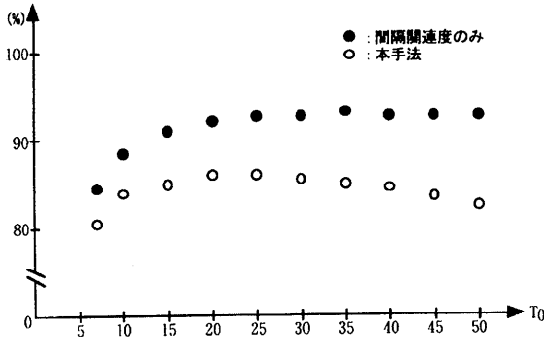


図7 T_0 と 2語から成るグループの割合の関係
Fig. 7 T_0 vs. 2 words group ratio.

今後の検討課題である。

4.2 情報ニーズの抽出例

「桜」に関する解析結果(図8)を例に、本手法の効果について考察する。図8のグラフは、3月1日から4月25日までの「桜」の1日の使用頻度(使用者数)の推移を示している。また、グラフの下の表は、「桜」と間隔関連度の高い語の上位3個を示している。たとえば、点線の枠で囲んだ部分は、「3月21日から27日までの1週間において、「桜」という検索語は「花見」「写真」「開花」という検索語と、この順序で間隔関連度が高かった」ことを示している。各区間での間隔関連度の高い語から、

- 3月中旬までは「桜前線」「開花」など、桜の咲き始める時期を求める要求が多かった。
- 3月下旬以降は、「高遠」「造幣局」「北海道」など、桜の名所に関する情報要求が多く、その場所は時とともに北上していく。

ことが分かる。なお、1997年は、東京における桜の開花宣言は3月21日に出され、また造幣局の通り抜けは4月21日から行われており、上記にほぼ一致している。

次に、図9に、 $T_0 \geq 5$ のときの T_0 の値と「桜」とグループ化された語との関係を示す。図9では、たとえば3月27日の週は、 $T_0 \geq 9.0$ で「桜」と「花見」が、 $T_0 < 9.0$ で「桜」と「花見」と「お花見」が、それぞれグループ化されていることを示す。図9より、3月から4月中旬までは、「桜」と「花見」はほぼ同義

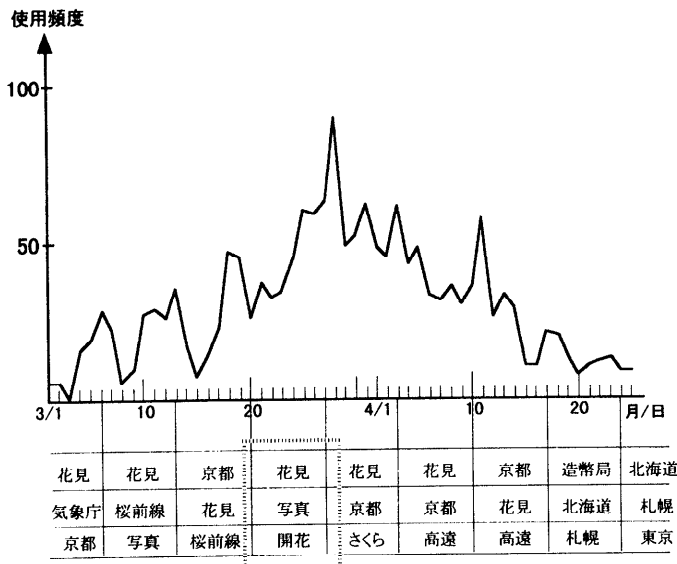


図8 「桜」の使用頻度の推移と、間隔関連度の高い検索語

Fig. 8 The number of uses of keyword “桜”, and a table of keywords related to it.

で用いられていることが分かる。 $T_0 = 20.0$ とすると、3月7日から4月10日までは「桜」と「花見」は同一の要求としてグループ化される。さらに、4月3日の週には、「桜」と「花見」と「さくら」がグループ化される。

各グループの使用頻度は、グループ内の各検索語の使用頻度の和から、同時に使用された回数を減じることによって求められる。図 10 のグラフの点線は、 $T_0 = 20.0$ で「桜」とグループ化された語の使用頻度を表している。たとえば、「桜」と「花見」がグループ化されている場合には、「桜」と「花見」の各使用回数の和から両者の同時使用回数を減じたものである。グループ化によって頻度を求めることにより、実際には、この期間に「桜の花見」に関して、「桜」単独の約 1.5 倍の要求があったことが分かる。

このように、間隔関連度の大きな語の提示と、検索語のグループ化により、

- ある特定の期間になぜ検索要求が増えたのか、利用者が何を求めているのかが分かる。
- 個々の検索語の使用頻度からは分からない、情報ニーズの真の強さが分かる。

といった効果がある。

5. おわりに

情報ニーズの抽出を目的として、WWW の検索ログから検索語間の関連度を計算しグループ化する手法を提案した。本手法では、検索語の使用頻度と使用された時間間隔とから、情報ニーズを直接反映したタイムリーな語関係を求めることができる。また、実際の検索ログを用いた評価実験によって、本手法の効果を示した。語の関連度を求めることにより、絞り込み検索のための検索語候補提示や、曖昧検索への応用も可能である。

情報検索では、これまでは対象となる情報集合は given として考えられてきた。しかし、インターネットのように、発信される情報が日々変化し、かつ急増している世界では、利用者の情報ニーズに適合した情報収集や情報提供インタフェースが重要であり、これが、ネットワーク上に散在する情報資源の有効活用につながると考えられる。

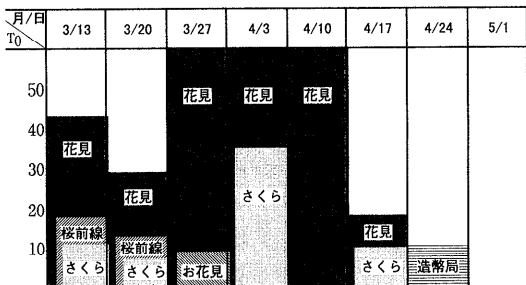


図 9 T_0 の値と、「桜」とグループ化された語との関係
Fig. 9 Keywords to be grouped with “桜” vs. T_0 .

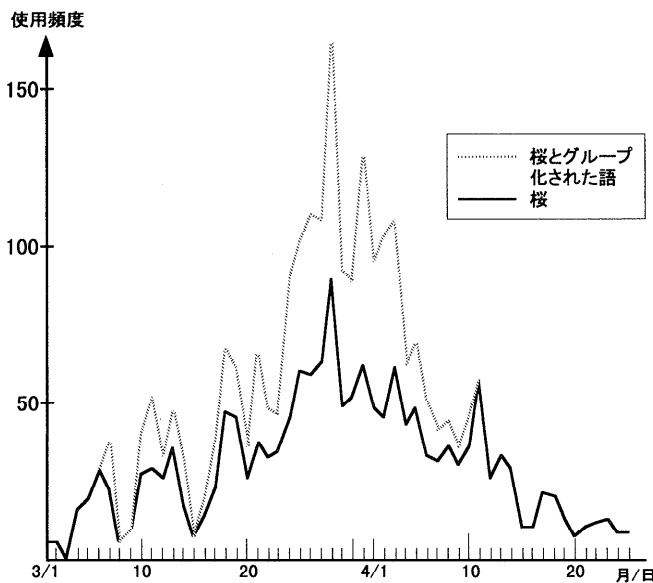


図 10 「桜」および「桜」グループの使用頻度の推移
Fig. 10 The number of uses of keyword “桜” and keywords grouped with “桜”.

参 考 文 献

- 1) Conte, R., Jr.: インターネット上の情報検索, *Internet Working*, Vol.2, No.7, pp.3-17 (1996).
- 2) 青木, こばやし, 根本, 乗越: 特集 検索エンジン使いこなしテクニック, *日経ネットナビ*, Vol.2, No.9, pp.68-89 (1997).
- 3) Srinivasanm, P.: *Thesaurus Construction, Information Retrieval: Data Structures & Algorithms*, Frakes, W.B. and Baeza-Yates, R. (Eds.), pp.161-218, Prentice Hall, NJ (1992).
- 4) 西村, 伊藤, 河野, 長谷川: 重み付き相関ルール導出アルゴリズムによる WWW データ資源の発見, 第 7 回データ工学ワークショップ (DEWS'96), 神戸, pp.79-84 (1996).
- 5) 丹羽: 動的な共起解析を用いた対話的文書検索支援, *情報処理学会研究報告*, 96-NL-115, pp.99-106 (1996).
- 6) 原田, 清水: WWW 検索システムにおける不特定多数の操作履歴の活用, *情報処理学会研究報告*, 97-DPS-81, pp.61-66 (1997).
- 7) 沼尾ほか: 特集 大規模データベースからの知識獲得, *人工知能学会誌*, Vol.12, No.4, pp.496-549 (1997).
- 8) 高橋, 出居, 小林, 小柳: *工科の数学 5, 統計・数値解析*, 培風館, 東京 (1969).
- 9) Aho, A.V., Hopcroft, J.E. and Ullman, J.D.: *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA (1974).
- 10) 田中: InfoBee 検索エンジンを用いたディレクトリ検索サービス, *NTT 技術ジャーナル*, Vol.8, No.8, pp.24-27 (1996).
- 11) Kristol, D. and Montulli, L.: HTTP State Management Mechanism, RFC 2109 (1997).

(平成 9 年 9 月 16 日受付)

(平成 10 年 5 月 8 日採録)



大久保雅且 (正会員)

1961 年生. 1985 年京都大学工学部情報工学科卒業. 1987 年同大学院修士課程修了. 同年, 日本電信電話(株)に入社. 現在, NTT ヒューマンインタフェース研究所主任研究員. 主に, 情報資源に内在する構造の発見とその応用技術の研究に従事. ACM 会員.



杉崎 正之 (正会員)

1970 年生. 1993 年東京理科大学理工学部情報科学科卒業. 1995 年同大学院修士課程修了. 同年, 日本電信電話(株)に入社. 現在, NTT ヒューマンインタフェース研究所勤務. 主に, テキスト情報の自動分類技術の研究に従事.



井上 孝史 (正会員)

1965 年生. 1990 年京都大学工学部電気系学科卒業. 1992 年同大学院修士課程修了. 同年, 日本電信電話(株)に入社. 以来 1998 年 4 月まで NTT ヒューマンインタフェース研究所で自然言語処理や情報検索の研究に従事. 現在, NTT プリンテック(株)に勤務.



田中 一男 (正会員)

1957 年生. 1979 年神戸大学工学部電子工学科卒業. 1981 年同大学院修士課程修了. 同年, 日本電信電話公社に入社. 1988~1990 年スタンフォード大学客員研究員. 現在, NTT ヒューマンインタフェース研究所主幹研究員. 情報検索等の情報資源活用技術の研究に従事. 人工知能学会, 日本認知科学会, AAAI 各会員.