論文

長期間利用される検索キーワード検出手法の提案

The proposal of the retrieval keywords discovery Method used for a long period of time

もっとも主張したい点

近年、Webのソーシャル化に伴い、情報の伝播速度が上がり、情報の流行り廃りが激しくなってきた。そのため、一時的には面白かったり有用であったりするが、長期的に見ると重要でない情報が増加してきた。

そこで、本研究では、長期間利用する情報という点に着目し、そういった情報を取得するための手法を提案する。また、どれだけ情報が長期間利用できるかを計算する手法を提案し、長期間アクセスされ続けている情報を取得するシステムを開発した。

要旨

近年、Webのソーシャル化に伴い、一部のWeb情報が急激に流行る傾向がでてきた。例えば、Twitterやソーシャルブックマークサービスなどの影響により、情報が一気に伝搬して流行ることが多くなってきた。こういった情報の中には、流行りが過ぎてしまうと、あまり有用ではなくなるものも多く、賞味期限の短い情報もどんどん増加してきている。

そのため、筆者らはその情報がどれだけ長期間利用されているかの指標、いわば情報の長期度のような指標が重要になってきたのではないかと考える。そこで、本研究では長期度の計算手法を提案し、長期間利用され続けている情報を取得するためのシステムを提案する。そして、検索キーワードがどれだけ長期間に渡って、検索され続けているかという時系列情報を利用して、Web上で長期間利用される検索キーワード一覧を取得するシステムを開発した。

ABSTRACT

In recent years, with socializing of Web, some part of Web information prevails rapidly. For example, by the influence of Twitter and social bookmark services, information comes through at a burst. Such information, when it goes out of fashion, most of it could be useless information. Information, the best-before date of which is short, is also increasing rapidly.

Thus, we think it becomes important that the information is used for a long period of time. Therefore we proposed a system that can get the information used for a long period of time. In addition, we developed search keyword retrieval system that can get used for a long period of time for search keyword. The system uses time information of how long keywords are searched.

キーワード:情報検索、検索キーワード、関連キーワード、長期間、ロングセラー

1 はじめに

1.1 研究の背景

近年、Web検索エンジンを利用して必要な情報を探すという行為が一般的に行われているが、Webの情報はどんどん膨大になってきているため、容易に必要な情報にたどり着けない場合もある。検索エンジンのアルゴリズムは、様々な手法が提案されているが（北 [2002], 兼宗 [2004]）、GoogleのPageRank（Page [1998]）が特に有名である。PageRankは主に被リンクを用いて、人気のページをランキングしている。また、この他にも流行の情報を発見する手法も多く提案されている。Kleinberg [2002] は、時系列データからburst[1]を検出する手法を提案しているし、こういった手法を応用して、実際に流行の情報を発見する研究も行われている（奥村 [2004]）。

このように人気の情報や流行の情報の発見する様々な手法が広く提案・利用されている一方で、長期間コンスタントに利用され続けるような情報に特化して取得する手法はほとんど存在しない。これに対して筆者らは、長期間コンスタントに利用されているということは、長期的に見て有用な情報である可能性が高いと考えた。また、近年では製品やコンテンツの寿命が短命化してきているというという背景があり、このことからも長期間利用できるモノや情報は見つけにくくなってきているのではないかと考えた。この考えに基づき、筆者らは2009年に長期間利用され続けるWebページを発見するシステム「セレクトブクマ」（上野[2010]）の開発を行った。

ただ、一口に長期間利用され続ける情報といっても、それはWebページだけにとどまらずに、多くの情報が考えられる。そこで筆者らは、本論文でどれだけ長期間利用されているかを表す指標である長期度という指標を提案し、これを検索キーワードに適用して、長期間コンスタントに検索され続けている検索キーワードを取得するシステムを提案する。本論文で、検索キーワードを対象とした理由と長期間利用する検索キーワードを取得する目的については、1.2に示す。

1.2 研究の目的

Web上からユーザが有用な情報を検索するためには、適切な検索キーワードを入力する必要がある。だが、誰もが適切な検索キーワードを選択できるとは限らないし、検索キーワードが思い浮かばないという問題も存在する。

こういった問題を解決するために、検索クエリを拡張する手法の研究（Qiu [1993] ）や検索キーワードに関連するキーワード一覧を提示する研究（大塚 [2005] ）が行われてきた。また、Googleサジェスト[2]などのサービスも提供されてきた。これらは、検索キーワードの類似性や検索回数の多さを利用して、推薦するキーワードを選出している。

こういった背景に対して、筆者らは長期間コンスタントに検索され続けている検索キーワードは定番の検索キーワードであり、有用な検索キーワードである可能性が高いと考えた。また、自分が不慣れな分野を調べるときに、その分野で長期間コンスタントに利用されている検索キーワードが分かれば、その分野を体系的に調べることが可能になるのではないかと考えた。

そこで本研究では、キーワードがどれだけ長期間コンスタントに検索され続けているかに着目し、調べたい分野の中で定番の検索キーワードを取得する手法を提案する。

2 提案手法

2.1 長期度の計算

情報がどれだけ長期間利用され続けているかの指標として、長期度という指標を定義し、長期度を計算する手法を提案する。長期度を計算するために、ここでは検索回数の時系列データを利用する。最初に、実際の時系列データを回数の多い順に並べる。次に、実際の時系列データの値の大きさ

から、べき乗則に基づくデータを生成する。そして、実際の時系列データとべき乗則に基づくデータの差分の大きさを計算し、これを長期度とする。実際に数式で示すと以下のようになる。

$$長期度=\sum_{k=1}^{n}(\alpha_k-\beta_k)$$

α_k：単位時間ごとの検索回数やアクセス回数

β_k：べき乗則に基づくデータ

ここでは、両対数グラフにおいて縦軸の最大値と横軸の最大値を結ぶ直線を描くようなデータとし、べき乗則の指数係数の値を設定した。

ここで、このような手法で長期度を求める理論を説明する。まず、複雑な条件に基づいて形成されるデータは、べき乗則に基づくという仮定(Barabasi [2002])や単語の使用頻度は、べき乗則に基づく(Booth [1967])などの理論を利用して、多くのキーワードは、単位時間ごとのアクセス数がべき乗則に基づく可能性が高いと仮定する。ここで、説明のため、横軸を単位時間ごとのアクセス数の順位、縦軸をアクセス数として、グラフを書く(図 1)。べき乗則に基づくデータは、図の実線のようになり、長期間利用されるデータの分布は、破線のようになる。べき乗則に基づく分布は、ロングテールとなる。このため、単位時間ごとのアクセス数順に並べた場合、長期間利用される分布は、べき乗則に基づく分布と比較して、大きな値となる期間が長いということが分かる。逆に、急速に流行って廃れていくものは、グラフの落ち方が急激になるため、べき乗則に基づく分布と比較して小さな値となる期間が長い。

以上のような法則を利用して、実際のデータとべき乗則に基づくデータの差分が大きければ大きいほど、長期度が高いとすることができる。また、分野によって、べき乗則との差分の値も変わってくる。そのため長期度は、べき乗則との差分の値が同じ分野においてどの程度大きいかで計算する。

なお、指数係数の値を変数とし、この値を小さな値にすることで長期間利用されるデータの分布を近似し、指数係数の大きさを長期度の指標とすることも考えられる。だが、今回の場合は長期間検索され続ける検索キーワードの中には、平均値に近い検索回数が多くなり、べき乗則に従わないものも存在するため、べき乗則に基づく分布との差分とした。

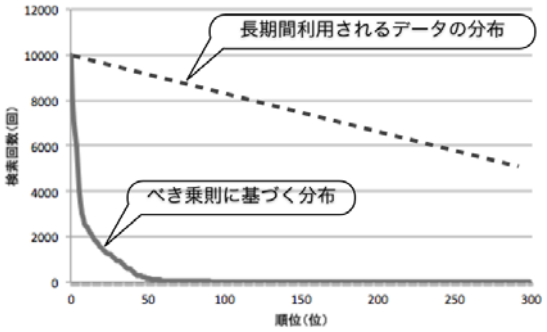

図 1 べき乗則のデータと長期間利用データ

2.2 提案するシステム

2.2.1 関連キーワード収集

関連キーワード取得は、既存の手法である、Lingua-JA-Expand[3]を利用する。

Lingua-JA-Expandは、以下の手順で関連キーワードを取得している。

・キーワードを受け取る

・Yahoo Search APIを利用してYahoo検索結果のスニペット[4]を取得

・TF-IDFによる計算を利用して、関連キーワードと関連度を取得

この手法で様々なキーワードで試してみたところ、関連キーワードの精度に不足を感じた。具体的に言うと、スニペットを利用しているため、どうしても取得したキーワードの特徴を表す単語が関連キーワードとして多く出てきてしまう。そのため、この関連キーワード一覧からキーワードを取得する逆引きによって、関連キーワードを取得した方が、より本システムに向いている関連キーワードが取得できるのではないかと

考えた。

そのため、Wikipediaから取得したキーワードとLingua-JA-Expandを利用して、関連キーワードデータベースを作成した。関連キーワードデータベースのデータ量は以下の表1のようになっている。

表1 関連キーワードデータベースのデータ量

キーワード	関連キーワード
約120万キーワード	約2500万キーワード

2.2.2 検索キーワードの時系列データ収集

検索キーワードの時系列データは、Google Insights for Search[5]というサービスのデータをクローリングする。Google Insights for Searchでは、2004年以降のGoogle検索におけるキーワードの1週間ごとの検索量を提示している。

この検索キーワードの時系列データに2.1で示した長期度の計算方式を適用して、長期度を計算する。そして、長期度が高い順にランキングする。

2.2.3 システム表示

本システムでは、ユーザがキーワードを入力するとその分野に関連するキーワード一覧が表示される。以下の図2は、例として「yahoo」というキーワードを入力して関連キーワード一覧を取得した画面である。画面には、長期度が高い順から上位10件のキーワードを提示する。また、検索キーワードは、リンクとして、リンク先はそのキーワードでの検索結果とする。図の横棒は、長期度の大きさを表している。

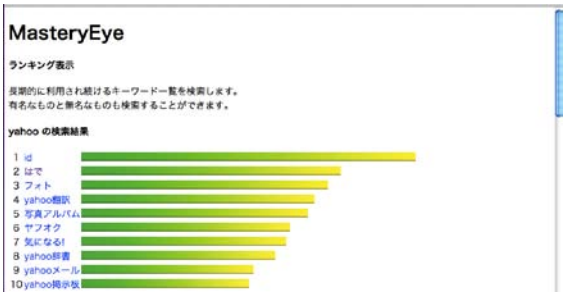

図2 「yahoo」での検索結果画面

3 評価実験

3.1 実験目的

評価実験は、以下3つを示すことを目的として行った。

① 長期度の計算手法の正当性

② 長期間利用するものは重要である可能性が高いこと

③ 開発したシステムの有用性

3.2 実験方法

評価実験では、以下の3種類の手法で関連キーワードを取得して比較実験を行った。

手法1:提案手法で長期度が高かったもの

手法2:提案手法で長期度が低かったもの

手法3:既存手法(reflexa[6])

手法1と手法2については、提案手法で長期度が高かった関連キーワード、低かった関連キーワードをそれぞれ上位10件ずつ利用した。また、手法3のreflexaとは、連想検索エンジンで、入力したキーワードと関連の深いキーワードを提示するシステムである。既存手法にreflexaを選んだ理由は、一般公開されているシステムであるからと、非常に多くのキーワードに対して関連キーワードを取得できるからである。

以上の3種類の手法それぞれについて、Googleで検索回数の多い上位4位のキーワードを利用して実験を行った。検索回数の多いキーワードは、Google Insights for Searchの情報を参照した(2012年4月5日)。

実験を行うにあたって、Googleでの検索回数が多いキーワードを選んだ理由は、本システムを利用するユーザ層として想定しているのが、検索キーワードがあまり思い浮かばないようなユーザを想定しているからである。そういった、検索リテラシーがあまり高くないユーザは、少しマイナーなキーワードよりも、Googleでの検索回数が多いようなメジャーなキーワ

ードの方が思い浮かぶ可能性も高いであろうと考えた。そして、こういったメジャーなキーワードから関連するキーワードを取得して、それらの元のキーワードに関連していて、かつ、長期間利用され続けているかを評価したいと考えた。例えば、以下の検索回数上位2位の「動画」というキーワードに関して、本システムで「動画」と入力すると動画に関連する検索キーワードの中で、長期間検索され続けているキーワードが提示できれば、動画に関連する定番の検索キーワードを知ることができるのではないかと考えた。

1位:Yahoo

2位:動画

3位:YouTube

4位:画像

以上の4つのキーワードの関連キーワードを10キーワードずつ3種類の手法で、合計120キーワード取得した。重複したキーワードを1つにまとめて、合計116キーワードをランダムに並べて被験者に提示した。被験者の属性は、表2に示す。被験者に合計116キーワードそれぞれに対して、以下の3つの質問項目に当てはまるものを選んでもらった。

質問1:この中であなたが知っているキーワードを選んで下さい

質問2:この中であなたが長期間利用してきたキーワードを選んで下さい

質問3:この中であなたが重要だと思うキーワードを選んで下さい

これらは複数選択可とし、選択数に制限は設けなかった。

表2 被験者属性

人数	20名
性別	男性:12名、女性:8名
年代	20代:15名、30代:2名、40代:2名、 50代:1名

3.3 実験結果

評価実験のために取得した関連キーワード一覧は、表9に示す。取得した関連キーワード一覧を見てみると、提案手法で長期度が高いものは、一部関連性が低そうなキーワードもあるが、ほとんどのキーワードは、元のキーワードに関連していることが分かる。提案手法で長期度が低いものは、関連性が高いキーワードが多いが、あまり多くの人から利用されていないキーワードも多いことが分かる。また、少し詳しく見てみると、例えば「動画」に対しての「アメーバビジョン」や「YouTube」に対しての「字幕.in」など、以前少し流行ったが、現在はあまり利用されていないサービスなども多い。relexaについては、元のキーワードと関連していて、なおかつ一般的なキーワードが多いが、特に「Yahoo」に関連するキーワードで、あまり利用されていないキーワードが多いことが分かる。

これらに対して、評価実験を行った結果を以下に示していく。評価実験で各質問項目について、ユーザが選択したキーワード数の評価を行った。3種類の手法に対して、選択されたキーワード数の合計数を表3に示す。ここで、被験者のうち誰か1人でも選択したキーワードに対して、選択されたキーワードとした。

表3 選択されたキーワードの数

	手法1*	手法2*	手法3*
質問1: 知っている	35	24	24
質問2: 長期間利用	32	14	21
質問3: 重要である	29	12	18

*手法1 提案手法で長期度の高いもの

*手法2 提案手法で長期度の低いもの

*手法3 既存手法(reflexa)

これらの選択されたキーワードの数が確率分布に基づく期待値と有意差があるかどうかカイ二乗検定[7]によって評価を行った。

最初に、3種類の手法に対して、各質問で選択されたキーワードが期待値と有意差があるかどうか、カ

イ二乗検定を行なった。質問1の「知っているかどうか」、質問2の「長期間利用しているかどうか」、質問3の「重要かどうか」の3種類の質問項目に対して行った(表4～表6)。

その結果、「知っているかどうか」については、p値は0.233(小数点第4以下四捨五入)となり、有意差は求められなかった。「長期間利用しているかどうか」については、p値は0.025(小数点第4以下四捨五入)となり、有意水準5%以下で有意差を求めることができた。「重要かどうか」については、p値は0.023(小数点第4以下四捨五入)となりこちらも有意水準5%以下で有意差を求めることができた。

表4 カイ二乗検定(知っているかどうか)

	手法1	手法2	手法3	計
観測度数	35	24	24	83
期待度数	27.67	27.67	27.67	83
p値	0.233			

表5 カイ二乗検定(長期間利用かどうか)

	手法1	手法2	手法3	計
観測度数	32	14	21	67
期待度数	22.33	22.33	22.33	67
p値	0.025			

表6 カイ二乗検定(重要かどうか)

	手法1	手法2	手法3	計
観測度数	29	12	18	59
期待度数	19.67	19.67	19.67	59
p値	0.023			

次に、長期度の計算手法の正当性を示すため、提案手法で長期度が高かったもの(手法1)と低かったもの(手法2)で取得したキーワードに対して、「長期間利用しているかどうか」に選択されたキーワード数の比較を行った(表7)。

その結果、選択されたキーワード数は手法1の方が多くなった。また、p値は0.0004となり、有意水準1%以下で有意差を求めることができた。このことから、システムで取得した長期度が高いものの方が、長期度が低いものよりも長期間利用されているキーワードであることが分かる。

表7 カイ二乗検定(長期間利用かどうか:長期度高と長期度低の比較)

	手法1	手法2	手法1	手法2	計
	選択あり	選択あり	選択なし	選択なし	
観測度数	32	14	7	26	79
期待度数	23	23	16.5	16.5	79
p値	0.0004				

また、長期間利用されているものが重要であるかどうかを調べるために、質問2の「長期間利用しているかどうか」と質問3の「重要であるかどうか」の質問項目に回答されたキーワードの関係性を調査した。質問2のみ回答されたもの、質問3のみ回答されたもの、質問2と質問3に重複して回答されたもの、どちらにも回答されなかったものの4種類に対してカイ二乗検定を行なって、有意差を求めた(表8)。

その結果、「長期間利用している」かつ「重要である」キーワード数が期待値より多くなった。また、p値は3.04×10^{-10}となり有意水準1%以下で有意差を求めることができた。この結果から、長期間利用しているものは重要である可能性が高いということが言える。

表8 カイ二乗検定(長期間利用と重要なものの関係)

	長期	重要	長期&重要	選択なし	計
観測度数	15	7	52	44	118
期待度数	33.5	25.5	33.5	25.5	118
p値	3.04×10^{-10}				

表9 取得した関連キーワード一覧

元キーワード	長期度の高いもの	長期度の低いもの	reflexa
Yahoo	id	リローンチ	OL 蔡桃桂
	はで	月刊 4b	ポアロのあと何分あるの?
	フォト	アリババグループ	すときゃ!
	yahoo 翻訳	東京めたりっく	漫畫
	写真アルバム	津乃村真子	関戸優希
	ヤフオク	キャロル・バーツ	Pheonix
	気になる！	リアルタイム検索	新浪
	yahoo 辞書	ヤフコメ	FQDN
	yahoo メール	みんなの検定	寶輔
	yahoo 掲示板	ポケモンガーデン	蔡桃
動画	動画サイト	アニタン	請
	動画編集	いじめ動画	原画
	サンプル動画	佳山三花	MPEG-4
	YouTube	日本動画協会	作画監督
	3gp	車載動画	YouTube
	デジタル動画	ニコニコ動画物語	東映動画
	ようつべ	動画大陸	H.264
	動画プレーヤー	ニコニコ組曲	コーデック
	日本動画	アメーバビジョン	MPEG
	ニコ動	グロ動画	DivX
YouTube	orbit	the 八犬伝	Stage6
	動画	楽珍トリオ	ニコニコ動画
	話	ヒピラくん	Google
	ようつべ	ユーチューブ xl	DivX
	3d 動画	陳士駿	PayPal
	ビデオソフト	ろうきゅうぶ	MOCO
	c-8	アキラブ	チャド・ハーリー
	fooooo	著作権管理	GUBA
	woopie	字幕.in	Veoh
	動画投稿サイト	私が恋愛できない理由	政見放送
画像	画像編集	死亡時画像診断	Ferrari
	画像掲示板	磁気共鳴画像	darkgreen
	おもしろ画像	衛星画像	ビットマップ
	画像診断	蓮画像	トランスミッション
	デジタル画像	綾波セナ	ビットマップ画像
	画像安定装置	ビットマップ画像	ピクセル
	画像処理	バカ画像	ストラット
	おすすめ画像	グロ画像	JPEG
	画像ビューア	レタッチソフト	クロスオーバーSVC
	画像認識	画像作成ソフト	トールワゴン

4 関連研究

4.1 情報の賞味期限

情報のライフサイクルや賞味期限に関する研究としては、例えば、竹井［2004］は、情報の価値と情報の時間的変化に基づくライフサイクルについて検討し、文書のライフサイクル管理と情報ライフサイクル管理の融合を検討している。また、Takahashi［2008］らは、情報の鮮度に着目し、ソーシャルブックマークの時間データを利用して、鮮度が高く賞味期限の切れていない Web ページを取得する手法を提案している。

これらの研究は、長期的な視点で情報の利用のされ方に着目しているため、本研究と近い。だが、竹井［2004］の研究は、Web 上から情報を発見するための研究ではなく、Takahashi［2008］の研究は、長期間利用される情報というより、情報の鮮度に着目している点が本研究と異なる。

4.2 関連キーワードの取得

関連キーワード取得に関する研究もさかんに行われてきた。今井[2010］らは、入力した検索クエリが多義語の場合、その後、選択された Web ページに基づいてクエリを推薦する手法を提案している。

また、Roy［2002］らは、大量のテキストデータから流行の単語を抽出する研究を行なっている。さらに、Rech［2007］では、Google Trend[8]という検索キーワードの検索回数の時系列データを取得できるサービスを利用して、流行りのキーワードを取得している。

これらの研究は、キーワードを取得する点で本研究と関連が深い。特に、Rech［2007]の研究では、検索キーワードの時系列データを利用してキーワードを取得している点で本研究に近い。これらと比較して、本研究は、関連キーワードの中でも長期間利用される検索キーワードに着目してキーワードを取得しているところが特徴的である。

4.3 ロングセラー

長期間利用されるという考え方とロングセラーという考え方は類似している。ロングセラーに関する研究は、主にマーケティング分野で古くか行われている。Hermann［1979］は、ブランドのライフサイクルとロングセラー化によるメリットについて述べている。また、寺本［2009］では、最寄品ブランドの小売店頭での販売展開方法とロングセラー化の関係について定量的に捉え，最寄品ブランドのロングセラー化に向けた店頭展開要件を抽出している。

本研究では、こういったロングセラーの考え方も参考にしつつ、長期間利用される、いわば情報のロングセラーに近い概念の研究を行なっている。

5 考察

実験結果から、本研究で提案した長期度計算方式で長期度が高かったキーワードは、長期度が低かったキーワードと比較して長期間利用されていることが分かった。また、長期間利用されているキーワードは重要であるキーワードと重複している確率が有意に高かったため、長期間利用されているものは重要である可能性があることが分かった。ただし、ここで言う重要という指標は、あくまで個人の主観であり、被験者各々によって重要の価値基準が異なる可能性もある。このような定性的評価手法には、限界があるため、今後は重要性や有用性の判断基準をもっと明確にし、評価していく必要がある。

また、実験対象とした元の4つのキーワードに対して関連したキーワードが多く取得できていたため、本システムを利用することによって、指定したキーワードに対して、それに関連する長期間利用され続けているキーワード一覧が取得できることが分かった。

6 課題と展望

6.1 課題

検索キーワードは、主に必要であったり有用であ

ったりする情報を検索するためのものだが、長期間利用される情報と言った場合、様々な種類の情報が考えられる。本論文では、検索キーワードのみを対象としているが、そのキーワードを利用して、本当に長期間利用できるWebページを発見できる可能性が高くなるのか評価を行う必要がある。さらに、システムで取得したキーワードを元にして長期間利用できる、Web上の情報を提示するように改良していくことも考えられる。そして、実際にユーザが長期間利用され続けている検索キーワード一覧を取得する利用シーンに合わせて、ユーザインタフェースを整え、Web上にサービスとして公開するべきである。

また現状のシステムでは、入力するキーワードに対しても、提示するキーワードに対しても、キーワードのゆらぎの問題が存在する。この問題にも対応していく必要がある。

さらに、今回の実験では、被験者20人に対して、Googleで検索数が多い4つのキーワードに関連するキーワードの評価実験を行った。だが、本来はもっと多く被験者に対して、もっと多くのさまざまなキーワードを選定して実験を行うことが望ましい。そこで、今後はユーザインタフェースを整えWebサービスとして公開した上で、実際に多くの人に利用してもらい評価を行なっていく。

6.2 今後の展望

今回は1つの実装例として、検索キーワードの関連キーワードを取得するシステムを実装して実験した。だが、世の中にはさまざまな種類の情報があふれている。本研究の最終目的は、長期間利用される情報を取得する手法を確立し、さまざまな情報に対して応用可能にすることである。

そのため今後は、他にもさまざまな情報に対して、提案した長期度計算手法を適用して、長期間利用される情報を取得するシステムを開発していく。これは例えば、長期間多くのユーザから再生され続けている動画や長期間レビューされ続けている商品などを取得するシステムである。こういったさまざまなシステムを統合して、最終的に長期間利用する様々な情報を取得するシステムを開発していく。

7 まとめ

情報検索手法や情報フィルタリング手法として、多くの手法が提案されているが、長期間利用されるという利用のされ方に着目して情報を探す手法はほとんど存在しない。また、Web上の情報を検索する上で、適切な検索キーワードが思いつかないという問題もある。

そこで、本研究では、長期間検索され続ける検索キーワードを取得するシステムを開発した。開発したシステムに対して評価実験を行った結果、提案手法により長期間利用され続けるキーワードを取得することができた。

注

1 爆発する、破裂する、急に起きる、勢い良くでるなどの意味を持つ。

2 Google検索の補助機能で、ユーザが検索キーワードを入力しているときに、よく検索されるキーワードの候補を提示する機能である。

3 任意のキーワードを送ると関連語セットが帰ってくるモジュール。以下からダウンロードできる。

Lingua-JA-Expand.

http://search.cpan.org/~miki/Lingua-JA-Expand/ (2012年5月31日参照)

4 スニペットとは、一般的には「切れ端」「断片」といったような意味の英語であるが、IT用語としては、検索エンジンの検索結果の一部として表示されるWebページの要約文のことを表す。

5 Googleでの検索回数の時系列を公開しているサービスである。以下参照。

Google Insights for Search.

http://www.google.com/insights/search/ (2012年5月31日参照)

6 連想検索エンジン。入力したキーワードに関連するキーワードを提示する。
連想検索エンジン reflexa.
http://labs.preferred.jp/reflexa/ (2012年5月31日参照)
7 カイ二乗検定とは、帰無仮説が正しければ検定統計量がカイ二乗分布に従うような統計学的検定法の総称である。
8 Google検索において、特定のキーワードの検索回数が時間経過に沿ってどのように変化しているか参照できるサービス。
Googleトレンド. http://www.google.co.jp/trends/ (2012年5月31日参照)

参考文献

今井良太, 戸田浩之, 関口裕一郎, 望月崇由, 鈴木智也, 今井桂子 (2010) Web 検索サービスにおける多義的なクエリ推薦手法, 日本データベース学会論文誌 Vol.9, No.1, pp.7-11.

上野大樹, 樋口文人, 安村通晃 (2010), ソーシャルブックマークの時間スケールに着目した長期間利用するWebページ収集支援システムの研究, 情報社会学会誌, Vol.5 No.1, pp43-52.

大塚真吾, 豊田正史, 喜連川優 (2005) 大域ウェブアクセスログを用いた関連語の発見法に関する一考察, 情報処理学会論文誌:データベース, Vol.46, No.SIG8(TOD26), pp.82-92.

奥村学, 南野朋之, 藤木稔明, 鈴木泰裕 (2004) blog ページの自動収集と監視に基づくテキストマイニング, 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01.

兼宗進 (2004), 検索エンジンの検索アルゴリズム, 情報の科学と技術, Vol.54, No.2, pp.78-83.

北研二, 津田和彦, 獅々堀正幹 (2002), 情報検索アルゴリズム, 共立出版.

竹井潔 (2004), 情報の価値とライフサイクル管理, 聖学院大学論叢17(1), pp11-31.

寺本高(2009) 最寄品ブランドの小売店頭での販売展開方法とロングセラー化の関係, 流通研究, Vol.12, No.2, pp.59-73.

Albert-laszlo Barabasi, Jennifer Frangos (2002) Linked: The New Science Of Networks Science Of Networks, *Basic Books*, 288pp.

Andrew D. Booth (1967) A Law of Occurrences for Words of Low Frequency, Information and Control Vol.10, No.4, pp.386-393.

Jon Kleinberg (2002) Bursty and hierarchical structure in streams, In *Proc. the 8th ACM SGKDD International Conference on Knowledge Discovery and Data Mining*, pp91-101.

Jorg Rech (2007), Discovering trends in software engineering with google trend, *ACM Sigsoft Software Engineering Notes*, vol.32, No.2, pp.1-2.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd (1998) The pagerank citation ranking:Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Simon Hermann (1979) Dynamics of Price Elasticty and Brand Life Cycle, *Journal of Marketing Research*, 16(4), pp.439-45.

Soma Roy, David Gevry, William M. Pottenger (2002) Methodologies for trend detection in textual datamining. In *the Textmine'02 Workshop, Second SIAM International Conference on Data Mining*.

Tsubasa Takahashi, Hiroyuki Kitagawa (2008) S-bits: Social bookmarking induced topic search. In *Proceedings. 9th International Conference on Web-Age information Management (WAIM2008),* pp.25-30.

Yonggang Qiu, Hans-Peter Frei (1993) Concept Based Query Expansion. In *Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp.160-169.