第1章　長期的な検索キーワードの提示

概要

本章では，長期的な検索キーワードを提示するシステム「MasteryEye」について述べる。

1.1. 背景

1.1.1. 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（[?]、[?]）、Google の PageRank[?]）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。[?] は、時系列データから burst*[?] 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（[?]）。

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。

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。

1.1.2. 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（[?]）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（[?]）が行われてきた。また、Google サジェスト *[?] 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。

1.2. MasteryEye の提案

1.2.1. 長期度の計算手法

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。

$$\text{長期度} = \sum_{k=1}^{n} (\alpha_k - \beta_k) \tag{1.1}$$

$$\alpha_k : \text{単位時間ごとの検索回数やアクセス回数} \tag{1.2}$$

$$\beta_k : \text{べき乗則に基づくデータ} \tag{1.3}$$

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 ([**?**]) や単語の使用頻度は、べき乗則に基づく（[**?**]）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 1.1)。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。

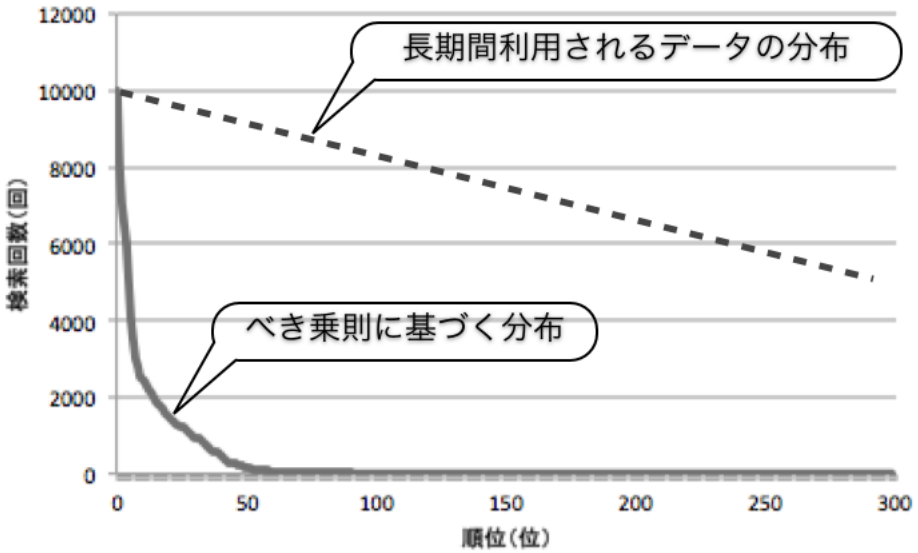

図 1.1: べき乗則のデータと長期間利用データ

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。

1.2.2. 関連キーワードの収集

関連キーワード取得は、既存の手法である、Lingua-JA-Expand[*?] 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 [*?] 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 1.1 のようになっている。

表 1.1: 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

1.2.3. 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[*?] 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。

1.2.4. 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 1.2 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。

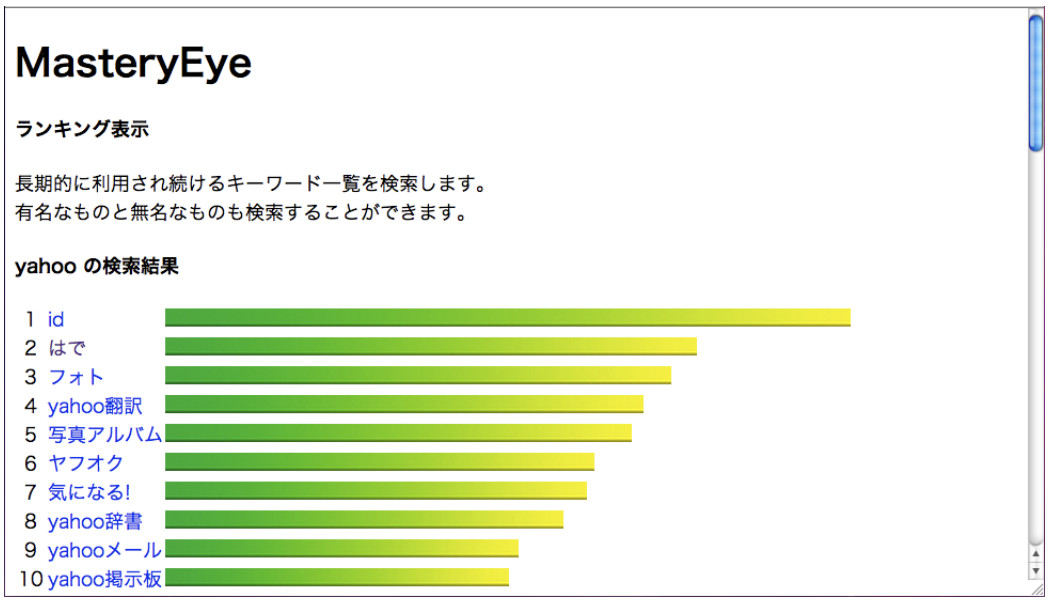

図 1.2: 「yahoo」での検索結果画面

1.3. 評価実験

1.3.1. 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

1.3.2. 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[*?]）

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 1.2 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 1.2: 被験者属性

人数	20 名
性別	男性：12 名、女性：8 名
年代	0 代：15 名、30 代：2 名、40 代：2 名、50 代：1 名

1.3.3. 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 1.9 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 1.3 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 1.3: 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

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 1.4〜表 1.6)。

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 1.4: 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 1.5: 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			

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 1.7)。

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 1.6: 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			

表 1.7: 長期間利用かどうか：長期度高と長期度低の比較（カイ二乗検定）

	選択あり :手法 1	選択あり：手法 2	選択なし：手法 1	選択なし：手法 2	計
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 1.8)。

表 1.8: 長期間利用と重要なものの関係（カイ二乗検定）

	長期	重要	長期かつ重要	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}				

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。

1.4. 課題と展望

1.4.1. 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

表 1.9: 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	竇輔
	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

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。

1.4.2. 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。

1.5. 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、ここでは、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。