2I1-1

検索頻度推定のための Wikipedia ページビューデータの分析

Wikipedia Page View Analysis for Search Trend Prediction

吉田光男 *1 Mitsuo Yoshida　荒瀬由紀 *2 Yuki Arase　角田孝昭 *3 Takaaki Tsunoda　山本幹雄 *3 Mikio Yamamoto

*1豊橋技術科学大学 Toyohashi University of Technology　*2大阪大学 Osaka University　*3筑波大学 University of Tsukuba

The frequency of a web search query generally reflects the degree of people's interest in the subject matter. Search logs are therefore a useful resource for trend analysis. However, accessing search logs is typically restricted to search engine providers. In this paper, we investigate whether search frequency can be estimated from another resource, namely, Wikipedia page view of open data. As a result, frequently searched queries revealed remarkably high correlations against Wikipedia page view. This fact suggests that Wikipedia page view is effective for understanding popular web search trends happening around the world.

1. はじめに

ある話題がいつ注目されたかを知るにはどうすれば良いだろうか。例えば，アメリカ合衆国の女優である「アン・ハサウェイ」が日本のインターネットユーザに注目されたのはいつだろうか。その答えを知る一つの方法は，ウェブ検索エンジンで「アン・ハサウェイ」というキーワードが頻繁に検索された時期を明らかにすることである。直近では 2014 年 12 月 12 日に映画「ダークナイト ライジング」がテレビ放送されたことにより注目され，頻繁に検索された。このように，ウェブ検索エンジンは詳細な情報を知りたいときに利用されると考えられ，そこに入力される検索キーワードはインターネットユーザの興味関心を反映している可能性が高い。実際，この仮説に基づいた研究開発が盛んに行われている [Choi 12, Radinsky 08]。しかしながら，通常，ウェブ検索エンジンの検索ログにアクセスすることはできず，ある話題がいつ注目されたかを知ることは困難である。検索ログに基づく情報提供サービスとしては Google Trends[*1] があるが，提供される情報は限定的である。例えば，特定の日時に流行した話題（キーワード）を知ることは困難であり，仮に知ろうとするならば，Google Trends に対して候補となり得るキーワードを大量にリクエストする必要がある。そのため，自由に利用できる公開情報（オープンデータ）から検索頻度情報を再現することが求められる。

本論文では，オープンデータを利用し，ウェブ検索エンジンに入力されたキーワードの頻度を推定するための基礎的な調査を行う。我々は，Wikipedia のページが検索結果の上位に表示されやすいことに着目した。また，過去の調査研究により，検索結果の上位ページをユーザが高い確率でアクセスする傾向があると明らかにされている [Jansen 05a, Jansen 05b]。さらに，ビットコインの価格変動特性を調査する中で，検索頻度と Wikipedia ページビューとの類似性を示唆した研究もある [Kristoufek 13]。同様に，我々も様々なキーワードについて検索頻度（Google Trends）と Wikipedia ページビューとのトレンドを比較したところ，図に示すように，互いに類似しているケースが多数見受けられた。これらより，Wikipedia のページビューデータを利用することにより，検索頻度を推定できると考えた。そこで今回，Wikipedia のページビューと検索頻度との類似性について調査し，高頻度アクセス帯のキーワードに関して高い類似性があることを示す。

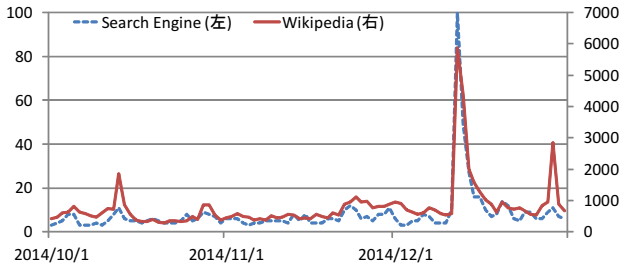

図 1: 「アン・ハサウェイ」のトレンド（日別）

2. 調査方法

本論文では検索頻度の推定を目指し，Wikipedia ページビューと検索頻度との類似性を調査する。検索頻度を活用するアプリケーションごとに利用する頻度データの粒度が異なることが予想されるため，本調査では，日別，週別及び月別のそれぞれの粒度における類似性を調査する。

類似性の評価にピアソンの積率相関係数（以下，単に「相関係数」という）を用いる。また，検索頻度のトレンドを利用する場合において，検索頻度が前月から上昇したか否かなど，値の上昇及び減少を捉えたい場合がある。しかしながら，相関係数では上昇及び減少の変動を捉え，類似性を評価するのは困難である。そこで，値の上昇及び減少を捉えるための指標，増減一致率を定義して調査に利用する。増減一致率（$UDCR$）は 2 つの時系列データ $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ 及び $\boldsymbol{Y} = \{y_1, y_2, ..., y_n\}$ が与えられたとき，次のように計算される。

$$UDCR(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\|\{t \in \{2, 3, ..., n\} | x'_t = y'_t\}\|}{n - 1}$$

$$x'_t = \begin{cases} 1 & (x_t - x_{t-1} > 0) \\ 0 & (x_t - x_{t-1} = 0) \\ -1 & (x_t - x_{t-1} < 0) \end{cases}$$

y'_t に関しても x'_t と同様に計算する。

連絡先: 吉田光男，豊橋技術科学大学，愛知県豊橋市天伯町雲雀ヶ丘 1-1, yoshida@cs.tut.ac.jp

*1 http://www.google.co.jp/trends/ (viewed 2015-03-27)

3. 頻度データの収集方法

通常，ウェブ検索エンジンのログにアクセスすることができず，検索頻度データの入手は困難である。本研究では，検索頻度の実データの代わりに，Google Trends が出力する頻度データを検索頻度データと見なして利用する。Google Trends が提供する頻度データは，頻度データ出力期間における最大値を 100 とする整数値である。そのため，検索頻度の実際の値を知ることはできない。また，低頻度クエリに関しては，頻度データを出力しないため，データの入手はできない。以上のような制約があるものの，本研究では，類似度を調査するクエリごとに Google Trends に問い合わせ，日別，週別及び月別の頻度データを収集する。

ウィキメディア財団は財団が運営するウェブサイトのページビューデータを配布しており *2，そのデータには Wikipedia のページビューデータも含まれる。配布されているデータ（ファイル）には言語，ファイルパス，アクセス回数及びデータ転送量が記述されており，1 時間ごとに分割されて提供されている *3。本研究では，このページビューデータと Wikipedia のダンプデータから調査に利用するデータを生成する。まず，ダンプデータを利用し，対象とする言語の見出し語リストを作成する。次に，言語で絞り込んだ上で，ページビューデータに含まれるファイルパスと見出し語リストに含まれる見出し語とを関連付ける。その際，URL エンコードされたファイルパスをデコードする。また，見出し語に含まれるスペース文字（空白）はファイルパスではアンダースコア（_）で表記されているため，ファイルパスに含まれるアンダースコアをスペース文字に置き換える。このように，一定の正規化処理を行うため，ページビューデータの各行と見出し語リストの各行とは，多対 1 の関係になる。これらの処理により，見出し語と時間別アクセス回数との対応が取れている。最後に，時間別アクセス回数を日別，週別及び月別のアクセス回数として集計処理を行う。この際，必要に応じてタイムゾーンの変換処理を行う。

4. 調査結果及び考察

4.1 頻度データの収集状況

本調査では，Wikipedia にアクセスされた上位の見出し語（キーワード）を調査対象クエリとする。具体的には，日本語版 Wikipedia に含まれるアニメ，漫画，映画，人名に関係する *4 見出し語に関し，2008 年から 2014 年にかけて 1 日平均 1,000 件以上のアクセスがあった見出し語を調査対象とする。これらの条件に基づき，頻度データを収集できたキーワード数を表 1 のキーワード数及び収集数に示す。日別の頻度データは 2014 年 10 月から 12 月の期間，週別の頻度データは 2012 年から 2014 年の期間，月別の頻度データは 2008 年から 2014 年の期間を対象として収集した *5。Google Trends は低頻度クエリの頻度データを出力しないため，該当クエリ全ての検索頻度データの入手が困難であることがわかる。

*2 http://dumps.wikimedia.org/other/pagecounts-raw/ (viewed 2015-03-27)

*3 pagecounts-20150101-120000.gz のように，ファイル名にはデータが生成された日時（協定世界時）が含まれる。このファイルであれば，2015 年 1 月 1 日 11 時台のページビューデータが格納されている。

*4 見出し語が属するカテゴリをもとに抽出する。

*5 Google Trends は指定する期間の長さによって頻度データの粒度が異なるが，日別，週別及び月別の頻度データの量が最大になる期間を設定している。Google Trends にクエリを送信する際には，見出し語末尾に含まれる「(女優)」など，曖昧性の回避に使われる文字列を除去している。

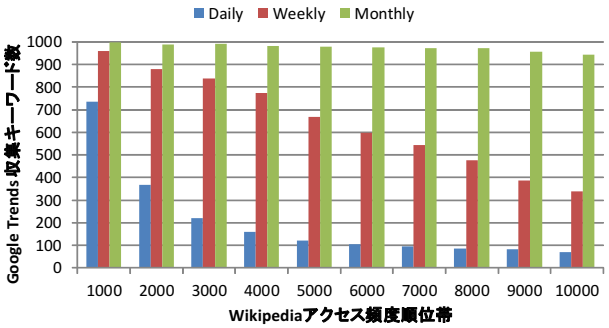

図 2: 収集キーワード数の分布（人名）

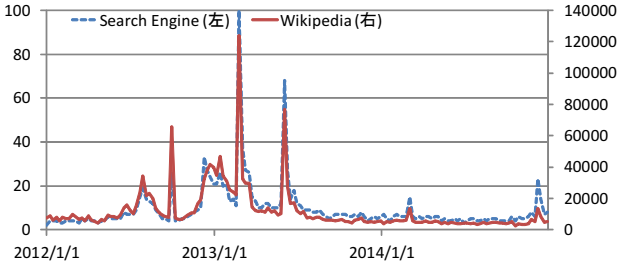

図 3: 「アン・ハサウェイ」のトレンド（週別）

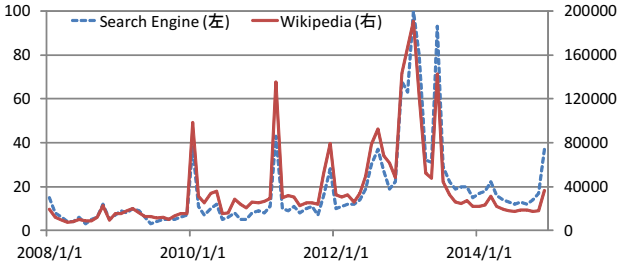

図 4: 「アン・ハサウェイ」のトレンド（月別）

人名に関しては，Wikipedia でのアクセス頻度が高い傾向があるため，先のデータに加え，アクセス頻度の上位 10,000 件までを調査対象とし，頻度別の類似度を調べることとした。収集できた日別の頻度データは 2,039 件，週別の頻度データは 6,457 件，月別の頻度データは 9,757 件であった。図 2 に示すように，Wikipedia でのアクセス頻度（順位）が低くなるほど，収集可能なキーワード数が減少する。

本研究の調査のために収集したデータは，Zenodo を通じて全て公開している *6。

4.2 トレンドの例

図 1，図 3 及び図 4 は「アン・ハサウェイ」に関する日別，週別及び月別の検索頻度（Google Trends）と Wikipedia ページビューのトレンドである。相関係数はそれぞれ 0.92，0.93，0.93 であり，増減一致率はそれぞれ 0.54，0.67，0.77 である。いずれの値も高い値を示しているものの，日別の増減一致率は少し低い値を示しており，増減一致率は相関係数とは異なる観点で類似性を捉えられていることが確認できる。

*6 http://dx.doi.org/10.5281/zenodo.14539

表 1: カテゴリ別の収集数及び類似度（1,000 アクセス/日）

		収集数			相関係数			増減一致率		
	キーワード数	日別	週別	月別	日別	週別	月別	日別	週別	月別
アニメ	163	122	153	159	0.43	0.71	0.78	0.53	0.57	0.65
漫画	182	134	170	178	0.37	0.70	0.75	0.53	0.56	0.63
映画	48	35	45	46	0.49	0.74	0.76	0.51	0.57	0.63
人名	837	645	810	835	0.57	0.73	0.72	0.52	0.66	0.70

4.3 カテゴリ別の類似度

カテゴリ別の類似度を表 1 の相関係数及び増減一致率に示す。日別の類似度は若干低い傾向にあるものの，週別及び月別の類似度は高い傾向を示している。このことから，人気のあるキーワードに関し，少なくとも週別又は月別であれば，Wikipedia のページビューデータから検索頻度を予測できる可能性が高いことがわかる。

週別及び月別と比較し，日別の類似度が低くなる原因として，2 つの原因が考えられる。まず，ウィキメディア財団によって提供されているページビューデータは，単純なアクセス回数であるため [*7]，ユーザによるリロード処理（F5 処理），ボットによる一時的なアクセスの増大などの影響を受けやすいことである。実際，データを検証したところ，1 時間でアクセス回数が 100 倍以上になるなど，所々で異常値のようなものが見受けられた。また，Google Trends のデータのタイムゾーンが不明であるため，Wikipedia のデータのタイムゾーン（本調査では日本標準時に変換）と一致しなかった可能性もある。いずれも，週別又は月別の粒度にすることで，データが平滑化されたと考えられる。

人名と比較し，アニメ，漫画，映画の類似度が低くなる原因として，それらの Wikipedia の見出し語には副題などが含まれる傾向があり，見出し語とその記事に到達する検索キーワードが一致していない可能性が挙げられる。例えば，見出し語「学園黙示録 HIGHSCHOOL OF THE DEAD」というアニメのページに到達する検索キーワードは「学園黙示録 HIGHSCHOOL OF THE DEAD」よりも，短縮された「学園黙示録」である可能性が高く [*8]，類似度を算出するキーワードの対応を調整する必要があったと考えられる。検索頻度の予測という側面では，見出し語に対して推定したいクエリを部分一致すれば十分である可能性もあり，今後，Wikipedia の見出し語に限定されないキーワード（特に正式名称を短縮したようなキーワード）での調査を進めたいと考えている。

4.4 アクセス頻度別の類似度

図 5 及び図 6 は人名に関し，アクセス頻度順位帯における相関係数及び増減一致率の平均値を示す。これらより，アクセス頻度が高いほど検索頻度と Wikipedia ページビューとの類似性が高いことがわかる。

高頻度帯（アクセス頻度上位）と比較し，低頻度帯の類似度が低くなる原因として，2 つの原因が考えられる。まず，アクセス回数が少ないことにより，4.3 節で述べたような異常値の影響を受けやすいことが挙げられる。また，Google Trends が提供するデータは最大値を 100 とする整数値であり，一時の

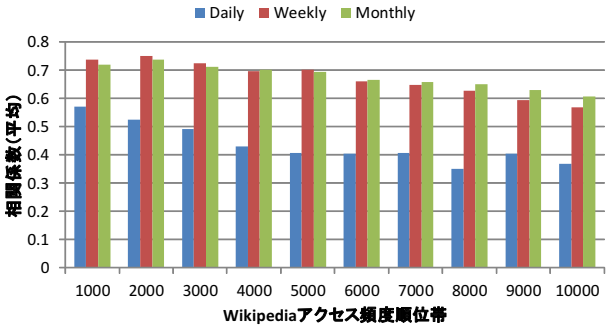

図 5: アクセス頻度別の平均相関係数（人名）

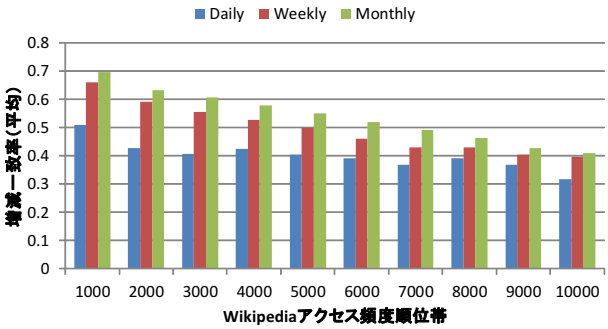

図 6: アクセス頻度別の平均増減一致率（人名）

検索頻度スパイクに引きずられ，頻度データ出力期間のうち低頻度期間が 0 で埋まる傾向があることも挙げられる。図 7 はアクセス頻度順位帯ごとに Google Trends の頻度データが 0 で占める割合の平均値を算出したものである [*9]。低頻度帯ほど 0 で占める割合が増加しており，図 6 と一定の負の関係が見られる。検索頻度データが実データであれば，低頻度期間であっても 0 にはならず，より正確な類似度を算出できると考えられる。そのようなデータを利用しての検証は今後の課題である。なお，Wikipedia 側の頻度データを 100 段階の値に調整した場合，相関係数にほとんど変化は見られない一方，増減一致率は図 8 のように平均して 0.04 ポイント（月別）改善することを確認した。

5. 関連研究

検索頻度データを分析することで，トレンドの予測を試みる研究開発が盛んに行われている。Radinsky らは，ニュー

*7 配布されているデータの説明では「the number of non-unique views」と記載されており，同一 IP アドレスからの連続的なアクセスも単純に加算されていると考えられる。

*8 Google Trends によれば，「学園黙示録」の検索頻度は「学園黙示録 HIGHSCHOOL OF THE DEAD」の 5 倍にのぼる。

*9 例えば月別の頻度データに関し，12ヶ月分が 0 で埋められている場合，その割合は 14.29%となる。

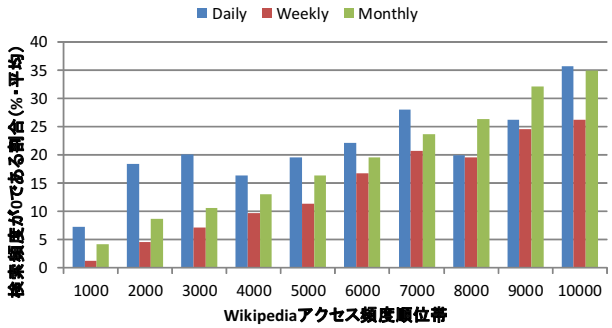

図 7: Google Trends のデータが 0 で埋まる割合（%）

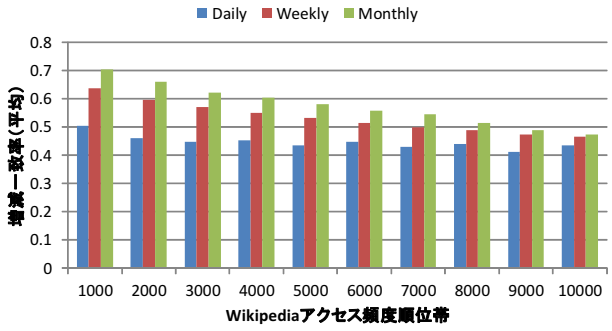

図 8: 最大値 100 に調整した場合の平均増減一致率（人名）

スの流行を予測するために，検索頻度データを利用している [Radinsky 08]。Choi と Varian は，自動車販売台数や失業率などの経済指標の予測を試みている [Choi 12]。検索クエリの推薦精度改善を目指し，検索クエリの季節性を発見するために検索頻度データが利用される場合もある [Shokouhi 12]。これらは，いずれも検索頻度データとして Google Trends を利用している。検索事業会社であるヤフー株式会社は，自社で保有する検索頻度データをもとにした「Yahoo! JAPAN ビックデータレポート」を公開している [*10]。ここでは，選挙の議席数や経済指標，感染症患者数などの予測及び分析が公開されている。

検索頻度データの代わりに，Wikipedia ページビューデータを利用し，トレンドの予測を試みる研究も行われている。Mestyan らは，映画の流行予測を試みている [Mestyán 13]。Moat らは，株式市場全体のトレンドを予測するために，Wikipedia のデータを利用し，編集回数よりもアクセス回数の方が有用であることを報告している [Moat 13]。

検索頻度データと Wikipedia ページビューデータの双方に着目した調査も行われている。Huss らは遺伝子工学に関するコミュニティ活動のために，その分野がどれほど需要があるかを双方のデータで確認している [Huss 10]。Kristoufek はビットコインの市場価格を予測するための基礎的な調査として，ビットコインの市場価格と双方のデータとの類似性を調査している [Kristoufek 13]。いずれの調査も検索頻度と Wikipedia ページビューとの間に類似性があることを示唆するような結果が見受けられるものの，明示的な調査及び類似する条件の検討などは行われていなかった。

*10 http://docs.yahoo.co.jp/info/bigdata/ (viewed 2015-03-27)

6. おわりに

本研究では，Wikipedia のページビューで検索頻度を予測できるかどうかを検証するために，それぞれの類似性を調査した。その結果，Wikipedia におけるアクセス回数が多いキーワードに関しては，ページビューと検索頻度（Google Trends）との間に高い類似性を確認できた。例えば，1 日平均 1,000 アクセス以上のキーワードに関し，週別及び月別トレンドにおける相関係数はそれぞれ 0.73，0.72 であった。このことから，人気のあるキーワードに関しては，Wikipedia のページビューで検索トレンドを推定できる可能性が高いと推察される。

参考文献

[Choi 12] Choi, H. and Varian, H.: Predicting the Present with Google Trends, *Economic Record*, Vol. 88, pp. 2–9 (2012)

[Huss 10] Huss, J. W., Lindenbaum, P., Martone, M., Roberts, D., Pizarro, A., Valafar, F., Hogenesch, J. B., and Su, A. I.: The Gene Wiki: community intelligence applied to human gene annotation., *Nucleic acids research*, Vol. 38, No. Database issue, pp. D633–9 (2010)

[Jansen 05a] Jansen, B. J. and Spink, A.: An analysis of Web searching by European AlltheWeb.com users, *Information Processing and Management*, Vol. 41, No. 2, pp. 361–381 (2005)

[Jansen 05b] Jansen, B. J., Spink, A., and Pedersen, J.: A temporal comparison of AltaVista Web searching, *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 6, pp. 559–570 (2005)

[Kristoufek 13] Kristoufek, L.: BitCoin meets Google Trends and Wikipedia: quantifying the relationship between phenomena of the Internet era., *Scientific Reports*, Vol. 3, p. 3415 (2013)

[Mestyán 13] Mestyán, M., Yasseri, T., and Kertész, J.: Early prediction of movie box office success based on Wikipedia activity big data., *PloS one*, Vol. 8, No. 8, p. e71226 (2013)

[Moat 13] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T.: Quantifying Wikipedia Usage Patterns Before Stock Market Moves, *Scientific Reports*, Vol. 3, No. 1801 (2013)

[Radinsky 08] Radinsky, K., Davidovich, S., and Markovitch, S.: Predicting the News of Tomorrow Using Patterns in Web Search Queries, in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 363–367 (2008)

[Shokouhi 12] Shokouhi, M. and Radinsky, K.: Time-sensitive query auto-completion, in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pp. 601–610 (2012)