

商用検索エンジンにランキングされたサイトの ランク変動パターン解析

吉田 泰明[†] 平手 勇宇^{††} 山名 早人^{‡,‡‡}

[†] 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

^{††} 早稲田大学メディアネットワークセンター 〒169-8050 東京都新宿区戸塚町 1-104

[‡] 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

^{‡‡} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: {yoshida,hirate,yamana}@yama.info.waseda.ac.jp

あらまし 現在、ウェブ上の情報量が膨大になった結果、商用検索エンジンが広く用いられるようになった。しかし、ユーザは一般に上位にランキングされたページしか見ないため、サイトの運営者にとって、商用検索エンジンのランキングが与える影響は極めて大きい。そこで本稿では、事前に決定した 1000 のクエリを用いて、日本で広く利用される Google, MSN, Yahoo! JAPAN にて検索を行い、ランキングデータを定期的に収集した。収集されたデータを用いて、ランキングされたサイトのランク変動パターンやサイトの特徴について調査を行った。その結果、ランキングの変動に一定のパターンを発見し、パターンごとに生成日時や更新頻度で異なる特徴を発見することができた。

キーワード 検索エンジン, ランキング, 特徴解析

Analyzing the Transition Patterns of Sites Ranked by Web Search Engines

Yasuaki YOSHIDA[†] Yu HIRATE^{††} and Hayato YAMANA^{‡,‡‡}

[†] Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

^{††} Media Network Center, Waseda University 1-104 Totsuka-cho, Shinjuku-ku, Tokyo, 169-8050, Japan

[‡] Faculty of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

^{‡‡} National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

E-mail: {yoshida,hirate,yamana}@yama.info.waseda.ac.jp

Abstract Many people use search engines to retrieve documents from the web. The influences of search engine rankings are significant, since people usually refer only to high-ranked web pages. In this paper, we gathered ranking data regularly by submitting 1000 pre-selected queries. Then we investigated the transition patterns of rankings and the feature of ranked sites by major search engines, Google, MSN and Yahoo! JAPAN. As a result, we have confirmed that there exist both some transition patterns of rankings, and the different features of ranked sites.

Keyword Search Engine, Ranking, Feature

1. はじめに

近年、ウェブ上の情報量が膨大になってきた結果、商用検索エンジンが一般に広く用いられるようになってきた。しかし、ウェブ上の情報量が膨大になりすぎた結果、商用検索エンジンを用いることなくユーザが欲しい情報を得ることは困難になっており、さらに、一般にユーザは商用検索エンジンによって上位にランキングされたサイトしか閲覧しないため、サイトの運営者にとって、商用検索エンジンのランキングが与える影響は極めて大きいと考えられる。

そこで本稿では、事前に決定した 1000 のクエリを

用いて、Google, MSN, Yahoo! JAPAN のランキングデータを 2007 年 6 月から 2008 年 12 月に渡って定期的に収集した。収集したデータを用いて、ランキングされたサイトについて、ランクの変動パターンやサイトの特徴について解析した。

本稿は以下の構成をとる。2 節で関連研究について紹介し、3 節でデータの収集方法について説明し、4 節で解析結果について説明し、5 節でまとめを行う。

2. 関連研究

本節では関連する研究について述べる。

ニューヨーク市立大学の Mowshowitz らは、2005 年に検索エンジンの検索結果の傾向を Bias という指標で評価している[1]。Bias とは基準となる検索結果と、ある検索エンジンの検索結果がどの程度異なるかという指標である。Mowshowitz らは Bias を用いて、50 のクエリで 16 の検索エンジンの上位 30 件のランキングを比較している。しかし、調査規模は 50 クエリ、上位 30 件と小さく、また、ランキングされたサイトについての特徴解析を行っていない。

Bar-Ilan University の Bar-Ilan らは 2006 年に複数の検索エンジンのランキングの時間変化について調査している[2]。Bar-Ilan らは Google, Yahoo!, Teoma に対して 3 つのクエリ、Google Image, Yahoo! Image, Picsearch に対して 2 つのクエリの上位 10 件を毎日検索してデータを収集している。収集データに対して、Overlap, Spearman's footrule[3][4], Fagin's measure[5], M measure[2]を用いて検索エンジンのランキングの時間変化の解析と検索エンジン間の比較を行っている。しかし、使用しているクエリは計 5 つ、取得する検索結果も上位 10 件と規模が小さく、調査としては不十分である。また、ランキングされたサイトの特徴解析も行っていない。

Old Dominion University の McCown らは 2007 年に Google, Yahoo!, MSN の API 上のランキングと一般に利用される WUI(Web User Interface)上のランキングの差異を調査している[6]。一般的な語として 50 クエリ、Computer Science に関する語として 50 クエリを対象として上位 100 件を毎日 5 か月間検索している。検索した結果に対して Overlap, Fagin's measure[5], M measure[2]の 3 つの指標で比較している。さらに、ランダムに選出した 100 の URL に対して被リンク数、ランダムに選出した 100 のサイトのページが各エンジンで何件インデックスされているか、ランダムに選出した 100 の URL に対してインデックスの有無を WUI と API でどう異なるか比較している。しかし McCown らの研究ではクエリ数が 100 であり、規模が小さく、また調査目的も WUI と API の比較を主な目的としており、検索エンジン間の比較と特徴解析およびランキングされたサイトの特徴解析を目的としている本研究とは立場が異なる。

いずれの調査も調査対象として少数のクエリやホストを選択しているが、母集団（考えられる全てのクエリや存在する全ホスト）から偏りなく選択されているとは言えず、検索エンジンの一般的な特徴を解析する方法としては不適切である。また、いずれの研究もサイトの特徴について解析していないため、サイトの運営者が検索エンジンにどう対応すべきかが分からないという問題がある。この問題に対し、我々は以前に

[7]の研究で大規模なランキングデータの収集及び Bias 指標を用いた比較、ページやサイトの特徴解析として被リンク数とランキングの関連の調査を行った。その結果、ランキングは上位ほど変化が少ないことや、ランキングはページ単位ではなくサイト単位で行われている可能性が高いこと、および Google は被リンク数の影響を大きく受け、MSN はほとんど影響を受けないことが分かっている。しかし、[7]においてはランキングの時間変動の解析としては Bias 指標の変動量のみ調査であり、時間変動の調査としては不十分であった。また、サイトの特徴解析としては被リンク数とランキングの関連の調査のみであり、サイトの特徴調査としては不十分であった。

そこで本稿では、ランキングの時間変動をさらに詳細に解析するために、サイトのランク変動パターンを抽出し、サイトの特徴を解析するために、サイトの生成日時や更新頻度について調査した。

3. データ収集方法

本節では、解析に用いたデータの収集方法について説明する。なお、本節で説明する方法は、収集期間以外は我々の[7]の研究におけるデータ収集手法と同一である。

調査対象の商用検索エンジンは Google, MSN, Yahoo! JAPAN とする。ランキング取得のため、各検索エンジンの API を用いる[8][9][10]。

クエリを選択する方法について述べる。ランキングの傾向を測るためにはクエリ自体に偏りがあってはならない。そこで、偏りなくクエリを選ぶために、単語の属するカテゴリを 10 種類（動物、植物、企業、商標（固有名詞）、商品名（「車」、「本」など一般的な名詞）、大学名、日本人名、英語人名、地名、食品）選択し、各カテゴリから、知名度（Yahoo! JAPAN の検索結果数）を基準として分散するように 100 語を選択した。

選択されたクエリを用いて、2007 年 6 月から 2008 年 12 月に渡り、一週間に 1 回上位 250 件を検索する。

4. データ収集方法

本節では収集データを解析した結果を述べる。

4.1. 主要なランキング変動パターン

ランキングの変動パターンを知るため、収集したデータを月ごとに分割し、サイトごとに各月の平均ランクを計算した。さらに、サイトの各月の平均ランクをベクトルの要素と見なし、k-means クラスタリングを行った。ただし k=15 である。10 回クラスタリングを行い、多くの種類のクラスタがバランスよく出現した

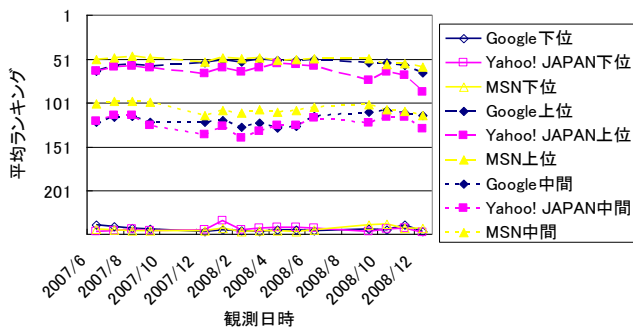


図 1: 定位置パターンに属するクラス

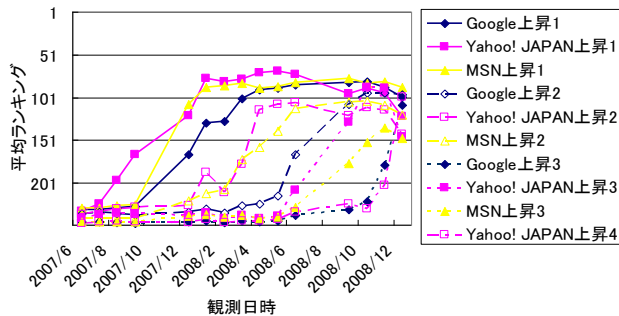


図 2: 上昇パターンに属するクラス

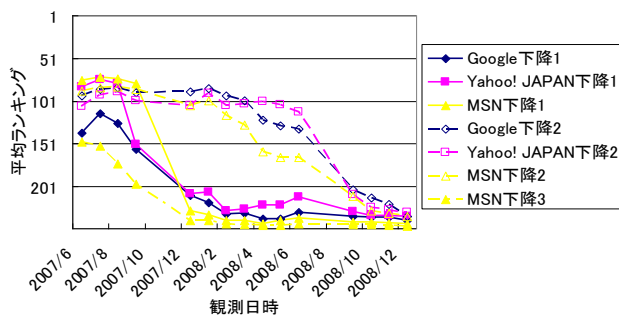


図 3: 下降パターンに属するクラス

結果を採用する¹.

サイトの平均ランクの求め方について述べる. 1000 のクエリで収集されたランキングすべてについて, サイト名 (ホスト名) を抽出する. 次に, サイトごとに 1 か月間で 1000 クエリのランクの合計を求め, ランキングの登場回数で割ることで, 平均ランクとする. クラスタリングのために, ベクトル作成する際の各要素 e を以下の式のように求める.

$$e = \begin{cases} m - \text{avg} + 1 & (\text{if ranked}) \\ 0 & (\text{if not ranked}) \end{cases} \quad (1)$$

ただし, avg は平均ランク, m は取得する上限の検索結果数 (本稿では $m=250$) とする.

こうして作成したベクトルを用いてクラスタリングを

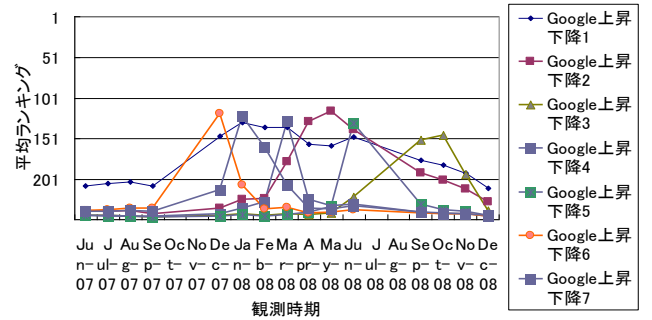


図 4: 上昇下降パターン (Google)

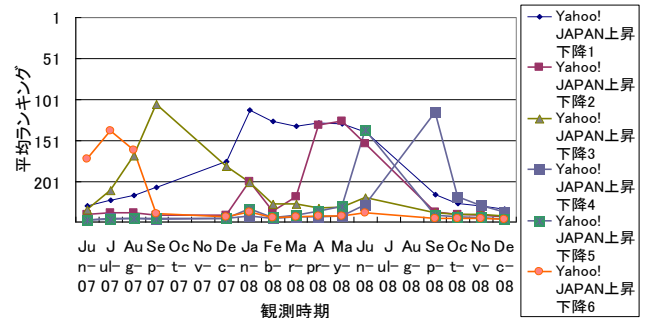


図 5: 上昇下降パターン (Yahoo! JAPAN)

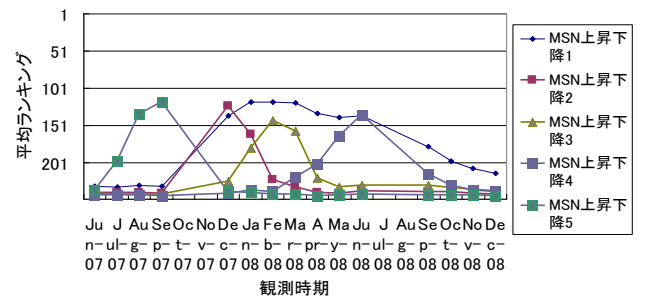


図 6: 上昇下降パターン (MSN)

行い, セントロイド c を算出して, c の各要素を $m+1$ から引いた値をクラスタの代表値とする. 結果として, 以下の 5 つのパターンが抽出された.

- 定位置パターン (図 1)
- 上昇パターン (図 2)
- 下降パターン (図 3)
- 上昇下降パターン (図 4 から図 6)
- その他のパターン (図 7)

また, 各クラスタに含まれるサイト数の割合を表 1 に示す. ランキングがほとんど変動しないパターン, ランキングが上昇していくパターン, ランキングが下降していくパターン, ランキングが上昇後に下降するパターンが抽出されるが, ランキングが下降後に上昇するパターンが抽出されることはなかった. このことはサイトのランキングが一度落ちた場合は上がってくることはほとんどないということを示しており, ランキ

¹ 自身でクラスタリング結果を比較し, 選択した.

表 1: 各クラスタに含まれるサイト数の割合

Google のクラスタ (202879 サイト)	割合	Yahoo! JAPAN のクラスタ (176611 サイト)	割合	MSN のクラスタ (281007 サイト)	割合
Google 下位	34.66%	Yahoo! JAPAN 下位	33.28%	MSN 下位	36.83%
Google 上位	4.65%	Yahoo! JAPAN 上位	4.21%	MSN 上位	3.24%
Google 中間	6.21%	Yahoo! JAPAN 中間	5.54%	MSN 中間	4.96%
Google 上昇 1	2.92%	Yahoo! JAPAN 上昇 1	2.41%	MSN 上昇 1	2.65%
Google 上昇 2	4.91%	Yahoo! JAPAN 上昇 2	2.91%	MSN 上昇 2	3.39%
Google 上昇 3	4.78%	Yahoo! JAPAN 上昇 3	4.32%	MSN 上昇 3	6.97%
Google 下降 1	5.53%	Yahoo! JAPAN 上昇 4	3.35%	MSN 下降 1	6.13%
Google 下降 2	2.97%	Yahoo! JAPAN 下降 1	3.73%	MSN 下降 2	2.61%
Google 上昇下降 1	4.67%	Yahoo! JAPAN 下降 2	3.50%	MSN 下降 3	8.48%
Google 上昇下降 2	3.57%	Yahoo! JAPAN 上昇下降 1	5.16%	MSN 上昇下降 1	3.59%
Google 上昇下降 3	7.73%	Yahoo! JAPAN 上昇下降 2	8.99%	MSN 上昇下降 2	4.92%
Google 上昇下降 4	3.56%	Yahoo! JAPAN 上昇下降 3	3.57%	MSN 上昇下降 3	4.10%
Google 上昇下降 5	5.91%	Yahoo! JAPAN 上昇下降 4	4.37%	MSN 上昇下降 4	5.63%
Google 上昇下降 6	3.35%	Yahoo! JAPAN 上昇下降 5	8.11%	MSN 上昇下降 5	2.92%
Google 上昇下降 7	4.58%	Yahoo! JAPAN 上昇下降 6	6.56%	MSN その他	3.57%

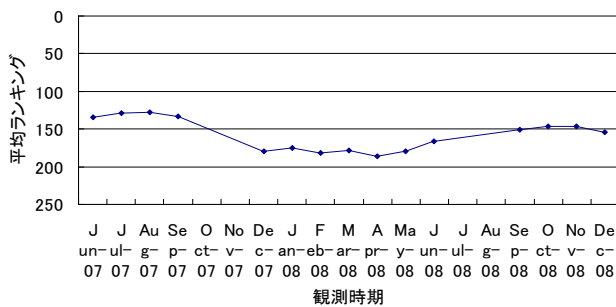


図 7: その他のパターン(MSN のみ)

ングが下がった場合は新しいサイトを作ったほうが早いということを示している。

4.2. 各パターンに属するサイトの生成日時

ランキングの変動パターンごとに、サイトの生成日時について調査した結果について述べる。サイトの生成日時を調査するために各パターンに属するサイトから 100 のサンプルを抽出し、Internet Archive[11]を用いて調査した²。ここで、Internet Archive に最初にアーカイブされた日時を生成日時と見なしている。

上位クラスタと下位クラスタにおける生成日時を図 8 に示す。ただし、Not Indexed とは Internet Archive 上にアーカイブされていなかったサイトである。上位クラスタでは生成日時が古く、下位クラスタでは新しい傾向があることが分かる。つまり、古いサイトがランキング上位に続ける傾向があるということである。

中間・上昇クラスタ 2・下降クラスタ 2・上昇下降

² ただし、1996 年から 1999 年及び 2008 年はアーカイブ数が少なかったため、精度が低い可能性がある。

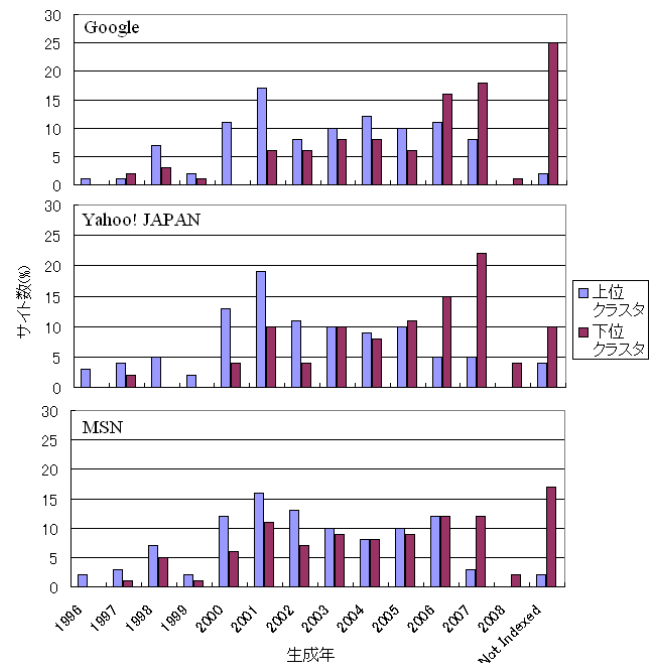


図 8: 上位・下位クラスタの生成日時

クラスタ 1 について調査した結果を図 9 に示す。上昇クラスタにおいては生成日時が新しい傾向があることが分かる。このことは、新しいサイトのほうが、ランキングが上昇しやすい傾向があることを示す。

4.3. 各パターンに属するサイトの更新頻度

生成日時と同様の方法で同じ 100 のサンプルサイトの更新頻度について調査した結果を述べる。ただし、ある年におけるサイトの更新頻度とは、そのサイトのトップページのアーカイブにおいて、Internet Archive

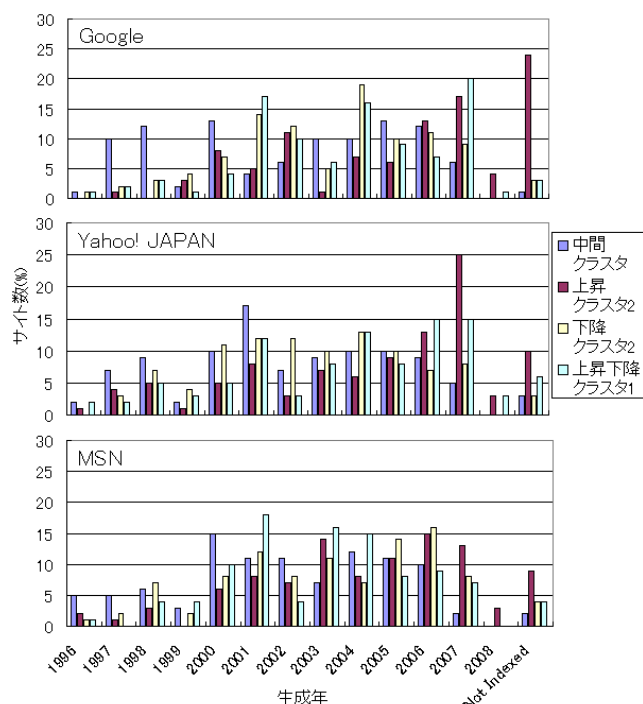


図 9: 中間・上昇 2・下降 2・上昇下降クラスタ 1 の生成年

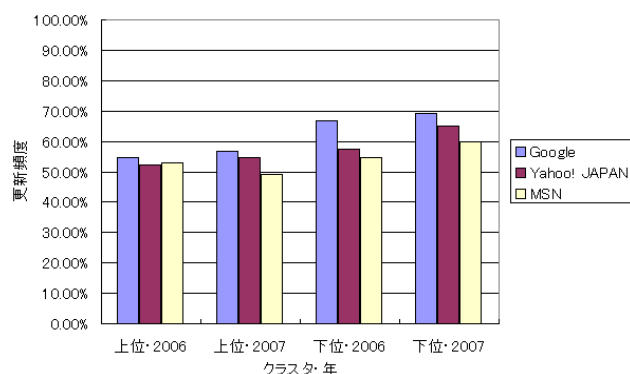


図 10: 上位・下位クラスタの各年における更新頻度

がそのページは更新されたと判定した回数を、その年に Internet Archive がアーカイブした回数で割ったものである。ここで、Internet Archive はアーカイブのリストにおいて、あるアーカイブとその前のアーカイブが少しでも異なる場合には “*” をつけているため (Internet Archive FAQs[12] の “What does it mean when a site’s archive data has been “updated”?” より), “*” の数を更新回数とする。ただし本項では、1996 年から 2005 年までに生成されたサンプルサイト数が少なく、かつ本研究で収集したランキングデータも 2007 年 6 月以降のものであるため、2005 年以前のデータは調査対象とはしない。また、Internet Archive 上のアーカイブ数も 2008 年分は調査時点 (2008 年 11 月) では少なかったため、2008 年分も調査対象からはずし、2006 年と 2007 年のみを調査対象とする。

上位クラスタと下位クラスタにおけるサイトの各

表 2: 上位・下位クラスタ間の平均差の p 値

	2006 年	2007 年
Google	0.008856	0.008128
Yahoo! JAPAN	0.23054	0.037529
MSN	0.734222	0.009438

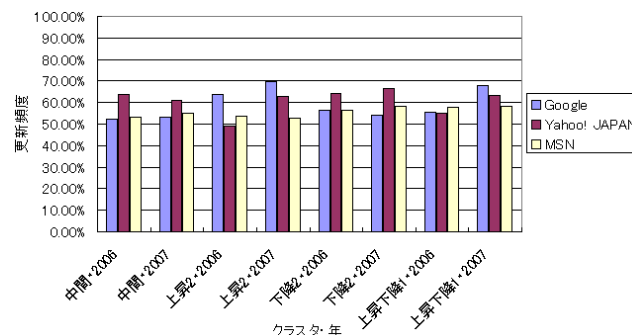


図 11: 中間・上昇 2・下降 2・上昇下降クラスタ 1 における各年の更新頻度

年の更新頻度の平均を検索エンジンごとに比較した結果を図 10 に示す。上位クラスタと下位クラスタを比較すると、いずれの検索エンジンでも更新頻度が少ないものが上位にランキングされる傾向があることが分かる。また、上位・下位クラスタにおいて共に Google は他の検索エンジンよりも最近の更新頻度が高いことから、Google は最も更新頻度を重視している可能性がある。この結果を検証するために上位クラスタと下位クラスタで平均値の差について t 検定を行った。仮説や有意水準を以下のように設定する。

- ・ 帰無仮説：各検索エンジンにおいて、上位クラスタと下位クラスタ間で更新頻度に有意な差はない
- ・ 対立仮説：各検索エンジンにおいて、上位クラスタと下位クラスタ間で更新頻度に有意な差がある
- ・ 有意水準：5%

検定した結果の p 値を表 2 に示す。いずれの検索エンジンにおいても 2007 年では有意水準より低い p 値となっており、帰無仮説は棄却され、対立仮説が採択される。また、p 値は Google が最も小さいことから、Google は最も更新頻度を重視している可能性がある。

中間・上昇クラスタ 2・下降クラスタ 2・上昇下降クラスタ 1 について調査した結果を図 11 に示す。上昇クラスタにおいて Google の更新頻度は高い数値を示している。このため、Google は更新頻度が多いとランキングが上がりやすい可能性が高い。Yahoo! JAPAN では下降クラスタにおける更新頻度がその他の検索エ

表 3: 各変動パターンにおける検索エンジン間の更新頻度の平均差の p 値

	中間		上昇 2		下降 2		上昇下降 1	
	2006 年	2007 年	2006 年	2007 年	2006 年	2007 年	2006 年	2007 年
Google & Yahoo! JAPAN	0.030525	0.109545	0.025139	0.291425	0.138788	0.017297	0.990601	0.350689
Yahoo! JAPAN & MSN	0.073751	0.19473	0.329837	0.069249	0.156271	0.045188	0.663759	0.32218
Google & MSN	0.62009	0.80214	0.203612	0.010097	0.893848	0.650193	0.655058	0.061644

エンジンの更新頻度より高く、更新頻度が高いサイトは他の検索エンジンと比較してランキングが下がる傾向が見られる。この結果を検証するため、各変動パターンにおける検索エンジン間の差について t 検定を行った。仮説や有意水準を以下のように設定する。

- ・ 帰無仮説：各変動パターンにおいて、検索エンジン間で更新頻度に有意な差はない
- ・ 対立仮説：各変動パターンにおいて、検索エンジン間で更新頻度に有意な差がある
- ・ 有意水準：5%

検定した結果の p 値を表 3 に示す。上昇パターンにおいては 2006 年の Google・Yahoo! JAPAN 間と 2007 年の Google・MSN 間で有意水準よりも低い p 値となっており、帰無仮説は棄却され、対立仮説が採択される。つまり、Google は最も更新頻度を重視している可能性が高い。また、下降パターンにおいても、2007 年の Google・Yahoo! JAPAN 間と Yahoo! JAPAN・MSN 間でも有意水準より低い p 値となっている。つまり、Yahoo! JAPAN では他の検索エンジンと比較して、更新頻度が高いとランキングが下がりやすい傾向があるといえる。

5. まとめ

本稿では、大量に収集したランキングのデータを用いて、サイトのランク変動パターンを解明した。さらに、Internet Archive を用いてサイトの特徴を解析し、更新頻度では検索エンジンごとに異なる特徴を発見することができた。

謝辞

本研究の一部は、科学研究費補助金「情報爆発に対応する高度にスケーラブルなモニタリングアーキテクチャ」によるものである。

参 考 文 献

- [1] Abbe Mowshowitz and Akira Kawaguchi, “Measuring

Search Engine Bias,” Information Processing and Management, Vol.41, pp.1193-1204, Sep. 2005.

- [2] Judit Bar-Ilan, Mazlita Mat-Hassan and Mark Levene, “Methods for comparing rankings of search engine results,” Computer Networks, Vol.50, pp.1448-1463, 2006.
- [3] P. Diaconis and R. L. Graham, “Spearman's footrule as a measure of disarray,” J. of the Royal Statistical Society, Series B (Methodological), Vol.39, pp.262-268, 1977.
- [4] C. Dwork, R. Kumar, M. Naor and D. Sivakumar, “Rank aggregation methods for the Web,” In Proc. of the 10th WWW Conf., pp.613-622, May 2001.
- [5] R. Fagin, R. Kumar and D. Sivakumar, “Comparing top k lists,” SIAM J. on Discrete Mathematics, Vol.17, pp.134-160, 2003.
- [6] Franc McCown and Michael L. Nelson, “Agreeing to Disagree: Search Engines and their Public Interfaces,” ACM IEEE Joint Conf. on Digital Libraries (JCDL 2007). 2007.
- [7] 吉田 泰明, 上田 高德, 田代 崇, 平手 勇字, 山名 早人, “商用検索エンジンのランキングに関する定量的評価と特徴解析”, 電子情報通信学会技術研究報告, Vol.107, No.131, pp.441-446, 2007.
- [8] Google SOAP Search API:
<http://code.google.com/apis/soapsearch/>
- [9] Live Search Developer Center:
<http://search.msn.com/developer/default.aspx>
- [10] Yahoo! デベロッパーネットワーク:
<http://developer.yahoo.co.jp/>
- [11] Internet Archive: <http://www.archive.org/index.php>
- [12] Internet Archive Frequently Asked Questions:
<http://www.archive.org/about/faqs.php>