

検索ヒット数のクラスタリングを用いた補正手法の検討

舟橋 卓也[†] 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: {takuya,hirate,yamana}@yama.info.waseda.ac.jp

あらまし 近年、検索エンジンのヒット数を利用した研究が数多く行われている。ここで検索エンジンのヒット数とは、検索エンジンが検索結果として返す、クエリに適合する Web ページ数を示す。ユーザは検索エンジンを利用することによって、ヒット数から手軽にクエリキーワードの利用頻度を知ることができる。ヒット数が広く用いられている一方、検索エンジンのヒット数の信頼性について疑問視をする声もある。そこで、筆者らはヒット数に対する信頼性の検証をこれまで行ってきた。本研究では、既存研究を踏まえて、クラスタリングを用いたヒット数の検証を行う。検索を開始するランキング位置の変化に伴い変動するヒット数の特徴をクラスタリングし、検索クエリとの関係性を調査した。調査の結果、利用頻度の高いと考えられるクエリはヒット数の変動が発生しにくいなどの特徴を認めることができた。

キーワード 検索エンジン、ヒット数、信頼性、情報検索

Reliability Verification of Search Engines' Hit Count by Clustering Analysis

Takuya FUNAHASHI[†] 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

^{†††} 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: {takuya,hirate,yamana}@yama.info.waseda.ac.jp

Abstract In this paper, we verify commercial search engines' hit count, a number returned as a search result count. Since many studies use search engines' hit count to judge the popularity of the query word, the reliability of hit count is indispensable to archive trustworthy researches. However, hit count will vary whenever a user clicks "Next" button on the searched result page, which results in unreliability of hit count. In order to verify search engines' hit count and to try to make clear the characteristics of hit count's transition, we have analyzed them. In this paper, we verify commercial search engines' hit count based on our previous research. By applying clustering analysis, we found search engines' characteristics. In conclusion, we have found that the hit counts of popular queries have tendency not to vary.

Keyword SearchEngine, Hit Count, Web Count, Reliability, Information Retrieval

1. はじめに

近年、検索エンジンのヒット数を利用した研究が数多く行われている。ここで検索エンジンのヒット数とは、検索エンジンが検索結果として返す、クエリに適合する Web ページ数を示す。検索エンジンのヒット数は、検索エンジンが収集した全 Web ページにおいてクエリキーワードが出現する Web ページ数の概算とみなすことができる。そのため、ヒット数を利用して、自然言語処理支援を行う研究[6][2] や、検索エンジンがインデックス化している Web ページ数

の推定を行う研究[4]、クエリ単語間の距離を定義する研究[8]などが存在している。

これらの研究は、検索エンジンが返すヒット数は正しいという仮定の下で行われてきた。しかし、実際には検索エンジンのヒット数は検索を実行した時間や検索開始のオフセットの変化、すなわち検索結果の何ページ目を見ているかにより変動する。検索を実行する時間や検索開始のオフセットの変化によってヒット数が増える理由は、検索エンジンがシステムの構成を明らかにしていないため断定することはできない。しかし、このようなヒット数の変動は現在の検索エンジンの構成上避けられないものだと推測される[1]。

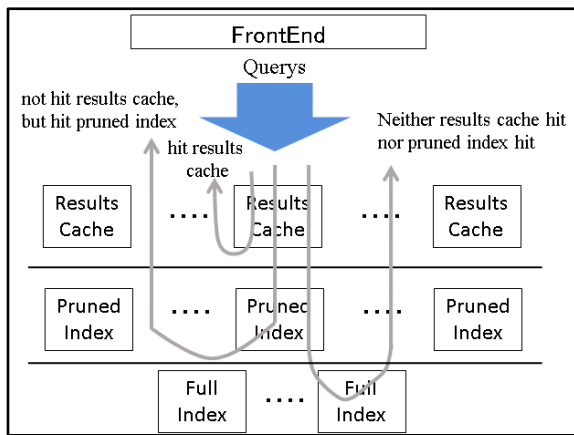


図 1. 検索システムの構成図

ヒット数の変動は、検索エンジンの構成上やむを得ないことであるとはいえ、ヒット数を研究に用いるためにはその特徴や信頼性を知る必要がある。なぜなら、ヒット数の信頼性が低いのであれば、ヒット数を用いた研究の信頼性も低くなるためである。そこで、これまでに筆者らは検索エンジンのヒット数の信頼性に対して検証を行ってきた[1]。これまでの検証の結果、ヒット数はGoogle, Live Search, Yahoo! JAPANでは5割から6割のクエリにおいて、実際に取得可能な結果数の100倍以上大きな値を取るなどがわかった。

本稿では、[1]の研究を発展させ、検索開始オフセットの変化に伴って変動するヒット数を、その変動の傾向によってクエリ毎にクラスタリングすることにより、ヒット数の変動傾向とクエリとの関係を調査する。また、ヒット数の変動傾向とクエリとの関係を用いて、検索ヒット数の補正手法の検討を行う。

以下、次のような構成をとる。まず、2節では一般的な検索エンジンの構成について述べる。次に3節では関連研究として著者らの過去の研究について述べる。続いて4節では、クエリ毎にヒット数の変動傾向をクラスタリングした結果をし、5節では、ラスタリングをした結果をもとに、検索ヒット数の補正手法の検討を行う。最後に6節においてまとめについて述べる。

2. 一般的な商用検索エンジンの構成とそのヒット数

一般的に、現在の商用検索エンジンは、転置インデックスを用いたブーリアン検索モデルで構築されている。検索エンジンは転置インデックスを用いることにより、クエリに合致する検索Webページを短時間で検索することができる。しかし、単位時間当たりのクエリ数が増加をすると、単純に転置インデックスを用いるだけでは高速な検索の提供が困難になる。そこで、検索エンジンは「転置インデックスの分散」や「転置

インデックスの削減」によって、高速な検索処理を実現していると推測される[7]。この構成の概略図を図1に示す。

図1に示した構成のように、検索エンジンが縮小された転置インデックス（Pruned Index）を利用していることを考えると、ヒット数に対して次のような考察を行うことができる。

『ヒット数は縮小転置インデックスにインデックス化されているWebページ、つまり静的なランキングが上位のWebページより統計的に算出されている』

これは、検索エンジンのヒット数で示される検索結果のWebページすべてを検索結果としては表示できないことを意味している。

また、Yahoo!が公開しているWeb検索APIリファレンスには、[ヒット数は検索結果を用いて算出しているため、検索開始オフセットと結果表示数（何件同時に表示するか）によって変化する可能性がある]

ことが記載されている[1]。

上記を踏まえた上で、本稿での検証に使用する3種類の検索エンジンのヒット数と、1つのベクトルを次のように定義する。

検索開始オフセットが1のときのヒット数	: h_f
検索開始オフセットが最大のときのヒット数	: h_l
検索開始オフセットを変化させ取得した	: $\vec{h}$
ヒット数のベクトル	
検索を行うことによって実際に取得できた結果数	: h_r

ここで、「検索を行うことによって実際に取得できた結果数」は、検索結果の「次へ」を順次クリックを行った際に、実際に検索結果として表示される検索結果（WebページのURL）の総数を示す。また、「検索開始オフセットを変化させ取得したヒット数のベクトル」 $\vec{h}$ は、検索開始オフセットを1から順に10ずつ増加をさせ取得を行い作成したヒット数のベクトル $\{h_f, \dots, h_l\}$ を示す。

検索エンジンの構成を考えると、Yahoo! APIのリファレンス[1]で説明されている通り、 h_l は h_f より正確なヒット数となることが推定できる。

3. 関連研究

著者らは、これまでに検索エンジンのヒット数の信頼性に対する検証を行ってきた。本節では、関連研究として筆者らのこれまでの研究について述べる。

筆者らは、検索エンジンが2節のような特徴を持つことを踏まえ、ヒット数の変動幅に対する検証を行ってきた。2節において、検索開始オフセットが大きいほどヒット数の値が正確であると考えられることについて述べた。そこで、より正確であると考えられる、

表 1. 検索エンジンの検索オプション設定

	Google	Yahoo! JAPAN	Live Search	Yahoo!
1度取得する結果数	10	10	10	10
類似フィルタ	off	off	off	off
アダルトフィルタ	off	off	off	off
検索対象とする言語	全言語	全言語	日本語	日本語

検索開始オフセットが最大のときのヒット数 h_l と、多くの研究で最もよく利用されているヒット数である検索開始オフセットが 1 のときのヒット数 h_f の比較を行い、これらの間にどの程度の差が存在しているのかを検証した。また、 h_l と実際に取得することができた検索結果数 h_r との比較も行い、実際に取得できた結果数とヒット数の間にどの程度の差が存在しているかについても検証を行った。

検証を行う際に、検索エンジンの検索結果のフィルタリングを極力避けるために、表 1 に示すように検索オプションの設定を行った。また、AND 検索による不定要因を極力排除するため、Wikipedia より検索エンジンが保有する転置インデックスのキーとなっていると考えられるキーワードを 500 個選出し、検索に用いた。

商用検索エンジンである Yahoo! JAPAN, Google, Live Search, Yahoo を対象に検証を行った結果、Yahoo! JAPAN, Google, Live Search では 5~6 割のクエリにおいて、 h_l は h_f に対して 100 倍以上大きな値を取るなどなどがわかった。

4. 検索開始インデックスの変化に伴うヒット数の変動傾向によるクラスタリング

本稿では、これまでの研究を踏まえ、検索開始オフセットの変化に伴って変動するヒット数を、その変動の傾向によってクエリ毎にクラスタリングすることにより、ヒット数の変動傾向とクエリとの関係を調査する。

そこで、本節では、本研究で用いるクラスタリング手法と、その結果について述べる。

4.1. クラスタリング手法

本研究におけるクラスタリングの目的は、ヒット数の変動傾向が似たクエリをグループ化することである。ここで、ヒット数の変動傾向が似ている、とは、2 節にて定義を行った $\vec{h}$ が似ていることを指す。

また、クラスタリングを行う際に用いる類似度の定義を次のように行う

$$(\text{ベクトル } a \text{ の積分値} - \text{ベクトル } b \text{ の積分値})^2$$

ここで、ベクトルの積分値とは、ベクトルの要素をすべて足し合わせた値を指す。この類似度は距離の公理を満たすため、距離として扱うことができる。

本研究では、上記の類似度を用いて k-means クラス

タリングを行った。なお、k-means クラスタリングを行う際に指定をするクラスタ数は、1-10 の値をそれぞれの検索エンジンにおいて指定した後に、著者がもっとも適当にクラスタ化されていると判断したクラスタサイズを用いた。

クラスタリングの対象としたデータは、関連研究[1]と同様のデータセットを用いた。ただし、本研究では 5 節にて、ヒット数の変動傾向より実際に検索エンジンが検索結果として表示できる検索結果数の推定を行う。そこで、本クラスタリングでは、 h_r が検索エンジンから取得可能な最大検索結果数 1000 件未満であるクエリのみを対象に、クラスタリングを行った。そのため、クラスタリングの対象となった要素数は検索エンジン毎に異なっている。

4.2. クラスタリング結果

4.1 で示した手法に基づいて、Yahoo! JAPAN, Google, Live Search, Yahoo に対してそれぞれクラスタリングを適用し、そのクラスタの平均値を図示した図を図 2~図 5 に示す。また、各検索エンジンについて、各クラスタの大きさをまとめた表を表 2 に、それぞれのクラスタに含まれるキーワードからサンプルを取り、リストアップした表を表 3~表 6 に示す。

クラスタリングを行った結果をみると、Yahoo! JAPAN, Google, Live Search のクラスタは、どの検索エンジンも次の 2 つの特徴を持つクラスタが存在することがわかる。

1. ヒット数が徐々に減少するクラスタ
2. ヒット数がほぼ一定のクラスタ

Yahoo! JAPAN, Google, Live Search において、最も大きなクラスタとなっているのは「2. ヒット数がほぼ一定のクラスタ」である。ヒット数がほぼ一定のクラスタには、Yahoo! JAPAN では「クロロフィル」や「西巢鴨」、Live Search では「ヘルシア」や「船員」といったキーワードが含まれている。これらのキーワードは、他のクラスタに含まれるキーワードと比較して、知名度が高いキーワードが多いように思われる。検索エンジンは一般的に、検索負荷を減らすため間近に検索されたクエリに対する検索結果をキャッシュしている。そのため、知名度の高いキーワード、つまり利用頻度の高いと考えられるキーワードをクエリとして用いた場合、その検索結果はキャッシュに残っている確率が高い。もし、キャッシュに残っている結果を表示しているのであれば、キャッシュに残っている結果を表示している間はヒット数は明らかに一定となる。

そこで、ヒット数がほぼ一定となったクラスタには本当にキャッシュに残っているような一般的な単語が多いのかどうかを確認するために、クエリとして用いた Wikipedia の記事名を持つ記事がどの程度のページ

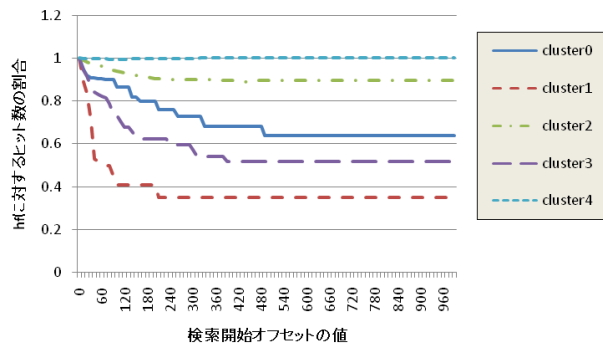


図 2. Yahoo! JAPAN のクラスタリング結果

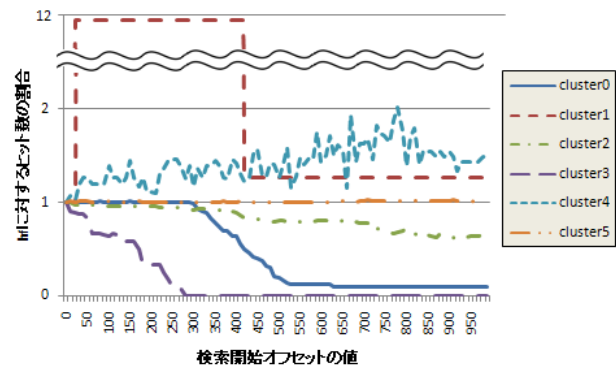


図 3. Google のクラスタリング結果

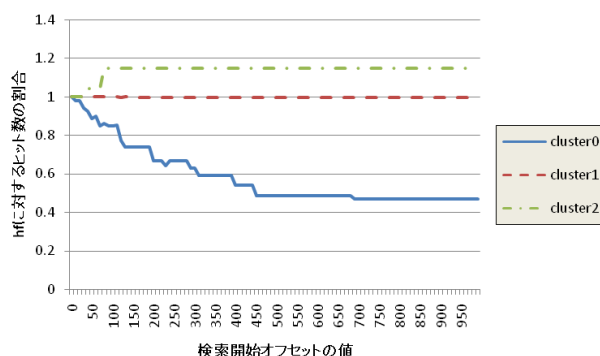


図 4. Live Search のクラスタリング結果

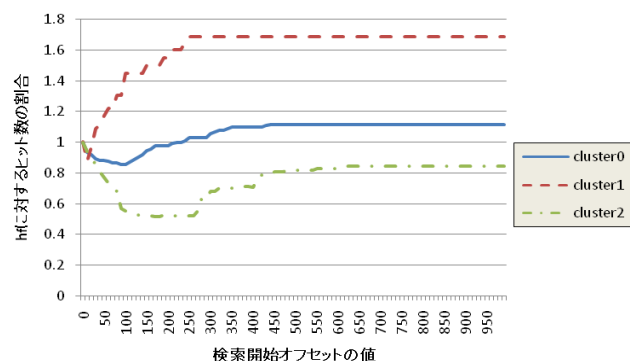


図 5. Yahoo! のクラスタリング結果

表 2. 検索エンジン毎のクラスタサイズ

	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	合計
Yahoo! JAPAN	6.1% (8)	4.5% (6)	18.2% (24)	13.6% (18)	57.6% (76)		100% (132)
Google	10.5% (42)	0.3% (1)	11.3% (45)	3% (12)	3.5% (14)	71.5% (286)	100% (400)
Live Search	4% (15)	95.5% (357)	0.5% (2)				100% (374)
Yahoo!	61.3% (65)	20.8% (22)	17.9% (19)				100% (106)

ビューを持っているのかを調べた。本論文ではページビューを確認するために Wikipedia article traffic statistics[9]を用いた。クエリとして用いた単語が記事名である Wikipedia 記事のページビューをすべてのクエリに対して算出したのち、先ほど求めた各クラスタに含まれる単語に対するページビュー数の平均値を算出した。その結果を表 7 に示す。なお表 7 では、検索結果を取得した時期と同じ、2008 年 11 月におけるページビューを示した。

表 7 を見ると、Yahoo! JAPAN、Live Search においては、最もページビューの平均値が大きなクラスタが「ヒット数がほぼ一定のクラスタ」となっており、他のクラスタと比べて知名度の高いキーワードが多いことがわかる。これより、Yahoo! JAPAN、Live Search においては、知名度の高いキーワードを用いたヒット数はほぼ一定となることが確認できた。これは、Yahoo! JAPAN と Live Search において、利用頻度の高いクエリに対してヒット数が最適化されているか、検索キャッシュを検索開始オフセットが最大の範囲まで持って

いることを示していると考えられる。

これに対して、Google においてページビュー数の最も大きなクラスタ (cluster4) は、ヒット数が上昇をするクラスタとなっている。また表 7 をみると、ヒット数の変動がないクラスタ (cluster5)、検索開始オフセットが 250 位付近で急激にヒット数が落ち込むクラスタ (cluster3) も同様にページビュー数が多い。この原因としては、Google は検索結果のキャッシュとしてランキングの最上位部分しか保持していないことが考えられる。もしキャッシュとしてランキング最上位部分しか保持していないと仮定すると、ランキング下位部分の取得を行う際には検索結果の検索を実行する必要がある、その結果を用いてヒット数を算出しているため、変動が発生するのではないかと考えられる。また、Google では検索を行った際に、 h_r が 500 を超えた値でないと、 h_l を h_r と等しい値を返す特徴がある。そのため、Google においてはヒット数の大幅な減少が発生するクラスタにおいても、平均ページビューが高くなっていると考えられる。

表 3. クラスタに含まれるクエリキーワードのサンプル (Yahoo! JAPAN)

Yahoo! JAPAN					
	cluster0	cluster1	cluster2	cluster3	cluster4
キーワード	カスタニャーロ	セキリュウヒ	ガルラ	ヴィアリージ	Nullpo
	クシャラ	モングエルフォ	巡撫	アルヴィート	クロロフィル
	リトヴィノフ	ヴァルネグラ	ベンゾキノ	ゾイガー	西巣鴨

表 4. クラスタに含まれるクエリキーワードのサンプル (Google)

Google						
	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5
キーワード	永字八法	ジャアファル	軟水	ミソ	ゲルニカ	組合
	麻酔		ペトリュラ	二科展	光沢	香辛料
	クイディッチ		古文	史学	ウォリアーズ	ゴージャス

表 5. クラスタに含まれるクエリキーワードのサンプル (Live Search)

Live Search			
	cluster0	cluster1	cluster2
キーワード	蛭	ヘルシア	リビジョニズム
	パワキャス	COUNTDOWN	ゾイガー
	リトヴィノフ	船員	

表 6. クラスタに含まれるクエリキーワードのサンプル (Yahoo!)

Yahoo!			
	cluster0	cluster1	cluster2
キーワード	五七調	素因数	竜骨
	ロディゴ	オムパズ	盗作
	生気	ナノティラス	ボドゾル

表 7. 各クラスタにおけるページビュー数の平均値 (2008.11.1 ~ 11.30)

	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5
Yahoo! JAPAN	143.1	23.5	205.6	148.8	1404.9	
Google	632.9	88	887.5	1420.7	3048.1	1502
Live Search	143.2	1679.1	7			
Yahoo!	684.2	117.4	520.7			

また表 7 より, Yahoo! JAPAN, Live Search において, 検索開始オフセットの値を増加させるにつれてヒット数が減少するクラスタのページビューは, ヒット数が変化をしないクラスタと比べて 1/10 程度の大きさとなっていることがわかる. これは, 知名度の低いキーワードで検索をした場合, h_f の値が h_l の値に比べて大きな値を返す傾向があることを示している. 特に Yahoo! JAPAN においては, ほぼすべてのクラスタにおいてページビューの値が低いほど, 検索開始オフセットを変化させたときのヒット数の減少の割合が高くなっている. そのため, Yahoo! JAPAN では知名度の高いクエリを利用した際のヒット数が正確になるようにカスタマイズされているのではないかと考えられる.

Yahoo! においては, 検索開始オフセットが 200~300 付近でヒット数の増加がみられる. そして, 増加をしたのちにヒット数の変動が収まり一定となる. この結果より Yahoo! では, Google 同様検索結果の最上位部分の結果しかキャッシュしていないため, 通常あまり検索されないランキングの低い結果を要求されると再検索を行い, ヒット数の更新を行っている可能性が考え

られる.

5. クラスタリングの結果を用いた検索ヒット数の補正手法の検討

5 節ではクラスタリングの結果を踏まえ, 各検索エンジンに対して検索ヒット数の補正手法を検討する. 4 節で述べたように, 検索エンジンのヒット数の変動の傾向はクエリの特徴によってある程度推測される. そこで, 本節では実際に取得することのできた検索結果数 h_r と, ヒット数の変動傾向を結びつけた, クラスタリングを用いた検索ヒット数の補正手法に対して検討を行う.

5.1. 各検索エンジンの特徴

4 節の結果から Google, Yahoo! JAPAN, Live Search, Yahoo! におけるヒット数に関して, 以下のような特徴が導かれる.

- Google

知名度の高いクエリはヒット数が「上昇する」, 「一定」となる, もしくは「検索開始オフセットが 300 の時まで大幅に減少する」. 知名度の低いクエリは,

表 8. 検討する補正手法

検索エンジン名	対象とするクエリ		補正手法
Google	知名度の高いクエリ	上昇	$hf \times 1.73$
		一定	hf
		大幅な減少	検索開始オフセットを 500 に設定し、取得したヒット数
	知名度の低いクエリ	徐々に減少	$hf \times 0.63$
		大幅に減少	検索開始オフセットを 500 に設定し、取得したヒット数
Yahoo! JAPAN	知名度の高いクエリ	一定	hf の値
	知名度の低いクエリ	徐々に減少	$hf \times 0.52(\text{cluster3})$
Live Search	知名度の高いクエリ	一定	hf
	知名度の低いクエリ	徐々に減少	$hf \times 0.47$
Yahoo!	すべてのクエリ		ヒット数が一度落ち込み、その後上昇をして一定となったヒット数

ヒット数が「徐々に減少をする」か、「検索開始オフセットが 500 の時まで大幅に減少する」。

- Yahoo! JAPAN

知名度の高いクエリはヒット数が「一定」となる。
知名度の低いクエリはヒット数が「徐々に減少する」。
その減少の大きさは、知名度の低さに応じて大きくなる。

- Live Search

知名度の高いクエリはヒット数が「一定」となる。
知名度の低いクエリはヒット数が「徐々に減少する」。
一部のクエリではヒット数が「上昇する」が、比較的早い段階で上昇は止まり、その後一定となる。

- Yahoo!

どのようなクエリも、一度ヒット数が減少に転じたのち、再び上昇に転じて一定となる。

なお、ここで用いているクエリの知名度は、検索エンジンにおけるそのクエリの利用頻度を指す。しかし、一般には検索エンジンにおけるキーワードの利用頻度は取得することは難しい。そのため、ここではクエリの知名度があらかじめ明らかとなっているという前提で議論を行う¹。

5.2. 補正手法の検討

前項で述べた検索エンジンの特徴を踏まえて、次のような補正手法の適用を検討する。すべての検索エンジンについて検討を行う補正手法を、表 8 に示す。表 8 に示した補正手法は、図 3～図 5 のヒット数の変動傾向と、クラスタリングに基づいて、数値を算出した。

6. おわりに

本論文では筆者らの既存研究を踏まえ、ヒット数の変動傾向に対してクラスタリングを行った。また、クラスタリングの結果を利用したヒット数の補正手法の提案と検討を行った。クラスタリングの結果、利用頻度の高いと考えられるクエリはヒット数の変動

が発生しにくいなどの特徴を認めることができた。また、その結果を基にしたヒット数の補正手法に関しても提案を行った。

今回の実験は、2008 年 11 月 10 日から 11 月 18 日にかけて収集をしたデータを対象に実験を行った。本稿で記載した結果は、今後検索エンジンのシステムの仕様変更などによって変化をする可能性がある。

今後の課題として、本論文で提案した補正手法が、どの程度の有効性を持つのかとともに、取得可能な限界数を超えるヒット数に関しても適用が可能かどうか検証を行っていく。

文 献

- [1] 舟橋卓也, 上田高德, 平手勇宇, 山名早人: “商用検索エンジンのヒット数に対する信頼性の検証”, 日本データベース学会論文誌, Vol.7, No.3, pp31-36(2008.13)
- [2] 大鹿広憲, 佐藤進, 山名早人: “Google を活用した英作文支援システムの構築”, DEWS2005, 4B-i8 (2005.3)
- [3] Google: <http://www.google.co.jp/>
- [4] K. Bharat and A. Border: “A technique for measuring the relative size and overlap of public web search engine,” Proc. of the 7th Int'l Conf. on WWW, pp.379-388 (1998.4).
- [5] Live Search: <http://www.live.com/>
- [6] M. Lapata, and F. Keller: “Web-based models for natural language processing,” ACM Trans. Speech and LanguageProcessing, Vol.2, No.1, pp.1-31 (2005.2).
- [7] R. Kumar, K. Punera, T. Suel, S. Vassilvitskii, “Top-k aggregation using intersections of ranked inputs,” Proc. of the Second ACM International Conference on Web Search and Data Mining, pp.222-231 (2009.2).
- [8] R. L. Cilibrasi, P. M. B. Vitanyi: “The Google Similarity Distance,” IEEE Trans. Knowledge and Data Engineering, Vol.19, pp.370-383 (2007.3).
- [9] Wikipedia article traffic statistics: <http://stats.grok.se/>
- [10] Yahoo!: <http://yahoo.com/>
- [11] Yahoo! JAPAN: <http://www.yahoo.co.jp/>
- [12] Yahoo! Search BOSS API Guide: [http:// developer.yahoo.com/search/boss/boss_guide/](http://developer.yahoo.com/search/boss/boss_guide/)

¹ 英語であれば、<http://freekeywords.wordtracker.com/> を参照することによりある程度の知名度を把握できる。