

## 商用検索エンジンのヒット数に対する信頼性の検証

舟橋 卓也<sup>†</sup> 上田 高德<sup>†</sup> 平手 勇宇<sup>††</sup> 山名 早人<sup>†††</sup>

<sup>†</sup> 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

<sup>††</sup> 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

<sup>†††</sup> 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,ueda,hirate,yamana}@yama.info.waseda.ac.jp

**あらまし** これまでに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。こうした研究では、当該クエリに対するヒット数を用いることにより、翻訳支援や自然言語処理支援など様々なアプリケーション構築を目指している。従来、検索エンジンのヒット数は信頼できるという仮定の下で用いられてきた。しかし検索エンジンが返すヒット数の信頼性に対する検証は筆者の知る限り行われていない。もしもヒット数が不正確な場合、ヒット数を利用した研究の信頼性は疑わしいものとなる。そこで本論文では、検索エンジンのヒット数に対してその信頼性の検証を行う。検証実験では、日本で広く用いられている商用検索エンジンである Google, Yahoo! Japan, Live Search が提供している検索 API を利用した。また実験に利用するクエリとして、日本語 Wikipedia からランダムに選択した 1,000 件の記事タイトルを使用した。検証実験の結果、ヒット数は 5%~50%の確率で、推定を行った理想的なヒット数に対して 5%以上大きさが異なることがわかった。

**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

Takuya FUNAHASHI<sup>†</sup> Takanori UEDA<sup>†</sup> Yu HIRATE<sup>††</sup> and Hayato YAMANA<sup>†††</sup>

<sup>†</sup> Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo,  
Shinjuku-ku, Tokyo 169-8555 Japan

<sup>††</sup> Media Network Center, Waseda University 1-104 Totsuka-cho, Shinjuku-ku, Tokyo, 169-8050, Japan

<sup>†††</sup> 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,ueda,hirate,yamana}@yama.info.waseda.ac.jp

**Abstract** A number of studies have been using Search Engines' hit count. The goal of these studies is to build applications for translation support or natural language processing support. These studies assume that the hit count is reliable. However, none of the studies have been verified the reliability of Search Engines' hit count. If the hit count is unreliable, studies using hit count become also unreliable. The purpose of this paper is to verify the reliability of Search Engines' hit count. In this experiment, we used Search APIs provided by Google, Yahoo! Japan and Live Search. Furthermore, we randomly randomly extracted 1,000 keywords from the titles of the articles on Wikipedia as queries. The verification experiment shows that hit count is depart from estimated ideal hit count in 5 - 50% of the time.

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

### 1. はじめに

これまでに、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。検索エンジンのヒット数とは、検索エンジンが返す、ユーザが入力したクエリに適合する Web ページ数の概算を指す。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[5]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[1]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によって検索エンジンのヒット数が変動する理由は、検索エンジンがシステム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、現在一般的に提案されている検索エンジンの構成より、ヒット数が変動する理

由として次のような原因を推定することができる。

- ・インデックスの不一致に起因  
一般的に検索エンジンは負荷を分散するため複数台のサーバを持ち、それぞれ個別のインデックスを持つ。分散されたサーバに存在するインデックスは、同期処理を行っているものの、完全な同期は難しいと考えられる。そのため、クエリがどのサーバに受け渡されるかによって結果が異なるのではないかと考えられる。
- ・AND 処理や OR 処理の仕組みに起因  
多くのユーザは検索結果の上位しか参照しない。また、Google, Yahoo! JAPAN, Live Search 共に最大 1,000 件の検索結果しか返さない。すなわち、AND, OR といった処理を行う際、クエリに該当する全ての Web ページに対して AND や OR を行う必要性はなく、最終的に上位にランキングされるであろう Web ページに対して論理演算を適用することにより検索効率が高くなる。このため、これらの検索エンジンでは何らかの方法を用いてヒット数の概算を求めているのではないかと考えられる。

このように、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ならば、ヒット数を用いた研究の信頼性も低くなるためである。しかし、これまでに検索エンジンのヒット数の信頼性に対して大規模な検証が行われていない。そこで本論文では、検索エンジンのヒット数に対してその信頼性の検証を行う。

これまでに行われた検索エンジンのヒット数を利用した研究は、主に複数のクエリによって得られたヒット数の大小関係に影響を受ける研究と、主に1つのクエリのヒット数の値に影響を受ける研究の2種類に分類することができる。それらを踏まえ、本稿ではその2種類の利用方法におけるヒット数の信頼性を検証する。なお、今回の検証では日時によるヒット数の違いについては、検証対象としない。また、本論文では、検索エンジンのヒット数が日時や検索開始オフセットによって異なる原因を解明することを目的とはしていない。

本稿では、まず複数のクエリ間において検索開始オフセットを変化させることにより、クエリ間のヒット数の大小関係が入れ替わるか調査を行った。続いて、1つのクエリにおける、検索開始オフセットの変化によるヒット数の変動幅の検証を行った。

以下、次のような構成をとる。まず、2節では関連研究として検索エンジンのヒット数を利用した研究と、

検索エンジンを利用した研究の是非について考察した研究について述べる。次に3節では、検索開始オフセットの変化によって複数のクエリ間にてヒット数が入れ替わるのか検証した結果について述べる。続いて4節では、検索開始オフセットを変化させることによって取得した単一のクエリに対する複数のヒット数は、どの程度の幅を持って変化をし、その変化がどの程度の確率で発生するのかについて検証を行う。最後に、5節においてまとめについて述べる。

## 2. 関連研究

2節では本論文の関連研究について述べる。まず2.1ではこれまでに行われてきた検索エンジンのヒット数を利用した研究について述べる。続いて2.2では検索エンジンを利用した研究の是非について考察した研究について述べる。

### 2.1. 検索エンジンのヒット数を利用した研究

これまでに検索エンジンのヒット数を利用した研究がおこなわれてきた。その中で、代表的な研究として、次のような研究が挙げられる。

Lapata ら[5]は、機械翻訳支援のために、翻訳を行う際に候補に挙げた複数のフレーズのうち、どのフレーズが最も一般的に用いられているかを検索エンジンのヒット数を用いることによって判定をしている。つまり、よりヒット数が大きいフレーズが最も一般的に用いられていると見なす。ここでフレーズとは連続する単語の集合を指す。

Gulli ら[1]は、検索エンジンからのヒット数と検索結果を利用して、検索エンジンがインデックス化している Web ページの数を算出している。また、複数の検索エンジンのインデックス化している Web ページ数を求めることにより、インデックス化可能な世界中の Web ページ数の予測を行っている。

これらの研究では、そのヒット数の用い方によって2種類の研究に分類することができる。1つは、主に複数のクエリによって得られたヒット数の大小関係に影響を受ける研究、もう1つは主に1つのクエリのヒット数の値に影響を受ける研究である。例えば、Lapata らの研究では、ヒット数が正確な値でなくとも、複数のフレーズ間におけるヒット数の大小関係が誤っていなければ研究結果に影響がない。対してヒット数を使用して Web ページの数を求めている Gulli らの研究では、ヒット数が信頼できない値であると、ヒット数を用いて計算を行った Web ページ数の信頼性も疑わしいものとなる。

### 2.2. 検索エンジンを用いた研究の是非を問う研究

Gilgarriff は検索エンジンの有用性を認めながらも、検索エンジンのシステムがブラックボックスであるこ

表 1. 検索エンジンの検索設定

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

とから検索エンジンを利用した研究に対して疑問を呈した[1]. [1]の中で Kilgariff は検索エンジンのヒット数に関する小規模な実験を行った. 1日に同じクエリキーワードを用いて 30回検索を行ったところ, 30回のうち 9回において, 1回目に取得したヒット数から 10%以上の変化が起きていることがわかった. Kilgariff は検索エンジンのシステムはブラックボックスであるので, 検索エンジンを用いた研究は行うべきではないと述べている. 本論文では, 検索エンジンが返すヒット数の特徴と信頼性について述べ, それを踏まえた上でどのようにヒット数を研究へ活用すればよいのか考察を行う.

### 3. 検索開始オフセットの変化による複数クエリ間のヒット数入れ替わりの調査

2節において, これまでに行われてきた検索エンジンのヒット数を用いた研究には, 主に複数のクエリによって得られたヒット数の大小関係に影響を受ける研究と, 主に1つのクエリのヒット数の値に影響を受ける研究の2種類があることについて述べた. 前者の研究においては, たとえヒット数の変動が発生してもすべてのクエリにおいて同様にヒット数の変動が発生していた場合には大小関係の入れ替わりは発生しない. そこで3節では複数のクエリによって得られたヒット数の大小関係が, 検索開始オフセットの変化によって入れ替わりうるのかどうか, 調査を行う. まず 3.1にて調査方法について述べる. 続いて 3.2にて調査結果について述べる.

#### 3.1. 調査方法

本調査の目的は, 検索開始オフセットを変化させることによって, 任意のクエリキーワード  $k_1, k_2, \dots, k_n$  による検索エンジンのヒット数  $h_1, h_2, \dots, h_n$  の大小関係が入れ替わりうるのかどうか調査をすることである. そこでまず次の (1) ~ (3) の手順で,  $k_i$  に対して検索開始オフセットを1から順に10ずつ増加させヒット数を取得する.

- (1) 検索開始オフセットを1に設定し<sup>1</sup>,  $k_i$ のヒット

<sup>1</sup> 実際に API に指定するオフセットとは異なる. 検

表 2. ヒット数の大小関係の入れ替わり調査結果

	クエリ数	大小関係を比較するクエリの組み合わせ数	大小入れ替わりの発生数	大小入れ替わりの発生率
Google	1,000	499,500	55,296	11.07%
Yahoo! Japan	1,000	499,500	12,744	2.55%
Live Search	952	452,676	26,609	5.88%

数  $h_{i-1}$  を取得する. ここで,  $_$ の後に続く数字は  $k_i$  に対して検索開始オフセットを変動させヒット数を取得した回数を表す.

- (2) 検索開始オフセットを前回の検索時から 10 増加させ, ヒット数  $h_{i,j}$  を取得する. ここで  $j$  は自然数を表し,  $k_i$  を用いて検索をした回数を表す.
- (3) (2) を検索結果が取得できなくなるまで繰り返す. 検索結果が取得できなくなったときは, 検索開始オフセットのランキングを持つ検索結果を取得できなくなった状態を指す. また,  $k_i$  に対して最後に検索結果が取得できたときの  $j$  の値を  $m_i$  とする.

上記の手順を  $k_1, k_2, \dots, k_n$  に適用し, それぞれのキーワードに対して検索開始オフセットを変化させヒット数を取得する. そして  $k_1, k_2, \dots, k_n$  の中から任意のキーワードペアを作り, そのキーワードペアにおいてヒット数の大小関係の入れ替わりが発生しているか調査する. 本論文においてヒット数の大小関係の入れ替わりが発生したとは, 任意のキーワードペアにおいて検索開始オフセットを変化させることにより一度でも他と異なるヒット数の大小関係が現れることを指す. また比較の際に同一となった場合は, 大小関係が入れ替わっていないとみなす.

本調査では, 日本において広く用いられている検索エンジンである Google[3], Yahoo! Japan[6], MSN Live Search[4]が公開している Web 検索 API を用いた. Web 検索 API を用いた場合, 直接 Web ブラウザからヒット数を取得する場合と異なった結果となる場合がある. しかし, 多くの場合検索エンジンを用いた研究は Web 検索 API を通して行われる. そこで本実験では同様に Web 検索 API を用いてヒット数の取得を実施した.

調査ではクエリとして日本語 Wikipedia からランダムに選択したリダイレクトでない項目名<sup>2</sup>1000件を使

検索エンジンによっては, 検索開始オフセットが 0 から始まる場合がある. 本論文では表記法を統一するため, 検索結果の  $n$  番目 (1~) からを出力する場合のオフセットを  $n$  として表記する.

<sup>2</sup> 通常 Wikipedia の項目名は正式名称を用いて作成される. しかしその項目名に別称が存在するような場合, 別称で記事検索が行われても正式名称の記事が閲覧できるよう, 別称の記事から正式名称の記事へと自動的に転送を行う機能が存在する. この機能をリダイ

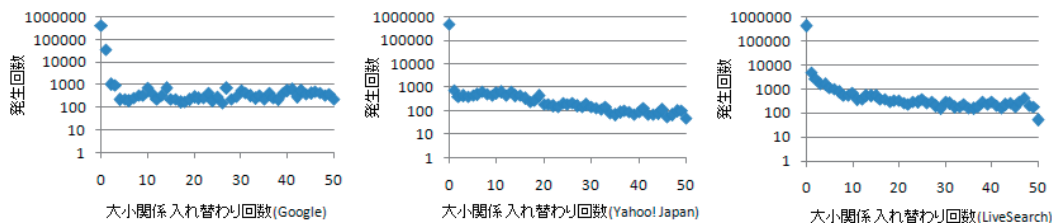


図 1. 大小関係の入れ替わり回数とその発生頻度

用した。また検索エンジンは、表 1 に示すように検索時のオプション設定を行った。Live Search では主に検索対象とする言語として日本語を指定している。これは、Live Search において検索対象とする言語に全言語を対象とすることができなかったためである。

### 3.2. 調査結果

調査を行った結果を表 2 に示す。表 2 のクエリ数はクエリとして使用したキーワード数を、組み合わせ数はヒット数の大小関係の入れ替わりを調べたキーワードペアの組み合わせ数を、発生数はヒット数の大小関係の入れ替わりが一度でも発生したキーワードペアの数を、発生率は発生数/組み合わせ数より求めた大小関係の入れ替わりが発生する確率をそれぞれ示している。Live Search においてクエリ数が 1,000 でないのは、検索時のエラーによって検索を行うことができなかったクエリが存在するためである。結果を見ると、検索開始オフセットを変化させることによって Google では 11.06%、Yahoo! Japan では 2.55%、Live Search では 5.88% のキーワードペアにおいてヒット数の大小関係の入れ替わりが発生している。この結果では、ヒット数の入れ替わりが 1 度でも発生したキーワードペア数を発生数として数えている。

次に、詳しく結果を分析するために、検索エンジンのヒット数の入れ替わりが発生したペアにおいて、何度ヒット数の入れ替わりが発生したのか測定した結果を図 1 に示す。図 1 は、任意のクエリキーワードペアにおいてヒット数の入れ替わりが発生した回数  $c$  を  $x$  軸に、ヒット数の入れ替わりが  $c$  回発生したキーワードペアの数を  $y$  軸に示した片対数グラフとなっている。ここで  $c$  は、任意のキーワードペアにおけるヒット数の大小関係の比較において少数派となったヒット数の入れ替わり発生回数とする。例えば、任意のキーワードペア  $k_1, k_2$  においてそれぞれ 100 回検索開始オフセットを変化させ取得したヒット数を比較した結果、 $k_1 > k_2$  となった結果が 70 回、 $k_2 < k_1$  となった結果が 30 回であるとき  $c$  は 30 となる。本実験では検索開始イン

デックスを最大で 100 回変化させているため、 $c$  は 0 から 50 までの自然数となる。図 1 を見ると、すべての検索エンジンにおいて、 $c$  回ヒット数の大小関係の入れ替わりが発生したキーワードペアの数は  $c$  が増えるにつれて減少する傾向があることが分かる。

### 4. 検索開始オフセットの変化によるヒット数の変動幅の検証

4 節では検索開始オフセットによって 1 つのクエリキーワードによるヒット数がどの程度の幅を持って変動するかについて検証する。

まず、4.1 にてヒット数の変動幅調査のための実験方法について述べる、次に、4.2 にて変動幅調査のために設定した仮定が正しいかどうか事前実験を行う。続いて 4.3 にて実験結果について述べる。最後に 4.4 にて 4 節のまとめを行う。

#### 4.1. 実験方法

本実験では 3 節で行った調査方法と同様の手法でヒット数を取得する。つまり、任意のクエリキーワードに対し検索開始オフセットを変化させることによって、それぞれの検索開始オフセットにおけるヒット数を取得する。これにより、検索開始オフセットの変化により変動するヒット数の変動幅を調査する。

具体的には次のような検証方法をとる。まず、検索エンジンがインデックス化しており、かつクエリキーワード  $q$  を含む Web ページ数、すなわち理想的なヒット数を  $I$  とする。次にこの  $I$  を、検索開始オフセットを変化させることによって取得したヒット数から推定する。最後に、推定した  $I$  と、検索開始オフセットを変化させて得たヒット数の差から、ヒット数の変動幅について検証を行う。

$I$  の推定方法として、次のような 2 つの方法が考えられる。

- (1) 検索エンジンは、最終的にランキング上位に表示される Web ページ群から、統計的な推定を行ってヒット数を求めていると仮定すると、推定の母体となる Web ページ数がより大きな場合に、より信頼できる結果が得られる。推定の母体となる Web ペー

レクトという。本論文ではリダイレクトを行わない項目名のみをクエリキーワードとして用いた。

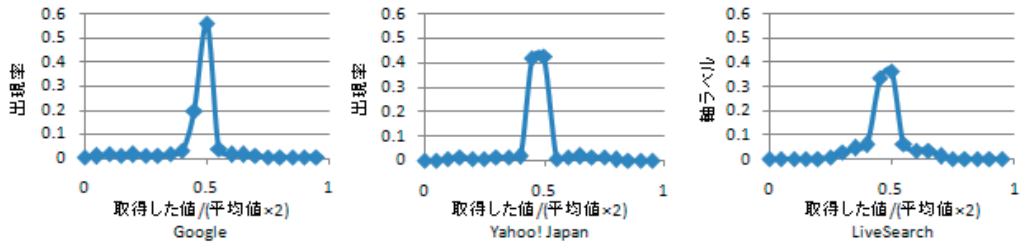


図 2.  $I$  の推定値に対する実際に取得したヒット数の誤差分布

は、検索結果として表示する数が多くなったとき、つまり検索開始オフセットが大きな値のときに、より大きくなるが予想される。そこで、検索開始オフセットが最も大きな値のときのヒット数が  $I$  であると仮定を行う。

- (2) 検索エンジンが返すヒット数は、 $I$  に対して正規分布に従う誤差が含まれたものであると仮定する。この誤差が正規分布に従うならば、 $I$  は検索開始オフセットを変化させて得たヒット数の平均値とみなすことができる。

本実験ではこの(1)、(2)の推定方法を両方ともを用いて  $I$  の推定を行う。

続いてヒット数の変動幅を求めるため  $I$  を推定した後、計測の結果得られたヒット数の変動幅と、変動する確率を求める。そして、変動の幅と変動する確率が小さいほど、信頼性が高いと考える。

実験において検索エンジンは Google, Yahoo! Japan, Live Search を使い、その設定は 3 節における実験と同様に表 1 に従うものとする。また使用するクエリも 3 節で使用したクエリと同様に日本語 Wikipedia からランダムに選択したりダイレクトでない項目名 1,000 件を使用する。

#### 4.2. 事前実験

4.1 で述べた  $I$  の推定方法のうち(2)では  $I$  の推定値に対するヒット数の誤差が正規分布に従うと仮定を行っている。そこで事前実験としてこの誤差が正規分布にしたがっているかどうか簡単な実験を行う。実験には、任意のクエリキーワードにおいて検索開始オフセットを変動させることにより取得したヒット数  $h_1, h_2, \dots, h_n$  と、その平均  $h_{ave}$  を用いる。ここで  $n$  は検索開始オフセットを変動させた回数を表す。 $n$  が十分大きいとき、つまり測定を行ったサンプル数が十分大きいときに、 $h_{ave}$  は  $I$  の推定値と等しいと考えることができる。そのため  $h_{ave}$  と  $h_1, h_2, \dots, h_n$  の差が  $I$  の推定地からの誤差となる。この誤差がどのように分布しているのか調べるためまず  $h_1, h_2, \dots, h_n$  を  $h_{ave} \times 2$  の値で正規化する。次にその値を 0, 0.05, 0.1, 0.15, ..., 0.9, 0.95 の 20 個の値にそれぞれ切り捨て、その出現

回数を数える。これを複数のクエリキーワードについて行い、すべてのクエリにおける 0 の出現回数の和、0.05 の出現回数の和、..., 0.95 の出現回数の和をそれぞれ求める。最後に出現回数の和よりそれぞれの値の出現率を求める。もしもヒット数を正規化し切り捨てを行った値と出現率の関係が正規分布に従っていれば、誤差は正規分布に従うと仮定を行っても問題がないこととなる。

この事前実験は 4.1 において説明した実験環境と同様のクエリ・環境を用いて実験を行った。実験を行い求めた、ヒット数を正規化し丸めこんだ値と出現率の関係を図 2 に示す。図 2 において平均値の 2 倍を超えた値の分布が表示されていないが、全ての検索エンジンにおいて平均値の 2 倍を超える値は全体の 1%未満の出現率であった。図 2 によると、Google, Yahoo! Japan, LiveSearch のすべての検索エンジンにおいて、ヒット数を正規化し丸めこんだ値と出現率の関係は正規分布に近い形をとっている。よって、誤差が正規分布に従うという仮定を用いても問題がないと考えられる。

#### 4.3. 実験結果

本実験では、 $I$  を 4.1 で示した(1)の方法で推定した検証と、(2)の方法で推定した検証を行った。まず、(1)の方法で  $I$  を推定し、検証を行った結果について述べる。つまり、 $I$  の推定値として検索開始オフセットが最も大きな値となったときのヒット数を用い、その変動幅と出現頻度を求める。求めた結果を図 3 に示す。図 3 では、 $I$  の推定値に対する取得したヒット数の比を求め、その比の値を 0, 0.1, 0.2, ..., 1.9 の値に切り捨てた値を  $x$  軸に、また  $x$  の値の出現頻度を  $y$  軸にプロットを行った。 $I$  の推定値に対してヒット数が 10%以内の変動である確率が Google では 75%, Yahoo! Japan では 91%, Live Search では 70%となる。つまり、それぞれの検索エンジンにおいて 9%~30%の確率で、 $I$  の推定値から 10%以上異なる変動が発生している。また、5%~10%の確率で  $I$  の推定値と倍以上異なるヒット数が得られることもあり、大きな変動幅をもっていることがわかる。

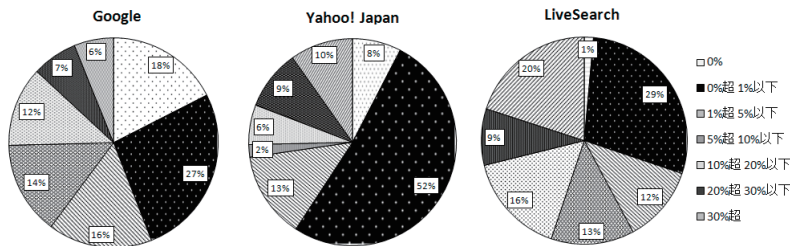


図 4. ヒット数の平均値に対する検索結果の信頼区間の変動幅

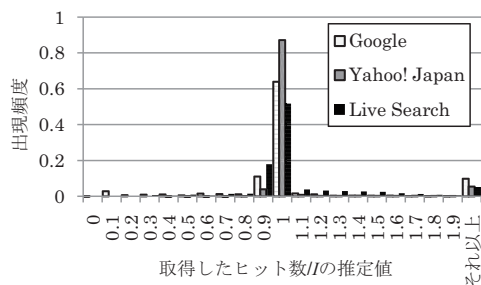


図 3. 最大オフセット値により推定した理想的なヒット数と、実際のヒット数の比の分布

続いて、(2)の方法で  $I$  を推定し、検証を行った結果について述べる。  $I$  の推定値と実際のヒット数の誤差が正規分布に従うと仮定を行った時に、それぞれのクエリについて  $I$  の推定値が 95% の確率で存在する信頼区間を求めた。続いて、それぞれのクエリについてその信頼区間の大きさが平均値の何%となるのか求めた。平均値に対する信頼区間の大きさがどのような分布になるか示した図を図 4 に示す。この結果を見ると、理想的なヒット数が存在する 95% 信頼区間の大きさが平均値の 10% 以内である確率が Google では 75%、Yahoo! Japan では 75%、LiveSearch では 55% であることがわかる。つまり、Google、Yahoo! Japan では  $0.95 \times 0.75 \approx 0.7125$ 、LiveSearch では  $0.95 \times 0.55 \approx 0.5225$  の確率で検索開始オフセットを変動させ取得した検索エンジンのヒット数の平均値の前後 5% 以内に理想的なヒット数だと考えられる値が存在していることとなる。

#### 4.4.4 節のまとめ

本節では、検索開始オフセットの変化によるヒット数の変動幅について検証を行った。理想的なヒット数  $I$  の推定方法によっても変化をするが、どの検索エンジンでも約 10%~50% の確率で、理想的なヒット数の推定値より 5% 以上変動することがわかった。この結果をまとめた表を表 3 に示す。ここで、表中の(1)、(2)はそれぞれ  $I$  の推定方法を表す。

表 3. ヒット数の変動幅の検証結果

	理想的なヒット数の推定値の前後の	
	10%以内に結果が収まる確率(1)	5%以内に結果が収まる確率(2)
Google	75%	71.3%
Yahoo! Japan	91%	71.3%
Live Search	70%	52.3%

## 5. おわりに

本論文では検索結果のヒット数に対する信頼性の検証を行った。その結果、ヒット数の大小関係は、検索開始オフセットを変動させ 3 回ヒット数を取得し、3 回の大小比較結果の多数決を用いることによって 95% の確率で信頼できる結果が得られることがわかった。また、1 つのクエリにおけるヒット数は、約 10%~50% の確率で、理想的なヒット数の推定値より 5% 以上変動することがわかった。この結果は、実験を行った 2008 年 7 月における結果であり、今後検索エンジンのシステムの変更により変化する可能性がある。

今後の課題として、理想的な  $I$  のヒット数が、今回使用した推定値とどの程度異なっているのか、調査を行う必要がある。またそれと同時に、どのようなヒット数を用いれば、より信頼性の高い結果が得られるのかについて、更なる考察が必要となる。

## 文 献

- [1] A. Gulli, and A. Signorini: "The Indexable Web is More than 11.5 Billion Pages", ACM WWW2005, pp.902-903, (2005.5)
- [2] A. Kilgarrieff: "Googleology is Bad Science," Computational Linguistics, Vol.33, Issue 1, pp.147-151 (2007.3)
- [3] Google: <http://www.google.co.jp/>
- [4] Live Search: <http://www.live.com/>
- [5] 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)
- [6] Yahoo! Japan: <http://www.yahoo.co.jp/>