検索エンジンのヒット数の信頼性に対する評価

佐藤 亘[1]　打田 研二[2]　山名 早人[3,4]

[1]早稲田大学 基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
[2]早稲田大学 基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
[3]早稲田大学 理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
[4]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: [1,2,3,4]{kohsatoh,k_uchida,yamana}@yama.info.waseda.ac.jp

あらまし 近年，自然言語処理をはじめとする数多くの研究が，検索エンジンから得られる検索結果数，すなわちヒット数を利用している．しかしながら，検索エンジンが返すヒット数は検索するタイミングによって不自然に変化し，研究のベースとして用いるには無視できないほどの大きな誤差が生じることがある．そのため，ヒット数の信頼性を評価，向上させる手法を考えることは，大きな課題である考えられる．本論文では，ヒット数の信頼性を評価する基準を設けるとともに，一定の水準以上でヒット数の信頼性を保証するヒット数取得条件の特定を試みる．実験の結果，ヒット数の大小関係は，大小関係が安定していることを観測する期間を長くとることによって信頼度の向上を図ることができると判明するとともに，一定の信頼度を保証するヒット数の取得条件を特定することができた．

キーワード 検索エンジン，ヒット数，信頼性評価

An Evaluation for Reliability of Search Engines' Hit Count

Koh SATOH[1]　Kenji UCHIDA[2]　and　Hayato YAMANA[3,4]

[1]School of Fundamental Science and Technology, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
[2]Graduate School of Fundamental Science and Technology, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
[3]Faculty of Fundamental Science and Technology, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
[4]National Institute of Infomatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
E-mail: [1,2,3,4]{kohsatoh,k_uchida,yamana}@yama.info.waseda.ac.jp

1. はじめに

近年，Web 上を流通するコンテンツは爆発的に増え，その増加はさらに加速する一方である．検索エンジンは，膨大な量の Web 上のコンテンツを網羅的に蓄えており，ユーザは検索エンジンを用いることで Web 上の幅広い情報に簡単にアクセスすることができる．このような Web 上の情報の網羅性の高さ，そして情報へのアクセス性の高さという検索エンジンの持つ大きな 2 つの特徴から，検索エンジンの検索結果を利用した研究が盛んに行われている．検索エンジンの検索結果を用いた研究の中でも，多くの検索エンジンがクエリに対する検索結果と共に出力する該当ページの概数，すなわちヒット数を利用した研究は数多い[1][2][3][4][5][6]．これらの研究は，検索エンジンによって得られるヒット数が，検索クエリに対する Web 上の文書集合における出現頻度とみなすことができるという前提のもとに行われている．ヒット数を用いた研究の例として，機械翻訳の支援を行う研究[1]，クエリ単語間の距離を定義する研究[2]，単語クラスタリングを試みる研究[3]などの自然言語処理に関する研究が多く挙げられる．近年では，その他にもセマンティック Web への応用のためのオントロジー構築[4]や，Web からの自動ソーシャルネットワーク抽出[5]にも用いられるなど，ヒット数の応用分野は増え続け，その重要性は日を追うごとに増している．

しかしながら，検索エンジンが返すヒット数は検索するタイミングによって不自然に変化する現象が見受けられるなど，様々な場合において誤差が生じることが知られており[7][8][9]，近年その信頼性が問題視されている．例えば 1 日，2 日といった短い期間でヒット数が 10 倍以上あるいは 1/10 倍以下に変化する現象がしばしば起こり，様々な研究やアプリケーションのベースとして用いるには無視できないほどの大きな誤差となっている．そのため，検索エンジンによって得られるヒット数の信頼性を明らかにすると共に，得られるヒット数に対する信頼性を向上させる手法を考えることは，大きな課題である．

これまで検索エンジンのヒット数に対する信頼性

の問題について，複数エンジン間のヒット数の違いや特徴について比較した研究[7]，検索エンジンのヒット数の正確性を比較した研究[8] などが行われてきた．これらに加えて，我々は信頼できるヒット数を得るための条件の特定を試みた研究[9]を行ってきた．しかしながら我々の知る限り，ヒット数の信頼性に対して定量的な評価を試みた研究は存在しない．[8]は，ヒット数の正確性を検索エンジン間で比較する際に，ヒット数の正確性に対する定量的な評価を行っているが，この評価手法が適用できるのは非常に限られた場合のみである．

以下，2 節において関連研究を紹介し，様々な研究におけるヒット数利用の問題点を指摘する．次に，3 においてヒット数が変動する原因を考察する．次に，4 節においてヒット数の信頼性に対する評価基準を定める．その後，5 節で定義された基準を用いた実験を行い，信頼性が一定の精度で保証できるヒット数の取得手法について論じる．

2. 関連研究

本節では，検索エンジンのヒット数に関連する研究についてまとめる．まず 2.1 節においてヒット数を利用した研究について紹介し，ヒット数が様々な研究においてどのように応用されているのかをまとめる．次に 2.2 において，本研究の類似研究として，ヒット数の信頼性を対象とした研究についてまとめる．

2.1. 検索エンジンのヒット数を利用した研究

本節では，ヒット数を利用した研究について紹介し，ヒット数が様々な研究においてどのように応用されているのかをまとめる．

2.1.1. ヒット数を機械翻訳支援に用いた研究

Grefenstette[1]は，ヒット数を機械翻訳支援に利用する研究を行った．この研究では，ある単語 A を別の言語の単語に置き換えるとき，A に対する翻訳語候補群に対する検索エンジンのヒット数を取得し，最も高いヒット数を得た単語が適切な翻訳語であるとしている．

この研究では，検索エンジンによって得られるヒット数の大小関係が入れ替わると，結果として得られる翻訳語が変わることがわかる．

2.1.2. ヒット数を用いて同義語抽出を行なう研究

Turney[6]は検索エンジンを利用した同義語抽出手法 PMI-IR を提案した．Turney は，TOEFL における問題に代表されるような，ある単語に対していくつかの同義語候補が挙げられたとき，どの単語が最も同義語としてふさわしいかを判別する手法を提案している．この手法では，問題語 *problem* に対して，同義語の候補となる単語 $choice_i$ に対し，

$$score(choice_i) = \frac{hits(problem\ AND\ choice_i)}{hits(choice_i)} \qquad (2.\ 1)$$

をそれぞれ算出して，最もスコアの高い単語が同義語としてふさわしいとしている．ここで hits(Q) は Q をクエリとしたときの検索エンジンによって得られるヒット数を示す．この手法では，候補語のヒット数の大小関係が入れ替わると，同義語として判断される語も変化することがわかる．

2.1.3. ヒット数を用いてクエリ単語間の類似度を定義した研究

Cilibrasi ら[2]は検索エンジンのヒット数を利用した単語間の類似度 Google Similarity Distance を提案した．検索エンジンにおいて AND 検索を利用することで，単語間の共起度を取得し，単語 x, y の類似度を次式のように定義している．

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \qquad (2.\ 2)$$

ここで $f(x)$ とは単語 x に対する Google 検索時のヒット数を表し，$f(x,y)$ とはクエリ「x AND y」に対する Google のヒット数を表す．また N は任意の x に対して $f(x)<N$ が成り立つような自然数であるとしている．式(2. 2)の右辺から，2 つのクエリに対するヒット数間の大小関係と，クエリに対するヒット数の絶対値の両方が，類似度に大きな影響を与えることが見て取れる．

2.2. ヒット数の信頼性を対象とした研究

本節では，本研究の類似研究として，ヒット数の信頼性を対象とした研究についてまとめる．

2.2.1. 複数の検索エンジン間でのヒット数を比較した研究

Thewall[7]は Google，Yahoo!， Live Search の 3 つの検索エンジンによって得られるヒット数と検索結果の比較実験を行った．Thewall は様々なヒット数をとる 2000 個のクエリを選出し，複数の検索エンジンによって得られるヒット数の相関を求めたところ，どの検索エンジンにおけるヒット数も高い相関があるという結果を得た．しかしながらヒット数の絶対値を比較すると，Yahoo!，Google が Live Search の 5,6 倍のヒット数を返していると指摘した．

Thelwall の研究は複数検索エンジン間のヒット数や検索結果の違いについて比較して論じているものであり，ヒット数の信頼性に対する定量的な評価を行っているものではない．また，どのようにして信頼性の高いヒット数を得るかについて論じているものでもない．

2.2.2. 各検索エンジンから得られるヒット数の正確性を比較した研究

Uyar[8]は，Google，Yahoo!， Live Search の 3 つの検索エンジンについてヒット数の正確性調査を行った．

これら 3 つの検索エンジンは検索クエリに該当する Web ページの上位 1000 件までを表示する．Uyar は，あるクエリに対する検索結果として取得した Web ページ総数が 1000 件以下のとき，実際に取得した Web ページ数がそのクエリに対するヒット数の正解値であるという仮定を行った上で，表示されるヒット数の正確性を調査した．

Uyar は，実際に取得した Web ページ数が 1000 件以下のとき，取得された Web ページ数 *ReturnedDocument*，表示されたヒット数 *Estimate* を用いてエラー率 *Percentage of Error* を次のように定義した．

$$Percentage\ of\ Error = \frac{Estimate - ReturnedDocument}{ReturnedDocument} \times 100 \quad (2.3)$$

Uyar は 1000 個のクエリについてエラー率を計算した．結果，エラー率が 10%以下となるクエリは，Google では 78%，Yahoo では 48%，Bing では 23%であると判明し，Google がもっとも正確なヒット数を返していると結論づけた．

このように Uyar は，取得した Web ページ数が 1000 件以下のとき，実際に取得した Web ページ数が正しいヒット数であるという仮定のもとにヒット数の評価を行なっている．そのため，この手法で 1000 件以上のページが返されたときのヒット数の信頼性評価が不可能であるという問題がある．

2.2.3. 信頼できるヒット数が得られる条件を考察した研究

舟橋ら[9]は Google，Yahoo!，Bing の 3 つの検索エンジンについてヒット数変動調査を行い，検索エンジンが信頼のできるヒット数を返す条件を考察した．最初に，舟橋らはヒット数の変動が観測できる場合が次の 3 ケースであると特定した．

Case 1. 短時間に繰り返し同じクエリを利用して検索した場合
Case 2. 短時間に繰り返し「次へ」ボタンをクリックした場合
Case 3. 検索を行う日時を変えた場合

その上で，それぞれのケースについて，検索エンジンが信頼できるヒット数を返す条件の特定を試みた．結果，次の 3 つの条件を満たしたときのヒット数は信頼できると結論づけた．

Case 1. 検索フィルタの影響を受けない場合
Case 2. 検索開始オフセットが最も大きい場合
Case 3. ヒット数が 1 週間以上にわたって観測開始時のヒット数から 30%以上増減していない場合

しかし，舟橋らはヒット数の信頼性に対する明確な定義を行っておらず，提案手法に対する評価がなされていないという点で不十分であるといえる．また，得られたヒット数が信頼できるか否かを判定するために最低でも 1 週間という長期にわたってヒット数推移を観測していないといけないという点も問題である．

3. ヒット数が変動する原因の考察

本節では，検索エンジンから得られるヒット数が検索するタイミングによって変動する理由について論じる．各検索エンジンのヒット数概算のためのアルゴリズムは公開されていないため，ヒット数が変動する確実な理由についてはわからない．そこで，本付録では一般的な検索エンジンの構成から，ヒット数が変動する要因を推測する．

3.1. インデックス更新によるヒット数変動

最近の検索エンジンは，ニュース検索[10][11][12]やリアルタイム検索[13]など，リアルタイム性の高い情報に対する検索結果を表示する機能を備えている．これらの機能を実現するために，検索エンジンは数分あるいは数秒といった短い間隔で更新頻度が高いと予想される Web ページに対するクローリングを行い，インデックスの更新を行う必要がある．[14]では，検索エンジン中には秒単位で頻繁に更新を行う小容量のインデックスと，日単位で更新を行う大容量のインデックスとが存在するとしている．検索エンジンのヒット数は，毎日若干の変動が見られるが，このような変動はインデックスの小規模な更新によるものだと推測される．

また舟橋らは，多数のクエリに対する時系列上のヒット数変動をクラスタリングしたところ，多くのクエリが大幅な変動を見せる期間が存在するという結果を得た[9]．このように多くのクエリに対する大幅なヒット数変動は，先述したニュース検索やリアルタイム検索などに使用される小規模なインデックスに対する更新に加えて，より大規模なインデックスに対する更新が行われている[14]ことが原因となっていると考えられる．

3.2. キャッシュヒット/キャッシュミスによるヒット数変動

検索エンジンは，Result Cache や Pruned Index などといったキャッシュ構造を持つことによって Full Index へのアクセスを削減し，高速化を図ることができる[15]．ここで Result Cache とはクエリログをもとにした検索結果のキャッシュを表し，Pruned Index とはページランクに基づいて縮小されたインデックスを表す．このように，多くのユーザが利用すると考えられるページのみを抽出して小さなインデックスとして保持することによって，サイズの大きい Full Index へのアクセスを減らし，高速化を図ることができる．

このとき，もし Result Cache/Pruned Index と Full Index の間に差異が生じてしまうと，キャッシュヒットした場合とキャッシュミスした場合とで得られるヒット数に変化が生じる可能性が考えられる．

3.3. 検索時に異なるデータセンターに接続した場合の変動

Google, Yahoo!, Bing などの大規模な検索エンジンは，世界中からの膨大な量のクエリに対応するため，インデックスを持つ多数のサーバから成るデータセンターを世界各地に配置している[16]．個々のデータセンターは，それぞれが独立して Web 検索を行えるよう，完全な検索クラスタを備えている[17]．各データセンターにおけるインデックスは基本的には一致しているが，インデックスの更新最中といった，インデックスがデータセンター間で異なる時期が存在すると考えられる．ユーザがブラウザ上で検索エンジンに対してクエリを入力すると，まずドメイン名(www.google.com 等)から IP アドレスへの名前解決が行われる．この際，ユーザと各クラスタとの位置関係，各クラスタの混雑状況に応じて，最も応答時間が短くなると予想されるデータセンターの IP アドレスが選択される．基本的にはユーザと地理的に最も近いデータセンターに接続されるが，データセンターにおける混雑状況等に依存して接続するデータセンターが変化する場合が考えられる．このように接続するデータセンターが変化した場合，かつデータセンター間でインデックスに差異があった場合，ユーザは異なる検索結果・ヒット数を取得してしまうことになる．

4. ヒット数の信頼性に対する評価基準の定義

本節では，ヒット数の信頼性をどのように定義することが妥当かについて論じ，そのうえで信頼性を評価するための基準を定める．

4.1. 大小関係の入れ替わりによるヒット数信頼性評価

2.1 にて述べたとおり，検索エンジンのヒット数を用いている研究の多くは，ヒット数の絶対値そのものを用いるのではなく，複数クエリに対するヒット数間の大小関係を用いている．たとえば先に紹介した同義語抽出の研究においては

$$score(choice_i) = \frac{hits(problem\ AND\ choice_i)}{hits(choice_i)} \qquad (2.\ 1)$$

という式によって得られるスコアの大小関係を用いて同義語を選出している．この例が示すとおり，「どちらのクエリがより Web 上での出現頻度が高いか」というヒット数の大小関係こそが，多くの研究において重要なファクターとなっている．つまり複数クエリに対するヒット数間の大小関係の入れ替わりが，ヒット数を用いた研究に対して大きな影響を与える．したがって，もし比較対象となるクエリにおけるヒット数間の大小関係の入れ替わりが頻繁に起こる場合，その期間におけるヒット数は信頼できないと考えられる．逆に，長期間にわたって同じクエリ同士で同一の大小関係が保たれている場合の，その期間におけるヒット数は信頼できる．すなわち，ある特定の期間 m 日間におけるクエリ A, B に対するヒット数 hit[A]，hit[B] の大小関係の信頼性 $reliability_1(hit[A]>hit[B], m)$ は次の確率関数によって評価できると考えられる．

$$reliability_1(hit[A] > hit[B], m) = \Pr(days(hit[A] > hit[B]) = m) \qquad (3.\ 1)$$

式中の関数 days(hit[A]>hit[B]) はヒット数の大小関係 hit[A] > hit[B]が保たれている日数を表す．つまりこの式は「クエリ A, B のヒット数の大小関係が m 日間入れ替わらない確率」を表している．得られる確率が大きいほど信頼性が高く，また式中の確率変数 m が大きいほど信頼性が高い．

4.2. 変動率によるヒット数信頼性評価

図 1 は，Yahoo におけるクエリ「東宝シネマ」に対する 2009 年 10 月から 2010 年 12 月の 1 年 2 ヶ月のヒット数変動を示している．この例に代表されるように，検索エンジンから返されるヒット数は安定した期間と値が大きくゆらぐ期間とに分けられる．全期間のなかで，大部分は値が安定しているのに対し，局所的に大きく揺らいでいる期間が確認できる．値が大きく揺らぐ期間にも，次のような様々な種類がある．

I. 数日にわたってコンスタントに変動が大きい期間
II. ２，３日といった短い期間の間に連続して急上昇と急降下が見られる期間
III. 一度急上昇(降下)し，しばらく安定してから急降下(上昇)してもとの値に戻る，という期間

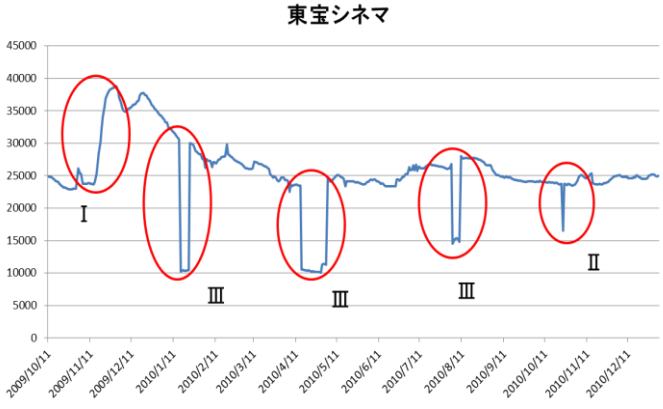

図 1. ヒット数変動の例

Web 上の文書は通常，インクリメンタルに追加されるものであり，ある任意のクエリを含む文書も同様に，

インクリメンタルに増減するものと考えられる．したがって，ヒット数の正しい変動は，ある程度なめらかな変動であると考えることが妥当である．図 1 に見られるような変動の大きいヒット数は，3 節における議論から，検索エンジンのインデックス更新等に起因するエラー値であるとみなすことができる．したがって，ヒット数の変動の中で，大きく変動する期間におけるヒット数は信頼できず，長期間にわたって一貫して安定した値をとっている期間におけるヒット数は信頼できると考えるのが妥当である．具体的には，次の 2 つの観点により，ヒット数の信頼性が評価できると考えられる．

I. 変動が小さいヒット数は信頼性が高い
II. ヒット数変動の小さい期間が長いほど信頼性が高い

すなわち，変動率に着目したある時点 *base* におけるクエリ *A* のヒット数の信頼性 reliability$_2$(hit$_{base}$[*A*], *m*, *R*) は次の確率関数によって評価できると考えられる．

$$reliability_2\left(hit_{base}[A], m, R\right) = \Pr\left(days\left(changeRatio_d\left(hit_{base}[A]\right) < R\right) = m\right) \quad (3.\ 2)$$

式中の関数 changeRatio$_d$(hit$_{base}$[*A*]) は，ある時点 *d* におけるクエリ *A* のヒット数の変動率を表し，観測基準時点におけるヒット数 hit$_{base}$[*A*]，ある時点におけるヒット数 hit$_d$[*A*]，絶対値関数 abs を用いて次のように定義される．

$$changeRatio_d\left(hit_{base}[A]\right) = \frac{abs\left(hit_d[A] - hit_{base}[A]\right)}{hit_{base}[A]} \quad (3.\ 3)$$

また days(changeRatio$_d$(hit$_{base}$[*A*]<*R*)) は変動率が *R* 以下に保たれている日数を表す．すなわちこの式は「ヒット数の変動率が *R* 未満に抑えられている日数が *m* 日である確率」を表している．得られる確率が大きいほど信頼性が高く，また式中の確率変数 *R* が小さいほど，また *m* が大きいほど信頼性が高い．

本評価方法では，観測基準時点を固定した場合の変化率を用いてヒット数の評価を行っている．変動率の基準時点を固定としたのは，ヒット数の変動パターンとして，数日にわたって大きな変化が観測される場合があるからである．例えば，直前の観測時点からの変化率 $\left(abs\left(hit_t[A] - hit_{t-1}[A]\right) / hit_{t-1}[A]\right)$ が小さい場合でも，数日間継続してヒット数を観測したとき，全体として大きく変化している場合がある．このような変動は，3 節の議論におけるインデックスの大規模な更新によって引き起こされるものだと考えられ，このような場合におけるヒット数の信頼性は低いと考えられる．そのため，観測基準時点を固定した場合の変化率を用いた．

この信頼性評価基準は，4.1 節の基準とは違い，クエリごとのヒット数の絶対値に対する信頼性指標となる．したがって，[2]などといったヒット数の大小関係に加えて絶対値が重要となる研究に有効となる．

5. ヒット数の信頼性評価実験

前節にて，ヒット数の信頼性に対する評価方法を定義した．本節では，いくつかの検索エンジンに対して行ったヒット数の信頼性に対する評価実験についてまとめる．

5.1. 大小関係の入れ替わりに着目した評価基準によるヒット数評価

5.1.1. 実験の目的

本実験は，前節の(3. 1)式で定義された 1 つ目の信頼性評価基準「クエリ *A*,*B* のヒット数の大小関係が *m* 日間入れ替わらない確率」を用いて，「どのような条件を満たしたヒット数は高い信頼性を保証するか」という点を明らかにすることを目的とする．すなわち，複数クエリのヒット数に対して様々な条件を設け，その条件を満たしたヒット数について信頼性を評価し，高い信頼性が保たれる必要条件の特定を試みた．複数クエリのヒット数に対する様々な条件とは，たとえば，「2 つのクエリに対するヒット数が 30%以上離れていたとき」などである．

5.1.2. 実験の手順

step1. 複数クエリのヒット数に対する条件の設定を行う．「ある観測日 *base* における 2 クエリのヒット数間の比率が *R*:1 であり，かつその 2 クエリのヒット数を *base* から *a* 日間観測したときの大小関係が変化しない」という条件を設定した．

step2. 大規模に収集したヒット数観測データに対し，step1 の条件中の各パラメータ *R*, *a* を様々に変化させ，条件にあてはまるヒット数のペアを抽出する．

step3. 各条件について抽出された *k* 組のヒット数列のペアに対し，「大小関係が保たれていた期間が *m* 日であったヒット数の組み合わせは *l* 個」に該当する *m* と *l* を計測する．これにより継続期間 *m* 日に対する信頼度 r = *l* / *k* を算出する．

以上の手順をまとめると，次のような関数 L$_1$ を求めることと同義となる．

$$function\ L_1 : (a, R) \rightarrow List\ of\ (m, r)$$
$$where\ r = reliability_1 \begin{pmatrix} hit[A] > hit[B], m \\ | a = observeddays(hit[A] > hit[B]), \\ hit_{base}[A] : hit_{base}[B] = R:1 \end{pmatrix} \quad (4.\ 1)$$

ただし observeddays(hit[*A*]>hit[*B*]) とは，クエリ *A* のヒット数がクエリ *B* のヒット数より大きいと観測した

日数を示す．すなわち，ある時点 *base* での 2 クエリ間のヒット数比率 *R* と，大小関係が保たれていると観測された日数 *a* を入力とし，その大小関係が *m* 日間続く確率は *r* であるという(*m*, *r*) のリストを出力する関数を求めることとなる．

5.2. ヒット数の安定性に着目した評価基準によるヒット数評価

5.2.1. 実験の目的

本実験は，前節の(3. 2)式で定義された 2 つ目の信頼性評価基準「ヒット数の変動率が *R* 未満に抑えられている日数が *m* 日である確率」を用いて，前節と同様「どのような条件を満たしたヒット数は高い信頼性を保証するか」という点を明らかにすることを目的とする．

5.2.2. 実験の手順

step1. クエリのヒット数に対する条件の設定を行う.「あるクエリのヒット数を *a* 日観測したとき，観測開始時点におけるヒット数から，変動率が *R* に抑えられている」という条件を設定した．

step2. 大規模に収集されたヒット数観測データに対し，条件中の各パラメータ *R*, *a* を様々に変化させ，条件にあてはまるヒット数を抽出する

step3. 各条件について抽出された *k* 個のヒット数列に対し，「変動率が *R*% 以内に抑えられていた日数が *m* 日であったヒット数は *l* 個」に該当する *m* と *l* を計測する．これにより継続期間 *m*，比率 *R* に対する信頼度 *r* = *l* / *k* を算出する．

以上の手順をまとめると，次のような関数 M_1 を求めることと同義となる．

$$function\ M_1 : (a, R) \rightarrow List\ of\ (m, r)$$
$$where\ r = reliability_2 \begin{pmatrix} hit[A], m, R \\ | a = observeddays\{ \\ changeRatio(hit[A]) < R\} \end{pmatrix} \quad (4.\ 2)$$

すなわち，観測日数 *a* と，その期間における観測開始日でのヒット数からの変化率の最大 *R* を入力とし，変化率が *R* 未満に保たれる期間が *m* 日間である確率は *r*%であるという(*m*,*r*)のリストを出力する関数を求めることと同義となる．

5.3. 使用したデータ

本実験においては，Yahoo! Japan の 2007 年 12 月のクエリログにおいて頻出順に並べて現れた上位 10,000 件をクエリとして利用した．頻出語は多くのユーザが検索を行うクエリであり，特に重要なクエリと考えられるため，頻出度をもとにクエリ選定を行った．Google, Yahoo!, Bing が提供している検索 API を用い，上記 10,000 個のクエリに対するヒット数を 2009 年 10 月から 2010 年 12 月にかけて観測し，得られたヒット数に対して実験を行った．

6. 実験結果

6.1. 大小関係の入れ替わりに着目した評価基準によるヒット数評価に対する結果

6.1.1. *r*-*m* グラフ

Google について，ヒット数間の比率 *R* を 2 に固定したときの安定期間 *m* に対する信頼度 *r* のグラフを図 2 に示す．

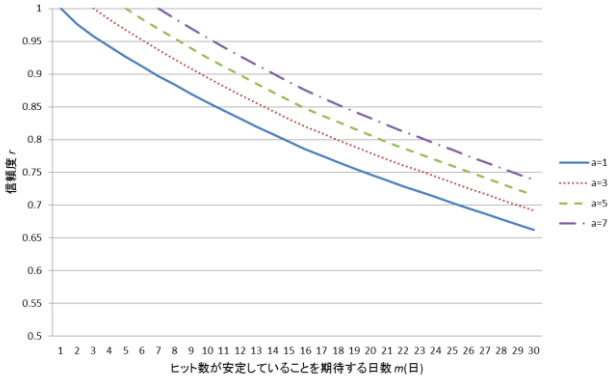

図 2. Google の *r*-*m* グラフ(R=2)

グラフから，ヒット数の大小関係が安定していることを期待する日数 *m* が増えるにつれて信頼度 *r* が減少していることが読み取れる．また，大小関係が安定していることを確認する日数 *a* を増やすことによって，信頼度の向上を図ることができるとわかる．

6.1.2. *a*-*R* グラフ

Google について，信頼度 r を 0.85 以上としたときの観測期間 a と 2 クエリのヒット数比率 R のグラフを図 3 に示す．

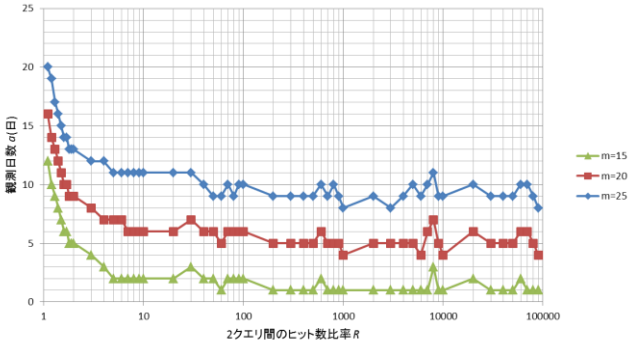

図 3. Google の *a*-*R* グラフ(r=0.85)

このグラフは「ある一定の信頼度を保証するヒット数の条件」を示している．例えば，ある観測時点でヒット数の比率が 10:1 であった 2 つのクエリの大小関係が 15 日間入れ替わらない確率を 85%で保証するには，2 日の間大小関係が保たれていることを確認すればよい．2 クエリ間のヒット数比率 *R* が大きいほど，観測日数は少なくてすむということがわかる．また，信頼性を保証したい期間 *m* が長いほど，観測すべき日数が増加することが見て取れる．

6.2. ヒット数の安定性に着目した評価基準によるヒ

ット数評価に対する結果

6.2.1. *r-m* グラフ

Google について，変動率 R を 0.2 に固定したときの安定期間 m に対する信頼度 r のグラフを図 4 に示す．

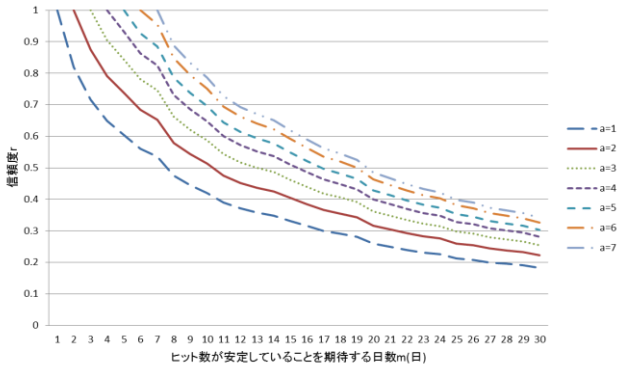

図 4. Google の *r-m* グラフ

グラフから，ヒット数が m 日間安定した値を取る確率は，m が増えるに連れて指数関数的に減少することが見て取れる．また，1 日だけ観測したヒット数の変動率が 10 日の間，0.2 以内に抑えられている確率は 40%程度という低い値であるのに対し，2 日，3 日と観測期間 a を増やしていくに連れて，50%，58%と，信頼性が大幅に向上していることがわかる．これらの結果から，ヒット数を研究に用いる際は，ヒット数が安定していることを可能な限り長期間確認することで，信頼できるヒット数を得ることができるということがわかる．

6.3. 実験結果のまとめ

6.1 の実験結果より，ヒット数の大小関係は，観測期間を増やすにつれて信頼度の向上を図ることができると判明した．さらに，図 3 の *a-R* グラフに示したとおり，一定の信頼度を保証するヒット数の取得条件を特定することができた．

また，6.2 の実験結果より，ヒット数の絶対値の信頼度は，安定していることを期待する日数 m が増えるに連れて指数関数的に減少するということがわかった．したがって，ヒット数の絶対値を研究として用いるのは望ましくないといえる．しかしながら，実験結果が示すとおり，値が安定していることを数日間観測することによって信頼度の向上を図ることができる．

7. おわりに

本研究では，ヒット数を研究に用いる場合の基盤となることを目指し，得られたヒット数の信頼性に対する評価基準を定義し，一定の信頼度を保証するヒット数の条件について考察をした．

現在，検索エンジンによって得られるヒット数は自然言語処理をはじめとする様々な研究で利用されている．しかしながら，ヒット数は様々な場合において値が揺らぐ現象が見受けられ，近年その信頼性が問題視されてきた．これまで検索エンジンのヒット数に対する信頼性の問題についていくつかの研究が行われてきたが，得られるヒット数について定量的な評価を行った研究は存在しなかった．そこで本研究では，様々な研究におけるヒット数の用いられ方を考慮し,「大小関係の入れ替わり」に着目した評価と,「変動率」に着目した評価の 2 つの基準によるヒット数の信頼性評価基準を定めた．その上で，大規模に収集したヒット数観測データを用い，ヒット数の信頼性評価実験を行い，一定の信頼度を保証するヒット数の条件の特定を試みた．その結果として，ヒット数の大小関係は，大小関係が安定していることを観測する期間を長くとることによって信頼度の向上を図ることができると判明し，さらに一定の信頼度を保証するヒット数の取得条件を特定することができた．例えば Google では，ある 2 つのクエリのヒット数の大小関係が 2 日間安定していることを観測した場合，その観測日におけるクエリ同士のヒット数が 10 倍以上離れていた際に 15 日間大小関係が入れ替わらないことを 85%以上で保証できる．

謝辞

本研究は，科学研究費補助金（基盤研究（B）21300038）の補助によるものである．

参　考　文　献

[1] G. Grefenstette, “The WWW as a resource for example-based MT tasks”, ASLIB Translating and the Computer Conference, London, 1999.
[2] R. L. Cilibrasi and P. M. B. Vitanyi, “The Google Similarity Distance”, IEEE Trans. on Knowledge and Data Engineering, Vol.19, No.3, pp.370 – 383, 2007
[3] Y. Matsuo, T. Sakai, K. Uchiyama and M. Ishizuka, “Graph-based Word Clustering using Web Search Engine”, Proc. of the Conf. on Empirica Methods in Natural Language Processing, pp.542-550, 2006.
[4] P. Cimiano and S. Handschuh, “Towards the self-anotating web”, Proc. WWW2004, pp462-471, 2004.
[5] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida and M. Ishizuka, “POLY-PHONET: An advanced social network extraction system”, Proc. WWW 2006, 2006.
[6] P. Turney, “Mining the web for synonyms: PMI-IR versus LSA on TOEFL”, Proc. of ECML-01, pp. 491-502, 2001.
[7] M. Thelwall, “Quantitative Comparisons of Search Engine Results”, J. of the American Society for Information Science and Technology, Vol.59, No.11, pp.1702-1710, 2008.
[8] A. Uyar, “Investigation of the Accuracy of Search Engine Hit Counts”, J. of Information Science, Vol.35, No.4, pp.469-480, 2009.
[9] 舟橋卓也，山名早人，“Hit Count Dance -検索エンジンのヒット数に対する信頼性検証-”，日本データベース学会論文誌，Vol.9, No.1, pp.18-22, 2010.
[10] Google News, http://news.google.com/, (2010.1.16 アクセス)
[11] The top news headlines on current events from Yahoo! News, http://news.yahoo.com/, (2010.1.16 アクセス)
[12] Bing, http://www.bing.com/?scope=news, (2010.1.16

アクセス)

[13] Google Realtime Search, http://www.google.com/realtime, (2010.1.16 アクセス)

[14] Challenges in Building Large-Scale Information Retrieval Systems, http://research.google.com/people/jeff/WSDM09-keynote.pdf, (2010.1.16 アクセス)

[15] G. Skobeltsyn, F. P. Junqueira, V. Plachouras and R. Baeza-Yates: “ResIn: A Combination of Result Caching and Index Pruning for High-performance Web Search Engines,” In Proc. of SIGIR’08, pp.131-138 (2008)

[16] L. Barroso, J. Dean, and U. Hoelzle: “Web search for a planet: the google cluster architecture,” IEEE Micro, Vol.23, No.2, pp.22–28, (2003)

[17] 西田圭介：“Google を支える技術”，技術評論社 (2008)