Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

検索エンジンのヒット数に対する信頼性評価指標の提案とその妥当性検証

佐藤亘[†]　打田研二[†]　山名早人[††]

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．我々はこの課題に対して，信頼できるヒット数を得ることができる条件の特定を試みた研究や，実際に得られたヒット数の信頼性を定量的に評価できる手法の提案を行ってきた．本論文では，後者の研究に追加して，信頼性評価指標の妥当性検証実験を行ったので結果を報告する．

A Proposal and Validity Inspection of Reliability Evaluation Method for Search Engines' Hit Count

Koh Satoh[†]　Kenji Uchida[†]　and　HayatoYamana[††]

Recently, there exit numerous researches based on the number of search results, or hit count. However, hit counts returned by search engines can fluctuate unnaturally when observed on different days, and may cause too large errors to be used in researches. Therefore, it is important to discuss on how we can evaluate and improve the reliability of hit counts. We have performed several researches about this problem such as a research to specify the conditions in which search engines can return reliable hit counts, and a research to define the reliability evaluation metrics. In this paper, in addition to the latter research, we'll report the result of validation experiments for the reliability evaluation metrics.

1. はじめに

近年，急速に増え続ける膨大な量の Web 上のコンテンツに簡単にアクセスする手段として，数多くの研究が検索エンジンの検索結果を用いている．そのような研究の中でも，クエリに対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]-[5]．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]，同義語抽出を行う研究[3]などの自然言語処理に関する研究が多く挙げられる．特に[2]や[3]を応用した研究は数多く，直近 2 年における[2]や[3]を参照している論文は合計で 428 件存在する．さらに，近年では，自然言語処理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[6][7][8]，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．

これらの議論を受けて，過去にヒット数の信頼性に関連していくつかの研究が行われてきたが，これらの研究は主に各検索エンジンのヒット数の変動傾向を特定するものであった[6][7]．

これまで我々は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，信頼できるヒット数を得ることができる条件の特定を試みた研究[8]，実際に得られたヒット数の信頼性を定量的に評価できる手法の提案[9]を行ってきた．本稿では，我々が行ってきたヒット数の信頼性評価手法の研究に加え，提案した信頼性評価手法の妥当性，すなわち過去のヒット数データを用いて計算された信頼性評価値が未来のヒット数遷移に通用するか否かの検証を行った結果を報告する．

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 節においてヒット数の変動傾向と変動原因について論じる．4 節において，我々の過去の研究において提案されたヒット数の信頼性に対する評価指

[†]早稲田大学大学院　基幹理工学研究科
Graduate School of Fundamental Science and Engineering, Waseda University
[††]早稲田大学理工学術院，国立情報学研究所
Faculty of Science and Engineering, Waseda University, National Institute of Informatics

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

標と，大規模なヒット数データを用いて行われた信頼性評価実験の結果についてまとめる．本研究では，上記評価指標に対する妥当性検証実験を行ったので，5 節にてその実験方法と結果を示す．6 節において議論をまとめる．

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報告する．まず 2.1 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説明する. 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 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 ヒット数を用いて同義語抽出を行なう研究

Turney[3]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．

2.1.2 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．

2.1.3 ヒット数を利用する研究のまとめ

これまでに例を示したように，ヒット数は Web 全体をコーパスとした自然言語処理に関する研究に多く用いられている．ヒット数を自然言語処理に用いる際，数多くの研究は 2.1.1 のように，複数クエリに対するヒット数の大小関係を用いている．すなわち, 複数クエリのうちどちらのほうがより多く Web 上に用いられているかを特定するために，ヒット数を用いているケースが多い[1][3]．この場合，ヒット数の大小関係さえ特定できれば，ヒット数の絶対値そのものは大きな意味を持たない．2.1.2 は，ヒット数の絶対値を用いて単語間の距離を定義しているが，2.1.2 を用いた数多くの研究は，結局求めた距離を大小関係問題に帰着させて用いている．

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[6]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[7]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．

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

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

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

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言える．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. ヒット数の変動傾向と変動原因の考察

本節では，ヒット数を長期的に観測した際に見受けられる変動傾向について論じた上で，検索を行うタイミングによってヒット数が変動する原因について論じる．

3.1 ヒット数変動傾向のパターン考察

本節では，ヒット数を長期的に観測した際に見受けられる変動傾向について論じる．

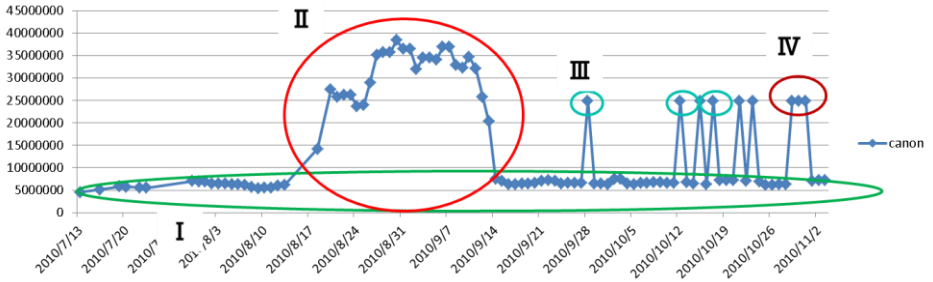


図 1. Google におけるクエリ”canon”に対するヒット数変動

図 1 は，Google におけるクエリ”canon”に対する各日にちにおけるヒット数変動を示している．このクエリのヒット数変動は，典型的なヒット数変動パターンをいくつか含んでいるため，この図を用いて変動傾向のパターンを考察する．

図 1 に示したとおり，典型的なヒット数の変動パターンは次の 4 種類である．

I. 長期間ヒット数を観測した際に大部分を占める，変動の少ない安定した部分
II. 数日間にわたってコンスタントに変動の大きい部分
III. 1 日のみ大きく外れ値をとる部分
IV. 数日間にわたって，比較的安定した外れ値をとる部分

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．Ⅱ～Ⅳで観測されるように，短期間の間にヒット数が大きく変動する場合，得られたヒット数は偶発的なエラー値であると考えられ，次節において考察する．図 1 に示したヒット数の値の変動幅は 5～7 倍の範囲で抑えられているが，他の多くのクエリではⅡ～Ⅳのような期間でヒット数の変動幅が 100～1000 倍ともなるクエリも数多く観測されており，Ⅱ～Ⅳのような変動の大きい期間に得られたヒット数を採用してしまうことによる悪影響は大きい．ヒット数を研究に用いる際には，Ⅱ～Ⅳにおけるヒット数を避け，安定した I におけるヒット数を採用すべきである．

対象となるクエリのヒット数を長期間観測し，図 1 のようなグラフを描くことができれば，上記の要求は簡単に実現できる．しかし，ヒット数を用いるたびに長期間の観測を行うことは現実的ではない．逆に，1 回のみヒット数を観測しただけでは，I～Ⅳのどの期間で得られたヒット数かを判断することができない．本研究は，大規模に収集したヒット数観測データを用いて「どのような条件を満たしたヒット数は，期間 I から得られたヒット数であると十分高い確率で保証できるか」を統計的に評価すること目的としたものである．

3.2 ヒット数が変動する原因の考察

本節では，前節にて論じたヒット数変動を発生させうる原因を考察し，まとめる．各検索エンジンにおける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原因についてはわからない．そこで，本節ではヒット数変動傾向と一般的な検索エンジンの構成を照らし合わせ，ヒット数が変動する要因を推測する．

3.2.1 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，検索エンジンは更新頻度が高いと予想される Web ページに対して，数分あるいは数秒といった短い間隔で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，前節の I のような期間においても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 8). 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. これは, 前節のⅡのような数日にわたるコンスタントな変動を説明しうる.

3.2.2 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. 前節のⅢのような変動パターンは，これによって引き起こされていると考えることができる.

3.2.3 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[16]. 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. 前節Ⅲ，Ⅳのような変動パターンはこの要因によって説明できる.

4. 検索エンジンのヒット数に対する信頼性評価指標

前節までに述べられたヒット数の信頼性に関する問題に対して，我々はヒット数の信頼性に対する定量的な評価指標を定義する試みを行ってきた[9]. もし，得られたヒット数に対して適切な信頼度評価値を得ることができれば，信頼性の低いヒット数の採用を避け，信頼性が十分高いヒット数のみを採用することが可能となる. したがってヒット数の信頼性に対して適切な評価指標を定義することは，ヒット数を研究で用いる際に非常に重要であると考えられる. 以下，4.1～4.4 において，我々がこれまで行ってきた検索エンジンのヒット数に対する信頼性評価指標定義とヒット数信頼性評価実験の概要をまとめる. 次に 4.5 において，4.4 で得られた信頼性実験結果を実際にヒット数ユーザに対して提供するシステムの概要を簡単に示す.

4.1 ヒット数信頼性評価指標の定義

2.1.3 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. 言い換えると「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(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.

$$\begin{aligned} &reliability\left(hit[A] > hit[B], m\right) \\ &\quad = \Pr\left(days\left(hit[A] > hit[B]\right) = m\right) \end{aligned} \tag{4.1}$$

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い.

4.2 ヒット数の信頼性評価実験

3.1 にて述べたとおり，本研究の目的は，「どのような条件を満たしたヒット数は安定した期間から得られたヒット数であると十分高い確率で保証できるか」を統計的に評価するというものである.

本実験では，前節にて定義された信頼度評価指標を用い，大規模に収集したヒット数データを用いて次のような関数（信頼度関数）を求めた.

$$\begin{aligned} &function\ L : (R, a, m) \rightarrow r \\ &where\ r = reliability\begin{pmatrix} hit[A] > hit[B], m \\ |\ hit_{base}[A] : hit_{base}[B] = R : 1, \\ a = observeddays\left(hit[A] > hit[B]\right) \end{pmatrix} \end{aligned} \tag{4.2}$$

ここで R はある時点 *base* におけるクエリ A，クエリ B のヒット数比率，a は 2 クエリのヒット数が安定していることを観測した日数，r は同一の大小関係が m 日続くことに対する信頼度を示している. m 日とは，ヒット数ユーザがヒット数に対して安定していてほしいと要求する期間である. つまりこの関数は，ある時点 *base* における 2 クエリのヒット数比率が R:1 であり，かつ同一の大小関係が a 日続いているという条

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

件のもとでの信頼度を，収集されたヒット数データを用い，*R*, *a*, *m* を様々に変化させて求めたものである．具体的な信頼度の算出方法は[9]に示しているのでここでは省く．

4.3 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．

4.4 ヒット数の信頼性評価実験結果

4.4.1 *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．

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．

4.4.2 *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．

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．

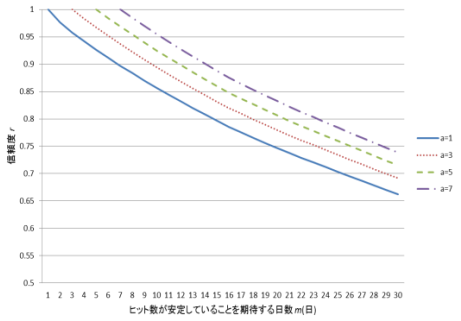


図 2. Google の *r*-*m* グラフ(R=2)

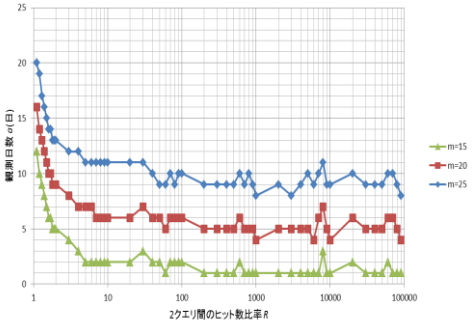


図 3. Google の *a*-*R* グラフ(r=0.85)

4.5 ヒット数信頼度提供システムの概要

本節では，4.1 にて定義された信頼度評価指標や 4.4 にて得られた信頼度関数を，実際にヒット数ユーザが用いる際にどのような手順を踏むかをまとめる．

図 4 は，ヒット数信頼度提供システムの概要を示している．システムは信頼度を提供するために「ヒット数収集」「信頼度関数算出」「信頼度提供」の 3 つのタスクを遂行する．

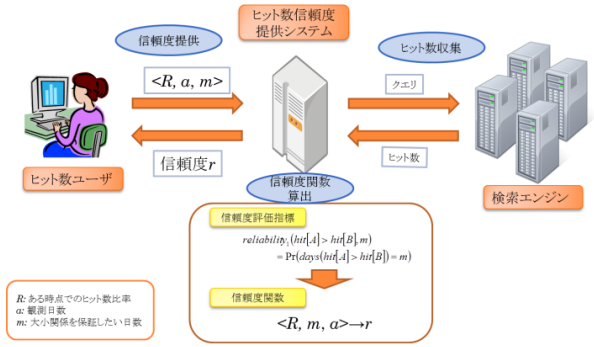


図 4. ヒット数信頼度提供システムの概要

4.5.1 ヒット数収集

システムは検索エンジンのヒット数をサンプリングするために，十分な数の固定クエリを保持する．システムはこれらのクエリ全てについて，毎日各検索エンジンに対して検索をかけてヒット数を収集し，推移を記録する．

4.5.2 信頼度関数算出

収集されたヒット数データに対して，先述したヒット数の信頼度評価指標を適用し，4.2 と同様に信頼度関数<*R*, *a*, *m*>→*r* を算出する．

4.5.3 信頼度提供

実際にヒット数ユーザが関わるフェーズである．ユーザは次のようなステップをたどる．ユーザは，ヒット数が一定期間 *m* 日間安定して欲しいという要求を持っていたとする．

step1. 大小関係を比較したい複数クエリについて，それぞれ検索エンジンを用いてヒット数を得る．ヒット数比率 *R* を得る．

step2. この時点での観測期間は 1 日なので，*a* = 1 となる．*R*, *m* は既に定まっているので，<*R*, *m*, 1>に対する信頼度 *r* をシステムに要求する．

step3. システムはあらかじめ算出してある信頼度関数に基づいて，ユーザの要求に対する信頼度 *r* をユーザに提示する．

step4. ユーザは提示された信頼度が要求に満たない場合，対象クエリに対するヒット数の観測を続ける．観測期間 *a* がインクリメントされていく．

step5. 4.4 で示したとおり，観測期間 *a* の上昇とともに信頼度 *r* も上昇していく．

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

ユーザは十分高い信頼度を得た時に観測をやめ，その時のヒット数の大小関係を採用する．

step6.　もし観測期間中に大小関係の入れ替わりが生じた場合，a= 1 として step1 に戻る．

5.　信頼性評価指標の妥当性検証実験

本節では，前節にて定義されたヒット数に対する信頼性評価指標の妥当性を検証する．以下，まず 5.1 において妥当性検証方法について述べ，次に 5.3 において検証実験の結果をまとめる．

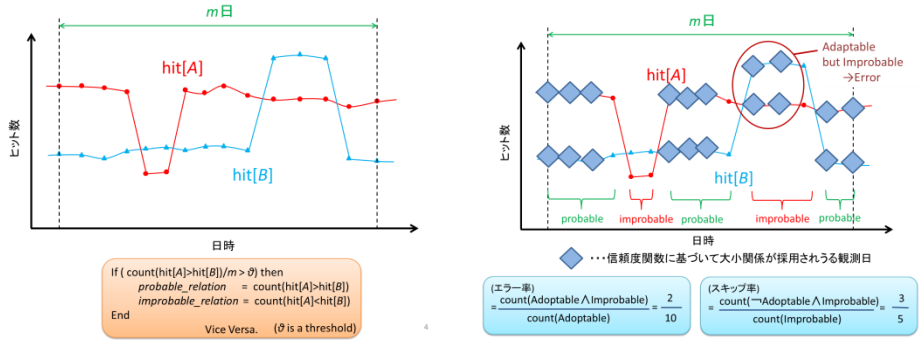
5.1　信頼性評価指標の妥当性検証実験方法

4.5 にて示したとおり，本研究において提案する信頼性評価指標を実際にヒット数ユーザが用いる際には，信頼度を提供するシステムがあらかじめ大規模に収集された過去のヒット数データに基づいてヒット数信頼性関数を算出しておき，ユーザはそのシステムに対し，<ある時点での複数クエリのヒット数比率 R, ヒット数を保証したい日数 m, 観測日数 a>の 3 つの変数を指定することによって，現在のヒット数の未来の変動に対する信頼度 r を得る．

すなわち，この指標の妥当性を主張するためには，「過去のヒット数データを用いて算出された信頼度関数が，未来のヒット数変動に通用する」ということを裏付ける必要がある．本実験ではこの点に着目し，過去のヒット数データを用いて算出された信頼度関数において，十分高い信頼度 r を保証するヒット数取得条件<R, a, m>に従ってヒット数を取得することで，実際に不安定な時期のヒット数の採用を避けることができているかを評価する．

5.1.1 妥当性検証に用いる用語の定義

妥当性検証にあたり，「正しいヒット数の大小関係」「エラー率」「スキップ率」を次のように定義する（図 5）．



(a) 正しいヒット数の大小関係の定義　　(b) エラー率，スキップ率の定義

図 5. 妥当性検証に用いる用語の定義

✧ **正しいヒット数の大小関係**：

ユーザが指定した期間 m 日間において，もし 2 クエリのヒット数間の大小関係がどちらかに十分偏っていれば，その大小関係を正しい大小関係と定義する．もし，大小関係がどちらにも偏っていない場合，その期間における正しい大小関係は不定とする．

すなわち，ある閾値 θ に対して，正しい大小関係を次のように定義する．

```
If ( days(hit[A]>hit[B])/m >  θ) then
    right_relation = hit[A]>hit[B]
Else If ( days(hit[A]<hit[B])/m >  θ) then
    right_relation = hit[A]<hit[B]
Else
    right_relation = UNDEFINED
End
```

days(hit[A]>hit[B])は，m 日間中にクエリ A のヒット数がクエリ B のヒット数を上回っていた日数を表す．

エラー率：

信頼度関数に基づいて採用したヒット数のうち，どれだけ誤りの大小関係を持ったヒット数を採用したかを示す指標である．次式によって表される．

$$(Error\ Ratio) = \frac{days(Adopted \wedge Wrong)}{days(Adopted)} \qquad (5.1)$$

days(*Adopted*)はヒット数を採用した日数を表し，days(Adopted ∧ Wrong)は大小関係が誤っているヒット数を採用した日数を表す．エラー率が低いほど，信頼性評価指標の妥当性が高いことを示している．

✧ **スキップ率**：

信頼度関数に基づいてヒット数を採用したとき，どれだけ誤りの大小関係を持ったヒット数を回避できたかを示す指標である．次式によって表される．

$$(Skip\ Ratio) = \frac{days(\neg Adopted \wedge Wrong)}{days(Wrong)} \qquad (5.2)$$

days(*Wrong*)はヒット数の大小関係が誤っている日数を表し，days(¬*Adopted* ∧ *Wrong*)は大小関係が誤っているヒット数を採用しなかった日数を表している．スキップ率が高いほど，信頼性評価指標の妥当性が高いことを示している．

5.1.2 妥当性検証の手順

妥当性検証実験では，次のようなステップをたどる．

step1.　観測期間中のある時点 D を固定し，D を仮想的に「現在」と見立てる．本検証の目的は，D より過去のデータを用いて算出された信頼度関数が，D 以降のヒット数変動に通用するか否かを確かめるというものである．

step2.　D より過去一定期間 T のヒット数データを用いて信頼度関数を算出する．本実験では T を 1 ヶ月とした．

step3.　ヒット数を保証したい日数 m を固定し，信頼度 r が一定値を超えるための

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

<*R*, *a*>の条件を，step1 にて算出された信頼度関数から特定する．本実験では *r*>0.9 と設定した．

step4.　観測を行った 10,000 件のクエリうち，*D* 以降のヒット数変動が step3 にて特定された<*R*, *a*>の条件を満たす 2 クエリを選出し，時点 *D* で得られた大小関係が *D* から *D*+*m* までの期間における「正しいヒット数の大小関係」と一致しているかをチェックする．本実験では「正しいヒット数の大小関係」を特定する際の閾値を θ = 0.85 と設定した．

step5.　様々な 2 クエリの組み合わせごとに step4 を行ってエラー率，スキップ率の集計を行い，時点 *D* における信頼性評価指標の妥当性評価値とする．

step6.　時点 *D* を様々に変化させ，信頼性評価指標の妥当性評価値の推移を得る．

5.2 使用したデータ

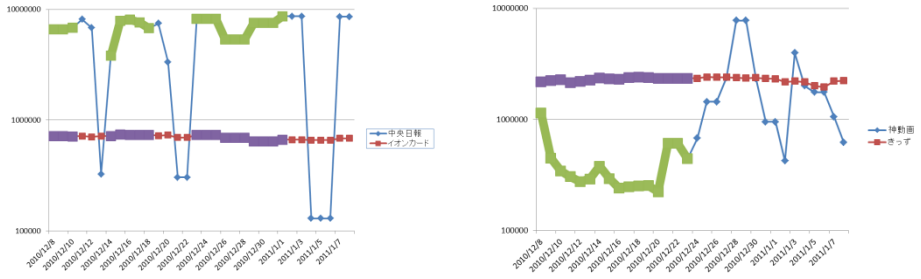
本実験において使用したデータは，Yahoo! Japan の 2007 年 12 月のクエリログにおいて頻出順に並べて現れた上位 10,000 件のクエリに対する Google，Yahoo!，Bing における 2010 年 9 月～2011 年 1 月のヒット数データである．

5.3 信頼性評価指標の妥当性検証実験結果

前節で説明した妥当性検証実験に対する結果を示す．まず 5.3.1 において具体的なヒット数実データを例にとって信頼度関数に基づいたヒット数の採用例を示した後，5.3.2 においてエラー率，スキップ率による信頼性評価指標の妥当性検証実験の結果を示す．

5.3.1 信頼度関数に基づいたヒット数の採用例

図 6 は，Google における 2010 年 11 月のヒット数データを用いて算出された信頼度関数を，2010 年 12 月のヒット数変動に適用し，*m* = 20 日に対する信頼度が一定値以上(*r*>0.9)となる条件を満たしたときにのみ大小関係を採用している図を表している．



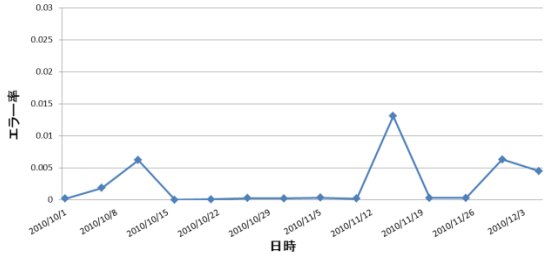
(a) 例 1　　(b) 例 2

図 6. Google における信頼度関数に基づいたヒット数の採用例(*m*=20, *r*>0.9)

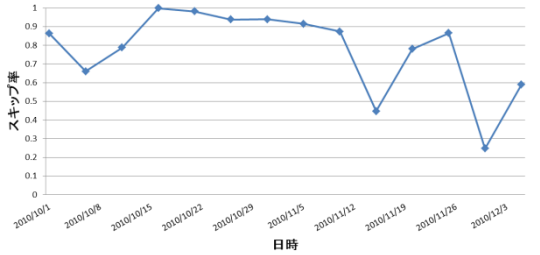
図で，ヒット数を採用した期間は太線によって表されている．不安定な時期における大小関係を効果的に避け，比較的安定した箇所の大小関係を採用していることがわかる．すなわちこれらの例は，本信頼性評価指標を裏付けている例であると言える．

5.3.2 信頼性評価指標の妥当性検証

図 7 の(a)，(b)にそれぞれエラー率による妥当性検証結果と，スキップ率による妥当性検証結果を示す．



(a) エラー率の推移（r>0.9，θ = 0.85）



(b) スキップ率の推移（r>0.9，θ = 0.85）

図 7. Google における 10/1～12/5 のエラー率とスキップ率の推移

図 7 より，エラー率は全体として低く保たれ，スキップ率は一部の期間を除いて高く保たれていることが見て取れる．これらの結果はともに，(4.1)に示したヒット数の信頼性評価指標の妥当性を強く裏付けている結果といえる．

ただし 11 月 15 日，11 月 30 日において，エラー率の小さな上昇，スキップ率の大きな下降が見られる．この原因を調査するため，いくつかのクエリに対する 11 月 20 日～12 月 10 日におけるヒット数変動を図 8 に示す．

図 8 で示したように，11 月 30 日付近のヒット数は互いに激しい入れ替わりを生じている．このような時期は Yahoo!, Bing においても時折見られる．[8]ではこのようなヒット数の変化が激しい時期を「変動期」と呼び，検索エンジンが内部のインデックスを大幅に更新する際に見られる現象であるとしている．

図 7 の結果より，変動期以前のデータを用いて算出された信頼度関数は，変動期で用いることができないということがわかる．しかし変動期において信頼度関数が使えない時，多くのクエリで同時に入れ替わりが発生するので，図 4 で示したような十分な数のクエリに対するヒット数を常に監視するようなシステムを実現することができれば，検索エンジンが変動期に入ったことをシステムが感知することができるため，ヒット数利用ユーザに対して警告するなどして対処することできると考えられる．

Vol.2011-DBS-152 No.8
Vol.2011-IFAT-103 No.8
2011/8/3

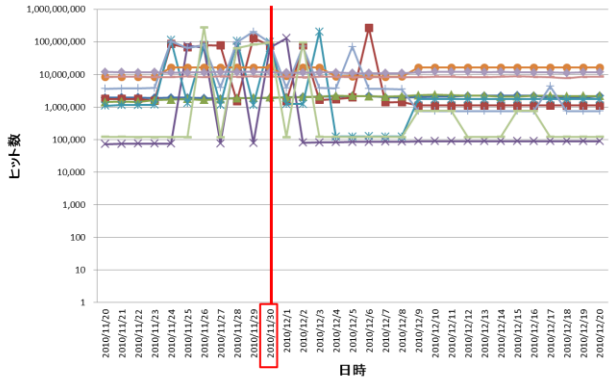


図 8. Google の 11 月 30 日付近のヒット数変動の例

6. 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指標を定義し，指標の妥当性評価を行った．

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．

そこで我々は，どのような条件を満たしたヒット数が，十分な信頼性を保証できているかを特定するため，ヒット数に対する信頼性評価指標の定義を行った．本研究では，ヒット数ユーザがヒット数に対してある一定期間(m 日間)安定していてほしいという要求があるという前提のもと,「複数クエリのヒット数の大小関係が m 日間変わらない確率」という信頼性評価指標を提案した．その上で，大規模なヒット数観測データをもとに各検索エンジンに対してヒット数信頼性評価実験を行い，<ある時点での複数クエリのヒット数比率 R, ヒット数を保証したい日数 m, 観測日数 a>と信頼度 r との対応付けを表す信頼度関数を算出した．

本稿では信頼性評価指標の提案，信頼性評価実験に加え，提案した信頼性評価指標の妥当性検証を行った．ここで，妥当性とは，過去のヒット数データを用いて算出された信頼度関数が，未来のヒット数変動に通用するか否かを意味するものである．妥当性を検証する実験として，過去のヒット数データを用いて算出された信頼度関数において一定以上の信頼度を保証するヒット数取得条件が，実際に未来のヒット数変動において大小関係が不安定な時期におけるヒット数の採用を効果的に避けているかを「エラー率」と「スキップ率」によって評価した．結果として，過去のデータを用いて算出された信頼度関数を未来のヒット数変動に適用した場合でも,エラー率は低く，スキップ率が高いということがわかった．したがって本評価指標は，高度に妥当性があると主張することができる．ただし，検索エンジンには短期間で多くのクエリが頻繁に入れ替わる時期（変動期）があり，このような時期においては過去のデータを用いて算出された信頼度関数を使うことができない．しかし変動期において信頼度関数が使えない時，多くのクエリで同時に入れ替わりが発生するので，図 4 で示したような十分な数のクエリに対するヒット数を常に監視するようなシステムを実現することができれば，検索エンジンが変動期に入ったことをシステムが感知することができるため，ヒット数利用ユーザに対して警告するなどして対処することできると考えられる．

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) Grefenstette, G.: The WWW as a resource for example-based MT tasks, ASLIB Translating and the Computer Conference, London (1999).
2) Cilibrasi, R. L. and Vitanyi, P. M. B.: The Google Similarity Distance, IEEE Trans. on Knowledge and Data Engineering, Vol.19, No.3, pp.370 – 383 (2007).
3) Turney, P.: Mining the web for synonyms: PMI-IR versusLSA on TOEFL, Proc. ofECML-01, pp. 491-502 (2001).
4) Cimiano, P. and Handschuh, S.: Towards the self-anotating web, Proc. WWW2004, pp462-471 (2004).
5) Matsuo, Y. et al.:POLY-PHONET: An advanced social network extraction system, Proc. WWW 2006 (2006).
6) Thelwall, M.: Quantitative Comparisons of Search Engine Results, J. of the American Society for Information Science and Technology, Vol.59, No.11, pp.1702-1710 (2008).
7) Uyar, A.: Investigation of the Accuracy of Search Engine Hit Counts, J. of Information Science, Vol.35, No.4, pp.469-480 (2009).
8) 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).
9) 佐藤亘，打田研二，山名早人: 検索エンジンのヒット数の信頼性に対する評価，第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM2011) (2011).
10) Google News, http://news.google.com/
11) The top news headlines on current events from Yahoo! News, http://news.yahoo.com/
12) Bing, http://www.bing.com/?scope=news
13) Google Realtime Search, http://www.google.com/realtime
14) Challenges in Building Large-Scale Information Retrieval Systems, http://research.google.com/people/jeff/WSDM09-keynote.pdf
15) Skobeltsyn, G. et al.: 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) Barroso, L.et al.: 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)