DBSJ Journal, Vol.12, No.1
June 2013

検索ヒット数の正確性評価：大規模クロールデータに対する文書頻度との比較

Accuracy Evaluation for Search Engine's Hit Count: Comparison with Document Frequency in Large-Scale Crawl Data

佐藤 亘♥　上田 高徳♦
山名 早人♦

Koh SATOH, Takanori UEDA
Hayato YAMANA

近年，自然言語処理をはじめとする Web 上の単語統計の指標として,検索エンジンから得られるヒット数を用いる研究が数多く行われている．しかし,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，研究の基盤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，ヒット数の正確性を評価し，ヒット数を利用する研究に対する影響の度合いを明らかにすることには重要である．本研究では，大規模な Web クローリングにより集められたデータにおける，あるワードの出現頻度とそのワードをクエリとした時のヒット数とを比較しヒット数の正確性評価を行った．結果として，これら二種類の値の間のピアソン相関係数は 0.807 であった．特に，時系列上で安定しているヒット数を持つクエリのみを用いると，相関係数が 0.897 に向上することから，時系列上で安定しているヒット数を用いることが重要であることがわかった．さらに,ヒット数が持つ誤差範囲を確率分布で表現することにより，二種類のクエリ間の大小関係を 99%以上の確率で保証するためにはヒット数が 31 倍以上離れていることが必要であることを明かにした．

Recently, there have been a number of studies that utilize search engines' hit counts as a reference index of word statistics. However, the hit counts can vary unnaturally when observed on different days, and may contain large errors that affect researches depend on those results. Thus, it is indispensable to evaluate the accuracy of hit counts and to clarify the degree of errors hit counts can cause. In this research, we performed large-scale Web crawling and evaluated the accuracy of hit counts by comparing hit counts and the document frequency among the crawled data. The results show that the Pearson correlation coefficient between hit counts and the number of words among the crawled data, where we assume they are accurate, is 0.807, while it improves up to 0.897 when selecting hit counts that are stable in time series. This result shows that it is better to use hit counts that are stable in time series. Moreover, we have made clear that we should confirm that two hit counts are different more than 31 times when we want to guarantee the relationship between the two hit counts' magnitude relation with 99% accuracy, by modeling their probabilistic distribution.

1. はじめに

膨大な量の Web コンテンツを活用した研究を実現するために，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]-[4]．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，Web 上の文書集合における検索クエリの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，クエリ単語間の距離を定義する研究[1]，同義語抽出を行う研究[2]などの自然言語処理に関する研究が多く挙げられるほか，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[3]や，Web からの自動ソーシャルネットワーク抽出[4]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，検索エンジンが返すヒット数には，検索するタイミングによって不自然に変化する現象や検索結果の表示開始ページに依存し大きく変動する現象が見受けられる[5]-[8]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，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 倍以下に変化することがしばしばあり，様々な研究やアプリケーションの基盤として用いるには無視できない誤差となっている．

これまで，検索エンジンの信頼性の問題についていくつかの研究が行われてきた[5]-[9]．しかし，これらの研究の多くは，複数の検索エンジンから得られるヒット数を比較したもの[5]や，各検索エンジンにおけるヒット数変動傾向を特定した研究[6]など，検索エンジンから得られる情報のみに基づいてヒット数の信頼性を議論する研究が主であった．しかし，ヒット数の正確性，すなわちヒット数が Web 上での当該語句の出現頻度とどれだけ一致しているかは，検索エンジンから得られる情報のみを用いて計量することはできない．ヒット数の正確性を確実に評価するためには，Web 上の網羅的な文書に対する単語統計とヒット数とを比較することが必要不可欠である．

そこで本研究は，大規模な Web クローリングを行い，集められたデータにおけるあるワードの出現頻度と，そのワードをクエリとした時のヒット数とを比較することによってヒット数の正確性評価を行う．ヒット件数の正確性評価によって，ヒット件数を用いる研究が，ヒット件数の誤差によってどれだけの誤差を生じうるかを特定することが可能となる．さらに，どのような条件下で得られたヒット数が正確な Web の単語統計と最も一致しているかを特定することができる．本稿では，Web クローリング方法，比較対象とするクエリの選び方を含めた，Web 上の網羅的な文書に対する単語統計の取得方法を提案すると共に，大規模なクロールデータにおける単語統計とヒット数とを比較した結果を示す．

以下では，第2節において関連研究についてまとめ，第3節にて用語を定義する．第4節においてWeb上の網羅的な文書に対する単語統計を取得するためのWebクローリング方

♥ 学生会員 早稲田大学 基幹理工学研究科
kohsatoh@yama.info.waseda.ac.jp
♦ 正会員 日本アイ・ビー・エム株式会社 東京基礎研究所（ただし本研究は，早稲田大学 基幹理工学研究科に在学中になされたものである）ueda@yama.info.waseda.ac.jp
♦ 正会員 早稲田大学 理工学術院，国立情報学研究所
yamana@waseda.jp

法と文書頻度のカウント手法について論じ，第5節にて，小規模なクロールデータにおける文書頻度と検索ヒット数とを多角的に比較する．

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 においてヒット数を利用した研究について紹介し，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

2.1.1. ヒット数を用いた同義語抽出

Turney[2] 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，ある単語に対して複数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，同義語の候補として与えられた単語 $choice_i$ のスコア

$$score(choice_i) = \frac{hits(problem\ AND\ choice_i)}{hits(choice_i)}$$

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．

2.1.2. ヒット数を用いたクエリ間類似度の定義

Cilibrasi ら[1]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．

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

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．

2.2. ヒット数の信頼性を対象とした研究

2.2.1. コーパスにおける単語の出現頻度とヒット数とを比較した研究

Keller ら[9]は，「形容詞+名詞」，「名詞+名詞」などいくつかの品詞の組み合わせで計 540 クエリを構成し，Google と Altavista におけるヒット数と，2 種類のコーパス(BNC[10], NANTC[11])における出現頻度とを比較し，結果として高い相関を得たとしている．例えば，「形容詞+名詞」のクエリに対する Google のヒット数と BNC における出現頻度を比較したとき，ピアソンの積率相関係数において 0.850 という値を得たと報告している．

この研究結果は，ヒット数を利用する研究に広く支持され，ヒット数を Web 上の文書集合における出現頻度とみなすことの論拠としてしばしば引用される[1]．しかし，Keller らの実験は比較に使用するクエリ数や比較対象とするコーパスの規模が小さいほか，ヒット数の時系列上の変動を考慮していないこと，さらに固定的なコーパスからは Web 特有の新語や固有名詞等の出現頻度を取得できないなどが問題として挙げられる．

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[6]は，Google，Yahoo!，Bing 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 1,000 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 1,000 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．

$$Percentage\ of\ Error = \frac{Estimate - ReturnedDocument}{ReturnedDocument} \times 100$$

1,000 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 が最も正確なヒット数を返していると結論づけた．この研究は，取得した Web ページ数が 1,000 件以下のときに取得できた Web ページ数が正しいヒット数であるという仮定に対する検証が十分なされていないという点と，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点に問題がある．

3. Web における単語統計取得手法

本節では Web 上の網羅的な文書に対する単語統計を取得するための手法を論じる．まず 3.1 において Web クローリング方法の概要を述べ，次に 3.2 においてクロールされたデータから文書頻度を取得する手法を論じる．

3.1. Web クローリング方法

本研究では，我々が開発した Web クローラ [12] を用いる．以下では，クロールのシード選出，クロールの範囲と制限，クロールの終了条件について論じる．

3.1.1. シードの選出

単語統計の取得に妥当なページを網羅的に収集するためには，多様かつ信頼性の高い Web ページをシードとして選出する必要がある．本研究では，Wikipedia の外部リンク集 [13] から政府や大学のページなど，経験的に信頼性が高いと考えられるリンク群を抽出する．

3.1.2. クロールの範囲と制限

本研究は，日本語と英語のクエリを対象として検索ヒット数の正確性評価を行う．このため，簡易的なフィルタとしてトップレベルドメインを日本(.jp)，英語圏(.uk, .ca 等)，ジェネリックドメイン(.com, .gov 等)に限定してクロールを行った．さらに，同一ホストから過剰にページを取得することや，広告系 Web ページへ頻繁にアクセスすることを避けるため，cgi や GET パラメータを含む URL を排除し，静的なページを示す URL のみをクロールの対象とした．

3.1.3. クロールの終了条件

終了条件を次のように定める．

1. あらかじめ定められたクエリ群 $\boldsymbol{Q}$ に対し，クロールデータに対する出現確率をリアルタイムに監視する(クエリ群の算出方法，クエリの出現頻度の取得方法については 3.2.2 で述べる)
2. $\boldsymbol{Q}$ のうち，閾値以上の割合のクエリに対する出現確率が

十分収束したとき，クロールを終了する．ここで出現確率の収束とは，クロールされた文書数の r_d%の増加に対し出現確率の変動が r_p%以内に収まっていることを指す．

3.2. 文書頻度のカウント手法

クロールデータから文書頻度を取得する方法を述べる．

3.2.1. 文書頻度を取得するクエリの選定

以下に示す2通りのクエリ群 Q_1, Q_2 を選出した．

- Q_1：Wikipedia のタイトルから，単語数・出現頻度・言語・トレンド性にばらつきがでるように 6,000 件選出したもの．一般的な検索エンジンのアーキテクチャ[14][15]を考慮したとき，検索クエリの単語数・出現頻度・言語・トレンド性の違いによって，検索エンジン内部でのヒット数概算手法や概算時に扱われるデータが大きく異なると考えられるため，この 4 つの観点から多様なクエリを選出した．詳細を表 1 にまとめる．
- Q_2：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．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を選定した．

表 1．クエリ群 Q_1 選出の基準

項目	選出した条件
単語数	1〜3 語
出現頻度	Yahoo! Japan Web 検索 API [19] で 10^3~10^7 の値をとるもの
言語	日本語，英語
トレンド性	Wikipedia のページアクセス数 [20] で上位 1,000 件に入るものとそうでないもの

3.2.2. 文書頻度の取得方法

収集されたクロールデータに対して，フィルタリングや本文抽出などいくつかの処理を施すフローと施さないフローそれぞれについて文書頻度をカウントすることで複数の単語統計データセットを取得する．以下，フロー中の各処理についてまとめる．

フィルタ層：
重複削除[16]を行い，スパムや重複ページによる単語統計への影響を排除する

ページランク層：
収集されたデータ内でページランクを計算し，その高低によって Web ページを分類する

本文抽出層：
Web ページ本文抽出[17]を適用し，サイドバーや広告などといった本文以外の部分に存在する語を排除する

ワードカウント層：
形態素解析 lucene-gosen[18]を用い，入力された Web 文書を形態素解析する．クエリを構成する各単語が一形態素として文書中に存在するときカウントアップする

このように複数の単語統計を取得することで，

1. ヒット数がどの単語統計と最も相関が高いかがわかり，ヒット数の特性を特定できる
2. 3.2.3 にて述べる文書頻度の妥当性検証によって，どのような処理を経て得られた文書頻度が Web 上の単語統計として最も妥当性が高いかがわかる

という利点が得られる．

3.2.3. 取得した文書頻度の妥当性検証

本研究では，ヒット数の正確性を評価するにあたり，3.2.2 の手順によって取得した単語統計を Web 全体の単語統計の正解セットとみなす．このため，得られた文書頻度の妥当性に対する強い裏付けが必要である．そこで次の 2 つの観点から取得した文書頻度の妥当性を裏付ける．

クロール時における出現確率の収束性
3.1.3 にて述べた通り，クエリ群の出現確率が十分収束するまでクロールを続ける．この収束性は，取得された文書頻度の妥当性に対する裏付けの一つと考えることができる．

固定的なコーパスにおける文書頻度との比較
関連研究[9]にならい，一般的な語に対する出現確率が，クロールデータに対するものとコーパス[10]とを比較して十分相関が高いことを確認する．

4. ヒット数の正確性評価

本節では，大規模なクロールデータに対する文書頻度を用いてヒット数正確性評価を行った結果を示す．

4.1. 使用したデータの概略

4.1.1. クロールデータ

Wikipedia 外部リンク集 [13] から政府 / 企業 / 大学系ページを抽出して得られた 7,882 個の URL をシードとし，2013 年 1 月 13 日〜16 日に収集できた 41,818,191 ページ (1.26TB) のデータを用いる．

4.1.2. ヒット数データ

Yahoo! Japan が提供する Web 検索 API [19] を用い，3.2.1 で述べたクエリ群に対しヒット数を収集した．

あるクエリに対するヒット数は検索オフセットと検索のタイミングによって変動する．本比較実験では，検索オフセットが 0 のときのヒット数を使用した．Wikipedia タイトルから選出したクエリ群 Q_1 については 2012 年 12 月 17 日〜2013 年 1 月 18 日の間，Yahoo! Japan 頻出クエリから選出したクエリ群 Q_2 については 2013 年 2 月 1 日〜同月 5 日の間に収集したヒット数データを用いる．検索する際には，クエリを構成する各単語を二重引用符で囲って[1]検索を行う．これは，検索エンジンに対して明示的に完全一致検索を行うよう指示するためであり，これによってクエリ拡張等，検索の前処理の影響を減らすことができる．

4.2. 文書頻度データの概略

クロールデータから抽出した文書頻度データセットの概略を示す．

4.2.1. 文書頻度の分布

クロールデータにおける選出したクエリの文書頻度の分布は図 1 のようになった．クロールデータ中に全く出現しなかったクエリはあったが，大きな偏りなくクエリを選定できていることが分かる．

4.2.2. 出現確率の推移

図 2 は Q_1 からランダムに選出したクエリについて，数え上げられた総文書数と出現確率との関係を示したものである．傾向として，出現頻度の高いクエリの出現確率は少ない文書数で収束するが，出現頻度の低いクエリの出現確率は本実験で用意した文書数では収束していないという特徴が見

[1] 例えば「the beatles」というクエリに対しては「“the beatles”」ではなく「“the” “beatles”」として検索を行う

受けられる．

4.3. ヒット数の正確性評価結果

4.3.1. 散布図と相関係数

Wikipedia タイトルから選出したクエリ群(Q_1)，Yahoo! Japan 頻出クエリから選出したクエリ群(Q_2)にそれぞれに対して，クロールデータ内での文書頻度を横軸，ヒット数を縦軸とした散布図を図 3 に示す．(a), (b)ともに正の相関が見て取れる．

次に，Q_1 と Q_2 それぞれの文書頻度についてヒット数との相関係数を表 2 に示す．ここでは相関係数はピアソンの積率相関係数とケンドールの順位相関係数を用いる．

4.3.2. 文書頻度の調整と誤差率分布

文書頻度によるヒット数の近似方法：

十分巨大な文書群 $\boldsymbol{D}$ に対して，クエリ q の出現確率 p(q) が文書群の選び方によらず一定であると仮定するならば，あるクエリに対する文書頻度は $\boldsymbol{D}$ 中の文書数にのみ依存する．すなわち検索エンジンがインデックスしている文書数 N, 本論文のクロールで収集した文書数 N'に対して，ヒット数 Hit(q)とクロールされた文書における文書頻度 DF(q)はそれぞれ

$$Hit(q)=p(q)\cdot N, \qquad DF(q)=p(q)\cdot N'$$

表 2．文書頻度とヒット数間の相関係数

	ピアソンの積率相関係数	ケンドールの順位相関係数
Q_1	0.807	0.798
Q_2	0.860	0.694

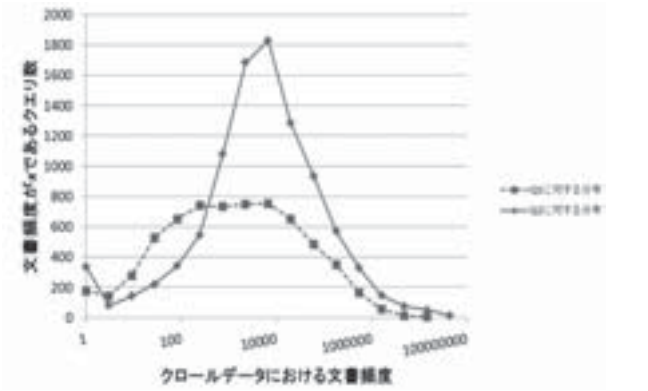


図 1．文書頻度の分布

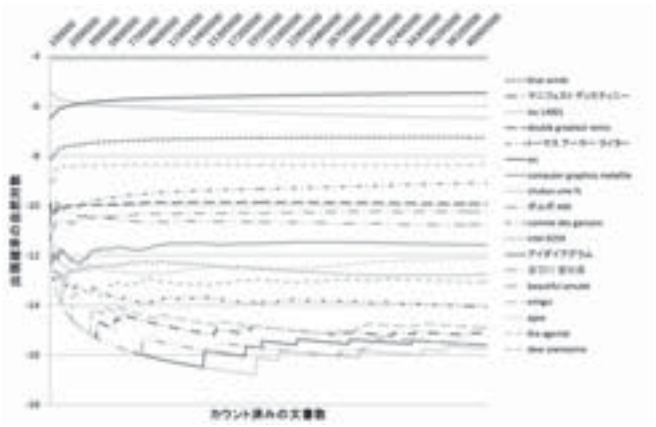


図 2．出現確率の推移

となり，

$$Hit(q)=k\cdot DF(q) \qquad (k=N/N')$$

を得る．したがって適切に k を選ぶことでヒット数をクロールデータにおける文書頻度で近似することができる．

そこで $q \in \boldsymbol{Q_1}$ に対して $k(q)=Hit(q)/DF(q)$を計算した．$k(q)$の自然対数 ln{$k(q)$} の分布を図 6 に示す．

$k(q)$の分布は対数正規分布によく近似しているという結果を得た．そこで k として，$k(q)$の中央値である k=164 を選んだ．

このときの Hit(q)と k･DF(q)の分布を図 4 に示す．分布が類似しており良く近似できていることが見て取れる．

ヒット数の誤差分布：

図 5 は，前段落にて述べた調整済み文書頻度を用い，ヒット数の誤差分布を示したものである．(ただし文書頻度が 0 のものを除いている)) ここで誤差を，

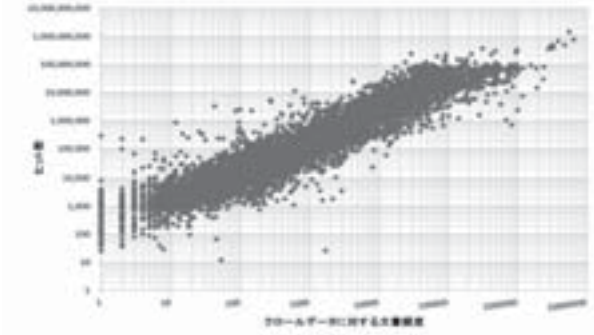
$$\frac{Hit(q)-k\cdot DF(q)}{\min(Hit(q),k\cdot DF(q))}$$

と定義する．

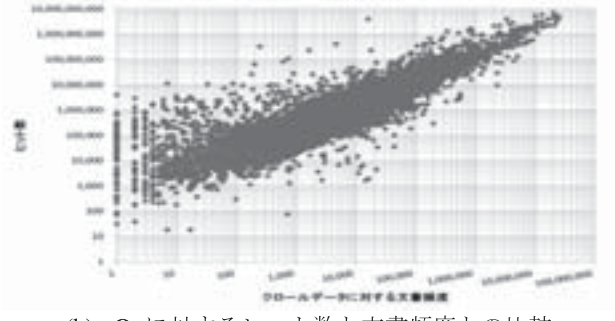
図 5 では, 誤差が 0 であるクエリが最も多く (2.05%)，誤差の絶対値が大きくなるに従ってその誤差を取るクエリ数が左右対称に減っていく傾向が見て取れる．調査の結果，誤差が-5.15～5.15 の範囲を取るクエリが全体の 90.0%を占めていた．つまり，検索エンジンからクエリ q に対するヒット数 $Hit(q)$を得たときに，クエリ q の正確な文書頻度が $Hit(q)$/6.15～6.15*$Hit(q)$の範囲に収まる確率が 90.0%であることを示している．図 5 は，ヒット数を用いる際に誤差の度合いとその発生確率を知る上で有効に活用できる．

大小関係を誤る確率：

図 5 の適用例の 1 つとして，複数のクエリに対するヒット数を取得したとき得られた複数のヒット数が互いにどれだけ離れていれば，そのクエリに対する正しい文書頻度の大小関係が十分高い確率で保証されるかを考える．



(a) $\mathbf{Q_1}$ に対するヒット数と文書頻度との比較



(b) $\mathbf{Q_2}$ に対するヒット数と文書頻度との比較

図 3．ヒット数と文書頻度との比較

ヒット数を利用する多くの研究[2]は複数のクエリに対して「どちらのクエリがより Web 上での出現頻度が高いか」を判定するためにヒット数を用いている．そのため，複数クエリに対して取得したヒット数が正しい文書頻度と比較してどの程度の確率で大小関係が誤っているのかを特定することは重要である．

いま，2 つのクエリに対するヒット数の大小関係が正しい文書頻度と比較して誤っている確率を考える．図 5 で示したように，あるヒット数に対する誤差の範囲とその発生確率は定まっているので，2 クエリに対するヒット数の大小関係が誤っている確率は，ヒット数の比にのみ依存する．図 7 はクエリ *a*, *b* に対するヒット数 *Hit(a)*, *Hit(b)* （*Hit(a)*<*Hit(b)*）を取得したときに，正しい文書頻度 *Df(a)*, *Df(b)*がとる確率分布を示したものである．それぞれの確率分布を $p_a(\cdot)$, $p_b(\cdot)$と表記する．クエリ *a, b* に対するヒット数 *Hit(a), Hit(b)*の大小関係が誤っている確率 *Prob.error* は次の式で表される．

$$Prob.error = \int_0^{\infty} p_b(n) \int_n^{\infty} p_a(m) dm dn$$

実測値を用いて *Prob.error* を計算し，「2 クエリに対するヒット数が *r* 倍離れていたときにその大小関係が誤っている確率は *p* である」を表したものが図 8 である．

95%の確率で保証したいならば 12 倍以上，99%の確率で保証したいならば 31 倍以上離れたヒット数を採用すべきであることがわかった．

正しい文書頻度の比を保証するヒット数比：

前段落と同様の考え方で，「2 クエリの正しい文書頻度に対する比が一定値 R_{real} 以上であることを確率 *p* で保証したいとき，2 クエリのヒット数間比は最低 R_{hit} 倍以上離れている必要がある」を示したのが図 9 である．自然言語処理等ヒット数の利用するいくつかのアプリケーションには，複数クエリに対して得られた文書頻度が互いに 10 倍以上離れている際に，取得した取得頻度を使用するという方針をとっているものがある．このようなケースには，図 9 を有用に用いることができる．

図 9 を見ると，例えば 2 クエリの正しい文書頻度間が 10 倍以上離れていることを 90%以上の確率で保証するためには，ヒット数の比が 85 倍以上離れてなくてはならないことがわかる．

4.3.3. 安定したヒット数のみを用いた比較結果

図 5 で，大きな誤差を取るクエリの時系列上でのヒット数推移を確認すると，その一部に大きな変動が確認された．

これを受けて，ヒット数が安定しているクエリのみを抽出してヒット数と文書頻度との比較を行い，全てのクエリで比較した場合と比べて類似性が高まるかを調査した．[7]で述べられた結論のひとつである，「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のヒット数は信頼できる」にならい，Q_1 からこれを満たす 1,655 クエリについて，ヒット数とクロールデータに対する文書頻度とを比較した．

比較の結果，ピアソンの積率相関係数で 0.897，ケンドールの順位相関係数で 0.800 であり，どちらの指標においても表 2 の値を上回っている．この結果は，ヒット数が数日にわたって安定していることを確認することでそのヒット数が正確である可能性を高めることができることを意味しており，既存研究[7][8]を支持する結果となっている．

4.3.4. その他の正確性評価結果

前小節までに述べた検証の他，ヒット数取得時のオフセットを変化させたときの正確性評価，3.2.1 で Q_1 に関して述べたクエリのタイプ別正確性評価，3.2.2 で述べた文書頻度の取得方法別の正確性評価を行ったが，明確な傾向は見られなかった．

5. まとめ

本研究では，検索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，ヒット数の正確性評価手法として大規模クロールデータに対する文書頻度と検索ヒット数との比較を多角的に行った．

本稿では，4,000 万の Web ページを収集し，計 16,300 件のクエリに対してヒット数と正確な文書頻度とを比較した．4,000 万の Web ページを収集したとき，出現頻度が e^{-14} 以上のクエリに対して，「カウント済みの文書数 10%の増加に対し， 99%以上のクエリに対する出現確率の変化が 5%以内に収まる」という極めて高い収束性が観測された．

クロールデータにおける文書頻度とヒット数との比較結果として，完全一致検索で得たヒット数はピアソンの積率相関係数において 0.807 という結果を得た．また，時系列上で

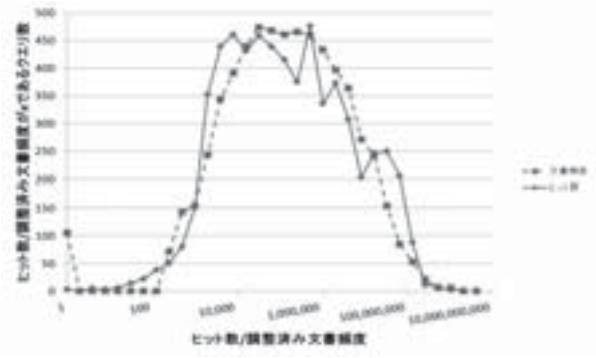


図 4. ヒット数と調整済み文書頻度の分布

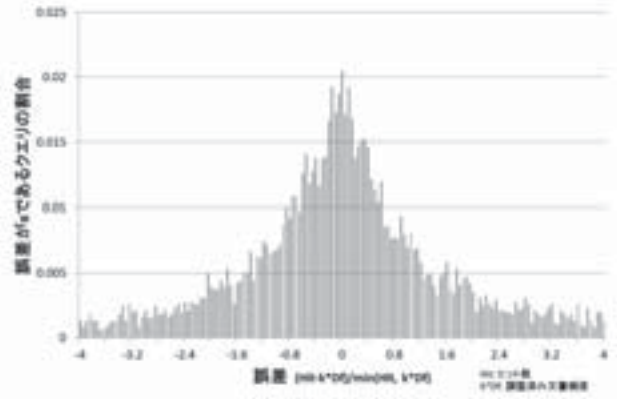


図 5. ヒット数の誤差分布

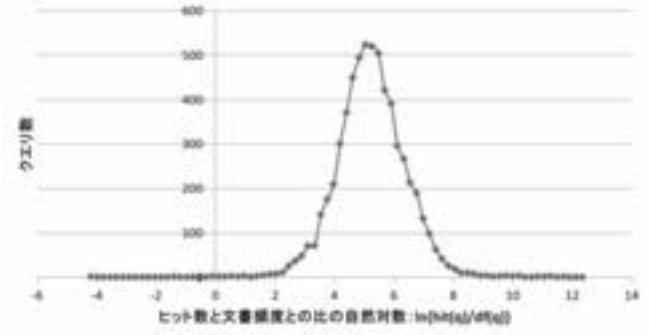


図 6. Q_1 に対するヒット数と文書頻度の比の自然対数の分布

安定しているヒット数のみを用いた場合，相関係数が 0.897 に向上し，複数日にまたがってヒット数が安定していることを確認することで取得したヒット数が正確である確率を高めることができることを確認した．さらに本研究ではヒット数の誤差の範囲とその発生確率を特定し，例えばヒット数が正確な文書頻度と比べて 1/6.15～6.15 倍の範囲を取る確率が 90%であることや，2 つのクエリに対するヒット数間が 31 倍以上離れていると，そのヒット数の大小関係が正しい文書頻度においても一致することが 99%以上の確率で保証できることなどが判明した．

[文献]

[1] 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
[2] P. Turney, “Mining the web for synonyms: PMI-IR versus LSA on TOEFL”, Proc. of ECML-01, pp. 491-502, 2001.
[3] P. Cimiano and S. Handschuh, “Towards the self-anotating web”, Proc. WWW2004, pp462-471, 2004.
[4] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida and M. Ishizuka, “POLY-PHO NET: An advanced social network extraction system”, Proc. WWW 2006, 2006.
[5] 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.
[6] A. Uyar, “Investigation of the Accuracy of Search Engine Hit Counts”, J. of Information Science, Vol.35, No.4, pp.469-480, 2009.
[7] Funahashi, T., Yamana, H.: Reliability verification of search engines' hit counts: How to select a reliable hit count for a query, Lecture Notes in Computer Science, 6385, pp.114--125, 2010.
[8] Satoh, K., Yamana, H.: Hit count reliability: how much can we trust hit counts?, Proc of APWeb’12, pp.751-758, 2012.
[9] F Keller, M Lapata, Using the web to obtain frequencies for unseen bigrams, Computational Linguistics, Vol.29, No.3, pp.459-484, 2003.
[10] B Lou.: Users Reference Guide, British National Corpus. British National Corpus Consortium, Oxford University Computing Services, Oxford, England, 1995.
[11] LDC Catalog, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T21,(2013.1.9 アクセス)
[12] 上田高徳，佐藤亘，鈴木大地，打田研二，森本浩介，秋岡明香，山名早人 : Producer-Consumer 型モジュールで構成された並列分散 Web クローラの開発，WebDB Forum 2012, 2012
[13] jawiki dump progress on 20121115, http://dumps.wikimedia.org/jawiki/20121115/, (2013.1.9 アクセス)
[14] Challenges in Building Large-Scale Information Retrieval Systems, http://research.google.com/people/jeff/WSDM09-keynote.pdf, (2013.1.9 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] Min-Hash LSH for Detecting Duplicate Documents, http://www.stanford.edu/~ashishg/amdm/handouts/scscrib-lec10.pdf, (2013.1.9 アクセス)
[17] Web ページの本文抽出， http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html, (2013.1.9 アクセス)
[18] lucene-gosen - Japanese analysis for Apache Lucene / Solr 3.6 and 4.0 , http://code.google.com/p/lucene-gosen/, (2013.1.9 アクセス)
[19] 検索:アップグレード版検索 API - Yahoo!デベロッパーネットワーク，http://developer.yahoo.co.jp/webapi/search/premium.html, (2013.1.9 アクセス)
[20] Page view statistics for Wikimedia projects, http://dumps.wikimedia.org/other/pagecounts-raw/, (2013.1.9 アクセス)

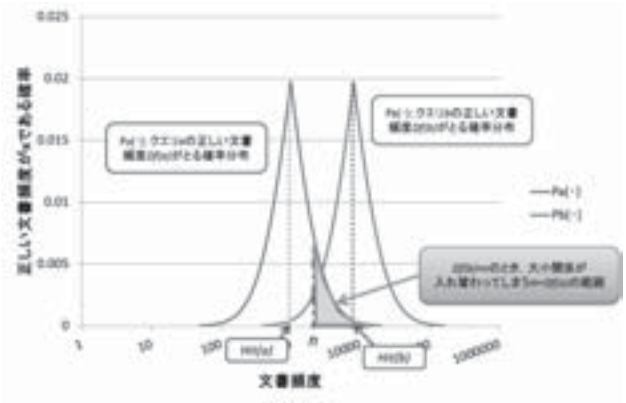


図 7. 2 クエリのヒット数を取得した際の正しい文書頻度の確率分布例

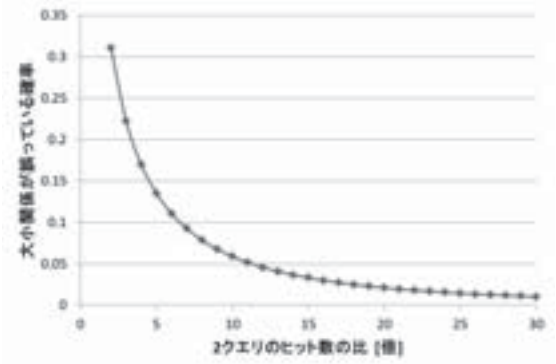


図 8. 2 クエリのヒット数比と大小関係が誤る確率の関係

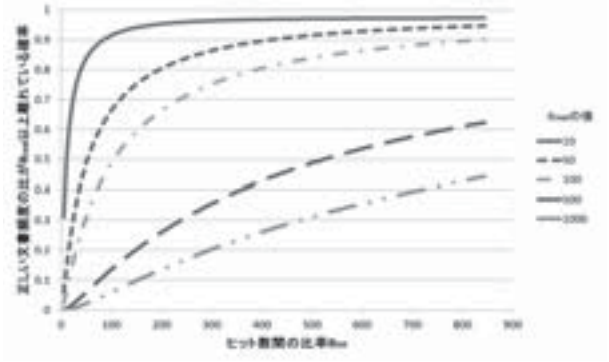


図 9. 2 クエリのヒット数比–正しい文書頻度の比が一定以上である確率の関係

佐藤 亘 Koh Satoh
早稲田大学大学院 基幹理工学研究科修士課程．検索エンジン，並列分散処理に関する研究に従事．DBSJ 学生会員．

上田 高徳 Takanori UEDA
2013 早稲田大学大学院基幹理工学研究科博士後期課程修了．博士（工学）．現在，日本アイ・ビー・エム株式会社 東京基礎研究所．IEEE, ACM, IEICE, IPSJ, DBSJ 各会員．

山名 早人 Hayato YAMANA
1993 早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．1993-2000 電子技術総合研究所．2000 早稲田大学理工学部助教授．2005 同大理工学術院教授，NII 客員教授．IEEE,ACM, AAAI, IEICE, IPSJ, DBSJ 各会員．