2012 年度 修士論文

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

提出日： 2013 年 2 月 1 日
指導： 山名 早人 教授

早稲田大学大学院 基幹理工学研究科 情報理工学専攻
学籍番号：5111B043-7

佐藤 亘

概　要

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．ヒット数は任意の単語・フレーズに対する Web 全体における単語の出現頻度を容易に取得する手段として有用であると考えられている．しかし，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など，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，ヒット数の正確性を評価し，ヒット数を利用する研究やアプリケーションに対する影響の度合いを明らかにすることには大きな意義がある．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ヒット数の変動傾向を特定することや，複数の検索エンジン間でのヒット数を比較することなどが目的であった．しかしヒット数の正確性を確実に評価するためには，大規模な Web データに基づいて正確な単語統計を取得した上でヒット数と比較を行う必要がある．そこで筆者らは，大規模に Web クローリングを行い，集められたデータにおける単語の出現頻度と，その単語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おこなった．本論文では，ヒット数の正確性に対する多角的な評価結果を示す．実験では，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結果，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%であることが判明した．

目　次

第1章　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．Web 上の膨大な情報の中から目的の情報を得たいとき，我々はしばしば検索エンジンを利用する．検索エンジンは，ユーザからの多様な要求に応えるために Web 上のコンテンツを網羅的に蓄えており，ユーザは 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, Google や Yahoo!や Bing など，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]-[7]．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][5]や，Web からの自動ソーシャルネットワーク抽出[7]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ページ番号によって大きく変動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 [8]-[11]，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 倍以下に変化することがしばしばあり，さまざま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正確性を評価し，ヒット数を利用する研究やアプリケーションに対する影響の度合いを明らかにすることには大きな意義がある．

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[8]-[13]．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[8]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[9]，検索エンジンが信頼できるヒット数を返す条件を特定した研究[11][12]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．しかし本研究で対象とするヒット数の正確性，すなわちヒット数が Web 上のクエリの出現頻度と比較してどれだけ正確であるかは，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が困難である．検索エンジン以外にコーパスを用意し，ヒット数とコーパスにおける単語頻度を比較した

研究もいくつか行われている[13]が，これらの研究には共通して比較実験が小規模であることや，ヒット数の時系列上の変動を考慮していないことなどの問題が挙げられる．

このようにヒット数の信頼性に関していくつかの研究が行われているが，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，ヒット数を研究に用いる場合の基盤となることを目的とし，大規模に 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 の単語統計ともっとも近似しているかを特定することができる．本論文では，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ともに，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．

本研究の意義は，検索エンジンのヒット数の正確性評価のみにとどまらず，Web から得られる単語統計の概略を明らかにするという意義も含まれる．例えば，多様なクエリに対して，その出現頻度が収束するにはどのくらい多くの Web ページを収集しなければならないのか，等を明らかにする．

本論文の構成は以下の通りである．まず第 2 章において本論文で使用する用語を定義し，第 3 章でヒット数の有用性と応用例を示す．第 4 章で関連研究についてまとめ，第 5 章で検索エンジンにおけるヒット数が誤差を生む原因を考察する．次いで第 6 章にて 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 クローリング方法と単語頻度のカウント手法について論じ，第 7 章にて取得された単語統計データの概略を示した上でクローリングデータにおける単語頻度と検索ヒット数とを多角的に比較する．

第2章　用語の定義

以下のように本論文で使用するいくつかの用語を定義する．

単語（単一語）：

一般的に使われる意味での単語を指す．本論文においては，形態素と同義である．

クエリ：

文書集合内あるいは検索エンジン内における出現頻度を観測するために用いる検索語であり，単語もしくは複数語を指す．

文書頻度：

文書集合 $\boldsymbol{D}$ に対するあるクエリの文書頻度とは，そのクエリを構成する単語すべてを形態素として含む文書 $d \in \boldsymbol{D}$ の個数である．本論文においては n 個の文書から成る文書群 $\boldsymbol{D}_n$ におけるクエリ q の文書頻度を $DF_n(q)$ と表記する．なお，クエリの出現確率とは，そのクエリの文書頻度を文書集合全体の文書数で割ったもの，即ち $DF_n(q)/n$ である．

ヒット数：

あるクエリに対するヒット数とは，そのクエリで検索エンジンに対して検索をかけたときに取得できる検索結果数の概算値を指す．複数語から成るクエリに対するヒット数は，本論文においては AND 検索，すなわちクエリを構成するすべての単語を含む検索結果に対するヒット数として扱う．あるクエリに対するヒット数は取得するタイミングや検索のオフセット[1]によって異なる．

[1] 検索オフセットとは，例えば検索結果ページにおいて「次へ」ボタンや検索結果ページ番号のリンクを押すことで指定することができる，ある検索結果ページの中で最も上位となるランキングの順位を指す．

第3章　ヒット数の有用性と応用例

本章では，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の有用性について論じる．3.1 においてヒット数を用いた研究やアプリケーションを紹介し，3.2 において単語統計の指標としてヒット数を用いることの利点をまとめる．

3.1　ヒット数を用いた研究・アプリケーション

本節では，検索エンジンから得られるヒット数を用いた研究について例を挙げ，ヒット数が様々な研究においてどのように用いられているかをまとめる．

3.1.1　ヒット数を機械翻訳支援に用いた研究

1999 年，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，ある語句 *A* を別の言語の語句に置き換えるとき，*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．

3.1.2　ヒット数を用いて同義語抽出を行なう研究

2001 年，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．Turney は，この手法をさらに発展させて，スコアの定義を

$$score(choice_i) = \frac{hits((problem\ NEAR\ choice_i)\ AND\ context\ AND\ NOT((problem\ OR\ choice_i)\ NEAR\ "not"))}{hits(choice_i\ AND\ context\ AND\ NOT(choice_i\ NEAR\ "not"))} \quad (2.2)$$

とすることによって，文脈を考慮に入れた同義語抽出が可能であるとしている．ここで検索クエリ“Q_1 NEAR Q_2”とは，単語 Q_1 と単語 Q_2 が近接している文書のみを取り出すクエリである．また，式中の *context* とは，*problem* が含まれる文書中の一語を選出したものである．

3.1.1 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

2007 年，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．

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (2.3)$$

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．

3.1.2 ヒット数を用いてソーシャルネットワーク抽出を行った研究

2006 年，松尾ら[7]は，検索エンジンから得られるヒット数を用いたソーシャルネットワーク抽出手法を提案した．この手法は，人物間のつながりの強さを求める際にヒット数を用いるものである．具体的には，ノードとして与えられた人名の集合から 2 つの人名 *X*, *Y* を取り出し，

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{\min(|X|,|Y|)} & if\ |X| > k\ and\ |Y| > k, \\ 0 & otherwise \end{cases} \quad (2.4)$$

を計算することによって人物 *X* と人物 *Y* の間のつながりの強さを算出する．ここで|X|

とはクエリ“X”に対するヒット数であり，$|X \cap Y|$とは，クエリ“X AND Y”に対するヒット数である．k は，X, Y それぞれに対するヒット数が小さすぎないことを保証するための定数である．この手法では，$R(X, Y)$の値が閾値以上であるならば，人名ノード間にエッジを張る，という手法を取る．また，松尾らは，$R(X, Y)$をエッジの長さの定義にも使用できるとしている．

3.2 ヒット数の有用性のまとめ

前節で述べたように，ヒット数は単語の出現頻度の取得手段として数多くの幅広い分野におけるアプリケーションに用いられている．以下，ヒット数が持つ特筆すべき特徴をまとめる．

コーパスの規模が膨大である：

Web 上に存在する文書はあらゆる分野・言語にまたがっている．検索エンジンは Web から網羅的に文書を取得していると考えられ，ヒット数は多くの分野・言語を網羅した膨大な規模のコーパスにおける出現頻度ととらえることができる．[1]は言語の網羅性を，[2]は分野の網羅性を利用したアプリケーションである．

コーパスが自動で更新される：

検索エンジンは，新しい情報を得たいという利用者の期待に応えるために，絶えず Web クローリングを行い，保持するデータとインデックスを最新の状態に保とうとしている．このためヒット数を利用すると，クエリに対して検索エンジンがインデックスしている最新の情報が反映された出現頻度を得ることができる．[3]に代表される Web からソーシャルネットワークを抽出する研究はヒット数が持つこの特徴を利用した例である．

即座に結果が取得できる：

通常，コーパスにおいてあるクエリの出現頻度を取得しようとした際には，クエリが発行されてからコーパスを走査してその出現回数をカウントするため，クエリが発行されてから結果を得るまでに時間がかかる．これに対して検索エンジンは，蓄積されたデータに対してインデックスを作成しており，ユーザから与えられたクエリに対して即座にヒット数を算出し，ユーザに返すことができる構成を取っている．前節で挙げたアプリケーションは全てヒット数が持つこの特徴を利用し，高い応答性を実現している．

幅広いクエリに対応している：

[3]のようなアプリケーションでは，個人名などの固有名詞をクエリとしている．このように，出現頻度の低い単語や，社会のトレンドなどによって新しく普及した単語な

どの出現頻度を取得が求められるアプリケーションは数多い．検索エンジンは新語や固有名詞などを含めた幅広いクエリに対してヒット数を返すことができ，ヒット数の利用者の多様な要求に応える貴重な情報資源といえる．

第4章　関連研究

前章にて述べたように，検索エンジンから得られるヒット数は Web におけるクエリの出現頻度の取得方法として有用性が高く，これまで幅広い研究に応用されてきた．しかし，ヒット数は様々な条件によって変動することが知られており，近年その信頼性が問題視されてきている．本章では関連研究として，ヒット数の信頼性を論じた研究をまとめる．まず 4. 1 において検索エンジンから得られるデータのみを用いてヒット数の信頼性を議論した研究を紹介し，次いで 4. 2 において他のコーパスにおける単語頻度とヒット数とを比較することによってヒット数の正確性を調査した研究をまとめる．

4. 1　検索エンジンから得られるデータを用いてヒット数の信頼性を議論した研究

4. 1. 1　複数の検索エンジン間でのヒット数を比較した研究

2008 年，Thewall[8]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 はさまざまなヒット数をとる 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．また，Yahoo!は Google と比べて，URL，サイト，ドメインなどの観点から見て，わずかに多様な検索結果を返しているという結果を得た．

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．

4. 1. 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

2009 年，Uyar[9]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 ページ数が正しいヒット数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．

$$Percentage\ of\ Error = \frac{Estimate - \mathrm{Re}turnedDocument}{\mathrm{Re}turnedDocument} \times 100 \qquad (2.5)$$

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%以下となるヒット数は，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．しかしこの手法では，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．5. 1. 2 にて論じるように，検索エンジンは高速化のためにインデックスを削減しており[17], 静的ランキングで下位となる Web ページはインデックスに含まれていない可能性がある．そのため，「取得した Web ページ数が 1000 件以下の場合，取得した Web ページ数がヒット数の正解値である」という前提も定かではない．実際，4. 1. 4 にて紹介する文献[10]において, 取得した Web ページ数が必ずしも信頼できる値ではないことが示されている．

4. 1. 3 正確なヒット数の算出を試みた研究

2007 年，松尾ら[14]はある単語に対する正確なヒット数を得たいときに，別の多くの単語群との共起ヒット数（AND 検索におけるヒット数）を利用し，統計的な処理を行うことによって正確なヒット数を推定する手法を提案した．この手法は，検索エンジンから得られるヒット数が k 件より少ない場合は値が正確であるという前提のもとに行われている．k とは検索エンジンが返す検索結果の最大数で，Google, Yahoo!, MSN などの主要な検索エンジンは k=1000 である．

松尾らは，ヒット数が k 件を超える検索クエリ a に対して正確なヒット数 n_a を推定するとき，次のステップを踏むことによって正確なヒット数を得ることができるとしている．

学習フェーズ

step1. 検索ヒット数を k 以下に抑えるような適切な語（以後，プローブ語と呼ぶ）x をクエリに加え，"x AND a"というクエリで絞り込み検索を行い，ヒット件数を得る．このときに得られたクエリ"x AND a"に対するヒット数を $n(x, a)$と表記する．

step2. 絞り込みに使う語 x が Web ドキュメントに含まれる確率 $p(x)$を求める．

step3. 語 a と語 x の Web 上での出現が独立同分布であると仮定すると，語 a の正確な検索ヒット件数 n_a は次式によって算出される

$$n_A = \frac{n(X, A)}{p(X)} \tag{2. 6}$$

松尾らは，この手法について，

- 確率 $p(x)$を計算するのが困難である
- 語 a とプローブ語 x の Web 上での出現が独立同分布であるという保証がない

という 2 つの問題を設定した．これらを解決するために，次に示す学習フェーズを踏むことによって適切なプローブ語 x と，$p(x)$の逆数の近似値 m_x（拡大率）を特定できるとしている．

学習フェーズ

step1. トレーニング語集合 $\boldsymbol{A}_{train}$ とプローブ語候補の集合 $\boldsymbol{X}_{cand}$ を用意する．

step2. トレーニング語 $a' \in \boldsymbol{A}_{train}$ に対するヒット件数 $n_{a'}$と，プローブ語候補 $x \in \boldsymbol{X}_{cand}$ との共起ヒット件数 n(x, a') を得る．

step3. プローブ語候補 x の拡大率 m_x，分散 σ^2_x，およびプローブ語選択の指標値 s_x を求める．ここで，拡大率 m_x とは各トレーニング語 $a' \in A_{train}$ に対する $n_{a'} / \mathrm{n}(x, a')$の平均値を指す．また指標値 s_x とは，拡大率の分散，χ^2 値，Kullback-Leibler 係数，コサイン距離の 4 つの値を用いたプローブ候補語 x に対する適切性評価値である．

step4. 拡大率の閾値 m_{thre} 以上の拡大率 m_x を持つプローブ語候補 x を指標値 s_x の小さいものから選び，プローブ語 x_{probe} とする．

step5. （回帰係数による統合の場合）$x \in \boldsymbol{X}_{probe}$ と a'の共起ヒット件数 n(x, a')を基に $n_{a'}$の線形回帰を行い，プローブ語 x の係数 k_x を計算する．

導出されたプローブ語 x，拡大率 m_x，分散 σ^2_x を用いて，正確なヒット数は次のように推定する．

推定フェーズ

step1. クエリ a を決める．

step2. プローブ語 $x \in \boldsymbol{X}_{probe}$ と語 a の共起ヒット数 n(x, a)を検索エンジンから得る．

step3. （単一係数による統合の場合）以下の式を計算する．

$$\hat{n}_a = \frac{\sum_{x \in x_{probe}} \left(m_x \times n(x,a) / \sigma^2{}_x \right)}{\sum_{x \in x_{probe}} k_x \times n(x,a)} \qquad (2.7)$$

step4. （回帰係数による統合の場合）以下の式を計算する．

$$\hat{n}_A = \sum_{x \in x_{probe}} k_x \times n(x,a) \qquad (2.8)$$

松尾らの研究は，様々なクエリとの共起ヒット数を利用することで，単一のクエリに対するヒット数の揺れを解消しようと試みたものである．しかし松尾らの研究では複数のクエリのヒット数を数日にわたって観測したとき，各クエリのヒット数が独立に大きく変動する現象が見られることを考慮していない．例えば，学習フェーズにおいてトレーニング語 a'に対するヒット数が極端に変化してしまうと，プローブ語 x に対する拡大率に大きな誤差を与える．さらにこの研究は，ヒット数が k 件以下の場合のヒット数は信頼できるという前提をおいているが，4. 1. 2 でも述べたとおり検索エンジンは検索の高速化のためインデックス削減を行っていると考えられるため，この前提が確実に正しいということはできない．実際 4. 1. 4 で示すように，値が 1,000 未満のヒット数でも信頼できるとはいえないという検証結果がある[10]．さらに図 1 に見られるように，ヒット数が 1000 件以下の場合も検索するタイミングによって大きく値が異なる現象が見られる．このような変動の大きい時期に得られたヒット数推定値は著しく不正確であるといえる．

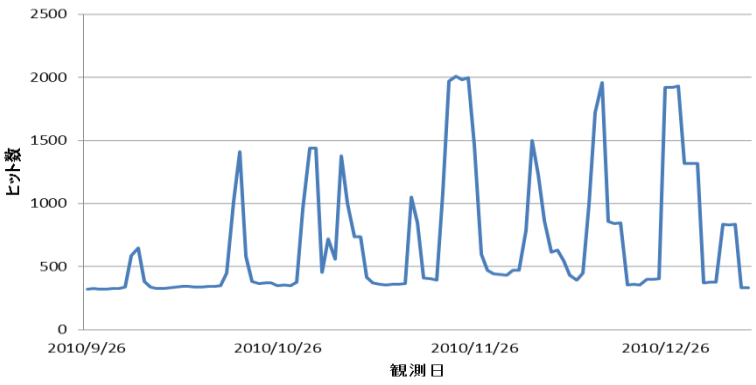


図 1. クエリ” Ochiai-minami-nagasaki Station”に対する Google のヒット数

4.1.4 ヒット数の変動幅を検証した研究

舟橋ら[10]は 500 個のクエリ群に対し，Google, Live Search, Yahoo! Japan, Yahoo!の 4 つの検索エンジンから得られるヒット数を収集した．このとき，各クエリに対してオフセットを 10 ずつ変化させながら複数回ヒット数を収集し，

(1) 検索オフセットが 1 のときのヒット数 h_f と検索オフセットが最大のときのヒット数 h_l の変動幅 $r_d = |h_1-h_f|/h_1$

(2) 1 クエリに対する一連の検索で実際に取得できた検索結果数 h_r が 1,000 件未満のときに，h_r と h_l の変動幅 $r_g = h_l/h_r$

を検証した．

その結果，(1)に関しては，r_d の平均で 4.7%(Live Search)～63.7%(Yahoo!)の変動があることや，Google，Yahoo!に対する r_d の分散がそれぞれ 0.609, 1.761 と大きな値になっていることがわかった．また(2)に関して，Google, Live Search, Yahoo! Japan では 5～6 割のクエリにおいて，h_l がと h_r 比べて 100 倍以上の値をとることがわかった．

これらの検証結果から，検索エンジンが返すヒット数は，その値が 1,000 件未満の場合であっても一概に信頼できるとはいえないこと，さらに実際に取得できた検索結果数が 1,000 件未満であってもその値が正しいともいえないことが読み取れる．

この研究はヒット数の信頼性に関するいくつかの先行研究が前提とする命題(「実際に取得できた検索結果数が 1,000 件未満であってもその値は正しい[9]」「1,000 件未満のヒット数は信頼できる[14]」)を覆したという点で意義のあるものだが，結論としてどのヒット数が最も信頼できるかという点は明確に述べられていない．

4.1.5 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[11]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．

まず，舟橋らはヒット数の変動が起こるきっかけが，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

舟橋らは, 頻繁に検索される 10,000 件のクエリについて 2 ヶ月の間ヒット数を観測した．

その上で，先述した 3 つのケースについて，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．その結果，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

舟橋らの研究では，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，ヒット数が 1 週間以上にわたって安定しているヒット数が，Web 上に存在する網羅的な文書におけるクエリの出現頻度と比べて正確であるという根拠を示していない．

4.2 他のコーパスを用いてヒット数の正確性を調査した研究

Keller ら[13]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[19], NANTC[20])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[2]．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．

4.3 関連研究のまとめ

ヒット数は幅広い研究に応用されているが，その一方で，ヒット数は様々な条件によって値が変動する現象が知られている．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[8]-[13]が，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[8]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[9]，検索エンジンが信頼できるヒット数を返す条件を特定した研究[11][12]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．しかし本研究で対象とするヒット数の正確性，すなわちヒット数が Web 上のクエリの出現頻度と比較してどれだけ正確であるかは，検索エンジンから得られる情報のみを用い

て計量することが困難である．たとえば同一クエリに対してヒット数が長期にわたって安定している場合そのヒット数は信頼できるとした研究が存在する[11][12]が，安定したヒット数が Web 上のクエリの出現頻度と一致性が高いかどうかは検索エンジンから得られる情報のみからはわからない．検索エンジン以外にコーパスを用意し，ヒット数とコーパスにおける単語頻度を比較した研究もいくつか行われている[13]が，これらの研究には共通して比較実験が小規模であることや，ヒット数の時系列上の変動を考慮していないことなどの問題が挙げられる．

以上のヒット数の信頼性に関連する研究を表 1 にまとめる．

ヒット数の信頼性に関連するこれらの研究の流れを受けて，本研究では，大規模に 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．

表 1. ヒット数の信頼性を対象とした研究のまとめ

	研究目的	問題点/本論文との差異
Thelwall の研究[8]	検索エンジン間でヒット数と検索結果の比較を行う	・ヒット数の正確性について論じられた研究ではない
Uyar の研究[9]	検索エンジン間でヒット数の正確性を比較する	・得られた文書数が 1000 件場合のみしか評価ができない ・前提となっている「実際に取得できた数=正確なヒット数」が信頼できるのかどうか疑問が残る
松尾らの研究[14]	統計的な処理によって正確なヒット数を算出する	・各クエリのヒット数が独立に大きく変動することを考慮していない ・前提となっている「値が k 件以下のヒット数は正確である」が信頼できるのかどうか疑問が残る
舟橋らの研究[10][11]	検索オフセットを変化させた時のヒット数の変動幅を検証する	・結論としてどの値が信頼できるヒット数であるかが明確に述べられていない
	検索エンジンが信頼できるヒット数を返す条件を特定する	・手法を適用して得られたヒット数に対する評価を行っていない ・得られたヒット数が信頼できるか否かを確かめるために 1 週間ヒット数を観測しないといけない
Keller らの研究[13]	ヒット数と固定的なコーパスに対する単語の出現頻度とを比較する	・比較に使用したコーパス・クエリ群が小規模である ・ヒット数の時系列上の変動を考慮していない

第5章　ヒット数の誤差原因に対する考察

本章では，ヒット数が Web 上の単語出現頻度と比べて誤差を生む原因を考察する．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，本章では一般的な検索エンジンの構成からヒット数概算方法や誤差の原因を考察する．5. 1 では検索エンジンの基本構成・高速化手法について述べる．次に 5. 2 においてヒット数が誤差を生む原因について論じる．

5. 1　ヒット数の概算方法に対する考察

5. 1. 1　検索エンジンの基本構成

Arasu ら[15]が調査報告を行った一般的な検索エンジンの基本構成を図 2 に示す．大きく分けて，Web 上から大量のデータを集めて蓄積し，データを整理してインデックスを作成するバックエンドと，ユーザからのクエリを受け付けてインデックスを走査し，検索結果を返すフロントエンドとが存在する．

バックエンドにおける処理：

バックエンドにおける処理は，クローリングとインデキシングに大別される．クローリングは，シードとして与えられた Web ページ群から，各ページが持つリンクを順次たどっていくことで Web 上の文書を網羅的に収集する処理である．またインデキシングとは，収集されたデータに対して効率よく要求したデータを取り出すことができるよう整理する処理である．より具体的には，与えられた単語に対してその単語を含む Web ページを紐付けた索引（転置インデックス：以下単に“インデックス”と表記する）を作成する．

フロントエンドにおける処理：

検索エンジンは，世界中からの検索要求に応えるため，世界中に検索ユニットを点在させている[21][23]．個々の検索ユニットは，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[21]．検索エンジンの利用者が Web ブラウザから検索要求を行うと，DNS サーバが名前解決を行う．この際，DNS サーバはユーザと検索ユニットとの距離，各検索ユニットのトラフィック状況等を考慮し，最も適切な検索ユニットの IP アドレスを返す．これによってユーザは最も応答時間が短いと判断された検索ユニットに接続する．次に，検索クエリは図 2 における Query Engine において，lemmatize，stemming などといった語幹処理やクエリ拡張など，クエリに対する前処理が行われる．次に検索エンジンは，処理済みのクエリを用いてインデッ

クスを走査し，クエリを含む Web ページのリストを取得する．最後に，取得したページのリストをランキングし，上位となったページ群を利用者に検索結果として提示する．

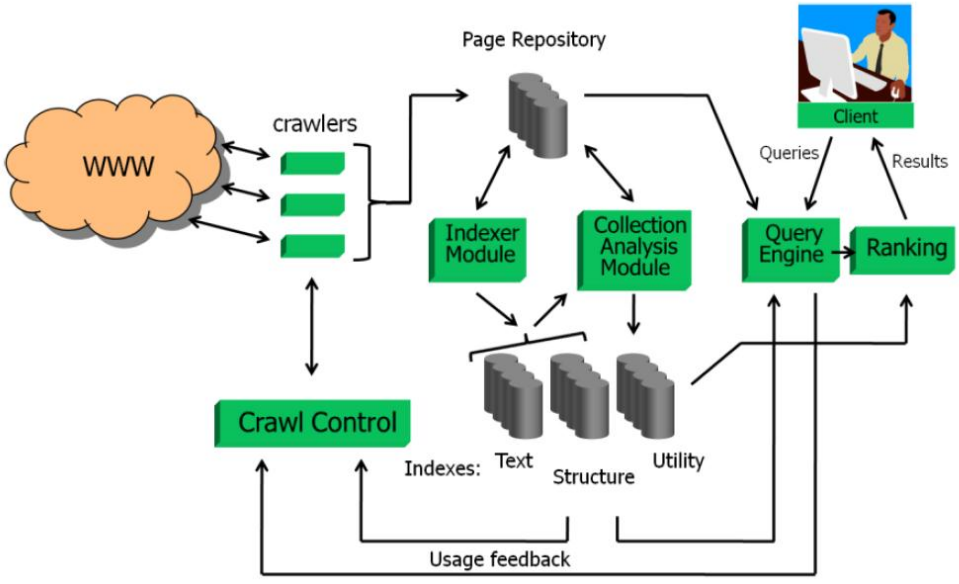


図 2. 検索エンジンの基本構成([15]より引用)

5. 1. 2 検索処理の高速化手法

Web 上を流通するコンテンツは日々増加しており，検索エンジンが扱うデータもそれに応じて膨大になってきている．さらに検索エンジンの利用者も増えてきたことから，検索が行われる度にインデックス全てを走査したのでは処理が追いつかなくなっていきている．そこで，検索処理の高速化手法が数多く提案されている．

Jansen ら[16]は，ほとんどの検索エンジンの利用者が検索結果に対して上位 20 件までしか閲覧しないという報告をしている．そこで，クエリに対して検索結果をランキングし，ある一定数のページを取得した時点で検索処理を打ち切る Early Termination と呼ばれる高速化手法が提案されている．

また，Ntoulas ら[17]や Skobeltsyn ら[18]は，よく参照される検索結果をキャッシュとして保持したり(Result Cache), ページランク等のクエリに依らない静的な重要度に基づいてインデックスを圧縮したりする(Pruned Index)など，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，インデックスの走査量を減らし，高速化を図る手法を提案している．

5.2　ヒット数が誤差を生む原因の考察

5.2.1 検索の前処理による誤差

検索エンジンはユーザからクエリを受け取ると,検索処理の前段階として図2中のQuery Engineにおいて lemmatize やクエリ拡張，スペルミスチェックなど，クエリに対する処理を行う．クエリ拡張やスペルミスチェックは検索エンジンの本来の用途である情報検索を行う際には有用である場合が多い一方で，単語の出現頻度の指標としてヒット数を用いる際には意図しない文書も結果に含めてしまうなどの理由から望ましくない場合があると考えられる．なお，Google や Yahoo! Japan 等，代表的な検索エンジンは，クエリやクエリを構成する単語を二重引用符（“”）で囲って検索をかけることにより，入力したクエリに対して前処理を行わせず検索を行うことができる完全一致検索機能を有している．

5.2.2 クエリのタイプによる概算方法の違い

検索エンジンの構成や高速化手法を考慮すると，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．以下それぞれについて詳しく述べる．

クエリを構成する単語数：

一般的に転置インデックスは，ある単語に対してその単語が紐付いている文書のリストとそのリスト長を保持している．単一語のクエリに対しては，あらかじめ保持しているリスト長をヒット数として返すことができるが，複数語のクエリに対してはサンプリングに基づいたより複雑な処理が必要である．このような処理の違いによって，ヒット数の正確性に差異が出る可能性が考えられる．

クエリの出現頻度：

Early Termination や Pruned Index のような検索処理の高速化手法を考えたときに，検索エンジンが検索結果を返すまでアクセスするデータの範囲は，クエリの出現頻度に依存すると考えられる．例えば，極度に出現頻度が小さいクエリは，Pruned Index に含まれないため，より大きなインデックスへ走査が及ぶ可能性が考えられる．このように，扱うデータの幅の違いによってヒット数の正確性に差が現れることが考えられる．

クエリの言語：

5.1.1，5.2.1 で述べた語幹処理をはじめとする検索の前処理は，言語によって大きく異なる．このためクエリの言語の違いがヒット数の概算に影響を与える可能性が考え

られる．

クエリのトレンド性：

近年の検索エンジンは，日単位で更新を行う大容量のインデックスとは別に，ニュース検索やリアルタイム検索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実現するために秒単位で頻繁に更新を行う小容量のインデックスを保持している[21][22]．トレンド性の高いクエリに対しては，上記のような高頻度に更新が行われるインデックスを利用して検索結果が生成されると推察できる．

5.2.3 検索オフセットの変化による変動の原因

舟橋ら[11]が示すように，ヒット数は検索オフセットを変えて検索を行うと大きく変動する場合がある．5.1.2 で述べたとおり，検索エンジンでは検索処理の際，検索結果の上位のみを高速に提示するためにインデックスを走査する量に制限を設けている．このとき，検索オフセットが 0 の場合の（すなわち最初の検索結果ページにおける）ヒット数は，その検索結果ページを提示するために走査された限られた量の Web ページの情報から算出されていると考えられる．これに対し，検索を行うユーザが「次へ」ボタンを押すなどして検索オフセットを上げていった場合，検索エンジンはより深くインデックスを走査する必要があり，ヒット数を算出するための Web ページ数が高まると考えられる．このように検索オフセットを変えて検索を行うと，ヒット数を算出するために用いる情報量が増すことによって，提示されるヒット数が変動するのだと舟橋らは述べている．

5.2.4 ヒット数の時系列上の変動の原因

図 3 に示されるように，検索エンジンから得られるヒット数を時系列で観測すると値が変動している時期が存在することがわかる．検索エンジンの構成を考えたとき，時系列上でヒット数が変動する原因として次のような項目が挙げられる．

インデックス更新によるヒット数変動：

検索エンジンは，ユーザからの検索要求に対して新しい情報を提示できるように，絶えず Web クローリングを行い，インデックスを更新している．舟橋ら[11]は検索エンジンのヒット数を長期間にわたって観測し，検索エンジンが小規模にインデックス更新を行う期間（安定期）と，大規模に更新を行う期間（変動期）の２つの変動傾向を観測したと報告している．この現象は，検索エンジンが頻繁に更新を行う小規模のインデックスと，比較的更新頻度の低い大規模なインデックスとを保持しているという検索エンジン側が発表している事項と合致している[21][22]．このように，検索エンジン内におけるインデックスの更新は取得できるヒット数に影響を与えると考えら

れる．

キャッシュヒット/キャッシュミスによるヒット数変動：

5. 1. 2 で述べたとおり検索エンジンは処理の高速化のために，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 Result Cache や Pruned Index のような部分的なインデックスとして保持していると考えられる．このとき，検索エンジンのユーザが検索をかけた際にこれらの部分的なインデックスにヒットするか否かや，キャッシュが更新されるタイミング等により，取得できるヒット数に差異があると考えられる．

検索時に異なる検索ユニットに接続した場合の変動：

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検索ユニットを世界各地に配置している．各検索ユニット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間で異なる時期が存在すると考えられる．5. 1. 1 で述べたとおり接続する検索ユニットは，その時点でのトラフィック状況等に依存する．接続する検索ユニットが変化した場合，かつ検索ユニット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．

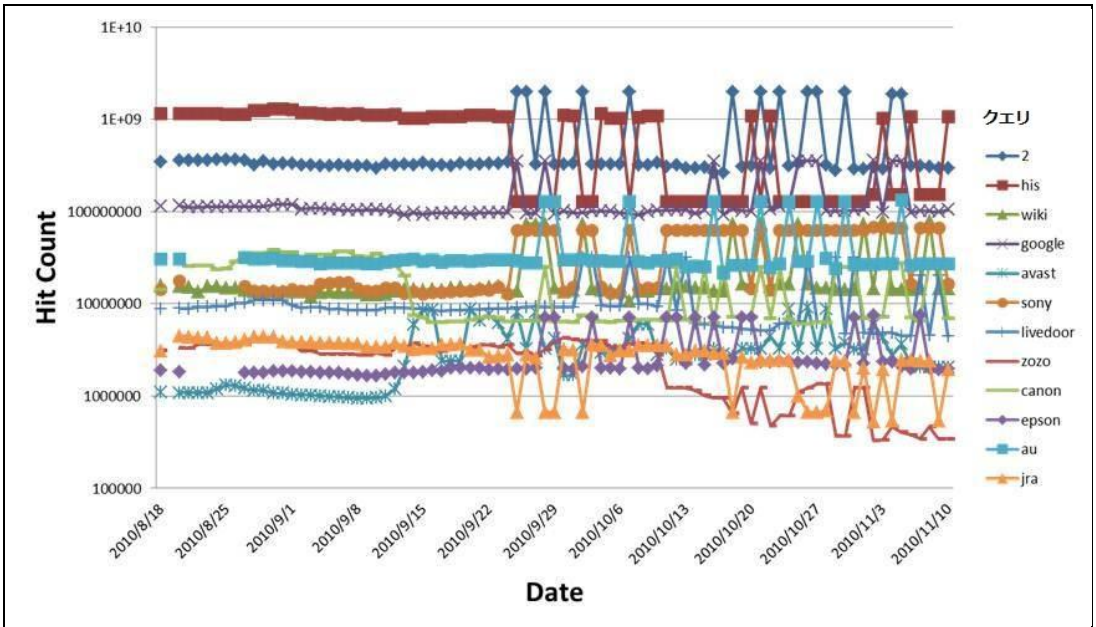


図 3. Google におけるヒット数推移の例

第6章　Webにおける単語統計取得手法

本章では 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 6. 1 において Web クローリング方法の概要を述べ，次に 6. 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．

6. 1　Web クローリング方法

本研究では，[24]にて提案されたクローラを用いる．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．

6. 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[25]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（付録 I 参照）．

6. 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おこなう．このため，日本語，英語以外の言語で書かれた Web ページを可能な限り収集しないようにすることを目的とし，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おこなった．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．

6. 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．すなわちクロール済み文書数が n である時点におけるクロール済文書群を $\boldsymbol{D}_n$ としたとき，$\exists q \in \boldsymbol{Q}$ に対して出現確率 $DF_n(q)/n$ を監視する．ここで，$DF_n(q)$とは $\boldsymbol{D}_n$ におけるクエリ q に対する文書頻度である（第 2 章参照）．なお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 6. 2. 2 で述べる．
2. $\boldsymbol{Q}$ のうち，ある閾値 $\theta (0< \theta <1)$に対して $\theta \cdot |\boldsymbol{Q}|$個以上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．

6.2 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．

6.2.1 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，5. 2. 1 で述べた 4 つの観点（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 6,300 件選出したもの．まず Wikipedia のタイトルを 50,000 件ランダムに抽出し，その中で単語数・出現頻度・言語が一様に分布するように 5,900 件のクエリを選んだ[2]．次に Wikipedia のページアクセス数を元にトレンド語を 400 件選出し，クエリ群に加えた．詳細を表 2 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おこなうクエリであり，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．

表 2. 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[29]で 10^3~10^7 の値をとるもの[3]
言語	日本語，英語
トレンド性	Wikipedia のページアクセス数[4][31]で 上位 1,000 件に入るものとそうでないもの

6.2.2 文書頻度の取得方法

文書頻度取得のフローを図 4 にまとめる．図のように，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文書頻度データセットを取得する．

[2] 合計で 6,000 件のクエリを選出する予定であったが，50,000 件のクエリ候補のうち [日本語/2, 3 語/10^6,10^7]を満たすクエリが計 100 件足りなかったため 5,900 件となった
[3] 2012/10/29 – 2012/11/4 の間で検索オフセットを 0 にして取得したヒット数の中央値を採用
[4] 2012 年 9 月のランキングから 200 件，同年 11 月のランキングから 200 件選出

複数の文書頻度を取得することで,

(i) ヒット数がどの文書頻度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

(ii) 6. 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 上の文書頻度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.

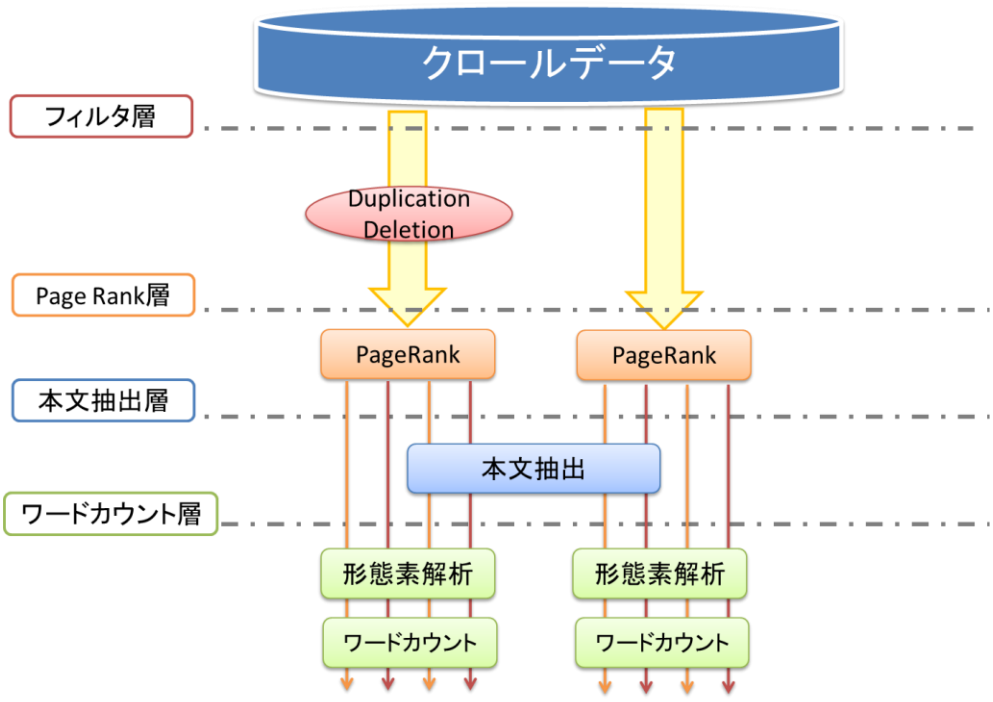


図 4. 文書頻度取得のフロー

以下，フロー中の各処理についてまとめる.

フィルタ層：

コンテンツの重複削除[26]を行い，重複ページによる単語統計への影響を排除する. “コピペサイト”に代表されるような重複を含むコンテンツは，その作成者が自分の運営するサイトへのアクセスを稼ぐことや他のサイトに誘導することを目的として作られたページであり，多くの場合そのページ自体に実質的な内容を含まない. したがって単語統計取得においては，重複ページは文書頻度のカウントに含めるべきではないと考えられる. 本実験ではこのような不正なページに対してフィルタをかけた場合とかけない場合とでヒット数との類似性に差が出るかを比較するためにこの処理を設けた.

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. 5. 1. 2 で述べたとおり，検索エンジンは通常，静的な重要度において上位の文

書群のみに基づいて検索結果を作成しており，ヒット数の算出もこれと同様の限られた文書群のみを用いて概算を行なっていると考えられる．本実験では静的な重要度としてページランクに着目し，ランクが高い文書のみを抽出して文書頻度を取得した際と，全ての文書を使用して文書頻度を取得した際とでヒット数との類似性に差が出るかを比較するためにこの処理を設けた．

本文抽出層：

Web ページ本文抽出[27]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．特にブログサービスを利用した Web ページに顕著な傾向として，同一サイト内の複数のページ間でサイドバーやヘッダー・フッター等に含まれる文章が共通しているページが数多く見られる．本文抽出を行わずページ全体に対して単純にカウントを行うと，サイドバー等の共通部分に含まれる単語の頻度が不正に高まる可能性が考えられるため，この処理を設けた．

ワードカウント層：

形態素解析 lucene-gosen[28]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．このとき同一文書中に複数回同一クエリが出現した場合も 1 とカウントする．一般的に，lucene-gosen をはじめとする形態素解析器は，“東京スカイツリー”などの未知の複合語等の入力に対して“東京”，“スカイ”，“ツリー”の 3 形態素に分けるなど，できるだけ細かい形態素に分割する傾向がある．そこで事前に，クエリを構成する単語を全て形態素解析器の辞書に登録しておく．これによって，形態素解析器が誤ってクエリを構成する単語をさらに細かい形態素へ分割してしまうことを防止する．

結果として，表 3 に示す 8 種類の文書頻度データを得る．表中の○は該当モジュールの処理を施す場合，×は施さない場合を意味する．

表 3. 取得する文書頻度データの種類

データラベル	モジュール		
	フィルタ	ページランク	本文抽出
1	×	×	×
2	○	×	×
3	×	○	×
4	○	○	×
5	×	×	○
6	○	×	○
7	×	○	○
8	○	○	○

なお，いずれのフローにおいても取得した文書に含まれる HTML タグやスクリプトはすべて排除している．

6. 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， 6. 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，得られた文書頻度が Web 全体の文書頻度と十分近似しているという妥当性に対する強い裏付けが必要である．そこで取得した文書頻度の妥当性を次の 2 つの観点から裏付ける．

クロール時における出現確率の収束性

6. 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[12]にならい，“it”や“an”など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固定的なコーパス[19]とを比較して十分相関が高いことを確認する．

第7章　ヒット数の正確性評価

本章では，前章で述べた Web における単語統計取得手法に基づき取得した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おこなった結果を示す.

7.1　使用したデータの概略

7.1.1 クロールデータ

Wikipedia 外部リンク集[25]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～同月 16 日の間に収集したデータを用いる．クロールの終了条件を，「クロール済みの文書数 10%の増加に対し，$\{Q_1, Q_2\}$に含まれる 95%以上のクエリに対する出現確率の変化が 10%以内に収まったとき」と定めた.

クロールされたデータの規模に関する基本的な情報を表 4 に示す

表 4. クロールデータ規模

収集ページ数	41,818,191 pages
データサイズ	1.26 TB
サイト数	3,232,525 hosts

7.1.2 ヒット数データ

Yahoo! Japan が提供する Web 検索 API[29]を用い，6. 2. 1 で述べたクエリ群に対しヒット数を収集した．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[5]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．ただし Q_2 については，前処理の影響を検証するために二重引用符で囲わない場合のヒット数も収集した.

なお，ヒット数は各クエリにつき検索オフセットを 1～400 の間で 100 刻みで変化させて収集した．検索オフセットの上限を 400 としたのは，オフセットを 500 以上としたとき，多くの場合使用した検索 API が検索結果を返さなかったためである.

[5] 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

7.2　ヒット数の正確性評価指標

7.2.1 ピアソンの積率相関係数とケンドールの順位相関係数

本研究では，ヒット数の正確性評価指標として，クロールデータに対する文書頻度と比較したときの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た．どちらの指標も，文書頻度との相関が高いほど，ヒット数の正確性が高いことを示すものである．以下，各々について定義式と特徴を述べる．

ピアソンの積率相関係数：

データ $\{(x_i, y_i)\}^n_{i=0}$ に対して次の式で定義され，$-1 \leqq r_p \leqq 1$ の値をとる．

$$r_p = \frac{\sum_{i-0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

ここで，$\bar{x}, \bar{y}$ はそれぞれ x, y の母集団の標本平均を指す．本研究においては，$q \in \boldsymbol{Q}$ に対して $DF(q), Hit(q)$ がそれぞれ x_i, y_i に相当する．

ピアソンの積率相関係数は2次元データの相関分析において最も一般的な指標であり，関連研究[13]においてもヒット数とコーパスにおける単語出現頻度との比較に用いられているが，x, y が正規分布に従うことを仮定している点や，外れ値の影響を強く受けるという点が欠点として挙げられる．

ケンドールの順位相関係数：

次の式で定義され，$-1 \leqq r_p \leqq 1$ の値をとる．

$$r_k = \frac{4P}{n(n-1)} - 1$$

ここで，n はデータの個数，P は 2 つのデータの順位を考えたときに大小関係が一致している組の個数である．

ピアソンの積率相関係数とは異なり，分析対象のデータにどのような分布も仮定しないノンパラメトリックな指標である．また，ピアソンの相関係数と比べて外れ値に対して頑健であるという性質を持っている．

7.2.2 ヒット数と文書頻度の比の標準偏差

以下の議論から，ヒット数と文書頻度の比を各単語に対して得たときに，その標準偏差を正確性評価指標として用いることができると考えた.

いま，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，クロールされた文書数 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

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

となり，

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

を得る．この k は理想的には $q \in \boldsymbol{Q}$ に依らず一定であることが期待される．したがって，各 $q \in \boldsymbol{Q}$ につき

$$k(q) = Hit(q) / DF(q)$$

を計算したとき，$k(q)$の標準偏差が小さいほどヒット数の正確性が高いと見なすことができる.

7.3 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.

7.3.1 文書頻度の分布

選出したクエリのクロールデータにおける文書頻度の分布は図 5 のようになった. $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$ 双方ともに，全く出現しなかったクエリが全体の 3%前後見られるが，その他に関しては幅広い出現確率をとるクエリ群を選出できていることがわかる.

$\boldsymbol{Q_1}$ は事前にヒット数を観測し，ばらつきがでるように選出したため $\boldsymbol{Q_2}$ と比べて分散の高い分布をとっている．また，$\boldsymbol{Q_2}$ は Yahoo! Japan における人気クエリから選出したため，文書頻度の平均が $\boldsymbol{Q_1}$ より高い分布を示している.

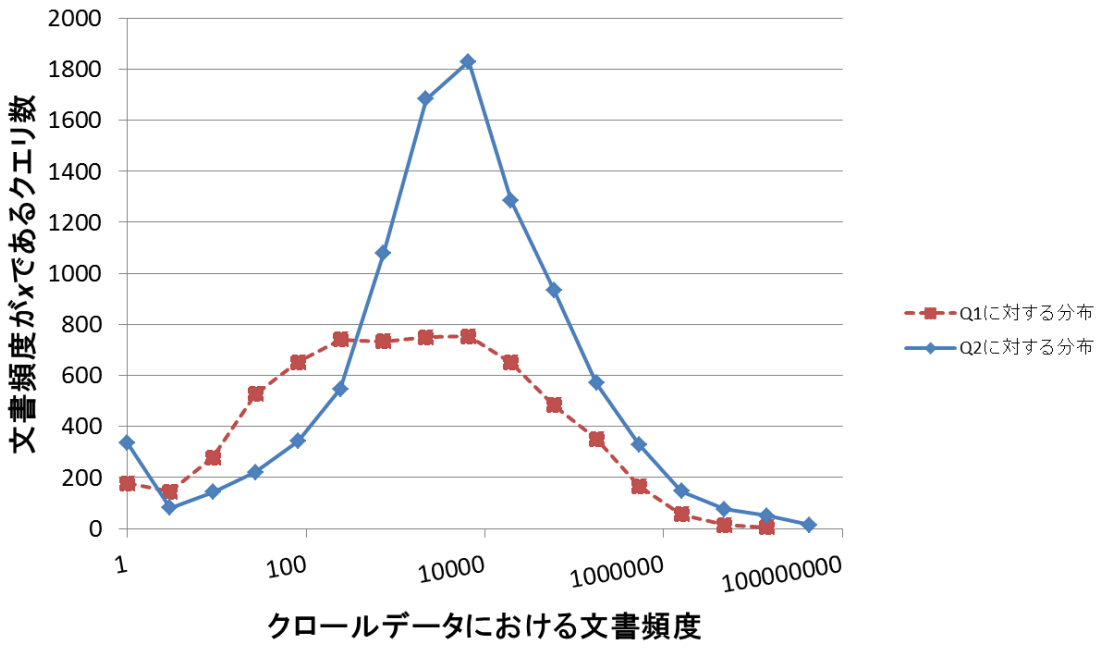


図 5. クロールデータおける Q_1, Q_2 の文書頻度の分布

7. 3. 2　出現確率の推移

図 6 は Q_1 からランダムに選出したクエリについて，カウント済み文書数と出現確率との関係を示したものである．すなわち，クロールの最中に得られる各クエリの出現確率の推移を表している．

傾向として, 例えば縦の点線で示したカウント済み文書数 n=200 万付近の時点での出現確率を見ればわかるように，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より多くの文書を収集しなければ収束しないという特徴が見受けられる．実際，今回使用した全てのクエリの出現確率に対する収束の程度は，クローリングの終了条件においても述べたとおり「カウント済みの文書数 10%の増加に対し， 95%以上のクエリに対する出現確率の変化が 10%以内に収まる」というものであったが，出現頻度が e^{-14} 以上のクエリ（これは使用したクエリをクロール終了時における出現頻度でソートした際に上位約半分に相当する） のみを抽出して収束性を調べると，「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．

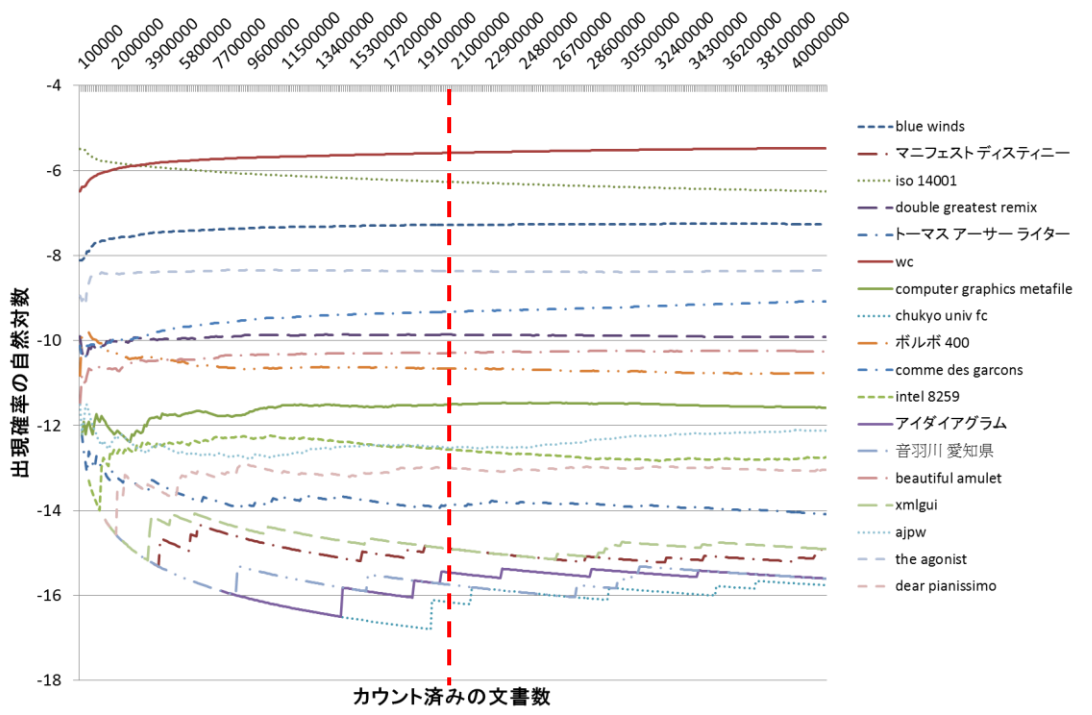


図 6. クエリの出現確率の推移

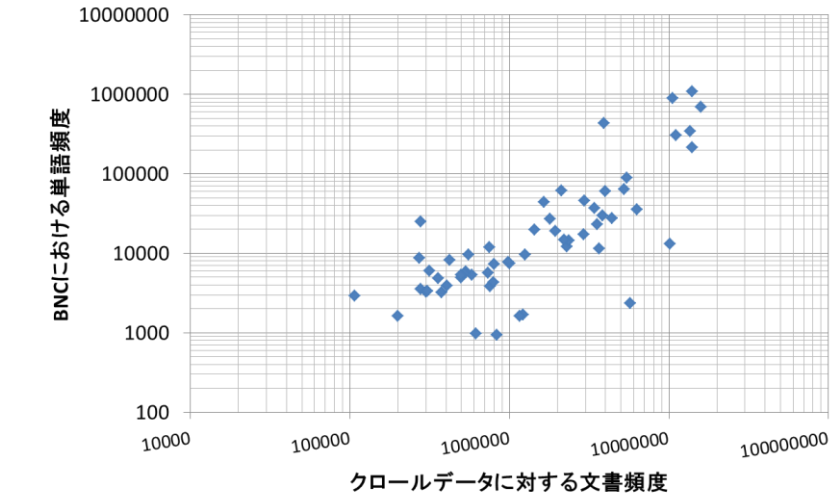
7.3.3 固定的なコーパスにおける文書頻度との比較

取得したクロールデータに対する文書頻度の妥当性を検証するために，取得した文書頻度と固定的なコーパスにおける文書頻度とを比較した．

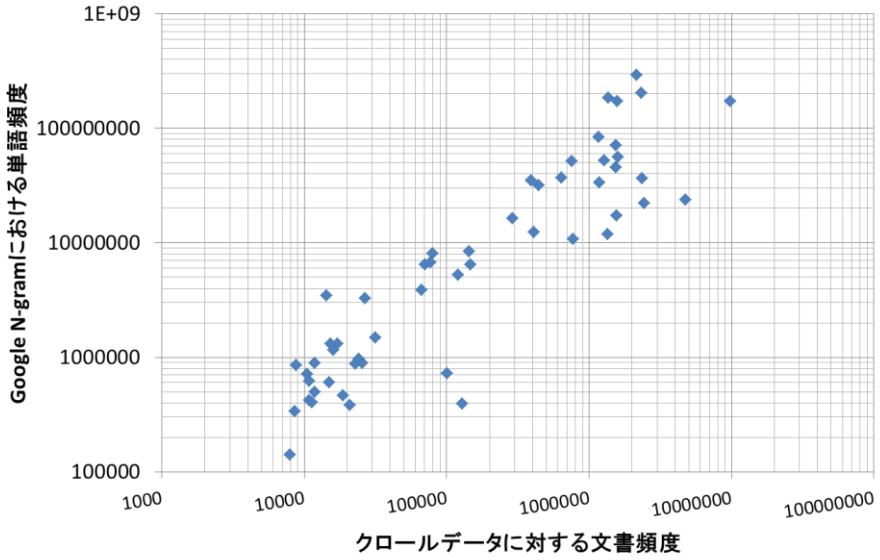
比較に使用したクエリは，Q_1，Q_2 から，出現頻度が文書の取得時期や分野に対して強い依存関係がないと筆者が判断した単一語から成るクエリを日本語，英語ともに 50 件ずつ選出したものである（付録 II 参照）．また比較対象のコーパスとして，英語のクエリに対しては関連研究[13]にならい BNC[19]を使用し，日本語クエリに対しては Google N-gram[32]を用いた．Google N-gram は 2007 年に Google によって公開された，約 200 億文の日本語データから作成した 1〜7 gram の出現文書数データセットである．ヒット数とは異なり，概算値ではなく正確に数え上げられた頻度であるため，Web 上の単語出現頻度の指標として信頼性が高いと考えられる[6]．

散布図を図 7 に示す．

[6] Google N-gram は Web 上の単語出現頻度の指標として信頼性が高いと考えられるが，本研究においてヒット数の正確性評価の基準として用いなかった理由として，データの作成時期が 2007 年と古いという点と，複数語から成るクエリに対する出現頻度を取得したいとき，N-gram モデルではクエリを構成する複数の単語が連なっている場合しか出現頻度に加算されないため，ヒット数から得られる頻度情報と意味合いが異なると判断した点が挙げられる．



(a) 英語のクエリに対する結果 (BNC との比較)



(b) 日本語のクエリに対する結果 (Google N-gram との比較)

図 7. 固定的なコーパスにおける文書頻度とクロールデータに対する文書頻度との比較

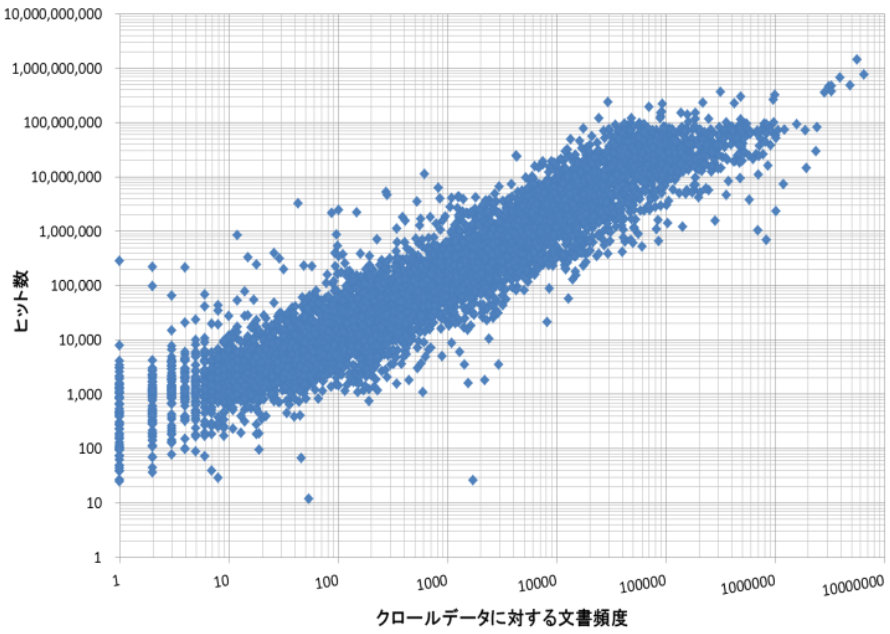
英語，日本語ともに強い相関が見受けられる．これは取得したクロールデータに対する文書頻度の妥当性を裏付ける根拠と見なすことができる．

7.4 ヒット数の正確性評価結果

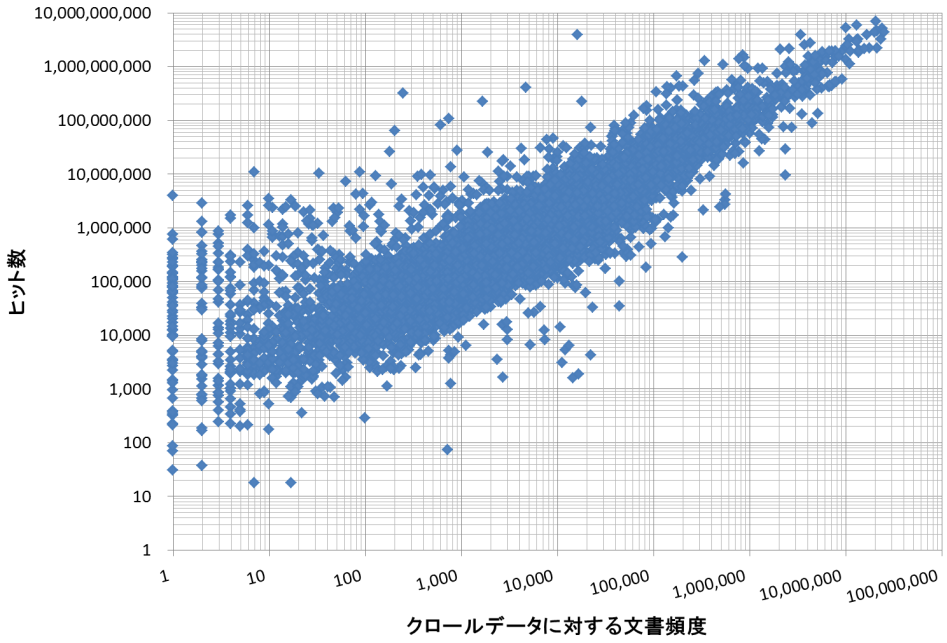
7.4.1 散布図と相関係数

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 8 に示す．Q_1 に対しては 2013/1/18 に取得したヒット数を，Q_2 に対しては 2012/9/30 に取得したヒット数を用いた．



(a) Q_1 に対するヒット数と文書頻度との比較



(b) Q_2 に対するヒット数と文書頻度との比較

図 8. ヒット数と文書頻度との比較

(a)については強い相関が見て取れる． (b)では，全体として正の相関が見受けられるが，

クロールデータに対する文書頻度が低いクエリに関してヒット数と相関が低い部分が目立つ.

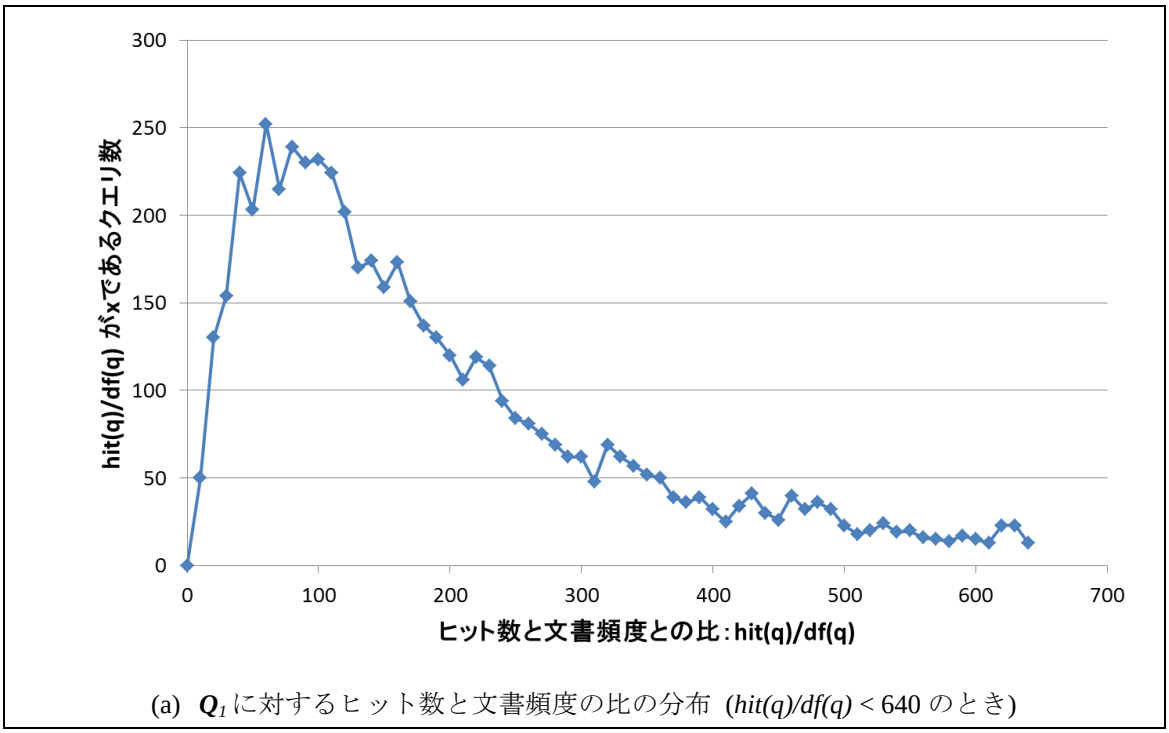
次に Q_1 と Q_2 それぞれの文書頻度についてヒット数との相関係数を示す.

表 5. ヒット数と文書頻度間の相関係数

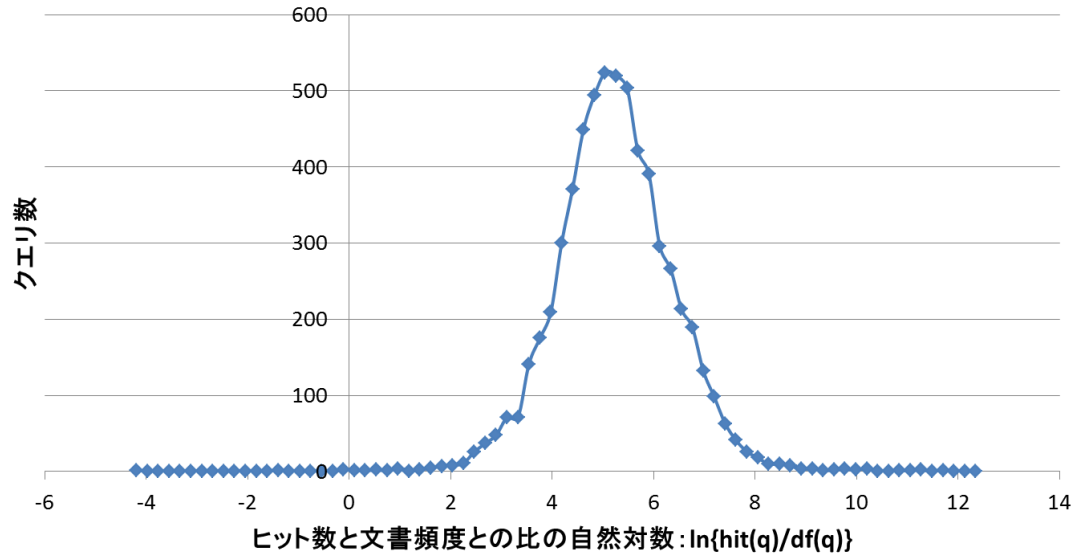
	ピアソンの積率相関係数	ケンドールの順位相関係数
Q_1	0.807	0.798
Q_2	0.860	0.694

7. 4. 2 ヒット数と文書頻度の比の標準偏差

7. 2. 2 で論じたように, $k(q) = Hit(q) / DF(q)$ を各 $q \in Q_1$ について計算した. $k(q)$の分布とその自然対数 $\ln\{k(q)\}$ の分布をそれぞれ図 9 に示す.



(a) Q_1 に対するヒット数と文書頻度の比の分布 ($hit(q)/df(q) < 640$ のとき)



(b) 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

図 9. Q_1 に対するヒット数と文書頻度の比の分布

興味深いことに，*k(q)*は対数正規分布によく近似しているという結果を得た．表 8 に *k(q)* に関するいくつかの統計量を示す．

表 6. ヒット数と文書頻度の比 *k(q)*に関する統計量

	平均	中央値	標準偏差
k(q)	446	164	4305
ln{ *k(q)*}	5.13	5.10	1.15

7. 4. 3　文書頻度の調整と誤差率分布

文書頻度の調整：

k の定義式

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

より，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，*k(q)*の中央値である *k*=164 を選んだ．この値を選んだ根拠は，ln{ *k(q)* }の平均値 5.13 を用いて *k* を復元した値 exp(5.13) = 168 とも近似しているからである．

このときの Hit(*q*)と *k*･DF(*q*)の分布を図 10 に示す．分布が類似していることが見て取れる．

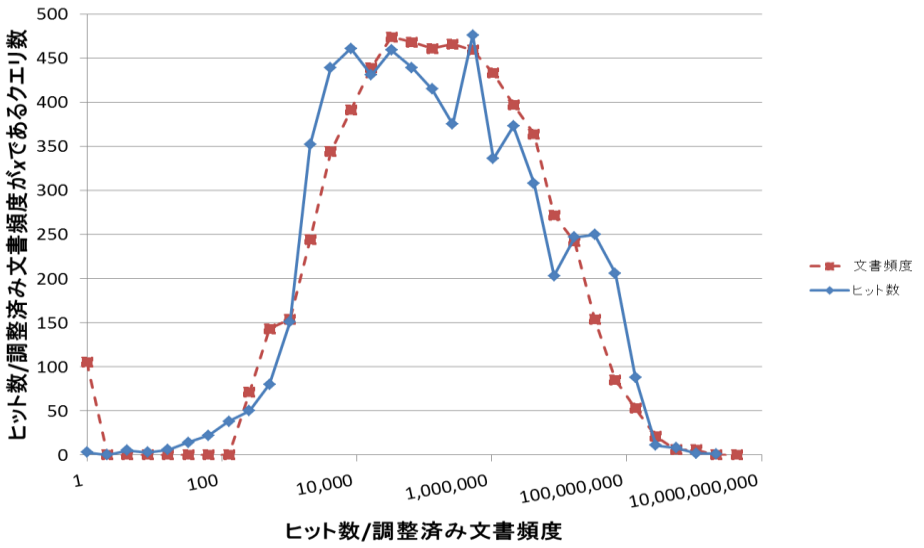


図 10. ヒット数と調整済み文書頻度の分布

誤差の定義

あるクエリ q に対する，クロールデータにおける文書頻度を基準としたヒット数の誤差を

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

と定義する．

一般的に誤差とは，“|測定値-真値|/真値”のように定義される場合が多く，これにならうとヒット数の誤差は“|*Hit(q)-k*DF(q)*|/ {*k*DF(q)*}”と定義すべきであるが，本比較においては誤差の分布が 0 を境に対象形に近い形をとるように *Hit(q)*と *k*DF(q)*のうち小さい値を分母に採用するようにした．例えば一般的な誤差率で計算した場合，*Hit(q)*が *k*DF(q)*と比べて極端に大きい場合，誤差率は絶対値の大きい正の値を取りうるが，逆に *k*DF(q)*が *Hit(q)*と比べて極端に大きい場合は絶対値が 1 以上の数値を取り得ない．一方，採用する誤差率を用いて計算すると *k*DF(q)*が *Hit(q)*と比べて極端に大きい場合でも，絶対値の大きい負の値を取ることができる．

調整済み文書頻度とヒット数の誤差分布：

ヒット数の誤差分布を図 11 に示す．ただし全データ 6,300 件のうち，文書頻度またはヒット数が 0 のもの(106 件)を除いている．また，グラフの可視性のために，外れ値を取るデータ(第 1 四分位点-1.5*IQR～第 3 四分位点+1.5*IQR [7]の範囲から外れるもの：877 件)も除去している．

[7] IQR (Interquartile Range): データのばらつきの指標のひとつで，“第 3 四分位点 - 第 1 四分位点”として計算される

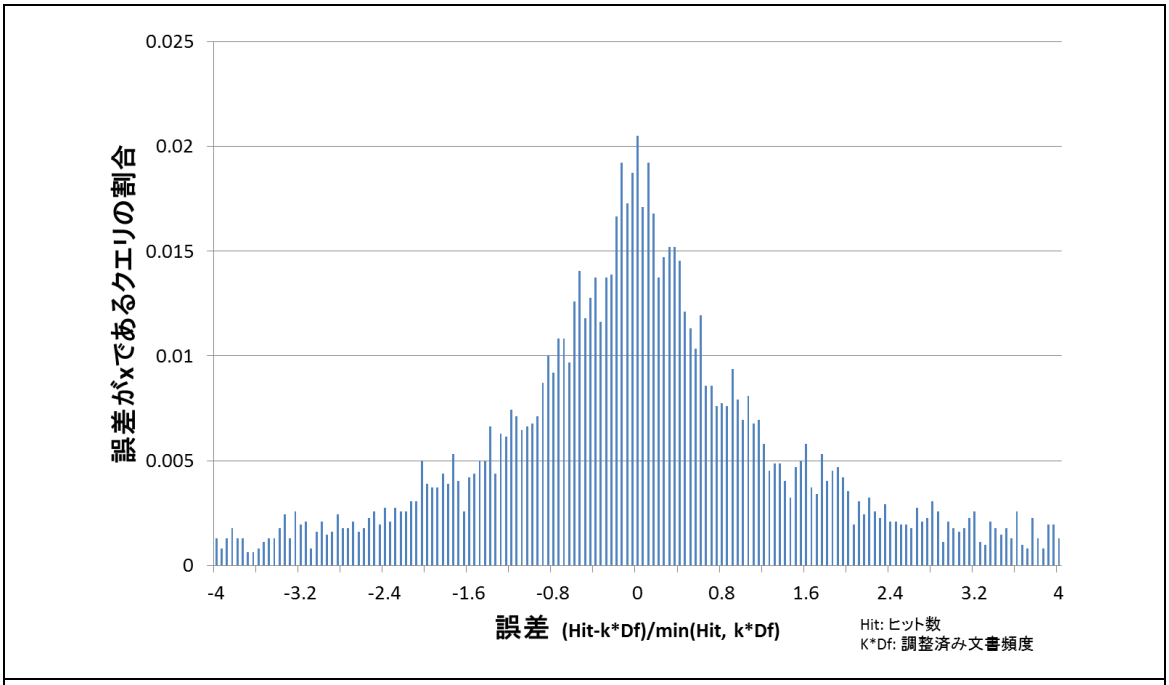


図 11. ヒット数の誤差分布

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 $Hit(q)/6.15 \sim 6.15*Hit(q)$の範囲に収まる確率が 90.0%であることを示している．

図 12 は,「誤差が$-x \sim x$ の範囲に収まるクエリ数の割合は y である」を示したものである．このグラフは，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．

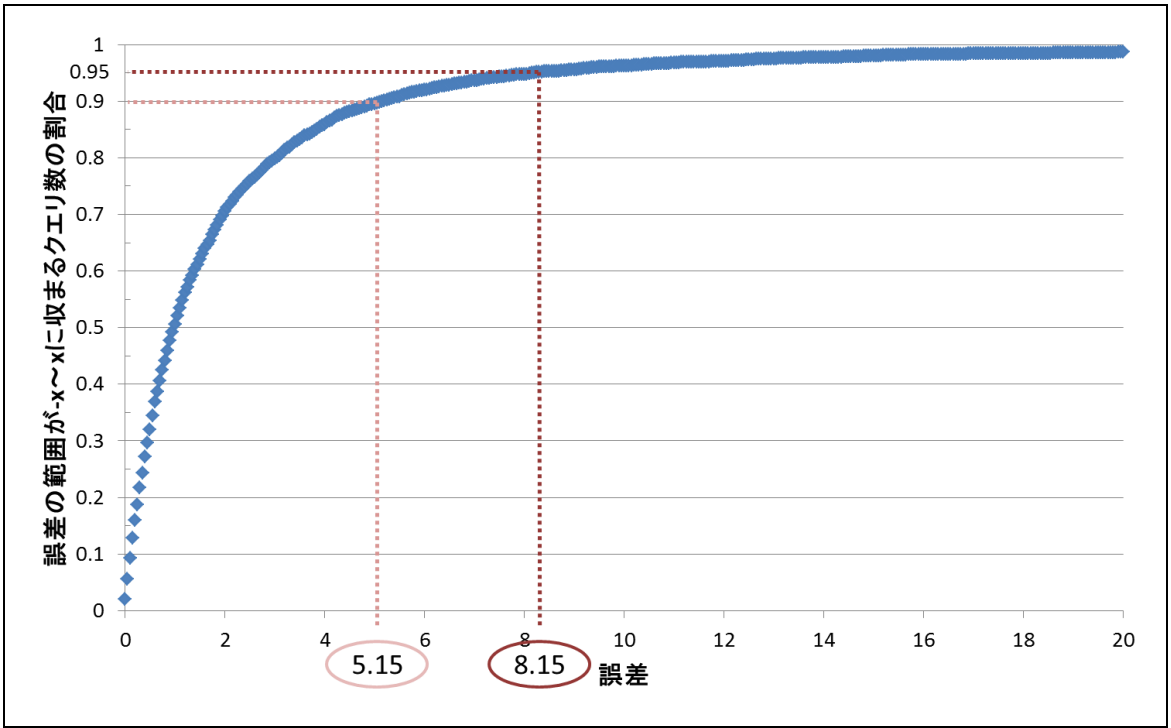


図 12. 誤差の範囲とクエリ数の割合の関係

次に，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．

3. 1 で例にあげたようなヒット数を利用する多くの研究[1][3][4][6]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」を知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 11 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 13 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．

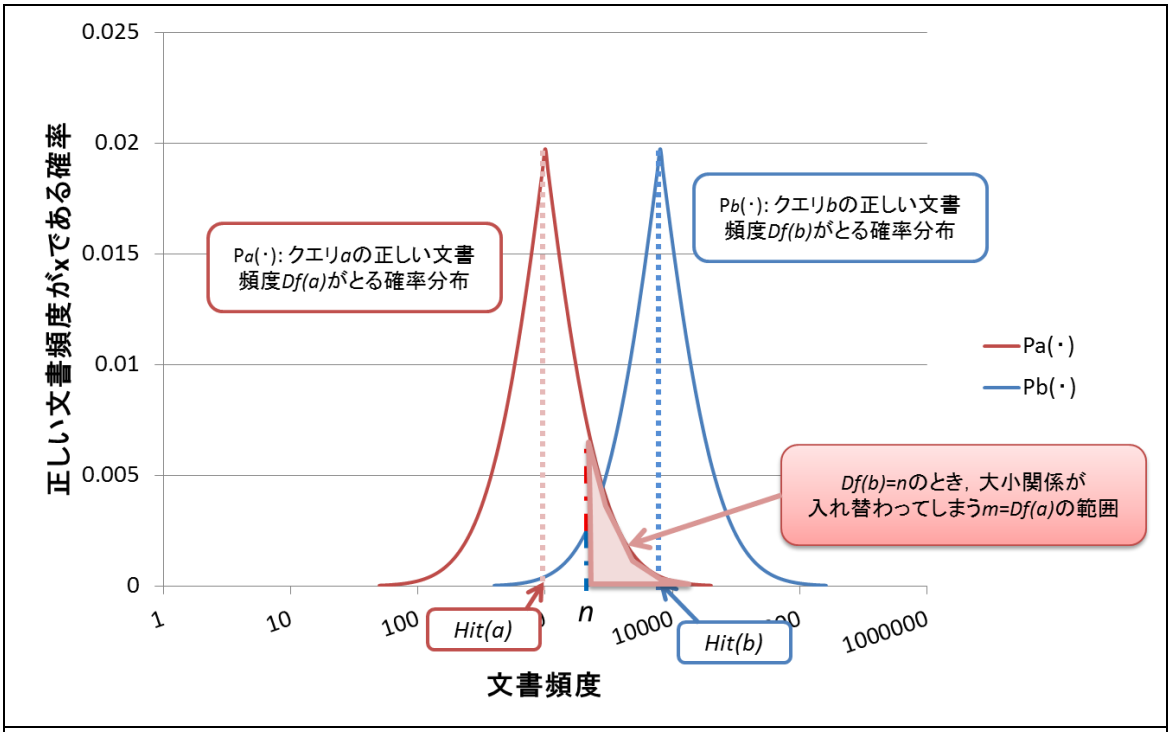


図 13. 2 クエリのヒット数を取得した際の正しい文書頻度の確率分布例

クエリ *b* に対してヒット数 *Hit(b)*を取得したとき，*b* の正しい文書頻度 *Df(b)=n* だったときのことを考える．大小関係が誤っているのは，クエリ *a* の正しい文書頻度 *Df(a)=m* が *m* ≧*n* となる範囲であるため，その確率は $\int_n^\infty p_a(m)dm$ である．クエリ *b* の正しい文書頻度が *n* である確率が $p_b(n)$であることを考えると，*Df(b)=n* という条件の元で誤った大小関係のヒット数を取得してしまう確率は $p_b(n)\int_n^\infty p_a(m)dm$ である．最後に，*n* の取る範囲は 0～∞ であるため，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．

$$Prob.error \quad = \quad \int_0^\infty p_b(n)\int_n^\infty p_a(m)dmdn$$

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 14 である．

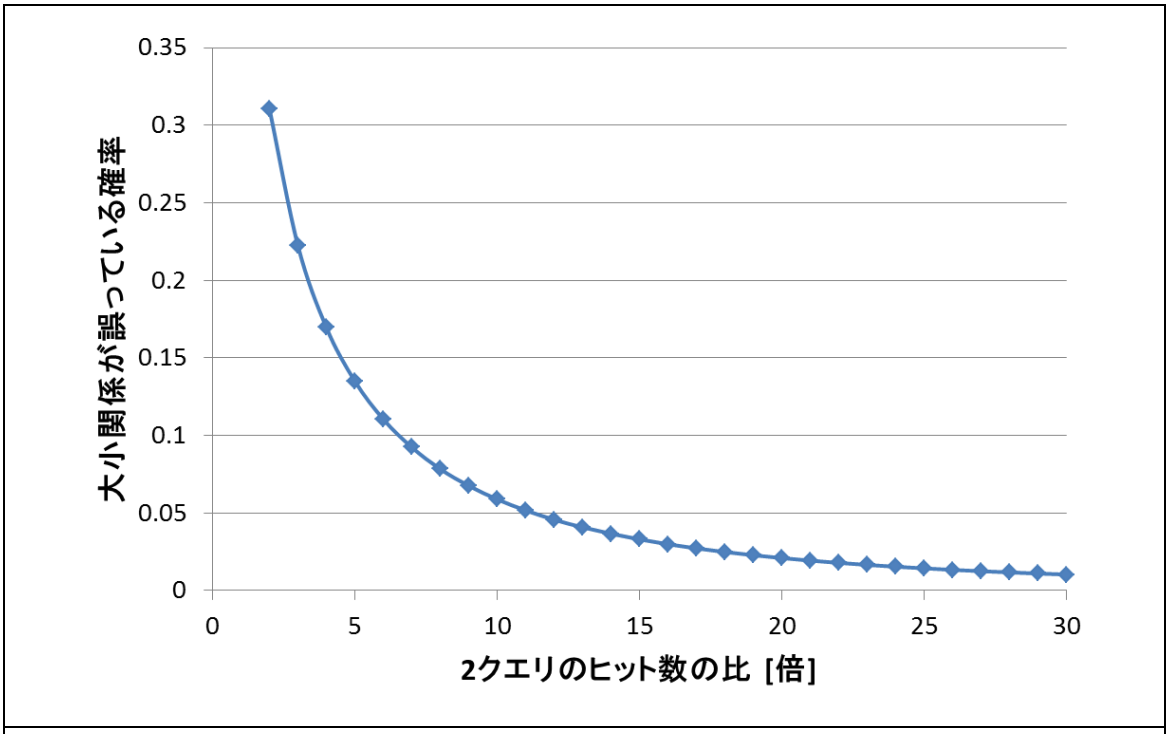


図 14. 2 クエリのヒット数比と大小関係が誤っている確率の関係

2 クエリのヒット数間の比が大きいほど，その大小関係が誤っている確率は低くなる．ヒット数比が 2 倍だと 31%, 5 倍だと 13.5%の確率で大小関係に誤りが生じる．また，大小関係が正しいことを 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．

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 15 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．このようなケースは，図 15 を有用に用いることができる 1 つの事例である．

図 15 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．

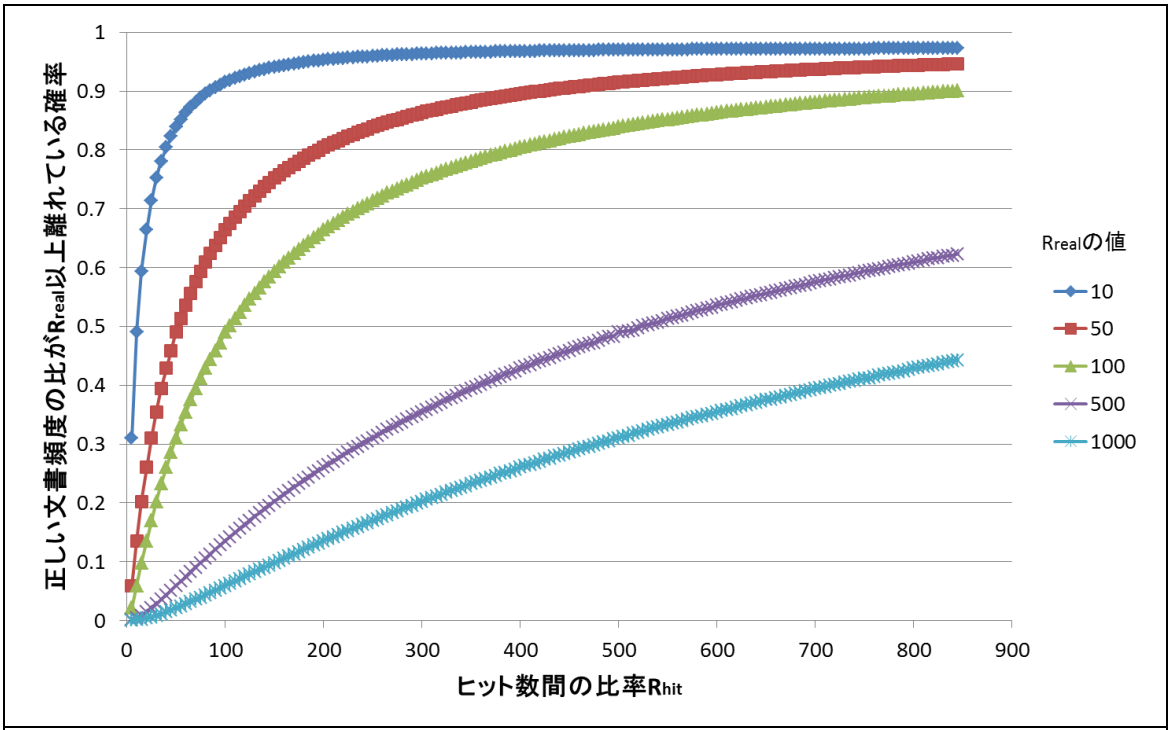


図 15. 2 クエリのヒット数比 – 正しい文書頻度の比が一定以上である確率の関係

7. 4. 4 検索の前処理におけるヒット数への影響の検証

クエリ拡張等，検索の前処理におけるヒット数への影響を調査するため，Q_2 についてクエリを構成する各単語を二重引用符で囲って検索を行った場合と，二重引用符で囲わずそのままの文字列で検索を行った場合の双方についてヒット数を収集した.

まず, 図 16 にクエリを二重引用符で囲って検索を行った場合と囲わない場合のヒット数の散布図を示す.

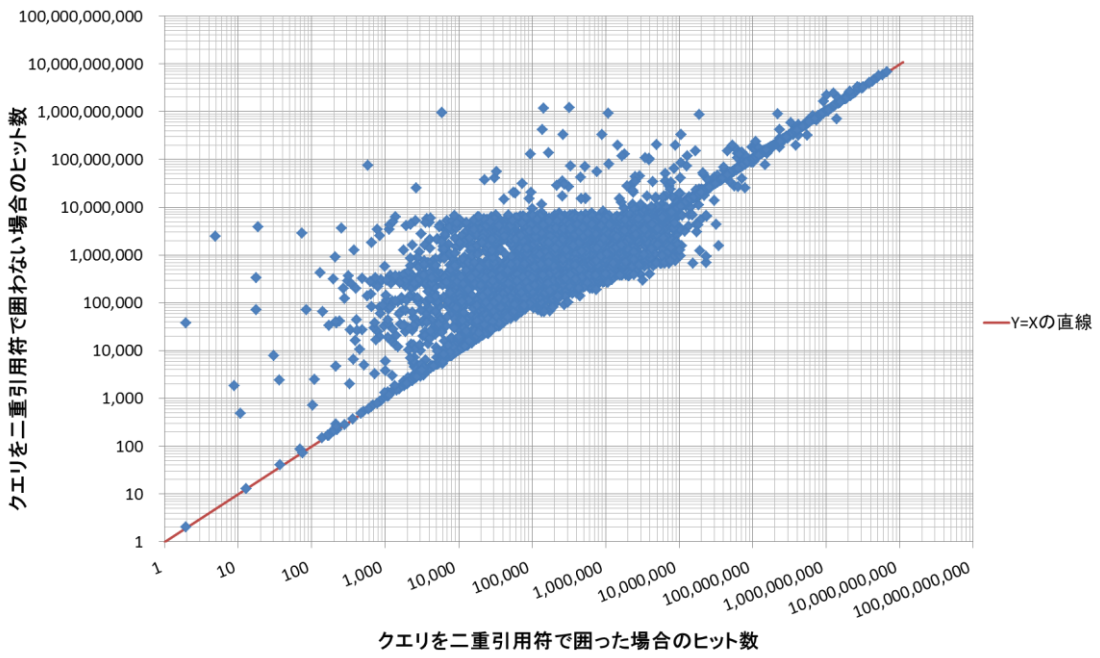
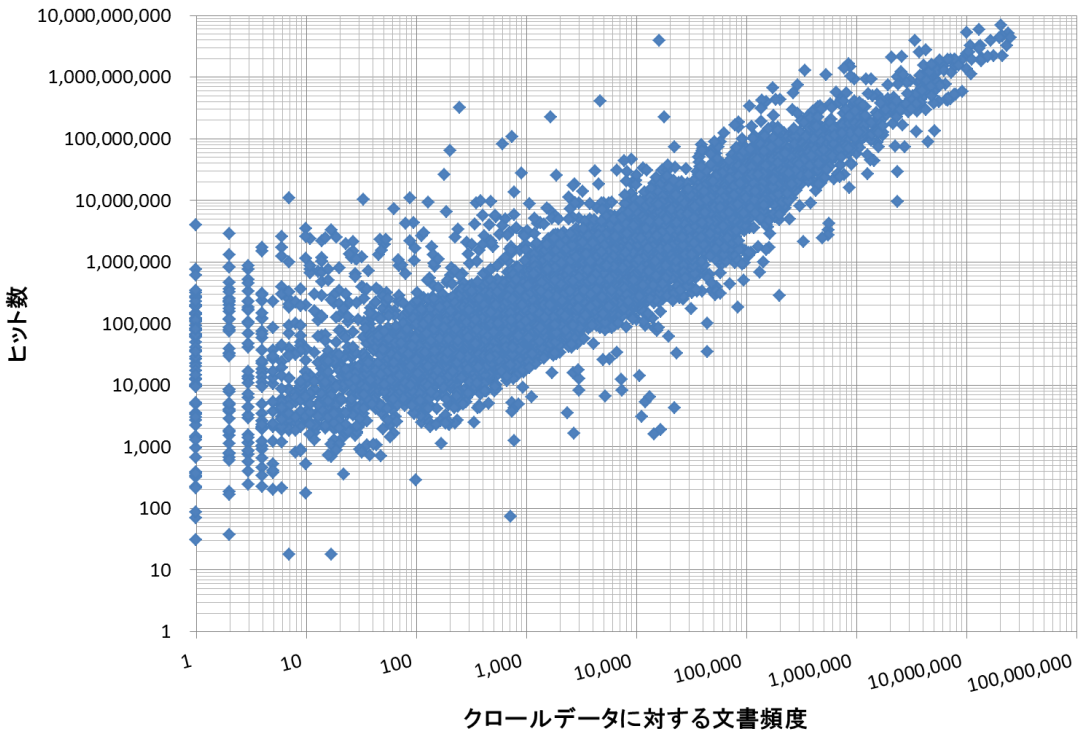


図 16. クエリを二重引用符で囲って検索を行った場合と囲わない場合のヒット数比較

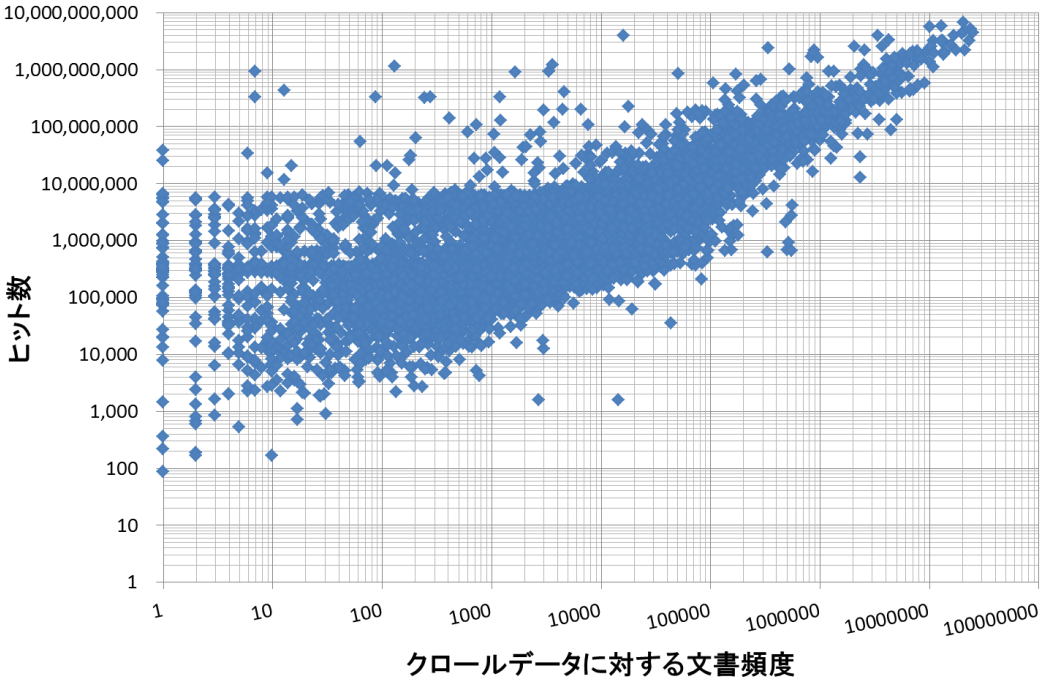
図中の直線は横軸と縦軸が一致している場合を示している．グラフを見ると，直線上に乗っているクエリが多数ある一方で，直線から大きく外れるヒット数をとるクエリも数多く見受けられる．顕著な例では，二重引用符で囲って検索を行った場合ヒット数が 10 以下の値をとるにもかかわらず，二重引用符で囲わない場合 100 万を超えるヒット数を返されるクエリも存在する．また，縦軸が 100,000 ~ 1,000,000，1,000,000 ~ 10,000,000 の間に不自然なプロットの偏りが見受けられる．

このように，検索の前処理におけるヒット数への影響は大きい．

次に，二重引用符で囲った場合と囲わない場合，どちらのほうがより正確な文書頻度に近いかを検証する．図 17 にそれぞれのヒット数とクロールデータにおける文書頻度との散布図を示す．



(a) クエリを二重引用符で囲った場合（図 8 の再掲）



(b) クエリを二重引用符で囲わない場合

図 17. ヒット数と文書頻度との比較

クエリを二重引用符で囲わない場合，グラフ中の左上に外れ値がでていることが確認できる．次に相関係数を示す．

表 7. クエリを二重引用符で囲って検索を行った場合と囲わない場合のヒット数と文書頻度間の相関係数

二重引用符	ピアソンの積率相関係数	ケンドールの順位相関係数
囲う(表 5 の再掲)	0.860	0.694
囲わない	0.842	0.540

どちらの相関係数で比較しても，二重引用符で囲ってヒット数を取得したときのほうがクロールデータにおける文書頻度に対して高い値をとっている．この結果から，Web 上の単語の出現頻度指標としてヒット数を用いる際には，クエリを構成する各単語を二重引用符で囲んだ完全一致検索を行うべきであるといえる．

7.4.5 クエリタイプ別の比較結果

6.2.1 で分類した Q_1 におけるクエリのタイプ別にヒット数と文書頻度との比較を行った．各々について比較結果を示す．

クエリを構成する単語数：

クエリを構成する単語数が 1～3 のものそれぞれについて相関係数を算出した[8]．結果を表 8 に示す．

表 8. クエリの単語数で分類した場合のヒット数と文書頻度間の相関係数

単語数	ピアソンの積率相関係数	ケンドールの順位相関係数
1	0.709	0.827
2	0.362	0.796
3	0.709	0.804

単語数 2 のクエリに対するピアソンの相関係数が著しく低い値をとっているが，これは図 18 に示すように絶対数が大きい少数のクエリが外れ値をとっていることが大きく影響しているためであると考えられる．実際，図中点線の丸で囲んだプロットのデータを除いてピアソンの相関係数を再度計算したところ 0.774 に上昇した．また，単語数 1 のクエリと単語数 3 のクエリとで相関係数が同等であることからも，単語数の違いによってヒット数の正確性に差異があるとは言えない．

[8] 各単語数ごとのクエリ数が等しくなるようにするため，後述するトレンド語に含まれるクエリを除いている

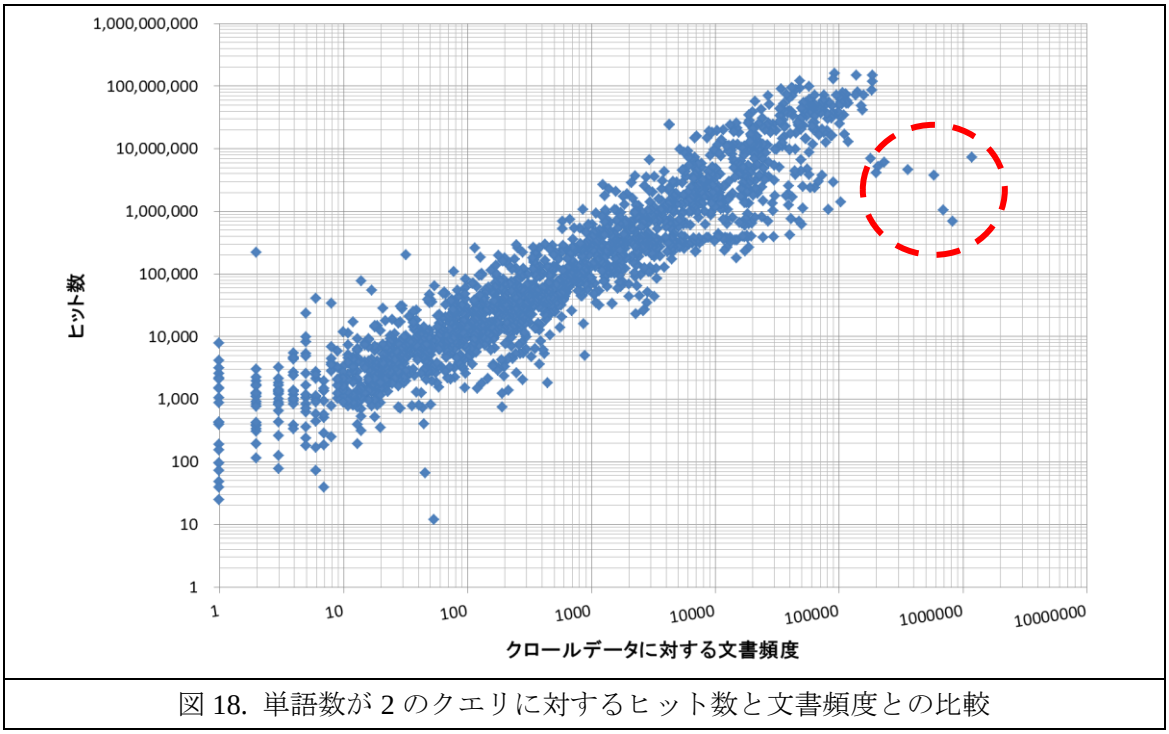


図 18. 単語数が 2 のクエリに対するヒット数と文書頻度との比較

なお，7. 4. 6 における議論と関連するところではあるが，図 18 で外れ値をとっているクエリのヒット数の時系列上での変化を見たとき，図に示すように大きな変動が確認された．このことから表 8 で単語数 2 のクエリの相関係数が低い原因は，複数単語から成るクエリに対する検索エンジン内でのヒット数概算方法に定常的な誤差があるというよりは，5. 2. 4 で述べた時系列上でのヒット数変動に大きく起因しているものだと考えられる．

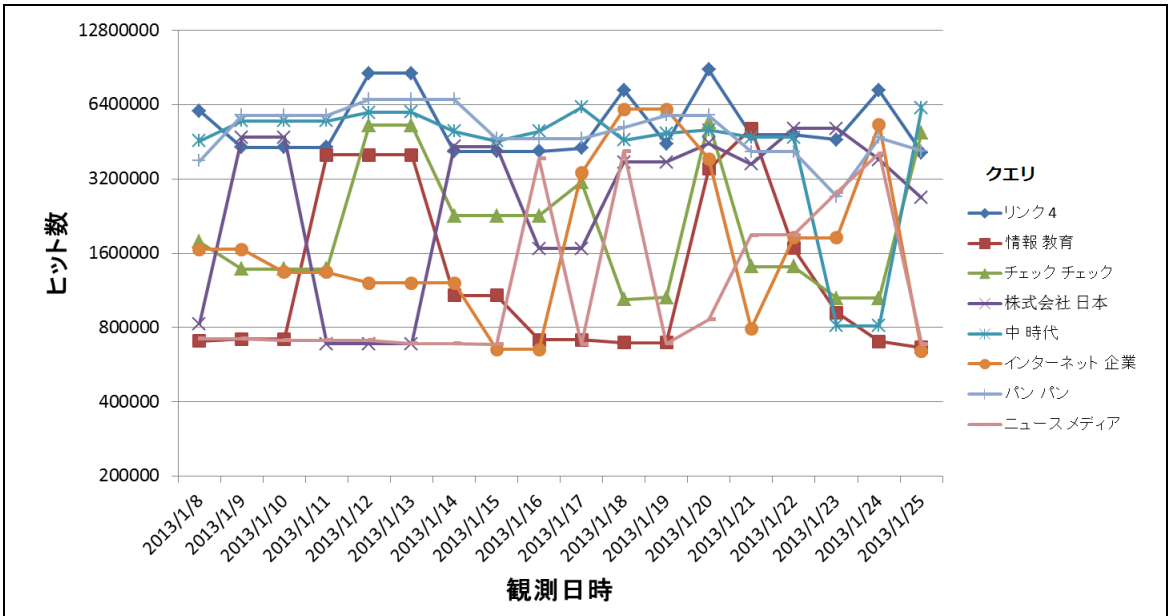


図 19. 外れ値をとるクエリのヒット数推移

クエリの言語：

クエリが英語のものと日本語のものそれぞれについて相関係数を算出した．結果を表 9 に示す．

表 9. クエリの言語で分類した場合のヒット数と文書頻度間の相関係数

言語	ピアソンの積率相関係数	ケンドールの順位相関係数
英語	0.578	0.818
日本語	0.717	0.808

散布図（図 20）を見ると，特にヒット数が大きいクエリに対する文書頻度に対して，日本語と比べて英語のクエリに大きなばらつきが確認できる．

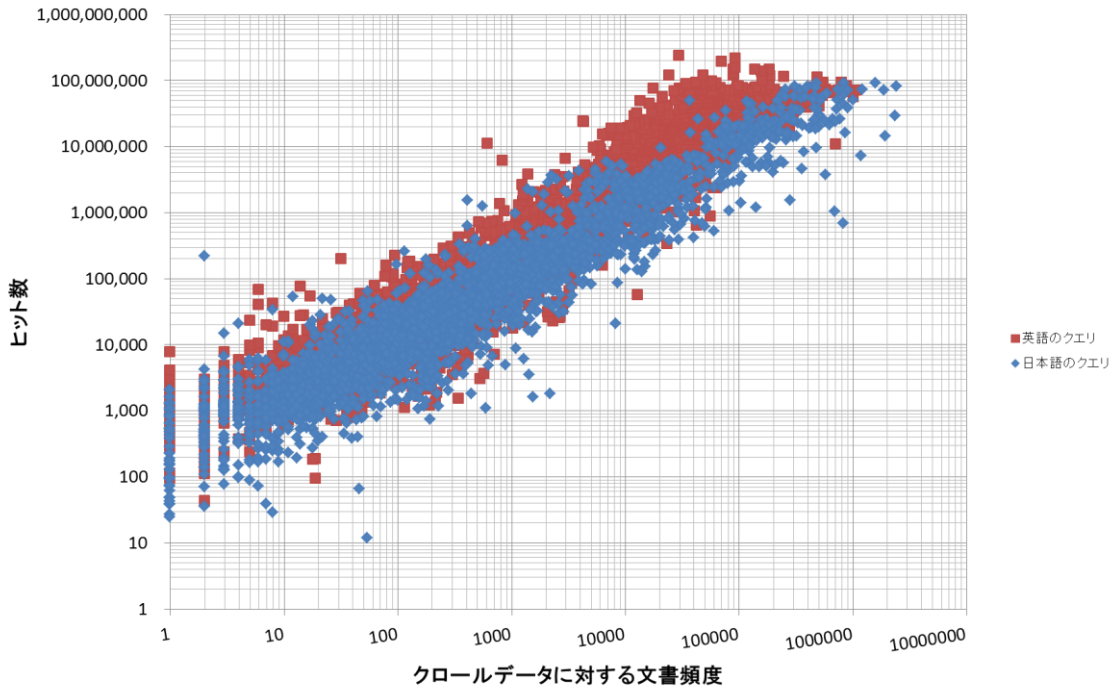


図 20. 英語，日本語のクエリに対するヒット数と文書頻度との比較

クエリのトレンド性：

Wikipedia のページアクセス数を元に選出されたトレンド語 400 件のクエリのみで比較を行い，相関係数を算出した．結果を表 10 に，散布図を図 21 に示す．

表 10. トレンド語のみ抽出した場合のヒット数と文書頻度間の相関係数

	ピアソンの積率相関係数	ケンドールの順位相関係数
トレンド語	0.909	0.602
Q_1 全クエリ（表 5 の再掲）	0.807	0.799

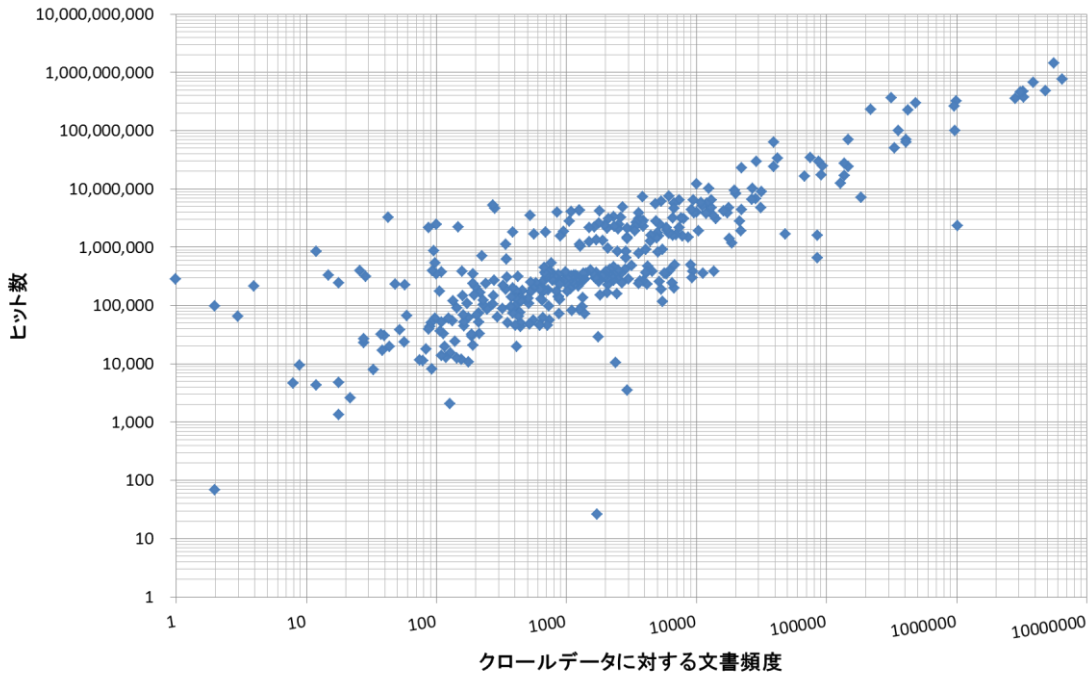


図 21. トレンド語に対するヒット数と文書頻度との比較

ピアソンの相関係数は高いものの，ケンドールの順位相関係数を比べると Q_1 全クエリで算出したものと比べて著しく低く，また散布図を見ても真に相関が高いとは言い切れない．

7.4.6 時系列上で安定したヒット数のみを用いた比較結果

[11]，[12]で述べられている「時系列上で安定しているヒット数は信頼できる」を検証するため，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．[11]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，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．図 22 に散布図を示す．

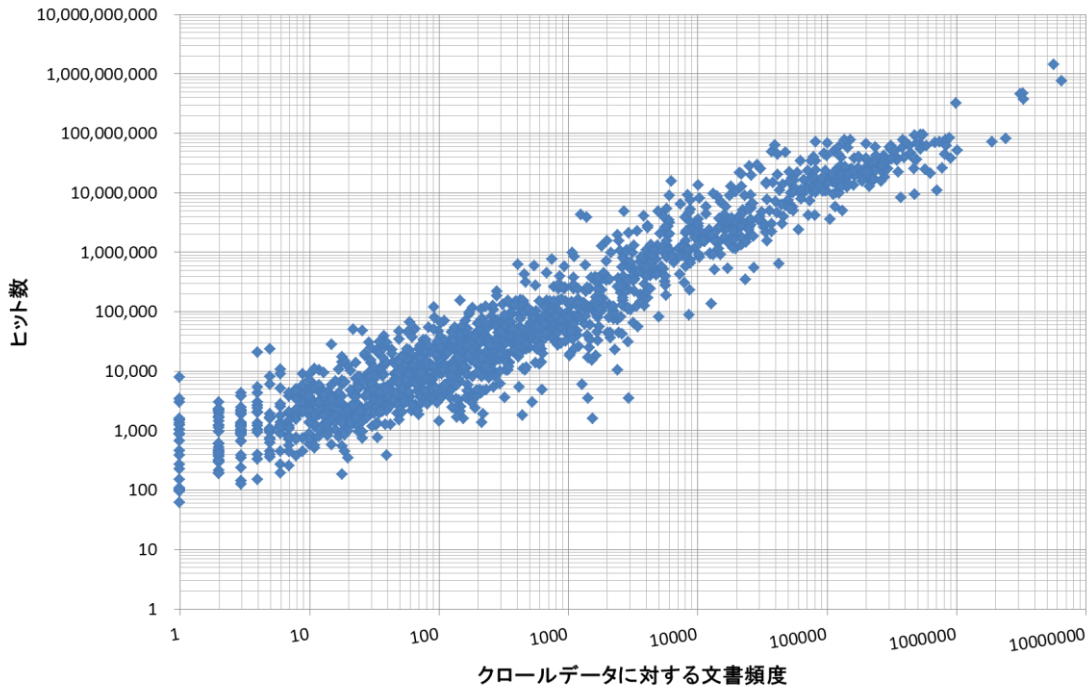


図 22. ヒット数が時系列上で安定したクエリに対するヒット数と文書頻度との比較

図 8 と比べてより強い相関が見て取れる．

表 11 に相関係数を示す．

表 11. ヒット数が安定したクエリに対するヒット数と文書頻度間の相関係数

	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 内の安定したクエリ	0.897	0.800
Q_1 全クエリ（表 5 の再掲）	0.807	0.799

ピアソンの積率相関係数においては，全クエリで比較した相関係数と比べ，安定したクエリのみで比較した相関係数が顕著に向上していることがわかる．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[11][12]を支持する結果となっている．

7. 4. 7 オフセットを変化させた場合の比較結果

オフセットを変化させた際に生じるヒット数の変動に着目し，ヒット数と取得した文書

頻度間の類似性に差が出るかを調査した.

Q_1 に含まれるクエリついて，オフセットを変化させたときのヒット数の変動について調査すると，舟橋ら[10][11]が指摘するように，最大オフセットにおけるヒット数が実際に取得できた検索結果数に調整される現象が見受けられた．例えばクエリ“ノロウイルス”に対して得られたヒット数は,

(offset, hit) -> (1, 1390000), (101, 1390000), (201, 1390000), (301, 1390000), (401, 416)

のように最終オフセットにおいて不自然に値の切り詰めが行われていた．そこで，先行研究[10][11]にならい，このような場合のヒット数は調整が行われる直前のヒット数を採用することとした.

まず，オフセットが 1 のときのヒット数 h_f と，オフセットが最大のときのヒット数 h_l の表した散布図を図 23 に示す.

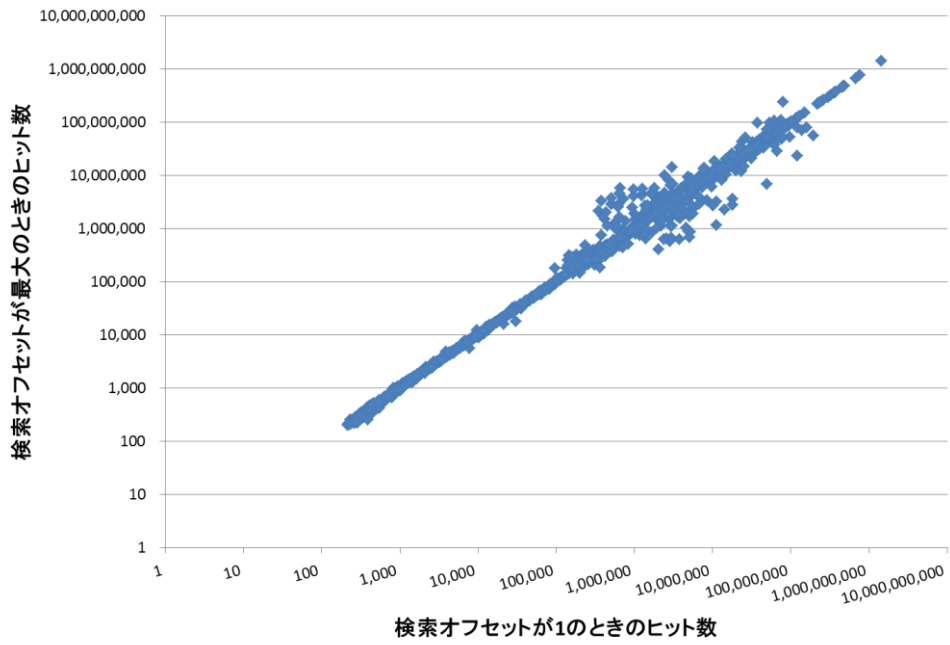


図 23. オフセット変化によるヒット数変動

基本的に高い相関があるが, 変動の大きなクエリもいくつか見受けられる. 調査の結果, h_f を基準として h_l が 50%以上増減しているクエリは 107 件(全体の 1.75%)存在した.

次に，7. 4. 1 と同等の比較を，最大オフセットにおけるヒット数を採用した場合について行った結果を示す.

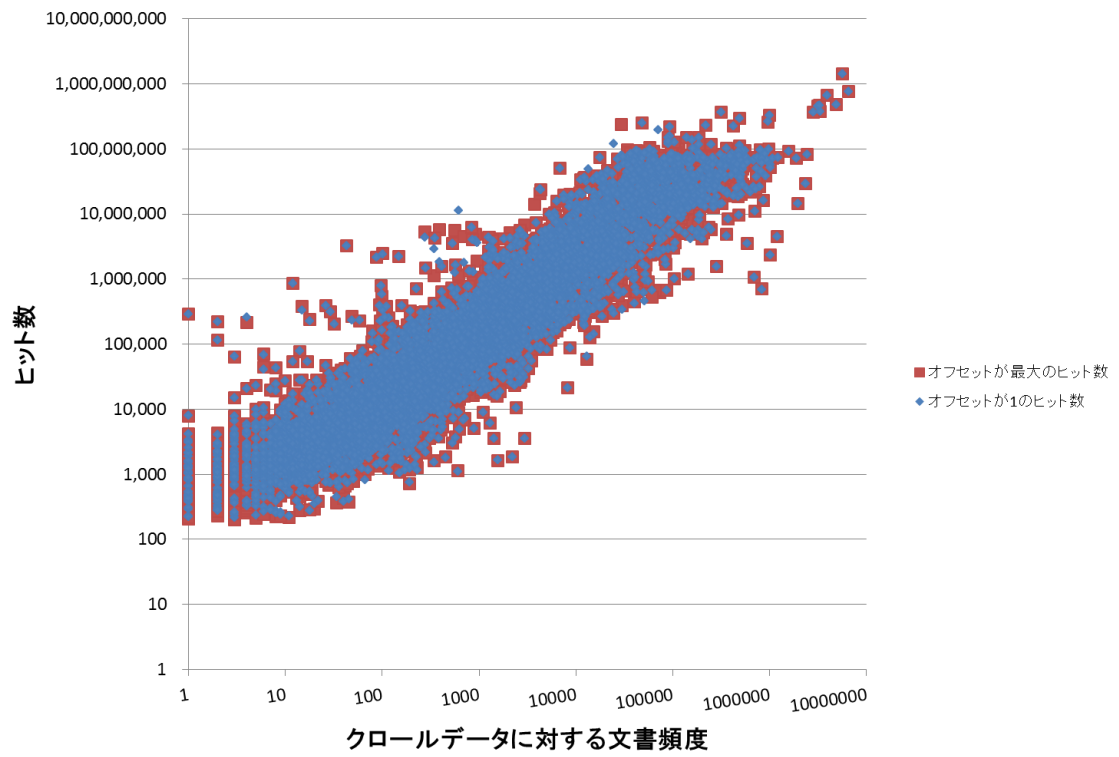


図 24. 最大オフセットにおけるヒット数と文書頻度との比較

そもそも h_f と h_l の差が大きいクエリの数が限られているため図 24 からは明確な差が確認できない.

次に，図 25 に h_f を基準として h_l が 50%以上増減しているクエリ 107 件を抽出して文書頻度との比較結果を示す.

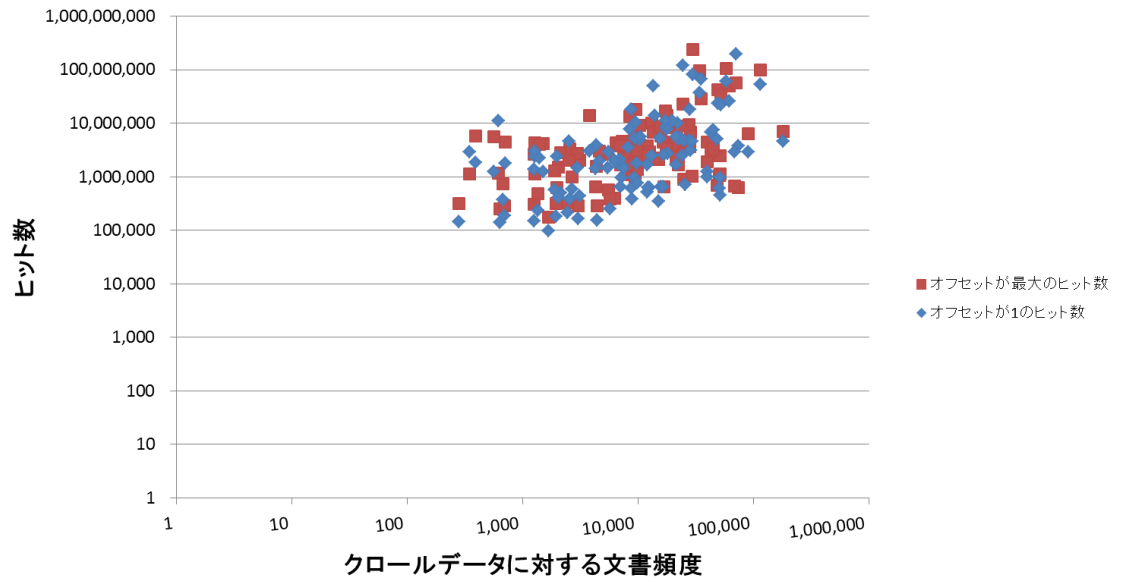


図 25. オフセット変化によるヒット数変動が大きいクエリの
最大オフセットにおけるヒット数と文書頻度との比較

図 25 から顕著な傾向は読み取れない．したがって，本検証からは検索オフセットの変化がヒット数の正確性に影響を与えるとはいえない．

本検証は，使用した API の仕様のために検索の最大オフセットを 401 として行ったものであり，これはいくつかの関連研究[9][10][11]が検証に用いているオフセットの値(1,000)より小さいものである．そのため先行研究[10][11]が報告しているほど，検索オフセットの変化によってヒット数が大きく変動するクエリを得ることができなかった．本比較結果は先行研究[10][11]における考察を否定するものであるが，上記の点で限定的であるといえる．

7. 4. 8　文書頻度の取得方法別の比較結果

6. 2. 2 で述べた文書頻度の取得方法に関する各フローで得られた文書頻度と，ヒット数とを比較した結果を示す．

ページランクによるフィルタリング：

ランクが高い文書のみを抽出して文書頻度を取得した際と，全ての文書を使用して文書頻度を取得した際とでヒット数との類似性に差が出るかを比較した結果を示す．

まず図 26 に，$q \in \{\boldsymbol{Q_1}, \boldsymbol{Q_2}\}$に対して，ページランクによるフィルタをかけた場合とかけていない場合とで取得した文書頻度の関係を表す散布図を示す．ページランクによるフィルタは，クロールされた Web ページ集合中で上位 25%, 50%, 75%に含まれるページを抽出して文書頻度を取得したそれぞれの結果を示す．

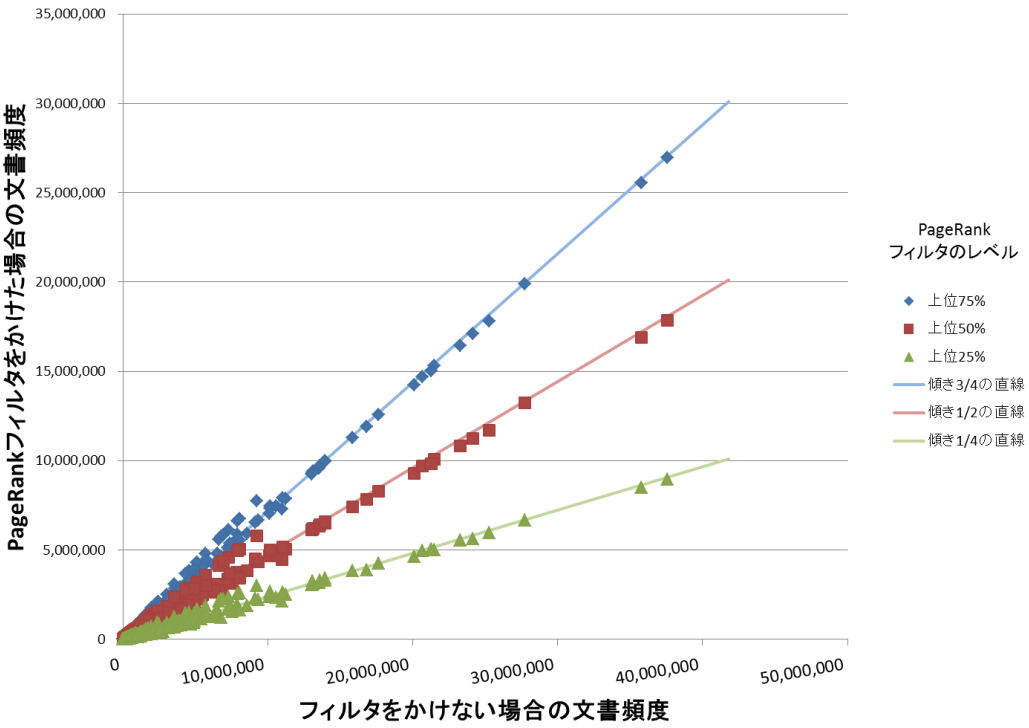


図 26. PageRank フィルタをかけた場合とかけない場合との文書頻度比較

グラフ中の直線は，文書群からランダムに 25%, 50%, 75%選出した場合に得られる文書頻度の期待値である(それぞれ傾きが 3/4, 1/2, 1/4 の直線に従う).

いずれのフィルタレベルにおいても，強い比例関係が見て取れ，その傾きは期待値と強く類似していることが見て取れる．このことから，特に文書頻度の高いクエリに対しては，ページランクの高い Web ページを優先的にサンプリングして文書頻度の概算を行った場合でも，ランダムにサンプリングされたときと比べて遜色のない正確性で概算値を算出することができると考えられる．

ただし，フィルタレベルが上位 50%のプロットに対して特に顕著な傾向として現れているが，フィルタをかけない場合の文書頻度が 10,000,000 以下の範囲で，フィルタをかけた場合の文書頻度が期待値よりも高く線形を成しているプロットがいくつか見られる．これらのプロットを表すクエリは，ページランクの高いページに偏って多く出現するクエリであることを意味している．いくつか例を表 12 に挙げる．「き」「お」「や」等，ひらがな一文字からなるクエリ（これらのクエリは Yahoo! Japan の人気クエリに含まれていた）のほか，「マップ」「平成」「株式会社」など，直感的に質の高い Web ページに含まれるクエリが含まれている．

表 12. フィルタをかけた場合の文書頻度が期待値よりも高いクエリの例

き	お	や	な	マップ
〒	株式会社	環境	概要	東京

次に，$q \in Q_1$ に対して，ページランクの高い Web ページ群における文書頻度とヒット数間の相関係数を表 13 に示す．

表 13. ページランクの高い Web ページ群における文書頻度とヒット数間の相関係数

PageRank フィルタレベル	ピアソンの積率相関係数	ケンドールの順位相関係数
上位 25%の Web ページ	0.770	0.769
上位 50%の Web ページ	0.774	0.781
上位 75%の Web ページ	0.795	0.790
全ページ（表 5 の再掲）	0.807	0.798

2 つの相関係数ともに，ページランクのフィルタをかけない場合のほうが，かけた場合よりも高い値をとっている．

先述したとおり，図 26 でフィルタをかけない場合の文書頻度が 10,000,000 以下の範囲でフィルタをかけた場合の文書頻度が期待値から離れているクエリがいくつか存在する．このように，ページランクによるフィルタが持つ影響は，特に文書頻度が小さいクエリに対して大きいと考えられる．そこで，フィルタをかけない場合の文書頻度が 10,000,000 以下のクエリを抽出して相関係数を算出した．結果を表 14 に示す．

表 14.文書頻度が低いクエリのみ抽出した場合の文書頻度とヒット数間の相関係数

PageRank フィルタレベル	ピアソンの積率相関係数	ケンドールの順位相関係数
上位 25%の Web ページ	0.658	0.708
上位 50%の Web ページ	0.668	0.725
上位 75%の Web ページ	0.696	0.736
全ページ	0.724	0.748

表 13 と同様に, 2 つの相関係数ともにページランクのフィルタをかけない場合のほうが，かけた場合よりも高い値をとっている．

これらの結果からは，ヒット数がページランクの高い Web ページのみを用いて算出されているとは言えない．

重複削除：

クロールデータに対して重複削除を行った場合とそうでない場合とで文書頻度とヒット数間の類似性に違いがでるかを調査した．重複削除の結果，全体の 14.4%に相当する 6,018,190 ページが削除された．この中には，アダルトサイトやコピペサイト等，本来の重複削除の目的に沿ったページの他，URL は異なるが同じページへリダイレクトされるコンテンツ(http://www.google.co.jp と http://google.com 等)，サイドバーやヘッダー・フッター等複数ページで出現する文字列が共通する箇所が大部分を占めるページ，アクセスクライアントの制限によるエラーメッセージ（「このサイトを表示するにはフレーム対応のブラウザが必要です。」など）が含まれた．

まず図 27 に，$q \in \{\boldsymbol{Q_1}, \boldsymbol{Q_2}\}$に対して，重複削除をした場合としない場合とで取得した文書頻度の関係を表す散布図を示す．

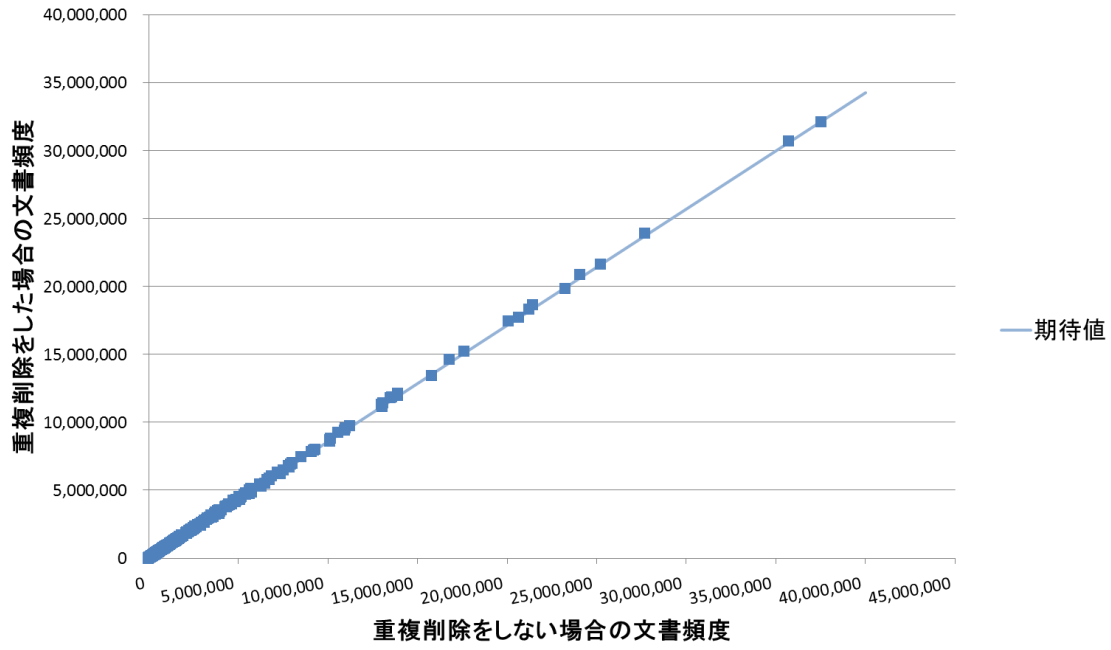


図 27. 重複削除をした場合としない場合との文書頻度比較

グラフ中の直線は，重複として検出された文書と同じ数だけ文書群からランダムに削除した場合に得られる文書頻度の期待値である．図 27 におけるプロットは，期待値とほぼ一致している．このため文書頻度を得るにあたっては，重複ページの影響は小さいものと考えられるが，重複削除後の文書頻度が期待値と比べて著しく低いクエリを並べると，アダルトワードや芸能人の名前などが数多く見受けられた．

次に，$q \in \boldsymbol{Q_1}$ に対して，重複削除前後の文書頻度とヒット数間の相関係数を表 15 に示す．

表 15. 重複削除後の Web ページ群における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
重複削除後の Web ページ	0.801	0.799
全ページ（表 5 の再掲）	0.807	0.798

結果からは，重複削除前後でヒット数の正確性に違いがあるかどうか，明確な傾向は見られない.

次に，先に述べた重複削除後の文書頻度が期待値と比べて低いクエリ 500 件を抽出して相関係数を計算した結果を表 16 に示す.

表 16. 頻度が期待値と比べて低いクエリに対する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
重複削除後の Web ページ	0.644	0.627
全ページ	0.652	0.629

表 15 における結果と同様に，重複削除前後でヒット数の正確性に違いがあるかどうか，明確な傾向は見られない.

これらの結果からは，検索エンジンが重複 Web ページを削除したあとのページ群を用いてヒット数を算出しているとは言えない.

本文抽出

クロールされた Web ページから，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し，本文のみ抽出を行った場合とそうでない場合とで文書頻度とヒット数間の類似性に違いがでるかを調査した結果を表 17 に示す.

2 つの相関係数ともに，本文抽出を行わない場合のほうが高い値を示している.

表 17. Web ページの本文抽出を行った場合の文書頻度とヒット数間の相関係数

本文抽出したか否か	ピアソンの積率相関係数	ケンドールの順位相関係数
本文抽出を行った場合	0.630	0.767
行わない場合（表 5 の再掲）	0.807	0.798

本節における検証の結果は，全てヒット数の傾向を示すに至らない結果であった．しか

し留意すべき点は，各種フィルタリングや本文抽出等の処理によって，文書頻度の取得対象とした文書量が減少してしまっていることである．本検証で用いたクロールデータセットは，全体として文書頻度の収束性を確認しているが，各処理後のデータに対する文書頻度の収束性は確認していない．このため，クエリの出現確率を算出するに充分な量の文書を集めていない可能性がある．

さらに大規模なクロールデータを用意し，各処理を行った後の文書頻度の収束性を確認した上で同様の検証を行うことが今後の課題である．

最後に，7.4 節で得られた知見を表 18 にまとめる．

表 18. ヒット数の正確性評価によって得られた知見

検証項目	得られた知見	節番号
全データを使った正確性評価	・ヒット数は正確な文書頻度と比較し相関係数 0.8 前後の相関がある ・ヒット数の誤差の範囲と発生確率が特定できた *e.g）ヒット数が正しい文書頻度の 1/6.15～6.15 倍範囲の値をとる確率は 90%* ・大小関係が正しいことを任意の確率以上で保証できる複数クエリのヒット数間の比率特定できた *e.g）大小関係が正しいことを 99%以上の確率で保証するには 2 クエリのヒット数間が 31:1 以上離れている必要がある*	7. 4. 1 ～ 7. 4. 3
検索の前処理における影響	・前処理の影響を排除したほうがヒット数の正確性が高い（検索するときはクエリを構成する各語を二重引用符で囲うべき）	7. 4. 4
クエリのタイプ別正確性評価	・明確な傾向は見られない	7. 4. 5
安定したヒット数の正確性評価	・時系列上で安定したヒット数は正確性が高い *i.e）ピアソンの積率相関係数において 9%の向上が確認できた*	7. 4. 6
検索オフセットを変化させた場合の正確性評価	・明確な傾向は見られない	7. 4. 7
文書頻度の取得方法別正確性評価	・明確な傾向は見られない	7. 4. 8

第8章　おわりに

本論文では，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．

ヒット数は現在幅広い研究に応用されているが，その一方で，ヒット数は様々な条件によって値が変動する現象が知られている．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が，ヒット数の正確性，すなわちヒット数が Web 上の網羅的な文書に対するクエリの出現頻度と比較してどれだけ正確であるかは十分に議論されていなかった．

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的とし，大規模に 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った．

本論文では，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 30 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．

本論文ではこの他に，ヒット数取得時の検索オフセットを変化や，クエリの単語数や言語の違いによってヒット数の正確性に違いがでるかどうかの検証や，収集された Web ページ群に対して重複削除やフィルタリングを行った場合の文書頻度とヒット数との比較も行ったが，これらの結果からは特筆すべき傾向を読み取ることができなかった．今後さらに大規模なクローリングを行い，より網羅的な文書群を用いてヒット数の正確性を評価することが課題である．

謝　辞

本学で研究を行うにあたり，様々なご指導，ご助言を頂いた山名早人教授に厚く御礼申し上げます．また，お忙しい中熱心に研究の助力をしてくださった上田高徳先輩，奥谷貴史君，本研究の基盤を整えていただいた OB の舟橋卓也先輩，山﨑邦弘先輩，そして実のある議論で研究の進展に貢献してくださった New Jersey Institute of Technology の Andrew Sohn 氏，同大学の James Geller 氏，Manhattan College の Tian Tian 氏，Melikşah University の Ahmet Uyar 氏に心より感謝申し上げます．

文献

[1] G. Gefenstette: "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," In 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," In Proc. WWW2004, pp.462-471 (2004)

[5] W. Hage, H. Kolb and G. Schreiber: "A method for learning part-whole relations," In Proc. ISWC2006 (2006)

[6] P. Turney: “Mining the web for synonyms: PMI-IR versus LSA on TOEFL,” In Proc. of ECML-01, pp. 491-502 (2001)

[7] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida and M. Ishizuka: "POLY-PHONET: An advanced social network extraction system," In Proc. WWW 2006 (2006)

[8] 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)

[9] A. Uyar: "Investigation of the accuracy of search engine hit counts," J. of Information Science, Vol.35, No.4, pp.469-480 (2009)

[10] 舟橋卓也，上田高徳，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, pp.31-36 (2008)

[11] 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)

[12] Satoh, K., Yamana, H.: Hit count reliability: how much can we trust hit counts?, Proc of APWeb’12, pp.751-758 (2012)

[13] F Keller, M Lapata, Using the web to obtain frequencies for unseen bigrams, Computational Linguistics, Vol.29, No.3, pp.459-484 (2003)

[14] Y. Matsuo, H. Tomobe and T. Nishimura: “Robust Estimation of Google Counts for Social Network Extraction,” In Proc. Twenty-Second AAAI Conference on Artificial Intelligence 2007, pp.1395-1400 (2007)

[15] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan, "Searching the Web", ACM Trans. on Internet Technology, Vol.1, No.1, pp.2-43 (2001)

[16] M. B. J. Jansen, A. Spink, T. Saracevic, "Real Live, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web", Information Processing and Management, Vol. 36, No.2, pp.202-227 (2000)

[17] A. Ntoulas and J. Cho: “Pruning policies for two-tiered inverted index with correctness guarantee,” In Proc. of SIGIR'07, pp.191-198 (2007)

[18] 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)

[19] B Lou.: Users Reference Guide, British National Corpus. British National Corpus Consortium, Oxford University Computing Services, Oxford, England (1995)

[20] LDC Catalog, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T21 (2013.1.9 アクセス)

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

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

[23] 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)

[24] 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)

[25] jawiki dump progress on 20121115, http://dumps.wikimedia.org/jawiki/20121115/ (2013.1.9 アクセス)

[26] Min-Hash LSH for Detecting Duplicate Documents, http://www.stanford.edu/~ashishg/amdm/handouts/scscrib-lec10.pdf (2013.1.9 アクセス)

[27] 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)

[28] 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)

[29] 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)

[30] 情報爆発時代に向けた新しい IT 基盤技術の研究, http://www.infoplosion.nii.ac.jp/info-plosion/ (2010.1.8 アクセス)

[31] Page view statistics for Wikimedia projects, http://dumps.wikimedia.org/other/pagecounts-raw/ (2013.1.9 アクセス)

[32] Google Japan Blog: 大規模日本語 n-gram データの公開,
http://googlejapan.blogspot.jp/2007/11/n-gram.html (2013.1.19 アクセス)

付録I　　クロールに使用したシードURLの例

属性	URL	
	日本語圏	英語圏
政府系	http://www.kantei.go.jp/	http://www.usa.gov/
	http://www.mof.go.jp/	http://www.treasury.gov/
	http://www.mext.go.jp/	http://www.ed.gov/
大学系	http://www.akita-u.ac.jp/	http://www.brown.edu/
	http://www.kansai.ac.jp/	http://www.caltech.edu/
	http://www.keio.ac.jp/	http://www.cmu.edu/
	http://www.kobe-u.ac.jp/	http://www.columbia.edu/
	http://www.kyokyo-u.ac.jp/	http://www.cornell.edu/
	http://www.kyushu-u.ac.jp/	http://www.dartmouth.edu/
	http://www.meiji.ac.jp/	http://www.duke.edu/
	http://www.nihon-u.ac.jp/	http://www.mannes.edu/
	http://www.osaka-u.ac.jp/	http://www.harvard.edu/
	http://www.tohoku.ac.jp/	http://www.juilliard.edu/
	http://www.tsukuba.ac.jp/	http://www.princeton.edu/
	http://www.u-tokyo.ac.jp/	http://www.upenn.edu/
	http://www.waseda.jp/	http://www.wustl.edu/
企業系	http://www.aica.co.jp/	http://www.diageo.com/
	http://www.barclayscapital.co.jp/	http://www.facebook.com/
	http://www.daido.co.jp/	http://www.fifa.com/
	http://www.denso.co.jp/	http://www.google.com/
	http://www.dentsu.co.jp/	http://www.hyatt.com/
	http://www.fancl.co.jp/	http://www.ibm.com/
	http://www.furukawa.co.jp/	http://www.microsoft.com/
	http://www.google.co.jp/	http://www.moodys.com/
	http://www.hitachi.co.jp/	http://www.nike.com/
	http://www.honda.co.jp/	http://www.oracle.com/
	http://www.isuzu.co.jp/	http://www.orange.com/
	http://www.itmedia.co.jp/	http://www.rbc.com/
	http://www.jomo.co.jp/	http://www.tel.com/
	http://www.kansai.co.jp/	http://www.tiffany.com/
	http://www.kose.co.jp/	http://www.ubs.com/
	http://www.makita.co.jp/	http://www.ubudhanginggardens.com/
	http://www.meiji.co.jp/	http://www.vodafone.com/
	http://www.mito.co.jp/	http://www.wired.com/
	http://www.nec.co.jp/	http://www.yahoo.com/

他，全 7886 件

付録II　　取得した文書頻度の妥当性を検証する際に用いたクエリ

7. 3. 3 にて述べた，BNC・Google N-gram と取得したクロールデータに対する文書頻度間の相関検証に用いたクエリ一覧を示す.

表Ⅰ. 1. 取得した文書頻度の妥当性を検証する際に用いたクエリ一覧

英語のクエリ(BNC との比較)			日本語のクエリ（Google N-gram との比較）		
shop	upper	supreme	夢	人物	写真
key	office	lock	免許	名字	地球
chip	share	it	明治	検索	休息
an	national	earth	法事	国旗	流星
ship	air	avenue	民法	電話	チェス
milk	ray	lie	咳	面接	こたつ
speed	sky	ice	時差	紅葉	東京
next	access	mix	時間	大学	めまい
rich	sweet	you	官報	皇居	自衛隊
humanity	red	more	絵文字	干支	電報
real	life	his	無料	環境	平成
hero	reality	beat	住宅	財布	家電
word	impress	fat	特番	稽古	イベント
number	birth	ask	建築	梅干し	天然記念物
news	wave	abroad	キリン	株式会社	物置
faith	ok	lost	ニュース	居酒屋	通販
I	always		手帳	日本	

付録III　　外部発表リスト

査読あり国際学会：

- Satoh, K., Yamana, H.: Hit count reliability: how much can we trust hit counts?, APWeb 2012, pp. 751-758 (2012.4)

査読なし国内学会：

- 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”, DEIM2011, F6-1 (2011.3)
- 佐藤亘，打田研二，山名早人：“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”，DBS152・IFAT103 合同研究発表会，2011-IFAT-103，Vol.8, pp.1-8 (2011.8)

受賞：

- 情報処理学会 情報基礎とアクセス技術研究会(SIG-IFAT) ヤング・リサーチャー賞 (2012.12)