様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月２１日現在

機関番号：３２６８９
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：21300038
研究課題名（和文）　検索エンジンの信頼性解析

研究課題名（英文）　Analyisis of Search Engines' Trustworthiness

研究代表者
山名　早人（ YAMANA　HAYATO ）
早稲田大学・理工学術院・教授
研究者番号：40230502

研究成果の概要（和文）：　検索エンジンは日常生活においても必要不可欠な存在となっているにも関わらず，その信頼性は不透明である．特に，検索結果として表示されるヒット数は，同じ検索語でも 100 倍，1000 倍と大きく変動する．本研究では，様々な指標として用いられているヒット数に着目し，ヒット数の変動傾向を 15 ヶ月に渡る調査から明かにした．さらに，信頼性の高いヒット数を得るための仕組みを考案し 99.5%の精度でヒット数の大小判定ができる仕組みを確立した．

研究成果の概要（英文）：Nowadays, search engines become indispensable for us to live a life; however, trustworthiness of search engines are unclear. Especially, the number of search results, i.e., hit-count, usually varies about 100 to 1000 times increase or decrease even if we put them the same query word. In this research, we have made clear the transition characteristics of hit-counts based on 15 months investigation for Google, Yahoo! JAPAN and Bing. Moreover, we have proposed a new method to choose trustworthy hit-counts, which results in 99.5% precision when we compare two hit-counts on the point which query word has larger number of search results.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	4,500,000	1,350,000	5,850,000
２０１０年度	5,000,000	1,500,000	6,500,000
２０１１年度	4,300,000	1,290,000	5,590,000
年度			
年度			
総　計	13,800,000	4,140,000	17,940,000

研究分野：総合領域
科研費の分科・細目：情報学，メディア情報学・データベース
キーワード：検索エンジン，信頼性，情報検索，ヒット数，信憑性，ランキング

１．研究開始当初の背景

研究開始当初(2008 年末)の Web ページ数は 1,000 億を超え，検索エンジンは日常生活に欠くことができない存在となった．しかし，我々はその性質をあまり理解していない．多くのインターネットユーザが検索エンジンを利用する現在，検索エンジンが社会に与える影響は計り知れず，検索エンジンの信頼性解析が急務となった．

検索エンジンの信頼性に関して，これまで

に「ランキングの変動解析」「ランキングの信頼性」「ヒット数の信頼性」の３分野で研究が進んでいる．世界ではインディアナ大学（米国）においてランキングの変動解析が実施され社会科学的な問題点「有名な情報ほど有名になり，有名でない情報が埋もれる」という事実が指摘された．ウォルバーハンプトン大学（英国），メルクッシュ大学（トルコ）では，商用検索エンジンが出力するヒット数の変動に対する解析が行われ，検索エンジンの挙動解析が行われた．また，ニュージャージー工科大学（米国）では，AND 検索時（２つの語句を同時に含むページの検索）における奇異なヒット数変動についての報告がある．

こうした中，検索エンジンの信頼性を第三者が評価し確認することが求められるようになってきた．

２．研究の目的

本研究では，利用者の立場に立ち，検索エンジンの信頼性として何が重要かを定義し，その信頼性を向上させる手法を検討し，世界中の誰もが安心して検索エンジンを利用できることを目指した．

具体的には，検索エンジンが持つべき信頼性を示す指標を考案し，Google，Yahoo! JAPAN，LiveSearch（本研究実施中に Bing に名称変更）等の商用検索エンジンが持つ信頼性を明らかにすることを目標とし，以下の３点の実施を目指した．

(1)信頼性判断のためのベンチマーク構築
(2)主要検索エンジンの信頼性評価
(3)提案する信頼性評価尺度の国際化

３．研究の方法

(1) 検索エンジンがインデックス対象としない SPAM 等のページ解析を行い，ヒット数の補正に用いる．
(2) テキスト検索だけでなく，画像検索の信頼性についても検討を行うために，画像検索で必要となる各種画像解析技術についても研究を進める．
(3) テキスト検索でのランキング変動，ヒット数変動を調査し，ベンチマーク作成の基礎データとする．
(4) 検索エンジンの信頼性として，どの部分の信頼性が最も重要かについて検討し，その対象に対して信頼性評価のための基準を策定する．
(5) 本分野において世界で活躍する研究者から適切なアドバイスを受け，真に役立つ信頼性指標の提案・普及を図る．

４．研究成果

(1)ワードサラダ型スパムの発見手法の開発

Web ページの爆発的な増加につれてスパム行為を行うページが増加し，インターネットから得られる情報の価値を下げている．このため，検索エンジンはこうしたスパムページをインデックスから外したりランキングを下げるという対応をとっている．ここでは，どのようなページがスパムとして対象となるかを調べ，ヒット数の補正に用いることを想定し，スパム発見手法について研究を実施した．

スパムの中でも従来の手法による判別が困難なワードサラダ型のスパムについて，その発見手法の研究を行った．ワードサラダとは，コンピュータにより自動的に文章が生成される手法（数語単位で組み合わせて文章を作成する手法）であり，人間が読めば意味が通じずスパムだと判定可能なのに対し，コンピュータにより自動的に発見するのは困難であると言われてきた．

本研究では，ワードサラダ型スパムを検出するため，n-gram と離散型共起表現を用いてカルバック・ライブラー情報量に基づく文章スコアを計算し，計算したスコアに基づき判定を行う手法を提案した．具体的には，ワードサラダ型スパムの特徴である「意味が通じない文」に着目し，「A→B→C→D→E→F」のような順番で単語 A から単語 F が一文に含まれる際，A→B や B→C といった近傍の単語間の共起確率ではなく，A→E や B→F といった離れた単語同士での共起確率を求めることにより，日本語としての存在しやすさを判定することによりスパム判定を行う手法である．

提案手法の評価実験を 2,000 件の Web ページ(半数にワードサラダ型 Web ページを含む)行った結果，従来手法(3-gram を用いた評価)の F 値（精度と再現率の調和平均）を 0.67 から 0.85 へ向上させることに成功した．これにより，本提案手法をワードサラダ型スパム検出に利用可能であることが確認できた．

(2) 画像検索結果を判定するための画像分類手法の開発

画像分類は，画像に対してカテゴリ（花，自動車，飛行機など）を付与することにより，コンピュータに画像の内容を理解させ，検索時に類似する画像を効率よくランキングするための手段として利用できる．近年では，分類対象のカテゴリに含まれている画像から特徴量を抽出し，特徴量の傾向を元に，カテゴリの集合を分割する操作を繰り返すことによって，階層的に分類を行う研究が進められている．しかし，既存の階層的な画像分類の手法には，複数の特徴量の重みを自動的に設定する枠組みは備わっていない．複数の特徴量を用いて分類を行う場合，分類に有効な特徴量は階層ごとに異なるため，特徴量の

重みを適切に設定する必要がある．

本研究では，MKL と呼ばれる学習器を既存の画像分類の枠組みに統合することで，分類対象の画像に合わせて，複数の特徴量を重み付けて統合しながら階層的に分類を行う手法を提案した．例えば，「色」「形状」「局所特徴量」という 3 つの特徴量が利用できる場合，ある階層では「色」が分類に有効であり，別のある階層では「形状」が分類に有効となることがある．これを自動的に判定することで分類精度向上を試みた．

Caltech256 の画像を分類する評価実験では，従来の階層的な分類方法に比較し平均 4%の精度向上が得られることを確認した．

(3) 検索エンジンの検索結果の収集と解析

検索エンジンのヒット数に対する信頼性の検証を行うため，2009 年 10 月～2011 年 1 月の 15 ヶ月間，Google，Yahoo! JAPAN，Bing の各検索エンジンに対して毎日 1 万種類の検索語（クエリ）によるヒット数の収集を各社が提供する API を用いて行った．

ヒット数を用いれば，Web 上において検索語がどの程度利用されているか容易に判断することができるため，ヒット数は様々な研究で利用されている．一方で，ヒット数は様々な状況において変化する．我々の調査によれば，ヒット数は，図 1 に示すように，短時間に繰り返し検索を行った場合，検索開始オフセットを変化させた場合，検索を行う日時が変わった場合に変化する．

このようにヒット数が変化した場合，どのような状況におけるヒット数に信頼があるかは自明ではない．そこで，まず，ヒット数が変化をするケースを図 1 の 3 つに分類し，それぞれのケースに対して信頼性を検証した．

図 1　ヒット数が変化する典型例

その結果，ケース 1 の場合の変動率は小さく，Google では 99.1%の検索語に対して変化せず 0.9%の検索語に対する変動率も 0.1%以下であった．Yahoo! JAPAN では 99.4%の検索語に対する変動率は 0.5%以内，最大でも 5%以下の変動率（例：最初のヒット数が 10,000 の場合，10,500 もしくは 9,500 に変動）である．一方，Bing では変動率が大きく 97.5%の検索語について変動率が 0.5%以下であるが，変動率が 20%を超える検索語が 0.1%存在する．また，変動率が大きい検索語はスパム系の検索語であることも確認した．

ケース 2 の場合の典型例を図 2 に示す．図 2 は Yahoo! JAPAN において検索結果画面で「次へ」をクリックし続けた場合のヒット数の変動（最初に表示されるヒット数を基準）を示している．変動にはいくつかのパターンがあるが，ここでは主要な 2 パターンを示している．図中，cluster2 のグラフで示されるように，おおよそ 100 件の検索結果毎に一気に検索結果が減るという減少が確認できた．これは，検索エンジンがヒット数を 100 件毎に再計算しているためであると推測される．ケース 2 における最大の変動率は 0.8（最初のヒット数が 10,000 の場合，クリックすると 2,000 に変動）程度である．

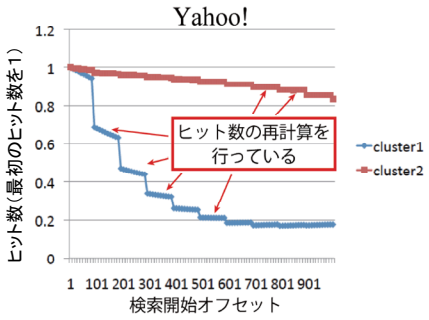


図 2 ヒット数変動（ケース 2）

ケース 3 の場合の典型的な変動例を図 3 に示す．図 3 は検索語「canon」に対する 2010 年 7 月から 11 月の Google におけるヒット数の推移を示しており，そのヒット数は 500 万～4000 万件の範囲で変動している．

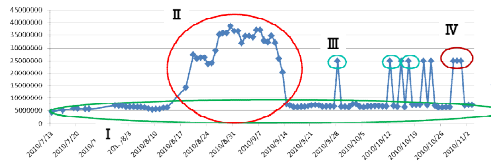


図 3　ヒット数変動（ケース 3）

図 3 に示す通り，ヒット数の典型的な変動パターンは以下の 4 種類に分類できる．

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．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となる（Ⅰの部分）．これに対して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，信頼性に欠けると考えることができる．

以上，長期間にわたる 1 万件の検索語を用いた調査から，ケース 1 やケース 2 の変動は小さく無視することが可能であるが，ケース 3 の変動は大きく，特にケース 3 のⅡ～Ⅳを避けてヒット数を得ることが重要であるとの結論に達した．

(4) 検索エンジンのヒット数に対する信頼性判定方法の開発（ベンチマーク開発）

検索エンジンのヒット数を用いるアプリケーションの多くは，ヒット数の絶対値を用いず，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．つまり「どちらの検索語（クエリ）がより Web 上での出現頻度が高いか」という大小関係こそが，重要なファクターとなる．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．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．

以上の考察に基づき，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定義した．

$$reliability(hit[A] > hit[B], m) = \Pr(days(hit[A] > hit[B]) = m)$$

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．

提案手法では，上記式で示される信頼度関数 Pr を 15 ヶ月にわたる検索エンジンのヒット数変動の統計から算出し「ベンチマーク」として備えた．この統計（ベンチマーク）に基づくヒット数信頼度提供システムを図 4 に示すように構築した．

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．利用者が指定するのは，大小関係が一定期間保たれる m 日であり，事前の大小関係の観測日数 α が長くなるほど精度よく信頼できるヒット数を出力できる．

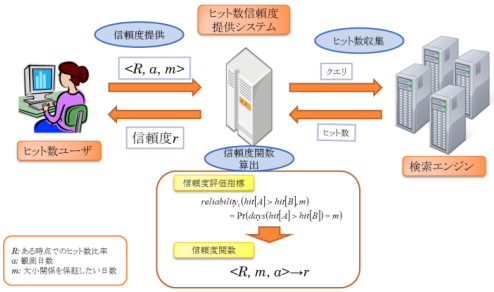


図 4 ヒット数信頼度提供システム

ここで，エラー率とスキップ率を図 5 に示すように定義する．エラー率は，信頼度関数 Pr に基づいて「信頼できるヒット数」として採用したヒット数における誤り率を示す．スキップ率は，信頼度関数 Pr に基づいてヒット数を採用したとき，「どれだけ誤った大小関係を持ったヒット数を回避できたか」を示す．すなわち，エラー率は 0 に近く，スキップ率は 1 に近い方がよい．

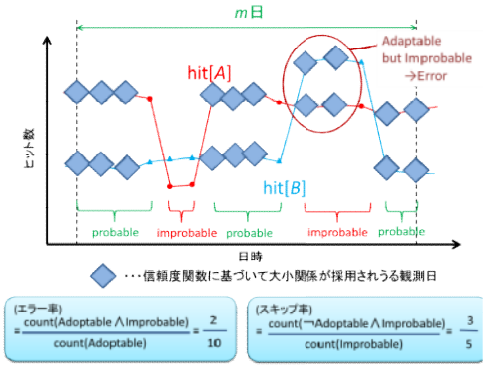


図 5 エラー率とスキップ率の定義

図 6 に Google を対象として「中央日報」と「イオンカー」をクエリとした場合のヒット数の採用事例（太線部分）を示す．図 6 に示されるとおり，急激にヒット数が変化している部分でのヒット数採用を避けることができていることがわかる．

図 7 は，2010 年 10 月～2010 年 12 月の Google を対象とした場合のエラー率とスキップ率である（m=20 日と設定）．このグラフは用いた 10,000 件のクエリの各組み合わせを用いた実験での平均値を示す．図から，エラー率は全体として低く保たれ(0.05 以下)，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わかる．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．これは，検索エンジンがインデックスを大規模に更

新した時期と一致しており，統計的な手法ではこうした大規模なインデックス更新の影響を避けることができないことを示している．一方で，このような大規模なインデックス更新は，他のクエリに対しても同様の傾向を示すため，別途，こうした事象を監視するシステムを追加することで対応が可能である．

以上，提案手法を用いることにより，2 つの検索語（クエリ）に対するヒット数の大小関係をエラー率 0.5%，すなわち精度 99.5%で判断できる仕組みの構築に成功した．

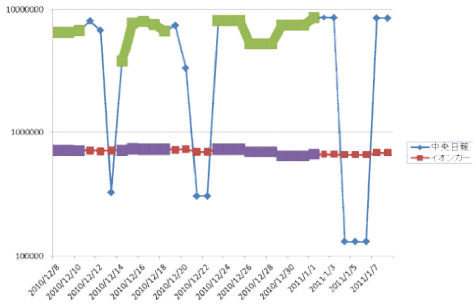
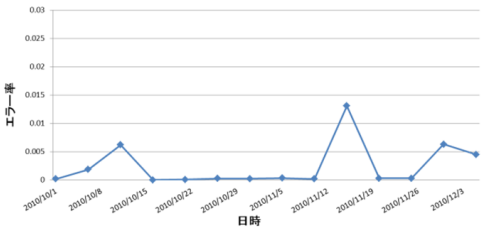
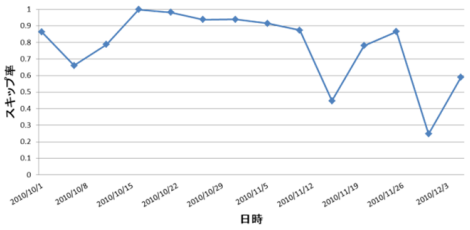


図 6 ヒット数の採用事例（太線部分）



(a) エラー率の推移



(b) スキップ率の推移

図 7 エラー率とスキップ率の推移

(5) 国際連携

国際連携においては，本分野で活躍している米国，英国，トルコの各研究者とミーティングを行い，本研究課題に対してアドバイスを受けると共に，我々の信頼解析方法の広報に努めた．さらに，信頼性解析方法については，商用検索エンジンの当事者である Google, Yahoo! JAPAN, Microsoft の研究者とも議論を行うことにより，評価手法として適切であることを確認している．

①提案手法の改善

トルコのメルクッシュ大学で検索エンジンのヒット数に対する信頼性を研究する Urt 准教授，英国ウォルバーハンプトン大学で検索エンジン間のヒット数の変動動向を調査した Thelwall 教授との連携では，単にヒット数を解析するだけでなく，検索エンジンのインデックス方式に依存するヒット数の「ゆれ」について助言を受けることができ，提案方式にその一部を組み込むことができた．具体的には，検索エンジンのインデックスでは，n-gram や形態素を用いたインデックスが複合的に用いられており，英単語であっても一語として扱われない場合が発生しヒット数の変動に影響を及ぼすというものである．

②社会科学分野での利用価値

米国インディアナ大学で「検索エンジンと社会科学との関係」について研究を行っている F. Menczer 教授との連携では，本研究が社会科学分野で行われる調査等に有効（精度向上に寄与）であるとの指摘を受けると共に，Web だけでなく昨今の SNS データや検索を対象とした研究の発展が重要であるとの指摘を受けた．この指摘を受け，Twitter データ検索におけるランキングについての研究を本研究の派生として開始することができた．

③提案手法の外部利用

本研究により提案した信頼できるヒット数の判定方法（2010 年時点の手法）がニュージャージー工科大学の Geller 教授のもとで研究を進める Tian 氏の「オントロジーを用いた検索エンジンの研究開発」において利用されることとなり，我々の成果の有効性が確認できた．

④提案手法の国際化

Web 分野での著名な国際会議の一つである Asia Pacific Web Conference での成果発表，欧州での情報検索に関する著名な ECIR 会議併設ワークショップでの成果発表，さらには，米国インディアナ大学，米国ニュージャージー工科大学，英国ウォルバーハンプトン大学，トルコ・メルクッシュ大学との連携を通し，本研究の研究成果について広めることができた．また，Google，Yahoo! JAPAN，マイクロソフトの研究者とのディスカッションにより検索エンジン運営会社に対して信頼性の重要性を訴えることができた．今後もホームページ等を通して本研究成果の発信を継続的に実施する．

５．主な発表論文等

〔雑誌論文〕（計１２件）

①Koh SATO and Hayato YAMANA, Hit Count Reliability: How Much Can We Trust Hit Counts?", Proc. of the 14th Asia Pacific Web Conference, LNCS, Vol.7235, 査読有,

2012, pp.751-758
②山名早人, ウェブサーチエンジンに見る統合検索, 情報の科学と技術, 査読有, Vol.61, No.9, 2011, pp.343-348
③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, 査読有, Vol.9, No.1, 2010, pp.18-22
④ Takuya FUNAHASHI, Hayato YAMANA , Reliability Verification of Search Engines' Hit Counts: How to Select a Reliable Hit Count for a Query, Proc. of 1st International Workshop on Quality in Web Engineering, LNCS, Vol.6385, 査読有, 2010, pp.114-125
⑤T.Kato, S.Honma, Y.Matsuyama, T.Yoshino and Y.Hoshino, Sensibility-aware image retrieval using computationally learned bases: RIM, JPG, J2K and their mxtures, LNCS, Vol.5506, 査読有, 2009, pp.620-627

〔学会発表〕(計２０件)
①山名早人, ソーシャルメディアからの情報抽出, シンポジウム『ソーシャルネットワークとソーシャルテレビサービス』,映像情報メディア学会年次大会, 2011/8, 成蹊大学
②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, 情報研報, Vol.2011-DBS-152, No.8, pp.1-8, 2011.7, 立命館大学
③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, E6-1,2011.2, 静岡県伊豆市
④小林大輔, 相川直視,山名早人, Localized Multiple Kernel Learningを用いた画像分類, MIRU2010, IS2-43, 2010.7, 釧路市
⑤ Hayato YAMANA, Search Engines' Trustworthiness - Current Status, Proc. of the 5th Korea-Japan Database Workshop, pp.219-240, 2010.5, 韓国
⑥新井啓介,森本浩介,山名早人,特徴領域の位置関係に着目したテンプレートマッチングによる類似物体検出,情処研報, Vol.2010-CVIM-172, No.4, pp.1-8, 2010.5, 名古屋工大
⑦舟橋卓也,山名 早人, Hit Count Dance -検索エンジンのヒット数に関する信頼性検証-,DEIM2010, 2010.2, 淡路島
⑧曽根広哲, 山名早人, ウィキペディア記事閲覧回数の特徴分析, Wikimedia Conference Japan 2009, SIG-SWO-A901-03, 2009.11, 東京大学
⑨舟橋卓也,曽根広哲,山名早人,複数キーワードクエリに対する検索ヒット数の信頼性検証,信学技報, Vol.109, No.153, pp.19-24, 2009.7, 神戸
⑩森本浩介, 片瀬弘晶, 山名早人, Ngramと離散型共起表現を用いたワードサラダスパム検出手法の提案, 情報研報, Vol.2009-DBS-148, No.24, pp.1-8, 2009.7, 神戸

〔その他〕
取材協力
①山名早人,ワールドビジネスサテライト, 日本からグーグルに挑戦, テレビ東京, 2011/5/18
②山名早人, 検索寡占進む@日本 ヤフーとグーグル提携発表, 朝日新聞, 2010/7/28,2面
③山名早人,検索サービスは Google を超えた!?,日経トレンディ 2009 年 9 月号, pp.134-139
④山名早人,検索新技術「巨人に挑む」, 朝日新聞, 2009/7/17, 18面
⑤山名早人,近未来型「計算知識エンジン」ウルフラムアルファの頭脳, R25.jp, 2009/6/18版

アウトリーチ活動
①山名早人, 検索エンジンは信頼できるか?, 早稲田大学理工展サイエンスカフェ, 2011/11/6

ホームページ
http://www.yama.info.waseda.ac.jp/~yamana/

６．研究組織
(1)研究代表者
山名 早人 (YAMANA HAYATO)
研究者番号：40230502

(2)研究分担者
無

(3)連携研究者
松山 泰男 (MATSUYAMA YASUO)
研究者番号：60125804