

ほんど?サーチ: 検索結果の集約とページ生成時間分布解析による Web 情報の信用度評価

Honto? Search: Measuring Trust of Information in the Web and its Evolution in Time

山本 祐輔^{*} 手塚 太郎^{*}
 アダム ヤトフト^{*} 田中 克己^{*}

Yusuke YAMAMOTO Taro TEZUKA
 Adam JATOWT Katsumi TANAKA

「日本の首相は小泉純一郎である」という命題の真偽に関わる質問は、既存の Web 検索エンジンでは受け付けられない。そこでユーザが真偽を調べたい命題の信用度の判定を Web 検索エンジンを用いて支援するシステムを提案する。本稿では、Web 上にある情報を集約し、さらに Web ページの生成時期を考慮することで命題の真偽を評価する。具体的にはある時期においてある命題がどの程度 Web 上で言及されているか、どの程度継続して言及されているかを分析することで、時間的な観点から命題の真偽を評価する。結果としてユーザは特定の時期における命題の真偽、命題の真偽に関する時間的な一貫性という観点から命題の真偽を判断することができる。

In this paper, we propose a system that helps users determine the trustworthiness of propositions they may be unsure about. For example, a user may want to know if a proposition, such as "the number of countries in European Union is 25" is true or not. Conventional search engines cannot directly provide answers to such questions. The trustworthiness of propositions is estimated in our method by aggregating knowledge from the Web and by analyzing creation time of pages. We also propose a method for estimating changes in the popularity of propositions in time by analyzing how many pages contained the proposition in selected time periods and how continuously they appeared in the Web. Thanks to this, users can know whether the proposition has been true or false during specific periods.

1. はじめに

Web ブラウジングをしている際、例えば「日本の首相は小泉純一郎である」や「恐竜は6500万年前に絶滅した」とい

^{*} 学生会員 京都大学大学院 情報学研究科 博士前期課程
yamamoto@dl.kuis.kyoto-u.ac.jp

^{*} 正会員 京都大学大学院 情報学研究科 社会情報学専攻
{tezuka, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

たように真実かどうか疑わしいフレーズに出くわすことは少なくない。そのような状況に遭遇したとき、現状ではユーザの多くはそのフレーズに関するキーワードを検索エンジンに入力することで Web 上でその話題がどれほど一般的に言われているかを調べたり、そのフレーズに関する真の解答を調べようと試みる。しかし実際にはこのような作業はキーワードを繰り返し入力したりするなどユーザにとって負担の大きい作業を伴う。

そこで我々は真偽の疑わしい話題の信用度をユーザが判断するための支援を行うシステム、「ほんど?サーチ」を提案する。我々はあくまで情報の信用性の判断支援を行うことに焦点を当てており、フレーズの真偽を判定することは視野に入れていない。真偽とは最終的にはユーザが判断するものであると考えるからである。

今回、我々の研究では多数派度を情報の信用度の基準として用いることにする。ほんど?サーチは、1. ユーザが真偽が疑わしいと思ったフレーズが WWW 内でどの程度言及されているかを評価するとともに、2. ユーザの疑問に思うフレーズに対する別の解となりうる候補を探しそれらに関して WWW 内でどの程度言及されているかを測定する。また本システムは、これらのフレーズに関するページのある時期における発生頻度の変化を提示する機能も有する。これはあるフレーズが「ある期間において多数派であった命題なのか」や「長期間言及されてきた命題なのか」をユーザに提示することで、より厳密にフレーズに対する信用度判定の支援を行うことができるからである。

これらの機能を実現するために、ほんど?サーチは次のようなステップを踏む。最初に、信用度判定を行いたいフレーズを分割し、その一部を検索エンジンに投げる。次に得られた検索結果から入力したフレーズに対する解候補となる部分を抽出する。さらに得られた解候補と初めに入力したフレーズから復元した解候補フレーズを再び検索エンジンに投げることで、それらフレーズの現時点での Web 上での頻度を測定する。また Web アーカイブを利用することで、それら解候補フレーズの頻度の時間的変化を分析する。最終的に、結果として解候補フレーズとそれらの発生頻度の時間的変化を分析したグラフをユーザに提示する。

本稿は以下のように構成されている。2章では Web から解の候補を抽出する方法を述べる。3章では Web アーカイブを用いた時間的分析の方法を述べる。4章では提案手法の効果を測定するための実験結果をのべ、最後にこれらを結論としてまとめる。

2. 解候補命題の収集

本章ではユーザがほんど?サーチに入力したフレーズから解候補となる命題を発見する方法について述べる。まず本システムはユーザが入力したフレーズのある特定の部分と交換可能な語を収集する。例えば「恐竜は6500万前に絶滅した」というフレーズの場合、ユーザは「6500万年」が正しいかどうかを知りたいとする。この時、システムは Web から「6000万年」「7000万年」「10,000年(これは事実と反する)」といったような表現を抽出する。我々はこのように抽出された語を「解候補語」、解候補語と元の入力された文から復元されたフレーズを「解候補命題」と呼ぶ。

ほんど?サーチを使用する際、ユーザはフレーズから真偽が疑わしいと思う箇所を指定する。この部分を本稿ではターゲットと呼ぶ。入力されたフレーズのターゲットと置き換え

られる語が解候補語である。ユーザがターゲットを指定しない場合は、システム側がフレーズを構成する語全てに対して同じ手順を適用する。

システムは次の手順で解候補命題を発見する。

Step1:

入力されたフレーズをターゲット T によって二つに分割したものの P_1 および P_2 を検索エンジンに投げるクエリとする。

Step2:

システムはクエリ " P_1 & P_2 " を Web サーチ API に投げる。次に、検索結果のスニペット中で正規表現 $/P_1 (.*) P_2/$ にマッチする部分を解候補語として抽出する。この正規表現の中央にマッチする部分が解候補語として抽出される。

Step3:

解候補語を検索結果から抽出された頻度順にソートする。ある解候補語が現れるページが多ければ多いほど、解候補語の頻度は高くなる。この段階でその頻度が閾値を超えなかった解候補語は、解候補として相応しくないとし除外する。

解候補命題は解候補語を元のフレーズを分割した2つ、" P_1 & P_2 " の間に挿入することで作成される。

システムは各解候補命題を再び検索エンジンに投げ、検索件数を調べる。このキーワードに基づく検索方法では本来表現していた意味を無視して検索してしまう。つまりフレーズが否定的な意見、期待を込めた意見、疑問を込めた意見を意味している場合はこの検索方法は脆弱なものとなる。この問題に対応するため、ほんと?サーチでは結果として返される解候補命題からは集約前の Web ページにリンクを張られており、ユーザがクリックすることで Web ページを確認することができるように実装している。

最終的に解候補命題の発生頻度がユーザに提示される。入力されたフレーズの頻度と解候補命題の頻度を比較することで、ユーザはどの命題がどの程度 Web 上で支持されているかを知ることができる。

3. 命題の生成時期の分析

2章で提案した手法によって収集された Web ページはそれぞれ生成時期が異なる。よってそれを考慮せずに多数決によって求められた解候補命題の発生頻度を真偽判定に用いるのは適切ではない。例えば、「次の夏季オリンピックの開催都市は北京である」という命題について考える。この命題は夏季オリンピックが北京で行われるまでの4年間に限っては正しい命題である、しかし北京オリンピックが終わってしまった後に関しては正しくない。

この命題のターゲット T として「北京」をユーザが指定し、ほんと?サーチが解候補語を抽出することを考えよう。システムは「アトランタ」「シドニー」「アテネ」「北京」といったような解候補語を Web から抽出する。これら解候補語から解候補命題を表現するクエリが作成され、検索エンジンに投げられる。そして検索エンジンから得られた検索結果の上位 1000 件を分析する。オリンピックに関する話題は全世界のあらゆる人々が興味を持っていると考えられるので、検索エンジンはこれらクエリを含む Web ページを大量に見つけることが可能である。アテネオリンピックは特に注目されたオリンピックであったので他のオリンピックよりもそれを言及する Web ページが大量に見見することができる。現階

階ではシステムは命題の真偽の評価方法として単純な多数決を用いているので、「次の夏季オリンピックの開催都市はアテネである」という命題が最も頻度の高い命題であると結果を返す。

しかし実際はどの解候補命題も誤りではない。ユーザが命題の真偽について調べたいと思った時点（例えば 2007 年 5 月）では次の夏季オリンピックの開催都市は「北京」である、とも十分考えられるからである。ユーザの調べたいと思った時期によっては他の解候補命題（アテネ、アトランタ、シドニーなど）は正しいと考えられる。この例から命題の真偽は時間に強く依存することが分かる。

時間的な観点から命題の真偽に関する評価を行う場合、さらに別の側面から命題の真偽を分析することが可能である。つまり、ある命題が過去から現在にいたるまで継続して正しいかどうかを分析できる。例えば「川端康成がノーベル文学賞を受賞した」という命題を考える。川端康成がノーベル文学賞を受賞したのは 1968 年のことであるが、この事実は 1968 年から現在にいたるまで受け入れられている事実である。この種の命題を言及する Web ページは現在に至るまで定期的に発生していると考えられる。

以上のように、時間的な観点で分析を行うことは命題の真偽を Web 上の知識の多数決により決定する上では重要となる。我々は命題の真偽を決定する上で二つの基準を定義する。図にこの二つの基準を用いた分析を行う手順を記す。

3.1 ある時期における命題の信用度

本章では、ある時期における命題の信用度を命題の時間分布を比較することで評価する方法を提案する。

命題に関連する Web ページの生成時期の分布を分析するためには、各 Web ページが発生した時間を決定する必要がある。このために我々は Internet Archive を用いる [1]。Internet Archive を用いると、ある URL に関する過去のページのスナップショットリストを得ることができる。最も古いタイムスタンプをその Web ページの生成時期として見ると、ある命題を言及している Web ページの生成時期を調べることが可能となる。

次に時期 t における命題 A の PF_A (PF : Proposition Frequency) を以下のように定義する。

$PF_A(t)$: 時期 t における命題 A を言及する Web ページの発生件数

PF を用いることで、ある時期において解候補命題の中でどの命題がもっとも信頼できるかを評価することができる。すなわちどの命題が最も信頼できるかを知るには、時期 t における命題の PF を比較すれば良い。例えば命題 A と命題 B の $PF_A(t)$ と $PF_B(t)$ を比較し、 $PF_A(t)$ が $PF_B(t)$ よりも大きかった場合、時期 t においては命題 A の方が命題 B よりもより正しいと言える。ユーザが疑わしいと思った命題の $PF(t)$ とほんと?サーチの抽出した解候補命題の $PF(t)$ を計算しそれらを比較することで、時刻 t において最も $PF(t)$ の値の大きかった命題をその時期に最も信頼できる命題とすることができる。

3.2 命題の一貫性

ほんと?サーチではある時期に最も言及された命題を抽出する目的だけでなく、その命題が十分長い時間 Web で言及

された命題であるかどうかを分析する目的で時間的な分析を行う。そのように分析された情報は命題の信用度を判定する上で有益な情報である。例えば「アルミニウムがアルツハイマー病の原因である」という命題はある時期においては最もメジャーな学説であり Web でも頻繁に議論された命題であったが、現在はそうではなくなっている。一方で「アルツハイマー病は認知症を引き起こす」という命題は依然として Web では頻繁に言及されている。二つの命題の頻度の変化を提示することは両命題の信用度を判断するためには有用である。

二つの命題の違いは以下のように形式化することができる：長い期間信用度が持続する命題の場合、それを言及する Web ページは定期的に発生する。一方ある時期に限定して信頼できる命題の場合、その命題を言及する Web ページの発生数は命題が信頼できなくなると減少する。これら 2 つのケースを分離するために、心理学の理論に注目する。

Hermann Ebbinghaus によると、人間の記憶量は指数関数的に減少する。時刻 t における記憶量 R は係数 γ を用いて以下のように定義することができる[2]:

$$\frac{d}{dt}R(t) = -\gamma R(t)$$

エビングハウスの理論に基づくと、次のようなモデルが考えられる。ある命題を言及する Web ページのある月の発生件数が前月の発生件数の λ 倍を超えた場合、その命題は依然として人々が関心をもっていると判断できる。一方 λ 倍を超えなかった場合、その命題は関心が薄れていっていると判断できる。つまり人々はその命題に注目しなくなり、その結果社会から忘れ去られることになると判断できる。 λ は閾値であり、実験的に調整される。

我々は命題がどれだけ長い期間人々の関心を惹き付けたかを示す指標として命題の一貫性を定義する。命題を言及する Web ページの発生件数が社会における命題への関心度を反映していると仮定する。

$PF_A(t)$ はある時期 t における命題 A の一貫性を表すものとして以下のように定義される (PC : Proposition Continuity)。

ある時期 t における命題 A の一貫性

$$PC_A(t) = \begin{cases} PC_A(t-1) + PF_A(t) & \text{if } PF_A(t) \geq \lambda PF_A(t-1) \\ \alpha PC_A(t-1) + PF_A(t) & \text{if } PF_A(t) < \lambda PF_A(t-1) \end{cases}$$

λ は命題が忘却される段階を発見するための閾値である。 α は PF が大幅に減少した際に PC の値を指数関数的に減少させるための係数である。ある時期に命題 A に関する新たに発生した Web ページに件数が前月のその値の λ 倍より大きい場合、 $PC_A(t)$ は $PF_A(t)$ だけ増加する。 λ 倍より小さい場合には、その命題は忘却される段階に入ると判断し $PC_A(t)$ の値を休息に減らしていく。

ほんとは?サーチでは、 PC の値を命題が社会でどの程度一貫したものとして受け入れられているかを表す指標として提示する。

4. 実験

本章では時間的観点から命題の信用度を評価する手法の効果測定の結果を記す。

4.1 解候補命題の発見と Web 情報の集約

解候補命題を得るために、Yahoo! Web Search API を用いて Yahoo! のインデックスを検索しその検索結果の上位 1000 件を分析した[3]。2 章で記した手法を用いて検索結果のスニペットから解候補語を抽出した。その際、検索結果中における各解候補語の頻度を求め、最も大きい頻度の 15% 以下の解候補語は除外した。これは既にこの時点で頻度の低い候補語は適切ではないと考えたからである。

解候補語から解候補命題を作成した後、形態素解析器 Mecab を用いて解候補命題から名詞、動詞、形容詞を抽出し解候補命題を表すクエリを作成した[4]。最終的にその各クエリを用いて検索し、得られた検索結果上位 1000 件を 2 章で述べた手順を用いて集約した。

実験例として「中国の胡錦涛国家主席」という命題を挙げると、ターゲット T としてはそれぞれ「胡錦涛」を指定した。

表 1: 命題の言及度評価

Table 1: Estimation of the Trustworthiness of Proposition

“中国の 胡錦涛 国家主席”	
解候補語	頻度
胡錦涛	589
江沢民	574

表 1 は入力された命題と Web 検索の結果から得られた解候補命題の頻度を示している。「中国の胡錦涛国家主席」という命題からは「中国の胡錦涛国家主席」「中国の江沢民国家主席」という解候補命題が得られる(胡錦涛は 2003 年まで国家主席)。表 1 から「中国の胡錦涛国家主席」という命題は最も頻度が高いのでその命題は信頼できる。これら 2 つの命題は時間的な観点から分析していない単純な評価である。

4.2 Web ページの生成時期の分析

3 章で提案した手法に基づき、ユーザが入力した命題と解候補命題の PF と PC を計算し、時間的観点から信用度を評価する。

Web ページの生成時期の測定には Internet Archive を用いた。実験では 1998 年から 2006 年に生成された Web ページを対象にした。Internet Archive では Web ページは収集して 6 ヶ月経過した後に閲覧可能となる。よって 2006 年 6 月までのデータを評価対象にした。 PC の計算には $\lambda = 0.8$, $\alpha = 0.5$ を用いた。

「中国の江沢民国家主席」という命題の PF 値はグラフの前半部分では「中国の胡錦涛国家主席」という命題の PF 値よりも高いが、2003 年頃に順位が入れ替わったことが図 1 より分かる。実際江沢民は 2003 年 5 月まで中国の国家主席であったが、それ以降は胡錦涛が国家主席である。

図 2 からは「中国の胡錦涛国家主席」という命題の PC 値は増え続けているものの、「中国の江沢民国家主席」の PC 値はある時点から減り始めており、国家主席が替わったことが分かる。

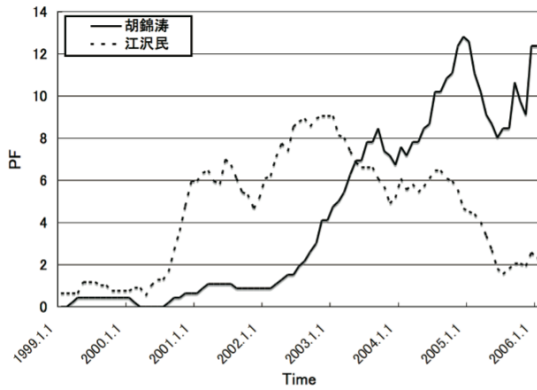


図 1 : PF の分布 : 命題「中国の胡锦涛国家主席」
Figure 1 : Proposition Freq. for the Proposition, “the President of China is Hu Jintao.”

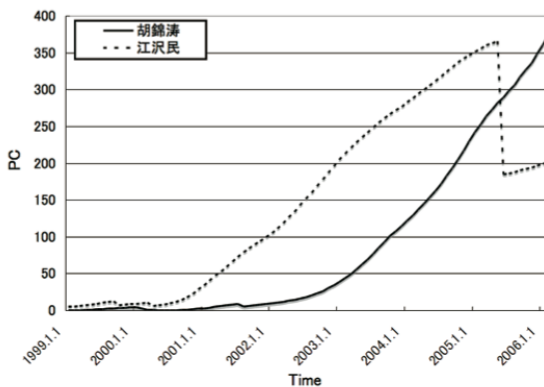


図 2 : PC の分布 : 命題「中国の胡锦涛国家主席」
Figure 2 : Proposition Cont. for the Proposition, “the President of China is Hu Jintao”

5. まとめと今後の課題

命題の真偽を調べたいときに、その信用度を評価するシステムは現状ではない。それゆえ我々は検索エンジンの結果を集約し、Web アーカイブを李様子津事で、Web 上にある情報を集約し時間的な観点から命題の信用度を評価する手法を提案した。

Web ページの発生した時期を分析することで、ある時期において命題が正しいか否か、過去から現在に至るまでその命題が正しいか否かを決定することができた。我々のアプローチの問題は Web ページを収集する時点で肯定的な意見か否定的な意見かを区別していない点、Web ページの生成時期の分析が Internet Archive に依存しており、それが一度停止するとページの発生時期を正確に決定することが難しくなる点が挙げられる。加えて 3 章で述べた解候補語自体も Yahoo! Search を行う時点で時間的な影響を受けてしまっており、検索結果上位 1000 件が最近の Web ページである場合はより古い解候補語を得ることができない点が挙げられる。

我々の提案手法は知識発見の一種である。Web 上にある知識の集約は命題の信用度を評価するだけでなく他の問題にも適用できると考えられる。今後の課題としてノイズを除去し、Web ページ自体の信用度を評価することが挙げられる。

【謝辞】

本研究の一部は、文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」(リーダー: 田中克己, 平成 14~18 年度), 文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」, 計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者: 田中克己, A01-00-02, 課題番号 18049041), 平成 18 年度科研費若手研究(B)「ウェブ活用のための情報統合による信頼性判断支援」(課題番号: 18700086, 代表: 手塚太郎), 文部科学省科学研究費補助金若手研究(B)「情報検索とウェブアーカイブにおけるマイニング」(課題番号: 18700111, 代表: Adam Jatowt) によるものです。ここに記して謝意を表すものとします。

【文献】

[1] Internet Archive. <http://www.archive.org/web/web.php>
 [2] Ebbinghaus, H., Memory: A Contribution to Experimental Psychology, Thoemmes Press, 1913.
 [3] Yahoo! Web Search APIs <http://developer.yahoo.com/search/web/V1/contextSearch.html>
 [4] 日本語形態素解析器 Mecab, <http://mecab.sourceforge.jp>
 [5] Andrenucci, A. and Sneiders, E., Automated Question Answering: Review of the Main Approaches, Third International Conference on Information Technology and Applications (ICITA 2005), pp. 514-519, 2005.
 [6] Kleinberg, J., Bursty and Hierarchical Structure in Streams, Data Mining and Knowledge Discovery, Vol. 7 Iss. 4, Kluwer Academic Publishers, 2003.

山本 祐輔 Yusuke YAMAMOTO

京都大学大学院情報学研究科博士前期課程在学中。2006 京都大学工学部情報学科卒業。情報検索、データマイニングの研究・開発に従事。日本データベース学会学生会員。

手塚 太郎 Taro TEZUKA

京都大学大学院情報学研究科社会情報学専攻助教。2005 京都大学大学院情報学研究科社会情報学専攻博士後期課程修了。博士(情報学)。主に地域情報検索システム、ウェブからの知識発見、検索システムの教育への応用の研究に従事。情報処理学会、日本データベース学会各会員。

アダム ヤトフト Adam JATOWT

京都大学大学院情報学研究科社会情報学専攻助教。2005 東京大学大学院情報理工学系研究科電子情報学博士後期課程修了。博士(情報学)。主にウェブ検索、ウェブアーカイブマイニングの研究に従事。ACM 会員。

田中 克己 Katsumi TANAKA

京都大学大学院情報学研究科社会情報学専攻教授。1976 京都大学大学院修士課程修了。博士(工学)。主にデータベース、マルチメディアコンテンツ処理の研究に従事。IEEE Computer Society, ACM, 人工知能学会、日本ソフトウェア科学会、情報処理学会、日本データベース学会等各会員。