3C2-03

Webページの信頼性の自動推定

Automatic Estimation of Web Page Credibility

福島 隆寛[*1] 内海 彰[*2]
Takahiro Fukushima Akira Utsumi

[*1]電気通信大学大学院電気通信学研究科システム工学専攻
Department of systems Engineering,The University of Electro-Communication

[*2]電気通信大学電気通信学部システム工学科
Department of systems Engineering,The University of Electro-Communication

This paper proposes an automatic estimation method of Web page credibility. The proposed method automatically judges whether a Web page satisfies each of 31 factors selected out of the 55 ones which determine Web credibility in Fogg et al's previous study, and then calculates the degree of credibility of that page using the result of judgment. This paper also reports a Web search support system which uses the proposed method, and the effectiveness of the system and the proposed method.

1. はじめに

一般的に Web の情報は書籍と比べ信頼性が乏しいと認識されているのが現状である．しかし，Web は情報の即時性という点で書籍に勝ることも事実である．そのため Web 検索によって得られる情報の信頼性を知ることは重要である．

しかしながら Web の信頼性を自動推定する研究は行われていない．Fogg ら [Fogg 01] の Web サイトの信頼性の調査研究や，ヴェラヤサンら [ヴェラヤサン 04] の Web ページの相対信頼度の提案はあるが，計算機への実装はされていない．現在，相対的に信頼できるサイトを並べて表示を行う検索エンジン TEOMA（http://www.teoma.com/）がある．これは Google （http://www.google.co.jp/）のページランクの手法を一歩押し進めたもので，検索テーマと同種のサイト間でのリンク数によって信頼性のページのランク付けを行っている．しかし，Web ページごとに信頼度を表示しないため，各々のページがどのくらい信頼できるか判断できないという問題点がある．そこで本研究では，文献 [Fogg 01] の改訂版である Fogg ら [Fogg 02] の調査結果で明らかになった Web サイトの信頼性を決定づける 55 尺度のうち自動推定が可能と思われる 31 個の尺度を用いた Web ページの信頼性を推測するシステムを提案する．

2. Webページの信頼性の定義

本研究では記載されている情報の真偽ではなく，ユーザが正しいと認識してしまう，あるいは鵜呑みにしてしまう情報を提供している Web ページを信頼できる Web ページと定義する．

3. 信頼性の推定手法

推定対象の Web ページに対して，表 1 に示す 31 個の尺度のそれぞれが成立するかどうかを，表 1 に示した処理により判定する．そして成立すると判定された尺度の加点の総和をその Web ページの信頼度とする．各尺度の加点は Fogg ら [Fogg 02] の調査研究で明らかにされた各尺度の重要度に対応している．

なお，Fogg ら [Fogg 02] はこれらの尺度を「専門的」,「信頼的」,「スポンサー的」,「その他」の 4 つの要因に分類しているが，本研究の検索支援システムでは信頼度計算の際にこれらの各要因を考慮するかしないかを選択することもできる．

連絡先: taka@utm.se.uec.ac,jp

4. WWW 検索支援システムの概要

3 章で提案した信頼性推定手法を用いた WWW 検索支援システムの概要を以下に示す．

1. ページの取得：ユーザからクエリの入力を受け取るとシステムは Google を利用して Web ページを取得する．その際，PDF ファイルは除かれる．
2. 本文の抽出：取得したページから改行，フォント等の文中に多く含まれるタグ，JavaScript およびスタイルシートを除去する．続いて，残ったタグに挟まれた文章の文字数を数え，一番多い文字数の文章を本文と見なす．
3. 解析：2 章で述べた方法を用いて各 Web ページを解析し，信頼度の推定を行う．
4. ジャンル分類：Web ページをテキスト系ページ，画像や FLASH 系のページ，ポータルサイトやリンク集系のページ，ウェブログや掲示板系のページ，オフィシャルページやデータベース系のページごとに分類する．
5. 出力：解析結果をもとにして信頼度が高い順にジャンルの名前とページタイトルを出力する．

5. 評価実験

5.1 WWW 検索支援システムの評価

WWW 検索支援システムを学生 6 人に使用してもらいアンケートに答えてもらった．アンケートでは，各クエリに対して表示された Web ページの信頼度全体に対して，5 段階評価（2：「信頼性が高そうなページから順に表示されている」，−2：「信頼性が高そうなページから順に表示されていない」）を行ってもらった．まず評価 1 として実験者が指定した「清涼飲料水」,「小泉純一郎」,「ミラクル」,「電波」,「すごい」を各クエリとした評価を行った．次に評価 2 として被験者自身が自由にクエリを 5 個選んで評価をしてもらった．続いて評価 3 では信頼度の計算にどの要因を考慮するかしないかを自由に

表 1: 信頼性の推定手法

尺度	加点	処理
専門的要因		
記事ごとに著者の表示がある	+1.3	本文を句点で分割したときの末文を茶筌を使用して形態素解析を行い，人名があれば成立とする．
引用や参考文献を表示している	+1.3	本文中で「引用」,「参考文献」，”『〜』によれば ”等の文字列が，1 つでも存在すれば成立とする．
検索機能がある	+1.2	検索，Search 等の名前のボタンを探し，1 つでも存在すれば成立とする．
記事にレイティングやレビューがついている	+0.7	コメント，トラックバック等の名前のリンクを探し，1 つでも存在すれば成立とする．また，Amazon 等のレイティングで有名なサイトの URI と，推定する対象の Web ページの URI がマッチングすれば成立とする．
以前訪問したことを覚えている	+0.4	ブラウザの履歴ファイルに残っている URI と，推定する対象の Web ページの URI がマッチングすれば成立とする．
情報ソースの根拠のない情報を提供している	-0.5	本文中で「引用」,「参考文献」，”『〜』によれば ”等の文字列が，1 つも存在しなければ成立とする．
タイプミスがある	-1.3	茶筌を使用し形態素解析を行い，未知語と判定されたものがあり，かつ一度しか出現しない語を探し，1 つでも存在すれば成立とする．
機能しないリンクがある	-1.4	Web ページ内の JavaScript 以外のすべての URI に対してリンク切れがあるかを調べ，1 つでも存在すれば成立とする．
サイトがアクセスできないときがある	-1.3	Web ページを取得するとき，最大で 2 度ダウンロードを行う．1 度目は失敗し，2 度目で成功した場合にのみ成立とする．
信頼的要因		
以前に利用して有益だと分かっていた	+2.0	ブラウザの Bookmark ファイルに含まれる URI と，推定する対象の Web ページの URI がマッチングすれば成立とする．
実世界の物理的住所を表示している	+1.7	郵便番号と都道府県名が連なっている文字列を探し，1 つでも存在すれば成立とする．
電話番号を表示している	+1.6	数字 10 桁，あるいは 11 桁とハイフンが 1，2 つの文字列を探し，1 つでも存在すれば成立とする．
電子メールアドレスを表示している	+1.5	HTML のリンクでメールを送ることができる mailto という文字列を探し，1 つでも存在すれば成立とする．
外部のサイトやリソースへリンクしている	+1.2	本文と本文抽出以前の段階の Web ページを比較し，本文の前後にサイト外へのリンクが存在すれば成立とする．
URL のドメインが「.org」で終わる	+0.3	対象とする Web ページの URI に「.org」が存在すれば成立とする．
スポンサー的要因		
各ページに一つ以上の広告がある	-0.6	対になるコメントタグを見つけ，その間に挟まれたリンクが画像であり，開くページの先がサイト外であるとき，あるいはリンク先がサイト外で別ウィンドウが開く仕組みが 1 つでも存在すれば成立とする．
自動的にポップアップ広告が表示される	-1.6	クリックするとポップアップするリンクがあり，その Web ページの URI がサイト内のものが 1 つでも存在すれば成立とする．
広告とコンテンツを区別できない	-1.9	ブラウザのステータス行に文字列を表示させる JavaScript の window.status が 1 つでも存在すれば成立とする．
その他の要因		
プロフェッショナルなデザイン	+1.5	FLASH が存在する，あるいは JavaScript とスタイルシートが共に存在していれば成立とする．
個人情報保護ポリシーが明示されている	+1.2	個人情報保護ポリシーのマークを発行しているサイトへ画像のリンクが貼られている，あるいはリンクの名前に個人情報保護等が含まれていれるものが 1 つでも存在すれば成立とする．
競合サイトへリンクしている	+1.0	Google での検索結果として得られる URI のどれかにリンクが貼られているかどうかを見て，1 つでも存在すれば成立とする．
印刷しやすいページデザインである	+1.0	フレームと FLASH が存在しなければ成立とする．
代表者とライブチャットができる	+0.6	チャット，chat という文字列が，1 つでも存在すれば成立とする．
検索結果の最初のページに表示される	+0.6	対象の Web ページが Google で検索された結果の上位 10 ページに含まれれば成立とする．
サーチエンジンのトップに表示される	+0.5	対象の Web ページが Google で検索された結果のトップに表示されれば成立とする．
登録やログインが必要	-0.1	Blog，CGI 以外の Web ページでパスワードを入力するフォームが 1 つでも存在すれば成立とする．
商業目的のサイトである	-0.3	取得したホスティングされていないページの URI が「.com」または「.co.jp」にマッチングすれば成立とする．
第 3 者にホスティングされている	-0.4	ホストの URI 辞書に存在すれば成立とする．
ダウンロードに時間がかかる	-1.0	Web ページを取得する直前から直後までの時間をはかり，1 分以上ならば成立とする．
会社の名前とサイトのドメイン名がマッチしていない	-1.1	URL からサードレベルドメインを取り出し，ローマ字読みの要領でカタカナに変換する．Web ページ内にサードレベルドメインかそれのカタカナに変換した文字列が存在しなければ成立とする．
ほとんど新しいコンテンツが追加されない	-1.7	ヘッダから最終更新日を取得し，新しい Web ページから順に並べ変える．その際，前後のページの更新日の差を計算する．対象とする Web ページが一番差の大きいページより古いページであれば成立とする．

設定することを許可した上で，自由に使用してもらい評価をしてもらった．なお，WWW 検索支援システムとして Web に公開したため「以前に利用して有益だと分かっていた」と「以前訪問したことを覚えている」の 2 つの尺度を信頼度計算から外した．

その結果，評価 1 の評定値の平均は「清涼飲料水」が 0.83,「小泉純一郎」が 1.67,「ミラクル」が 0.33,「電波」が 1,「すごい」が 0 となった．評価 2 は平均 0.79 となった．評価 3 は平均 0.67 となった．

5.2 各尺度の評価

茶筌の辞書からランダムに選んだ 500 個の語をクエリによる結果に対して検索を行い，4771 ページを取得した．その中から 1000 ページ以上で成立すると判断された尺度について，システムによる判断と人手による判断（正解）を比較した．評価には，以下の再現率および適合率を用いた．

$$\text{再現率} = \frac{\text{システムが正しく尺度を抽出したページ数}}{\text{尺度を含むページ数}}$$

$$\text{適合率} = \frac{\text{システムが正しく尺度を抽出したページ数}}{\text{システムが尺度を抽出したページ数}}$$

表 2: 頻出する尺度の再現率と適合率

尺度	再現率	適合率
引用や参考文献を表示している	0.31	0.83
検索機能がある	1	1
情報ソースの根拠のない情報を提供している	0.95	0.57
各ページに一つ以上の広告がある	0.74	0.93
機能しないリンクがある	0.88	1

6. 考察

システムの評価では，全ての評価において平均が 0 以上の結果となった．特に，どちらかといえば，抽象的なクエリよりも具体的な名詞のほうが高い評価を得られたようである．高評価と低評価のクエリを実際にシステムを使って検索してみると，高評価の場合に比べて低評価の場合のほうが得られた Web ページのジャンルが多様であり，それが実験参加者の評価に影響したと考えられる．しかしながら，本来の主旨である信頼性の自動推定を応用した検索支援の面から見れば，この結果はシステムの目的を十分に果たしているといえる．

一方，表 2 に示した各尺度の評価の結果を見ると全体的に再現率，適合率ともに高いが「引用や参考文献を表示している」という尺度のみ，再現率が低いことがわかる．これは抽出パターンでは抽出できない引用や参考文献の表示が多いためと考えられる．

7. おわりに

本研究では Web ページの信頼性の自動推定法とそれに基づく WWW 検索システムを提案し，おおむね良好な結果を得た．今後の課題として，Fogg らの調査研究の対象は欧米人であり，日本人には必ずしも当てはまらないと考えられるので独自に Web ページの信頼性の調査を行う必要がある．また，ウェブログやポータルサイト等のジャンルごとに信頼性の尺度を切り替える手法の提案を考えている．

参考文献

[Fogg 01] B.J. Fogg et al.:What makes a Web site credible? A report on a large quantitative study,*Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems*, Vol. 1, pp.61-68, New York: ACM Press (2001).

[Fogg 02] B.J. Fogg et al.:*Stanford-Makovsky Web Credibility Study 2002 Investigating what makes Web sites credible today*:
http://captology.stanford.edu/pdf/Stanford-MakovskyWebCredStudy2002-prelim.pdf

[ヴェラヤサン 04] ヴェラヤサン ガネサン, 山田 誠二:Web ページの相対信頼度:第 19 回人工知能学会全国大会，3F1-05 (2004).