3F1-05

Web ページの相対信頼度

A Computational Method to Detect the Credibility of a Web Page on the Internet

ヴェラヤサン ガネサン[*1 *2] 山田 誠二[*2]
Ganesan Velayathan Seiji Yamada

[*1] 総合研究大学院大学 [*2] 国立情報学研究所
The Graduate University for Advanced Studies National Institute of Informatics

In this paper, we will propose a method to calculate the credibility of a web page on the Internet. Due to information explosion and information disorder on the net, it is getting harder and harder to locate high quality information. However, Internet is one of the best ways to access newly fresh information. Over in this paper we will propose some rules to detect the information credibility of a web page.

1. はじめに

インターネットの普及に伴い個人でも簡単にWebページの作成し情報発信ができるようになってきた. Web サイトは誰でも閲覧できると同時に,誰もが公開できるため,情報爆発を引き起こしてしまっている.

インターネットにおいて, テレビや新聞とは異なり, 公開前に編集者によって内容の確認がされていないため,正しいものもあれば,誤ったものもある. インターネットで得た情報はすべて正しいと信じることはできない [AllAbout02] . インターネット上に提供されているどの情報を信用するべきかどの情報を信用に値しないかを的確に把握する必要がある.

現在の技術では,自然文で記述している文書を解析し信頼度を算出するのは困難であるが, インターネットの特徴を考慮に入れて,Web ページとその周りの属性を解析することによって信頼度を算出することができると考えられる.

ユーザの常識として,その情報がどれくらい信頼できるモノか, 最低限のチェックする必要がある. その部分を computational に行い,ユーザに一目で理解できるようなシステムを作成し提供するのが本研究の目的である.

2. 情報の信頼度

2.1 研究背景

情報の信頼度または信憑性(Information Credibility, Credibility of Information)とは, どれだけその情報が信用できるかを表す基準である. 現在において,被験者実験による情報の信憑性に関する研究[Fogg 01][Fogg 02a]がある. しかし,この問題に対して computational なアプローチを用いた研究や提案はされていない.

そこで,本研究では情報の信頼度を computational に行い,ユーザに一目で理解できるようなシステムを作成し提供することを目的にしている. 本報告において,computational に求める為のルールの設定について報告する.

2.2 出版物の評価方法

インターネットが登場する前に本や論文のような出版物の主な品質の評価方法として:

- 有名な出版社から出版されているかどうか
- 有名な人によって書かれているかどうか
- 第三者の評価が高かったかどうか

と言った方法が用いられてきた[Standler 03].

しかし,上記のような方法はインターネットのような膨大かつ管理されていないシステムには効果的ではない[Standler 03].

また,人間と人間の間の会話は,顔色・声のトーン・いわゆる,その場の雰囲気を分析しながら会話をしていくため,話を雰囲気と共に把握し,どの部分が重要かどの部分が冗談であるかを理解できる.

しかし,Web ページ上に存在する自然文から雰囲気など読み取るのは非常に難しい作業である.

2.3 関連研究

米国 Stanford 大学の Fogg ら[Fogg 01][Fogg 02a]がインターネットにおける情報の信憑性の要因について被験者実験により調査し, どのようなページや内容の表現が信憑性に影響を及ぼすかをクニカルレポートとして, 報告している.

さらに, Fogg らの研究によってまとめられた,「Stanford Guidelines for Web Credibility」[Fogg 02b]によって 10 のガイドラインが提案された. 主な内容として, どのように Web ページをデザインすれば,情報の提供の信頼度が上がるかに着目している.

この研究の根本的な問題点としては, 被験者実験による評価であるため, Web 環境全体を考慮していないのと, あまりにも一つのページに対する信頼度に偏っていることがあげられる.

3. インターネット上の情報の信頼度

インターネットのような中央管理者がいないシステムにおいて, 情報の管理を行うのは事実上不可能である. しかし,ある程度高品質の情報(信憑性が高い情報)を提供するようなサイトを把握することができる.

直接あるURLを見ているとき, もしそのURLは有名なサイトのものであれば(例えば, http://www.yahoo.co.jp/) 自然とそこに記載されている情報を我々は信じる. これは出版物の評価方法で言うと「有名な出版社から出版されている」という区分に分類できる. しかし, 情報によって必ずしも有名なサイト記載されているといえない. 有名ではないサイトに記載されている情報見ているとき,その情報をどれぐらい信じるべきかを判断するのは難しい.

連絡先:Ganesan Velayathan, 国立情報学研究所, 〒101-8430 東京都千代田区一ツ橋 2-1-2 大学院係

4. 本研究で提案する情報の信頼度の検出方法

Fogg[Fogg 02a]らが提案する情報の信憑性に関する研究の主な問題点は被験者実験で行うことを目的に設計されている為, 周辺属性をほとんど考慮に入れていないことにある. インターネットにおいて, ひとつのページを見るだけで,そのページの信頼度が得られると考えにくい.

そこで,本研究では情報の信頼度ページそのものとそのページの周辺属性を考慮に入れ, computational に求めることにする.

本研究で提案するシステムを簡略化するために, 一つのページを対象に説明する. そのページを Page A とする. Page A の信頼度を算出するためには以下のデータを分析する必要がある. なお, Fogg らが提案した Web Information Credibility Reports[Fogg 02a]の中から, computational で実装できる部分もルールとして加えている.

4.1 ページのトポロジーからの分析

インターネットの特徴を利用し, 分析を行う. 主にリンク構造の特徴を解析し,Webページの信頼度を算出する.

- PageA から信頼度を証明するページへリンクしている. (Verisign Secure Site シールプログラム)
- PageA は Blog ページからリファー・トラックバックされている.
- Page A は信頼できるサイトからリンクされている [Fogg02a][Page98] .
- PageA は信頼できる掲示板からリンクされる.
- PageA は信頼できるサイトにリンクしている [Page98] .

4.2 ドメイン名から得られる信頼度

Web ページが属しているドメイン自身が持っている信頼度を利用し, そのドメイン内に属している Web ページの信頼度を算出する.

- PageA は TLD(nTLD,gTLD,iTLD) (例, .com , .edu, .net, .gov, .mil, .jp, .co.jp, .ne.jp, .go.jp, .ac.jp) に属している. または,そこからリンクされている[DC03].
- 個人のフォルダにある.
- PageA は無料 Web サーバに提供されている.

4.3 ページの構成やデザインから得られる信頼度

タグ構造などページデザインに係るものが正しく利用されてデザインしているかどうかから Web ページの信頼度を算出する.

- Page A 内の全ての リンクが正しく機能している(リンク切れは無い)[Fogg02a] .
- PageA に「Sitemap」が存在している.
- PageA はひとつ以上の言語で用意されている [Fogg02a] .
- PageA Copyright 情報が存在している [Fogg 02a] .
- PageA に広告バナーが存在しない,または多くない.
- PageA にポップアップ広告がない.

4.4 ページの表現から得られる信頼度

ページ内の自然文の構成などを利用して信頼度を算出する.

- PageA の文書の中にタイプミスが無い.
- PageA の文書の中に単純な文法のミスは無い.

4.5 アクセス時間から得られる信頼度

ユーザの実体験するアクセス時間などから信頼度を算出する.

- PageA または Site A にはミラーが存在する.
- PageA は短時間で表示される.

4.6 提供しているサーバから得られる信頼度

提供しているサーバが正しくメンテナンスされていることを確認し, それを利用して信頼度を算出する.

- PageA のサーバのメンテナンスが正しく行われている.
- PageA のサーバのバグフィックスやパッチなど,セキュリティ面での対策が正しく行われている.
- PageA のサーバの Web ルートに robots.txt (/robots.txt)が存在していて, 正しく記述されている.
- PageA のサーバの whois が正しく機能している. (登録者情報が隠されていない)
- PageA のサーバが時間帯によってサービス停止していたりしない. (常に Web サービスが提供さている)
- PageA のドメインはバーチャルドメインではない,サーバはバーチャルサーバではない.

5. まとめ

本研究で,インターネット上の文献を読んでいるとき,その情報をどれだけ信用できるかを判断する基準の提案をした.

特に新技術の場合, 専門家であっても意見の対立が生じることがある. そのような場合情報の信頼度を一次元のものとして表すことができないのが本研究の問題点である.

今後, システムの作成, 被験者実験の実施と動作検証を行う.

参考文献

[Fogg 01] Fogg, B.J., Marshall, J. Kameda, T. Solomon, J., Rangnekar, A., Boyd, J. & Brown, B: Web Credibility Research: A Method for Online Experiments and Some Early Study Results, Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems, pp 295-296. New York: ACM Press,2001.

[Fogg 02a] Fogg, B.J., Kameda, T., Boyd, J., Marshall, J., Sethi, R., Sockol, M., Trowbridge, T. Stanford-Makovsky Web Credibility Study 2002: Investigating what makes Web sites credible today. A Research Report by the Stanford Persuasive Technology Lab in collaboration with Makvosky & Company. Stanford University, 2002.

[Fogg 02b] Stanford Guidelines for Web Credibility
http://www.webcredibility.org/guidelines/index.html

[Page 98] Lawrence Page and Sergey Brin and Rajeev Motwani and Terry Winograd: The PageRank Citation Ranking: Bringing Order to the Web.
http:// citeseer.ist.psu.edu/article/page98pagerank.html

[DC 03] How credible is the information from the Internet?
http://www.searchengines.com/credibleInfo1.html

[AllAbout 02] 情報の信憑性を意識していますか?
http://allabout.co.jp/career/soho/closeup/CU20020212E/

[Standler 03] Evaluating Credibility of Information on the Internet.
http://www.rbs0.com/credible.pdf