様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 6 月 26 日現在

研究種目： 若手研究（B）
研究期間： 2006～2008
課題番号： 18700086
研究課題名（和文） ウェブ活用のための情報統合による信頼性判断支援

研究課題名（英文） Decision Support for Reliability of Web Content using Information Integration

研究代表者
手塚 太郎（TEZUKA TARO）
立命館大学・情報理工学部・講師
研究者番号：40423016

研究成果の概要：
World Wide Web 上には大量の情報が蓄積されているが、その中には内容に関して信頼性が低いものも多く含まれている。そこで本研究ではコンテンツの内容自体に基づく信頼性評価手法の開発を行った。具体的にはテキスト情報の基本要素としての「文」に着目し、個々の文の信頼性を評価する手法を開発した。また、集約結果をページ評価に使用するだけでなく、それ自体を新たな情報として提供するシステムを開発した。

交付額

（金額単位：円）

	直接経費	間接経費	合 計
２００６年度	1,200,000	0	1,200,000
２００７年度	1,200,000	0	1,200,000
２００８年度	1,200,000	360,000	1,560,000
総 計	3,600,000	360,000	3,960,000

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：ウェブ，情報検索，信頼性，テキストマイニング

１．研究開始当初の背景

World Wide Web（以下、ウェブ）は急速に規模を拡大し、ウェブ上の情報を用いて日常生活に関わる判断を行うことも一般的となってきている。しかし、書籍やデジタルアーカイブを用いる場合と比較して、ウェブ上には不特定多数の記述者によって発信された情報が大量に存在し、信頼性の問題が常に伴う。ウェブ上に記述された情報に対する信頼性の問題は、いまだ解決されていない重要な課題である。多様な分野にまたがるウェブ上の情報の信頼性の判断をすべてシステムが完全に自動的に行うことは困難であり、最終的にはユーザ自身によって判断が行われる必要があると考えられるが、システムがその判断を支援することは可能である。現状では、検索エンジン（Google, Yahoo! 等）を用いてフレーズ検索（語彙列単位での完全一致による検索）を行った場合、キーワード検索（語彙単位での検索）と比較して、著しく少ない結果しか得られない。一方、ユーザクエリからいくつかの重要キーワードに抽出し、既存の検索エンジンに対して複数キーワード検索を行った場合、キーワードが文章中に離れて存在する場合も検索結果に含まれてしまい、適合率が著しく低下する。すなわち、どのような形で複数の検索語句が関連しているかの情報が考慮されていないため、命題の類似検索も容易ではない。

２．研究の目的

本研究では、命題に対する類似検索手法を発展させることで、ユーザによる信頼性の判断を円滑化するシステムを構築する。具体的には、ウェブコンテンツに対する信頼性の判断を支援するために、対象となる文章と内容的に類似する文章群をウェブ上から収集するメタサーチエンジンを構築する。

ユーザから送られたクエリを複数の検索エンジンに転送し、検索結果を集約して表示するメタサーチシステムは数多く実装されているが、信頼性判断のために命題レベルで集約を行うシステムはまだ一般的でない。本研究では、複数の検索エンジンから取得される文章に構文解析を行い、命題としての類似性を判定し、ユーザに提供するシステムを開発する。これらのシステムはメタサーチエンジンと自然言語処理モジュール、それを統括するコントローラの組み合わせによって実現される。

３．研究の方法

研究の遂行にあたって、実験の基礎となる自然言語処理 API の開発を行う。さらに、ウェブページを多数の検索エンジンから取得するメタサーチエンジンの実装を行う。

研究の遂行にあたり、基礎的な統計処理・自然言語処理を行うプログラム群を開発する。例として、単語の重要性に対する代表的な統計指標である tf-idf、係り受け解析、格フレーム判定等を行うモジュール群を作成する。

これらのプログラムを用いて多様な特徴量の抽出を実現し、その後のリランキングメカニズムの開発に利用する。

さらに、長期間に及ぶ研究期間に向けて、開発されたソフトウェアの再利用性を確保するため、ライブラリとしての整備を行う。

また、ウェブコンテンツの収集手法についても、メタサーチのみならず、多様な手法を検討する。特に、RSS フィード、検索エンジンが提供する API､独自クローラによる収集、ウェブアーカイブといった多様なソースの利用を検討する。

４．研究成果

平成 18 年度は、ウェブコンテンツに対する信頼性判断を支援するシステムの開発に向けて、基礎研究ならびにプロトタイプシステムの実装に取り組んだ。具体的には、ユーザが信頼性を判断したいと考えた自然言語文の入力に対し、クエリ自体ならびに類似するフレーズをウェブ文書中から収集し、その相対頻度や文書中での表現の不確実性を数値化して提示することで、ユーザによる信頼性判断を支援するシステムの開発を行った。また、システムの公開を通してログを収集し、求められる改善点等を明らかにした。

さらに、ウェブページの信頼性を評価する新しい尺度として、リンク元ページの空間的分布を利用する手法を開発した。具体的には、ユーザが指定したウェブページにリンクしているサイトの IP アドレスを地理空間上にマッピングすることで、その広がりや分散の度合いを計算し、信頼性判断の基準として提示する手法を提案した。多数のウェブページに対して実験を行い、ページの空間的支持の度合いに応じて異なる数値が得られることを確認した。加えて、提案手法によって得られた情報に基づき、リンク元ページの空間的分布を視覚的に提示するユーザインタフェースの実装も行った。

図 1：空間分布の視覚化インタフェース

平成 19 年度は、World Wide Web 上のコンテンツの信頼性を定量的に評価する手法の構築を目的として、テキストマイニング・ウェブ構造解析、ならびにウェブアーカイブを利用した研究を進めた。特に混合分布等の確率モデルを推定プロセスに導入することによって、学習に基づく新たな信頼性評価手法を構築した。具体的にはウェブから大量のテキストデータを収集し、集約によって概念レベル／命題レベルでの知識抽出を行った。さらに、集約されたデータの統計的頻度・空間的分布、ならびに時間変化等を特徴量として用い、得られた知識の信頼性を自動推定するシステムの開発を行った。確率モデルの導入はノイズの低減にも効果があり、従来手法に比べて評価尺度の精度向上が実現された。

ユーザが Web コンテンツの信頼性に関して的確な判断を行うためには情報源の多様性が重要であるという観点から、「情報ポートフォリオ」の概念を提唱し、プロトタイプシステムを実装した。開発された「コンフリクト・サーチ」においては、ユーザが入力した任意の命題に対し、Web 上でそれを肯定する主張と否定する主張とが併置して表示され、それぞれの主張を特徴づけるキーワードや両意見に共通するキーワードを一括的に閲覧することができる。

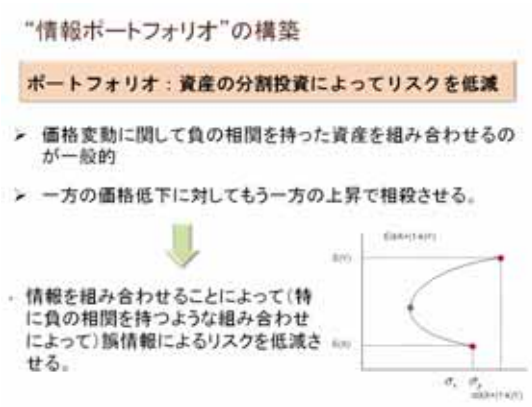

図 2：情報ポートフォリオ

図 3：コンフリクト・サーチ

平成 20 年度は、ユーザによるウェブコンテンツの信頼性判断を支援するため、メタサーチと集約を利用したプロトタイプシステムの開発を行った。ウェブコンテンツ・メタデータ・リンク構造等の解析によって得られる多様な尺度をユーザに提示することで、総合的な視点から信頼性の判断を促すシステムを実装した。

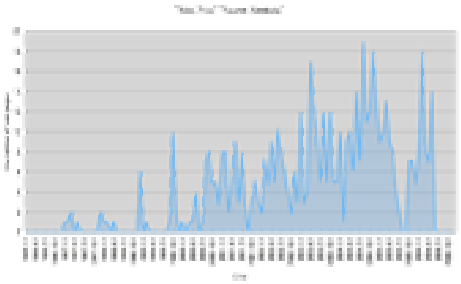

図 4：言及の時間変化の可視化

さらに、時間的および空間的視点を導入し、命題に関する言及の時間的変化、空間的分布を用いた信頼性評価手法を導入した。

また、今後ウェブの利用形態がページ単位から知識単位へと移行していくことが予想されるため、ページ単位での信頼性だけでなく、知識単位での信頼性を評価する手法に関して基礎研究を進めた。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

① 手塚太郎，近藤浩之，田中克己，混合ガウス分布を用いたウェブコンテンツの地域性推定とオブジェクトレベルローカルサーチ，情報処理学会論文誌（トランザクション）データベース，Vol.1，No.1，pp.13-25，2008 年 6 月，査読有り．

② 山本祐輔，手塚太郎，Adam Jatowt，田中克己，WebAlert:Web 情報の印象集約を利用した閲覧ページ内容に対する反対意見提示，日本データベース学会論文誌，Vol.7，No.1，pp.251-256，2008 年 6 月，査読有り．

③ 山本祐輔，手塚太郎，アダム ヤトフト，田中克己，ページ特性を考慮した Web 検索結果の集約とページ生成時間分析による知識の信頼性評価，電子情報通信学会和文論文誌 D「データ工学特集号」，Vol.J91-D，No.3，pp.576-584，2008 年 3 月，査読有り．

④ 服部俊，手塚太郎，田中克己，文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出，情報処理学会論文誌（トランザクション）データベース，Vol.48，TOD34，pp.69-82，2007 年 6 月，査読有り．

⑤ 山本祐輔，手塚太郎，アダム ヤトフト，田中克己，ほんと?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，日本データベース学会 Letters，Vol.6，No.1，pp.53-56，日本データベース学会，2007 年 6 月，査読有り．

〔学会発表〕（計４件）

① Taro Tezuka，Hiroyuki Kondo，Katsumi Tanaka，Estimation of Geographic Relevance for Web Objects using Probabilistic Models，Proceedings of the 8th International Symposium on Web and Wireless Geographical Information Systems，pp. 124-139，Shanghai，China，2008，査読有り．

② Yusuke Yamamoto，Taro Tezuka，Adam Jatowt，Katsumi Tanaka，Supporting Judgement of Fact Trustworthiness Considering Temporal and Sentimental Aspects，Proceedings of the 7th International Conference on Web Information Systems Engineering (WISE

2008), pp. 206-220, Auckland, New Zealand, 2008, 査読有り.
③ Shun Hattori, Taro Tezuka and Katsumi Tanaka, Mining the Web for Appearance Description, Database and Expert Systems Applications - Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007), Lecture Notes in Computer Science 4653, pp. 790-800, Springer-Verlag, August 2007, 査読有り.
④ Yusuke Yamamoto, Taro Tezuka, Adam Jatowt and Katsumi Tanaka, Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis, Proceedings of the Joint Conference of The 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management (APWeb/WAIM 2007), Lecture Notes in Computer Science 4505, pp. 253-264, June 2007, 査読有り.

６．研究組織

(1)研究代表者

手塚太郎（TEZUKA TARO）
立命館大学・情報理工学部・講師
研究者番号：**40423016**