様式 C-7-2

自己評価報告書

平成２１年５月２５日現在

研究種目：特定領域研究
研究期間：２００６～２０１０
課題番号：１８０４９０６９
研究課題名（和文）情報爆発時代の情報検索基盤技術
研究課題名（英文）Fundamental Technology for Linkage and Retrieval of Exploding Information

研究代表者
安達　淳（ADACHI JUN）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：８０１４３５５１

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，知能情報学
キーワード：情報検索，情報リンケージ，大規模コーパス，テキスト処理，機械学習

１．研究計画の概要

本研究は、インターネット上で公開される各種テキストや個人・組織が管理する文書を対象として、関連する情報を結び付ける「情報リンケージ」プラットフォームを構築することを目的としている。これにより、無秩序に拡大する大量情報の中からその場で自分の必要とする情報を的確に取り出し、わかりやすく提示する新たな情報検索の枠組みを考案することを目指している。

本研究では、特に本・人物・製品といった具体物（モノ）および台風や選挙といった事件（イベント）を対象とし、大量の情報の中に現れる同一のモノやイベントをリンクづけるとともに、それらのリンクづけられた情報を検索し提示するシステムの研究開発を進めている。具体的には、下記に示す課題に取り組んでいる。

（１）リンケージエンジンの構築
効率的に大量データのリンケージを実現するためのインデキシングおよび分散処理法の研究とそれらを統合したエンジンの開発。

（２）マッチングモデルの提案
多様な情報源から得られるモノやイベント情報を適応的にマッチングするためのモデルと効率的な処理アルゴリズムの考案。

（３）モノ・イベントリンケージ
人物や書誌といった具体的なモノや新聞記事等に現れるイベントを高精度にリンクづける手法の研究。

（４）アカデミックリンケージシステム
学術情報を対象とした大規模なリンケージシステムの研究開発および研究者を中心としたデータ再構築の手法提案と実装。

２．研究の進捗状況

（１）リンケージエンジンの構築
大規模データを処理するためには、個々のアルゴリズムの効率化に加えて分散処理により効果的に計算資源を活用することが必要になる。本研究では、計算資源を柔軟に活用するために、自律性の高い分散処理の枠組みである P2P システム上に情報検索を行う方法を提案した。提案手法は、符号理論の手法を用い、語の出現頻度に基づいてピアに文書を割り当てることによって、負荷分散をはかり、大幅に効率を向上させた点に特徴がある。

（２）マッチングモデルの提案
モノやイベントに関する情報は、文書やレコードなど様々な形式で記述されている。本研究では、このような状況に対応するために、木構造で表わされるデータのマッチングモデルの研究を行った。木構造データのマッチングはコストの高い計算を必要とするため、本研究では、従来のアルゴリズムと比較して同程度の近似度で計算量の低い近似マッチングアルゴリズムを考案した。また、データに即した高精度なマッチングを行うために、統計的な木マッチングモデルの学習法を提案した。

（３）モノ・イベントリンケージ
特に、論文情報や人物といったモノに関する研究を進めてきた。人物は、学術文献や Web などに現れる重要な情報であるが、そのリンケージのためには、同姓同名の人物を区別する必要がある。本研究では、各人物に関連するトピックをその人物の記述を含む文書中から抽出することによってリンケージ性能を高めることができることを示した。

（４）アカデミックリンケージシステム

大規模な書誌データベースのリンケージをオンラインで行うための要素技術を開発し、１千万件を超える論文の柔軟な同定をリアルタイムで実現する同定エンジンのプロトタイプシステムをウェブ上にて公開した。また、研究者を単位とする情報の再構築を目的として、開発したエンジンにより数千万規模の同定を行い、国立情報学研究所の新サービス（http://seikaplus.nii.ac.jp）の論文リンクを作成した。

３．現在までの達成度

全体としては、②おおむね順調に進んでいる。

（１）リンケージエンジンの構築

自律性の高い分散環境での情報検索技術を提案済みでおおむね順調に進んでいる。

（２）マッチングモデルの提案

順序木のマッチングモデルおよびアルゴリズムの構築済で計画どおりに進んでいる。

（３）モノ・イベントリンケージ

モノリンケージについては、すでにリンケージシステムとして実システムに組み込まれており計画以上に進んでいる。イベントリンケージについては、若干遅れぎみであるが、残りの期間で重点的に進める予定である。

（４）アカデミックリンケージシステム

書誌リンケージについて、明示的なフォーマットを持たない文字列や自動文字読み取り装置（OCR）の出力など誤りを含む文字列に対しても、柔軟で高精度な同定を高速で行うエンジンを実装し、複数のサービスや外部プロジェクトにデータ提供を行うなど、計画通りに進んでいる。

４．今後の研究の推進方策

下記のとおり各課題の研究を推進するとともに、全体を統合し情報リンケージ基盤として取りまとめる。

（１）リンケージエンジンの構築

大規模データを効率的に処理するためのインデキシングの技術を中心に研究を進める。

（２）マッチングモデルの提案

様々な情報源から得られる情報のリンケージには無順序木のマッチングが有効であるが、計算コストが高いため、効率的な無順序木のマッチングの研究を中心に進める。

（３）モノ・イベントリンケージ

イベントリンケージ問題を中心に取り組む。イベントの抽出およびリンケージを行うために、固有名抽出や隠れトピックモデルの技術を適用し特徴を抽出するとともにクラスタリングを行うことにより、イベントリンケージを実現することを計画している。

（４）アカデミックリンケージシステム

数万人、数千万件規模の情報に対するリンケージを目標として、これまで開発した要素技術を統合した大規模リンケージ基盤を構築し、その総合性能を評価する。また、リンケージにより得られる確信度つきの統合データを利用した、俯瞰的な統計分析の枠組を検討する。

５．　代表的な研究成果

（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 11 件）　（すべて査読有）

① A. Aizawa, A. Takasu, D. Fukagawa, M. Takaku, J. Adachi:” Academic Linkage: A Linkage Platform for Large Volumes of Academic Information”, Progress in Informatics, No.6, pp.41-47, 2009. 査読有

② T. Akutsu, D. Fukagawa, A. Takasu: “Approximate Tree Edit Distance through String Edit Distance”, Algorithmica, 掲載決定, 査読有

③ Q.M. Vu, A. Takasu, J. Adachi: “Improving the Performance of Personal Name Disambiguation Using Web Directories”, Information Processing and Management, Vol.44, pp. 1546-1561, 2008. 査読有

〔学会発表〕（計 30 件）

① H. Kurasawa, A. Takasu, J. Adachi: “Huffman-DHT: Index Structure Refinement Scheme for P2P Information Retrieval”, Intl. Symp. on Applications and Internet, 2008.8.1, Turku, Finland.

② V. B. Dang, Akiko Aizawa:”Multi-class named entity recognition via bootstrapping with dependency tree-based patterns”, Pacific-Asian Conf. on Knowledge Discovery and Data Mining, 2008.05.11, Osaka, Japan.

③ A. Takasu, D. Fukagawa, T. Akutsu: “Statistical Learning Algorithm for Tree Similarity”, Intl. Conf. Data Mining, 2007.10.29, Omaha, USA.

〔図書〕（計 0 件）

〔その他〕

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件

ホームページ

アカデミックリンケージ：

http://i-linkage.nii.ac.jp/cinii