様式 C-7-2

自己評価報告書

平成 22 年 3 月 31 日現在

研究種目：基盤研究（C）
研究期間：2007～2010
課題番号：19500098
研究課題名（和文）　Web アーカイブにおけるストリームマイニングに関する研究

研究課題名（英文）　Research of Stream Mining in Web Archives

研究代表者
河野　浩之（KAWANO HIROYUKI）
南山大学・情報理工学部・教授
研究者番号：70224813

研究代表者の専門分野：情報システム
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：デジタルアーカイブ，コンテンツ流通，評判モデル，Web アーカイブ，Web クローリング

１．研究計画の概要

本研究では，これまで研究計画者が国立国会図書館の非常勤調査員として関係してきたデジタルアーカイブ基盤の一部をなす WARP プロジェクトから得られた知見に基づき，データマイニングに関するアルゴリズムを応用し「デジタルアーカイブにおけるストリームマイニング指向ナビゲーション」技術確立を目指す．特に，デジタルアーカイブにおけるナビゲーション（閲覧）に関わる課題に焦点を絞り，アーカイブデータのアクセス傾向変化に着目したストリームマイニング技術を用いたナビゲーションインターフェース実装を行う．

２．研究の進捗状況

デジタルアーカイブに関わる研究の構成要素は，アーカイブ対象となるデータの「収集」「蓄積」「検索」である．以下，これらの項目に分けて研究実施計画を記す．これまで，平成 19 年度に解析したアーカイブシステムへのアクセス履歴や検索履歴パターンを利用していたが，2010 年 3 月に，国立国会図書館のデジタルライブラリーの利用状況を解析するログデータ解析を開始した．最終年度となる平成 22 年度は，解析により得られる参照特性に基づき，アーカイブ対象となる Web ページに対して参照特性の時間的変化に基づいたナビゲーションを行うアルゴリズムの実装を中心に研究の総括を行う．

(1) 収集：アーカイブ Web ロボットに関する課題

これまで，WARP システムならびにバルク収集環境において，アーカイブが困難なデータを保存対象に含めるための技術的可能性の検討を踏まえて，アーカイブデータ収集 Web ロボットプログラムである Hreritrix の利用を進めてきた．昨年度は，新規 WARP システムの研究調査で利用を検討したプロトタイプの貸与を受け，各種課題の検討を進めた．本年度は，引き続いてアーカイブアルゴリズムの改良を検討している．

(2) 蓄積：Web アーカイビングデータベースに関する課題

過去，参照頻度に基づく評価尺度を用いて，適切な保存デバイスとフォーマットを決定する階層型ストレージシステムにおけるファイル移動アルゴリズムを拡張してきた．本年度は，2010 年 3 月の近代デジタルライブラリー，WARP データへのアクセスログに基づいて，同様の評価予測を行い，長期アーカイブシステムの運用特性を議論する．

(3) 検索：Web アーカイブナビゲーションに関する課題

博物館データを実例に，デジタルアーカイブ用のデータ提供に有効なシステムの提案，その問題点を検証した．最終年度となる今年度も，引き続き，デジタルアーカイブデータのナビゲーション手法の研究を継続する．

３．現在までの達成度
②おおむね順調に進展している。
（理由）
単独の研究であり，本研究に直接する単著で査読付き論文を出版しており，また，同様に学会発表も期間中継続して実施している．なお，主たるデータ入手先である国立国会図書館デジタルライブラリ関連のシステム更新に伴うシステム開発状況ならびにデータ整備の時期の関係から，若干データ入手が遅くなったため分析に遅れている部分がある．

４．今後の研究の推進方策
これまで，平成 19 年度に解析したアーカイブシステムへのアクセス履歴や検索履歴パターンを利用していたが，2010 年 3 月に，国立国会図書館のデジタルライブラリーの利用状況を解析するログデータ解析を開始した．最終年度となる平成 22 年度は，解析により得られる参照特性に基づき，アーカイブ対象となる Web ページに対して参照特性の時間的変化に基づいたナビゲーションを行うアルゴリズムの実装を中心に研究の総括を行う．

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 6 件）

① H. Kawano, Towards Digital Archive Systems: Architecture and Design of Digital Museum Archives, Lecture Notes in Operations Research World Publishing Corporation 10, pp.14-21, 2009, 査読有

②伊藤 洋輔, 河野 浩之, P2P 環境における評判モデルを用いた公平性評価, 電子情報通信学会論文誌 D 電子情報通信学会, J91-D/3, pp.628-638, 2008, 査読有

③H. Kawano, Reputation-based Contents Crawling in Web Archiving System , Lecture Notes in Operations Research, Proc. of Operations Research and Its Applications, 7th International Symposium (ISORA'08), World Publishing Corporation, Vol.8, pp.317-324, 2008, 査読有

④H. Kawano, Strategy of Digital Contents Archive Based on Reputation Model , 19th International Conference on Systems Engineering 2008, IEEE ICSENG '08, pp.288-293, 2008, 査読有

⑤伊藤洋輔, 河野浩之, 信頼連鎖による P2P コンテンツ流通システムの提案と評価, 日本データベース学会 Letters, 日本データベース学会, Vol.6/No.1, pp.21-24, 2007, 査読有

⑥森下広史, 河野浩之, セマンティックな確率的 P2P ルーティングの提案, 第 21 回人工知能学会全国大会会議録, 人工知能学会第 21 回, pp.1G1-5, 2007, 査読有

〔学会発表〕（計 4 件）

①H. Kawano, Technical Aspects of Digital Archive Systems - How to transmit digital contents from generation to generation - , Japanese-Austrian Workshop on Natural Language and Spatio-Temporal Information , Tokyo University, 2009

②河野浩之, Web アーカイブ・クローラーの選択収集ポリシーについて, ネットワークと情報処理, 甲南大学知的情報通信研究所, 2008

③H. Kawano, Steps Towards Better Digital Archives, Digital Archive Workshop on Human Civilization Digital Archive Forum with APAN and APNG, 2007

④河野浩之, 伊藤洋輔, P2P コンテンツ流通における「ただ乗り」問題, ネットワークと情報処理, 甲南大学知的情報通信研究所, 日本 OR 学会「情報ネットワーク性能評価」研究部, 2007

〔その他〕

https://nzn.jim.nanzan-u.ac.jp/rd/search/researcher/048595/profile-j.html