様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月１７日現在

機関番号：３３９１７
研究種目：基盤研究（C）
研究期間：2007～2010
課題番号：１９５００００９８
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）　Studies on Stream Mining in Web Archive

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

研究成果の概要（和文）：

Web アーカイブサイズは加速度的に増大しており，図書館や IIPC によって長期保存にむけた努力がなされている. 本研究ではデータマイニングの観点から P2P 環境を含むコンテンツの評判分析などを行い単調増加するデータのアーカイブシステムに関する提案を行った．特に，ファイルアクセスパターンなどを含む検討に基づいて，RAM や HDD に加えて磁気テープなどニアリニア媒体を含む階層的ストレージシステムのアーキテクチャについて論じた.

研究成果の概要（英文）：

The size of the web archive is increasing exponentially, many national libraries and IIPC (International Internet Preservation Consortium) are making efforts to decide guidelines of long-term preservation of digital contents. In this research, from the view points of data mining techniques for reputation model, we reconsider a growth model of storage volume in web archive system. We discuss a basic architecture of hierarchical storage system based on characteristics of memory devices such as RAM, HDD, magnetic tapes and disks. We improve the file moving algorithm by using file retrieval patterns and access frequencies.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	900,000	270,000	1,170,000
２００８年度	1,100,000	330,000	1,430,000
２００９年度	700,000	210,000	910,000
２０１０年度	500,000	150,000	650,000
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：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．研究開始当初の背景

研究計画当初（2006 年），デジタルアーカイブに関わる代表的プロジェクトとして，全米デジタル情報基盤整備・保存プログラム (NDIIPP: National Digital Information Infrastructure and Preservation Program)

が注目されており，官民をあげてデジタルコンテンツを保存する幅広い取組みが開始されてきた．加えて，同じく米国においては，権利処理に課題を残すものの大規模サイトとして Internet Archive が運用されていた．その他，法改正により収集を行う英国図書館(Britain on the Web)，オーストラリア国立図書館，フランス国立図書館，オーストリア国立図書館，フィンランド国立図書館，スウェーデン王立図書館など欧州各国において，さらに，中国・韓国においても各国内のデジタルコンテンツを対象とした数多くのアーカイブプロジェクトが進行しつつあった．

また，IIPC (the International Internet Preservation Consortium) では，デジタルアーカイブに対する研究開発を積極的に推進しており，関連するソフトウェアとして，例えば，Heritrix, WERA, NutchWAX などが開発されつつあり，NDL におけるデジタルアーカイブプロジェクトにおいても，その開発動向に注目していた．

そこで，本研究は，国立国会図書館におけるデジタルアーカイブプロジェクトの基盤システムに関わる関西館電子図書館課における複数のプロジェクト（特に WARP プロジェクト）におけるシステム開発・運用に携わることで得られた知見に基づいて，コンピュータ周辺機器の短期間の開発ロードマップに比較し，アーカイブコンテンツの時系列ストリームを扱う問題に焦点を絞る部分に注目する研究計画となった．また，本研究を進めるにあたって，過去進めてきた検索エンジン開発における研究成果であるユーザの閲覧履歴などを利用する情報フィルタリングや情報視覚化技術が，大容量 Web アーカイブデータの効果的な長期保存，さらに，効率的な閲覧システムを開発する上で重要な基礎技術となると想定していた．

２．研究の目的

知識流通基盤となるインターネット上において，デジタル化され蓄積されつつあるコンテンツは，従来の出版物に比べ，空間的・時間的に安定しない問題が指摘されていた．特に，情報内容の更新・改変が容易であるため，安定した原本保存が難しく，また仮に同じ内容であっても，URL (Uniform Resource Locator)が変更になる問題がある．加えて，著者やサーバ管理者の都合により公開中止となる場合も頻繁に生じてきた．そこで，国の文化資産としてデジタルコンテンツのアーカイブが推進されつつあり，特に Web ページの体系的な蓄積，数十年・数百年といった長期保存を目指すウェブアーカイブ(Web Archive)が，世界各国の国立図書館などを中心に推進されてきた．

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 プロジェクトで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」技術確立を研究目的とした．本研究の推進によって得られる知見は，デジタルアーカイブポータル構築における技術的課題と密接に関係していることから，関係する諸課題を解決する上で直接的・間接的にフィードバックすることが可能である．

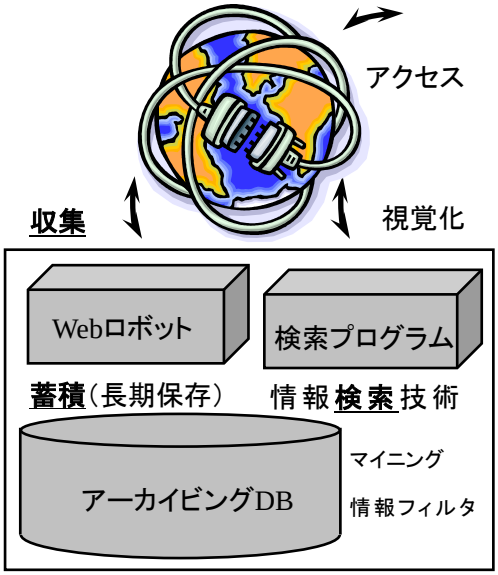
また，デジタルアーカイブの主要構成システムである Web アーカイブは Web ページを時系列順に長期間保存するため単調増加する点が，ページ更新時に上書きを行う既存のサーチエンジンと大きく異なる．そこで，本研究では，デジタルアーカイブにおけるアーカイブデータのアクセス傾向変化に着目したストリームマイニング技術開発を目的とした．

さらに，アーカイブされた Web ページに対するクリックストリームに基づくアルゴリズムを提案し，単調増加するアーカイブデータ群の効率良い操作が可能なアーカイブシステム技術の検討を行う．円滑なナビゲーションを支援するシステムアーキテクチャの性能評価も行う．

３．研究の方法

研究代表者は，国立国会図書館関西館電子図書館課の非常勤調査員として Web アーカイブシステム WARP の運用・関連開発技術などに関係していることから，デジタルアーカイブにおけるシステム開発技術にフィードバック可能な研究を行った．

以下，本研究の対象となるアーカイブシステムのアーキテクチャを述べ，図に示す主要構成要素に分けて，デジタルアーカイブに関わる研究計画を示す．主要構成要素は，データの「収集」「蓄積」「検索」であり，「収集」は，アーカイブ対象となるコンテンツ更新時期等に合わせたアーカイブ用 Web ロボット等の設計技術を必要とする．「蓄積」では，アーカイブされたデータを可搬性の高い形式で適切な記憶媒体に格納する技術を必要とする．「検索」では，単調に増加するアーカイブコンテンツに対してマルチメディアデータベース等を用いて一貫性の高いコンテンツ管理を必要とする．また，サーチエンジンと同様の全文検索機能などを含む情報検索機能が必要となる．



デジタルアーカイブシステムを構築するにあたって，WARP システム運用における問題と，2006 年度に向けて進行している国立国会図書館におけるデジタルアーカイブ構築，総務省によるアーカイブ実証実験の実証実験などで扱われる範囲を調査し，以下のような技術的課題を含む新たな問題の把握に努めた．

- データ容量，収集頻度，収集戦略等（Web アーカイブにおけるデータ増加傾向の分析）
- データ蓄積システム構築運用技術（階層型アーキテクチャ，冗長型記憶装置，追加型記憶装置）
- メタデータ，識別子（Dublin Core，さらに，MD5，SHA 等）
- 時間依存型ナビゲーション・検索（収集時間差によるデータ一貫性確保，全文検索，意味的検索等）
- 深層 Web，動的 Web に対するデータ収集（Focused Crawling 技術）

ただし，デジタルアーカイブには多くの課題があるため，「収集」に伴うコンテンツの品質管理と，「蓄積」に伴う長期保存を含む課題に絞る．

① 収集：アーカイブ Web ロボット
現在の 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，ハイパーリンクによる Web グラフ構造，Web アーカイブシステムにおいてクリックストリームとして記録されるデータ解析に基づくアーカイブアルゴリズムを提案し，その性能について考察を行う．

② 蓄積：Web アーカイビングデータベース
現状の WARP システムにおける，大容量ストレージシステムを用いた長期運用・保存に関する技術的課題に焦点を当てる．これまでに提案してきた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する．また，WARP データに基づく評価予測を行うことで，長期間のアーカイブシステム運用に適した性能が実現できることを明らかにする．加えて，Organic Storage や Autonomic Storage 等のストレージ・アーキテクチャに関わる技術を視野に入れ，データ保存形式やデバイス特性を考慮した階層型ストレージ・アーキテクチャを発展させる．

③ 評価：Web アーカイブシステム
「問答」検索システム，ならびに，ピア・ツー・ピア情報フィルタリングシステムの研究における研究成果の適用を考える．また，NDL で運用されている WARP システムの利用履歴に基づいて，ナビゲーションに伴うクリックストリームの特性を分析する．デジタルアーカイブシステムにおけるナビゲーションアルゴリズムの性能評価を行う．加えて，提案するデジタルアーカイブシステムにおけるクリックストリーム解析アルゴリズムの性能評価，アーカイブデータ収集 Web ロボットの性能評価，階層型ストレージのアーキテクチャの性能評価など，各種解析とシミュレーションを用いた評価研究なども進める．

４．研究成果

まず，多様性の高い膨大なデジタルコンテンツが氾濫する状況下において，どのような基準でコンテンツを収集し保存するための資源を割り当てるべきかという問題は重要である．そこで，デジタルアーカイブに関わる研究の構成要素を，アーカイブ対象となるデータの「収集」「蓄積」「検索」に分け，「収集」と「蓄積」に密接する研究を実施した．

研究発表「Web アーカイブ・クローラーの選択収集ポリシーについて」，ならびに，論文「Reputation-based Contents Crawling in Web Archiving System」では，アーカイブシステムの収集蓄積対象となるコンテンツを提供するシステムのもつ諸特性値に対してデータマイニングを行なうことで，どのような基準によって各種資源（ネットワーク帯域，ディスク容量，蓄積時間等）を割り当てれば良いかという格差サービスを行なう基本的指針を検討した．研究を進めるにあたって，P2P ネットワークを用いた各種リソースの共有環境において，各種リソースを一方的に利用のみを行う「ただ乗り問題」が，P2P シス

テム設計における重要な課題であることを指摘し， P2P システムの信頼機構によって，システムに提供する各種サービスの貢献度に応じた格差サービスが実現できることを論じた．提案したアルゴリズムは，コンテンツ流通の状況を，評判モデル(reputation model)を用いて拡張したものである．

論文「Towards Digital Archive Systems: Architecture and Design of Digital Museum Archives」では，デジタルコンテンツを保存するデジタルアーカイブシステム構築に関する技術的課題を，デジタルミュージアムアーカイブを一例にとりあげ「蓄積」技術と，基本的な検索機能について議論した．口頭発表では，より幅広い観点から，デジタルコンテンツの保存技術に関する課題を議論した．

論文「Management of Storage Devices and File Formats in Web Archive Systems」では，2009 年から 2010 年にかけての WARP 運用データに基づいて，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によるファイル圧縮などを再評価し，システム性能などを論じた．

なお，インターネット関連の先端技術を扱う国際会議に付随して開催された「Digital Archive Workshop on Human Civilization Digital Archive Forum with APAN and APNG」において，日本におけるデジタルアーカイブの問題点に関する招待講演を行ったこと，ならびに，オペレーションズリサーチに関する国際会議 ISORA2009 において，今後のデジタルコンテンツ増加に対応するシステム構成に対して解決すべき方向性を Plenary Speaker として提示したこと等，本研究に関わる技術的問題ならびに社会的な問題を多くの分野の研究者で共有できたことも有意義な点である．

最後に，本報告書作成時点においても，NDL の非常勤調査員を継続している．よって，現在進行形で構築されているデジタルアーカイブに関係して，本研究成果に基づいた技術的知見を提供することは，今後の研究成果の展望として重要な点である．特に，単調増大する収集データにより，ストレージの絶え間ない漸増的拡充が必要とされており，電子書庫におけるストレージ仮想化技術や階層的ストレージ採用の検討，クラウド環境下におけるデータセンター利用環境の課題など，本研究と強く関連する観点からシステム導入の技術的検討を続ける．

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６件）

1) H. Kawano : Management of Storage Devices and File Formats in Web Archive Systems, Lecture Notes in Operations Research World Publishing Corporation 12, pp.356-361, 2010.（査読有）
2) 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.（Plenary Speaker,査読有）
3) 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）
4) H. 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）
5) H. 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）
6) 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）

〔学会発表〕（計４件）

1) H. 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, Univ. of Tokyo, 2009.（日本・オーストリア修好 140 周年記念事業,口頭発表，査読無）
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研究会，甲南大学知的情報通信研究所,2009.（口頭発表，査読無）
3) H. Kawano:Steps Towards Better Digital Archives, Digital Archive Workshop on Human Civilization Digital Archive Forum with APAN and APNG, Xian, 2007.（招待講演,査読無）
4) 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）

６．研究組織

(1)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