

機関番号：12102

研究種目：若手研究(B)

研究期間：2008～2010

課題番号：20700076

研究課題名（和文）開放型情報空間ガバナンス実現のための技術基盤の研究

研究課題名（英文）Research on Technology for Governance of Open Information Spaces

研究代表者

森嶋厚行 (MORISHIMA ATSUYUKI)

筑波大学・大学院図書館情報メディア研究科・准教授

研究者番号：70338309

研究成果の概要（和文）：本研究では、既存の DBMS の管理下でない各種データを、柔軟に管理可能なソフトウェアエンジンの開発に取り組んだ。主な成果は次の通りである。(1) Web 情報空間をはじめとする開放型情報空間を管理するためのフレームワークの提案とそれに基づくシステム開発 (2) 開放型情報空間管理エンジン実現に関連する各種要素技術の開発。特に、移動先 Web ページの発見技術、Web 情報空間に存在するデータ一貫性制約の発見支援技術、大量のグラフデータに対するナビゲーション問合せの効率よい処理アルゴリズムの開発。

研究成果の概要（英文）：This research attempted to develop software engines that manage various types of data residing outside traditional DBMSs. The main results of the research are as follows: we (1) proposed an architectural framework for the management of open information spaces such as Web information spaces and developed systems based on it, and (2) developed technology components for the management systems for open information spaces. The components include algorithms for finding moved Web pages, those for finding data integrity constraints in Web information spaces, and those for efficiently processing navigational queries against large graph data.

交付決定額

(金額単位：円)

	直接経費	間接経費	合 計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,000,000	300,000	1,300,000
2010 年度	1,100,000	330,000	1,430,000
年度			
年度			
総 計	3,200,000	960,000	4,160,000

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：情報統合, Web データ管理, コミュニティデータ管理

1. 研究開始当初の背景

近年の情報利用環境では、各種プロジェクト、業務、リサーチ、コミュニティ等に関連するデータやファイルは、それに関わる複数の構成員によって複数の計算機で分散管理されることが多い。また、我々が日常作業する情報空間に含まれるデータの大部分は DB の管

理外にあることが報告されている。したがって、既存の DBMS の管理下でないデータ管理が大きな問題であるといえる。現在の計算機による情報管理のアプローチでは、データを管理するための情報システム(主に DBMS をバックエンドに持つ)をまず構築し、それらを用いて管理を行うことが一般的である。しかし、このアプローチでは、データ

を DBMS に格納することが前提となっており、また、専用の情報管理システムの構築が必要になる。このような、伝統的アプローチによる情報管理では、一般に要求の変化の対応するためにシステム再構築が必要であり多大なるコストがかかる。

2. 研究の目的

本研究では、これまでの DBMS 分野における研究成果を発展・適用させることにより、既存の DBMS の管理下でない各種データを、個別の管理システムを開発することなく柔軟に管理可能なソフトウェアエンジンの実現に取り組む。本研究の、最も抽象度の高いアイデアは次のようになる。すなわち、開放型情報空間の管理を行うために、DBMS 専用のストレージに格納すること無しに、管理を行う独立したエンジンを開発する。具体的には、このエンジンは、管理範囲の情報空間から管理に必要なデータ・メタデータを抽出し、管理に利用する。これらの情報やそれらが満たすべき制約や管理ポリシーは、これまではユーザの頭の中に存在していた。これらを明示的に計算機に保持するために必要な技術の研究開発がポイントである。制約や管理ポリシーは「高レベルの宣言型記述」によって簡単に指定可能にする。一般に、Web サーチエンジンの世界では、クローラというソフトウェアを用いてインデックスを構築し、Web の情報検索に利用するが、それに対し、情報空間管理エンジンは特に管理的側面に焦点を置いた機能を実現する。本エンジンは、情報検索のためのクローラとエンジンの組合せや既存の DBMS とは下記のような点で大きく異なる。(1) データだけでなく、人の操作履歴などのログ情報が管理には必要である。これらの「メタデータ」の利用が大きな焦点となる。(2) 情報空間の管理要求は多様である。これらの要求を容易に反映できるような仕組みが必要である。(3) 必要に応じて管理下にあるデータを更新したり、情報空間内の適切な管理者に通知を行う。(4) DBMS が管理する情報空間と比べて、対象となる情報空間の操作の自由度が高く、情報空間の参加者による自由な操作と全体的な統治の折り合いを見つける必要がある。

3. 研究の方法

本研究を推進するにあたっては、情報空間から抽出されたデータ・メタデータデータベースを中心とする各種要素技術の開発も重要であるが、それらを体系的にまとめるためのフレームワークや、そのフレームワークを実装したソフトウェア・エンジンを開発することも、重要な焦点となる。具体的には、次の点を中心に研究を推進した。

(1) 開放型情報空間統治エンジン実現のた

めのフレームワークの設計。各種データからなる開放型情報空間の統治エンジンを実現するために、その中心となるメタデータデータベースの全体像の設計、情報空間と接続するための API や、情報空間における操作フックのための仕組みなどを体系化する。具体的には、XML/RDB データ、ファイルシステムで管理されているデータ、World Wide Web データなどのいくつかの重要な情報資源において、それぞれ必要とされる要求を調査し、体系的な解となりうるフレームワークの検討および設計を行った。

(2) 実現のための要素技術の調査および開発。特にメタデータ関連技術の開発。上記(1)で設計したフレームワークの実現に必要な要素技術に関して、既存技術の調査、および新たな技術の開発を行った。特に、情報空間における制約やポリシーの自動発見手法、メタデータ自動抽出手法、アノテーション生成手法、データクリーニング手法、大規模メタデータデータベースの間合せ処理技術等の要素技術は重要である。また、これらの技術を用いて構築された情報空間を監視し、一貫性の管理を行うための技術の開発が必要である。このために、情報空間における制約記述手法や、制約違反の効率よい発見技術、自動メタデータメンテナンス技術、の開発を推進した。

4. 研究成果

(1) 開放型情報空間統治エンジン実現のフレームワークに関する成果。

(1.1) コミュニティ情報空間統治エンジン実現のフレームワークの基礎研究：まず、開放型情報空間統合エンジンを実現するためのアーキテクチャ設計の準備として、特にミドルウェアレイヤにおける API の設計などの検討を行った[28]。本ミドルウェアの開発においては、クライアントサイドのモジュールとサーバ側のモジュールとの実行時の連携に関する検討も行った[23]。また、情報空間統治に関する論理記述に従い処理を行うインタプリタエンジンの設計を行った[21]。さらに、これらの基礎研究で得られた成果に基づき、情報空間統治エンジンのプロトタイプシステムの開発を行った。このような、開放型情報空間におけるデータ指向アプリケーションは、計算機だけでなく人間の知識を利用した情報処理を考慮する必要があることから、その可能性を示す幾つかのアプリケーションを開発した[12][13]。

(1.2) Web 情報空間を対象とした情報空間統治エンジン実現のフレームワークの提案：Web 情報空間の統治に関しては、2つの点に焦点を絞り研究を推進した。すなわち、Web リンクの一貫性維持(リンク切れ防止)と Web コンテンツにおける内容の一貫性維持であ

る。本研究では、Web 情報を監視し、そのコンテンツの一貫性制約の違反を発見すると、その報告と修正を試みるというアーキテクチャの設計を行った[3][6][8][9][26]

(2) 実現のための要素技術に関する成果

(2.1) Web 情報空間におけるデータ一貫性管理のための制約発見と修正に関する基礎研究：まず、第一の要素技術として、Web リンク切れの修正を支援するために、リンクオソリティの概念を用いたページ移動先発見の仕組みを開発した。その結果、70%以上の発見率を実現し、Google, Yahoo, MSN といった商用サーチエンジンを利用した場合と比べて統計的にも高い割合でページの移動先を発見することを示した[7]。第二の要素技術として、Web 情報空間統治エンジンの入力として必要なデータ一貫性制約を、既存の Web データから発見するための支援手法を実現した。具体的には、包含従属性の効率よい発見手法に関する成果が得られた。包含従属性の発見手法としては、リレーショナルデータベースにおける包含従属性の発見手法を拡張し、包含率の概念と階層構造の存在を扱うための仕組みを導入したものを開発した。包含率の導入により、不完全・不規則な情報が多い Web 情報空間において包含関係を発見することができ、包含従属性発見の支援に役立つことが可能であることを示した[24]。さらに、そのような Web データを対象とした包含従属性の発見に関して、要素の特徴情報やヒューリスティクスを用いた手法[1][17]の他、ビットシグネチャを用いる効率化手法を開発した。また、実 Web データを用いた実験により、Web コンテンツを対象とした包含従属性の発見に関して、ビットシグネチャを用いた効率化手法の有効性の評価を行った。その結果、提案手法が大幅な効率化を実現することが実験的にも明らかになった[10]。さらに、大量に得られる包含関係のランキング手法を実現した[14]。

(2.2) 大量の情報を支えるグラフデータベースの効率的な処理の実現：情報空間のガバナンスを実現するためには大量のメタデータの効果的な管理が必要である。それに関しても研究を推進し下記のような成果が得られた。まず、大規模グラフメタデータベースの効率よい問合せ処理技術の開発を行った。問合せのクラスとしては、正規パス問合せの重要な構成要素である子供演算と子孫演算、およびその組合せからなる問合せを対象としている。本技術では、グラフデータを格納する際のデータ配置順序に工夫を行うことによって、データ全体のサイズに依存しない効率の良い処理を実現することを図った[2][25][27]。特に理論的な解析を進め、提案手法が理論上可能な最小の I/O コストによる処理を実現可能であることを示

した[5]。

次に、XML データ管理を支援する技術の開発を行った。特に、XML データの一貫性管理を支援するために RDB の機能を利用可能とするような XML-RDB マッピング技術について研究を行った。成果としては、既存のマッピング手法を包含し、かつデータ一貫性制約の維持を実現するようなマッピング手法を開発したことがあげられる[16][19]。

(2.3) 柔軟な情報資源ラッピングの実現機構の実現：開放型情報空間統治を効率よく行うために不可欠な情報資源ラッピング手法に関する研究を推進した。その成果としては、ラッピングを実際に行うソフトウェアモジュールであるラッパーを効率よく開発するための言語を設計・開発したことが挙げられる。特に、双方向のデータ変換を実現した言語である iWraplet、および、抽象度の高いラッピングルールを導入し、再利用性の高いラッパーを構築を実現する言語である Wraplet/A の設計およびそれら処理系を実現した[15][22]。また、MDL コストを利用したラッパー生成の自動化手法の開発[4]や、ブラウザの表示情報を用いたラッピング手法[18]の開発を行った。

(2.3) その他の情報空間管理要素技術：削除して良いファイルの発見技術[11]や人物検索システム[20]の開発を行った。

5. 主な発表論文等

〔雑誌論文〕(計 4 件)

[1] 高橋公海, 森嶋厚行, 弓矢英梨佳, 杉本重雄, 北川博之. 「ビットシグネチャを用いた Web ページの包含従属性発見の効率化」情報処理学会論文誌データベース, 査読有, Vol.3, No. 3, pp. 1-10. 2010 年 9 月

[2] 只石正輝, 森嶋厚行, 田島敬史. 「効率のよい子供／子孫問合せ処理のためのグラフデータ格納手法の提案」日本データベース学会論文誌, 査読有, Vol. 7, No. 3, pp. 55-60, 2008 年 12 月.

[3] 高橋公海, 森嶋厚行, 松本亜季子, 杉本重雄, 北川博之. 「Web コンテンツ管理のための一貫性制約発見支援」日本データベース学会論文誌, 査読有, Vol. 7, No. 3, pp. 25-30, 2008 年 12 月.

[4] 澤菜津美, 森嶋厚行, 杉本重雄, 北川博之. 「HTML ラッパ自動構築手法の提案」日本データベース学会論文誌, 査読有, Vol. 7, No. 1, pp. 263-268, 2008 年 6 月, 日本データベース学会.

〔国際会議論文〕(計 5 件)

[1] Atsuyuki Morishima, Keishi Tajima, Masateru Tadaishi. Optimal Tree Node Ordering for Child/Descendant Navigations. The 26th International Conference on Data Engineering (ICDE 2010), pp. 840-843, Long Beach,

California, USA, March 1-6, 2010, (査読有).

[2] Masami Takahashi, Atsuyuki Morishima, Shigeo Sugimoto, Hiroyuki Kitagawa. A Constraint-based Tool for Data Integrity Management on the Web. The 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010), pp. 144-146, January 14-15, 2010, SKKU, Suwon, Korea, (査読有).

[3] Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, Hiroyuki Kitagawa. Bringing Your Dead Links Back to Life: A Comprehensive Approach and Lessons Learned. The 20th ACM Conference on Hypertext and Multimedia (ACM Hypertext 2009), pp. 15-24, Torino, Italy, June 29th -July 1st, 2009, (査読有).

[4] Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, Hiroyuki Kitagawa. Why are Moved Web Pages Difficult to Find? The WISH Approach. The 18th International World Wide Web Conference (WWW 2009), pp. 1117-1118, Madrid, Spain, April 20-24, 2009 (poster), (査読有).

[5] Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, Hiroyuki Kitagawa. PageChaser: A Tool for the Automatic Correction of Broken Web Links. The 24th International Conference on Data Engineering (ICDE 2008), Cancun, Mexico, April 7-12, 2008. pp. 1486-1488, (査読有).

〔学会発表〕(計 19 件)

[1] 高橋公海, 森嶋厚行, 杉本重雄, 北川博之. 「ビットシグネチャを用いた Web ページの包含従属性発見の効率化」Web とデータベースに関するフォーラム(WebDB Forum) 2009, 8 pages, 慶應義塾大学, 2009 年 11 月 19 日. 学生奨励賞

[2] 伊藤弘人, 森嶋厚行. 「ファイルに関する多様な情報を利用したファイル削除候補の発見」Web とデータベースに関するフォーラム(WebDB Forum) 2009, 8 pages, 慶應義塾大学, 2009 年 11 月 19 日.

[3] 三津石智美, 望月祥司, 森嶋厚行. 「人と計算機を情報資源とする統合情報検索システム CySearch の提案」情報処理学会第 73 回全国大会講演論文集(第 1 分冊), pp. 683-684, 東京工業大学, 2011 年 3 月 2 日.

[4] 安永ゆい, 望月祥司, 森嶋厚行. 「GWAP によるオントロジ構築手法の提案」情報処理学会第 73 回全国大会講演論文集(第 1 分冊), pp. 765-766, 東京工業大学, 2011 年 3 月 2 日.

[5] 弓矢英梨佳, 高橋公海, 森嶋厚行, 杉本重雄, 北川博之. 「Web コンテンツにおける包含従属性発見支援のためのランキング手法」第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM 2011), 8 pages, 静岡

県伊豆市, 2011 年 2 月 27 日.

[6] 石井悠太, 森嶋厚行. 「抽象的な記述が可能な Web からのデータ抽出言語の提案」第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM 2011), 8 pages, 静岡県伊豆市, 2011 年 2 月 27 日.

[7] 太田壮祐, 森嶋厚行, 天笠俊之, 只石正輝. 「関数従属性と包含従属性を用いた XML-RDB マッピング」第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM 2011), 8 pages, 静岡県伊豆市, 2011 年 2 月 28 日.

[8] 弓矢英梨佳, 高橋公海, 森嶋厚行, 杉本重雄, 北川博之. 「ヒューリスティクスを用いた Web コンテンツの包含従属性発見の効率化」情報処理学会創立 50 周年記念(第 72 回) 全国大会講演論文集(第 5 分冊), pp. 237-238, 東京大学, 2010 年 3 月 10 日.

[9] 袖山広輝, 森嶋厚行. 「ブラウザでの表示位置に着目した Web コンテンツの書換えによる応用処理支援」情報処理学会創立 50 周年記念(第 72 回) 全国大会講演論文集(第 1 分冊), pp. 759-760, 東京大学, 2010 年 3 月 10 日.

[10] 太田壮祐, 森嶋厚行, 天笠俊之, 只石正輝. 「一貫性制約を考慮したデータ管理のための XML-RDB マッピング手法」情報処理学会創立 50 周年記念(第 72 回) 全国大会講演論文集(第 1 分冊), pp. 905-906, 東京大学, 2010 年 3 月 11 日.

[11] 山元潤, 森嶋厚行. 「非公開データを用いた人物検索システムの開発」情報処理学会創立 50 周年記念(第 72 回) 全国大会講演論文集(第 1 分冊), pp. 725-726, 東京大学, 2010 年 3 月 9 日.

[12] 安西則晃, 森嶋厚行. 「計算機と人との協調によるデータ管理のためのデータベース言語の提案」情報処理学会研究報告データベースシステム(DBS), 2009-DBS-149(12), pp. 1-8, 慶應義塾大学, 2009 年 11 月 20 日.

[13] 石井悠太, 森嶋厚行, 杉本重雄, 北川博之. 「ラッピング言語を用いた Web サイトの再構築手法の提案」情報処理学会研究報告データベースシステム(DBS), 2009-DBS-149(13), pp. 1-8, 慶應義塾大学, 2009 年 11 月 21 日.

[14] 三森祐一郎, 森嶋厚行. 「分散ファイル群高度管理のためのファイル関連の発見エンジンの開発」データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2009), 8 pages, 静岡県掛川市, 2009 年 3 月 8 日.

[15] 高橋公海, 森嶋厚行, 杉本重雄, 北川博之. 「Web ページを対象とした包含従属性の効率的な発見手法」データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2009), 6 pages, 静岡県掛川市, 2009 年 3 月 10 日.

[16] 只石正輝, 森嶋厚行, 田島敬史. 「任意のグラフノードを起点とする正規パス式の評価手法」データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2009), 8 pages, 静岡県掛川市, 2009年3月10日.

[17] 高橋公海, 森嶋厚行, 松本亜季子, 杉本重雄, 北川博之. 「Web コンテンツ一貫性管理のための制約発見支援」情報処理学会研究報告, vol. 2008, No. 146 (2008-DBS-146), pp. 127-132, 福島市, 2008年9月22日.

学生奨励賞

[18] 只石正輝, 森嶋厚行, 田島敬史. 「大規模グラフ構造に対する効率的な子供／子孫問合せの処理手法の提案」情報処理学会研究報告 Vol.2008, No.88(2008-DBS-146), pp. 229-234, 福島市, 2008年9月22日.

[19] 三森祐一郎, 森嶋厚行, 「分散ファイル群高度管理を目的としたファイル関連の発見支援機構」情報処理学会研究報告, vol. 2008, No. 146 (2008-DBS-146), pp. 319-324, 福島市, 2008年9月23日.

〔産業財産権〕

○取得状況 (計1件)

名称: リンクオーソリティ決定方法及び装置並びにプログラム

発明者: 森嶋厚行, 飯田敏成, 杉本重雄, 北川博之, 中溝昌佳

権利者: 国立大学法人 筑波大学

種類: 特許

番号: 第4496370号

取得年月日: 2010年4月23日

国内外の別: 国内

6. 研究組織

(1)研究代表者

森嶋厚行 (MORISHIMA ATSUYUKI)

筑波大学・大学院図書館情報メディア研究科・准教授

研究者番号: 70338309