インターネットアーカイブに基づくWebディレクトリの設計と構築

Design and Construction of Web Directory based on Internet Archive

名古屋大学情報連携基盤センター
Nagoya University Information Technology Center

名古屋大学大学院情報科学研究科
Nagoya University Graduate School of Information Science

松原　茂樹　　鈴木　祐介
MATSUBARA, Shigeki　SUZUKI, Yusuke

Abstract

This paper describes organization of a group of Web documents produced in a university. In this research, the organization is archived by constructing a Web directory according to substantial distribution of the documents. The Web directory is constructed by classifying the Web documents into the directories and making the connection with directories. This is intended to renew the arrangement of the documents in WWW that reflects the units in the university such as faculties and laboratories to arrangement that considers user's viewpoint. An experiment on Web directory construction was done by using Web documents of Nagoya University that collected with the internet archive. The Web directory was constructed with the design of the Web directory structure consisting of three hierarchies and the classification of the Web documents into each directory, and had the scale of 178 directories and 11243 documents. These can be used as a document retrieval interface and as the basic data to advance the techniques for the document classification and information retrieval, etc.

１．はじめに

大学においても文書のデジタル化が急速に進行し、大量の情報がWWWを通して発信されるようになった。一方で、大学情報を受信する利用者の目的は実に多様である。利用者が必要とする情報に容易にアクセス可能な環境を整備することは、開かれた大学としての責務であり、その重要性は今後益々増大すると予想される。

しかしながら、学術情報の発信環境として、上述の要求に十分に応えている大学は必ずしも多くない。その理由として、大学における情報生産過程の特異性が挙げられる。というのも、大学では多くの場合、大学全体の他に、学部や学科、研究室、個人など、様々な構成単位でWWW環境を設け、運用している。情報の生産者は、作成したデジタル文書を独自のWWW環境で発信することになり、それらの情報が機関全体の制御下におかれることは稀である。このことは、生産者が情報を容易に発信することができ、

ひいては、機関全体が高い情報発信力をもつことになる。しかし、その一方で、情報の発信スタイルは、WWW 環境間で一致しておらず、また、学内事情に精通していない利用者にとって、ある文書がどの組織のWWW環境に属しているかは明らかでなく、利用者が目的とする文書にたどり着くことは容易ではない。このために、機関として情報の発信を統制するよりも、むしろ、利用者のニーズを考慮して、改めて情報を組織化することが有用な方法である。

そこで本論文では、WWW上の大学情報を対象としたデジタル文書群の組織化について述べる。本研究では、文書群の内容的な分布にしたがって Web ディレクトリを構成することにより組織化を実現する。すなわち、文書をその内容にしたがってディレクトリに分類し、ディレクトリ間を意味的に関係づけることにより、Web ディレクトリを構築する。これは、学部や研究室等、大学内の組織を反映したWWW上の文書の配置から、利用者の視点を考慮した配置へと再構成することを意図している。

インターネットアーカイブにより収集した名古屋大学の Web 文書データ 1) を用いて、Web ディレクトリ構築実験を行った。実験では３つの階層からなる Web ディレクトリ構造を設計し、Web 文書を各ディレクトリに分類することにより、178 ディレクトリ、11243 文書の規模を備えた Web ディレクトリを構築した。これらは、文書検索インタフェースとしてはもちろん、文書分類や情報検索等の技術開発を進めるための基礎データとしても活用できる。

本論文の構成は以下の通りである。2 章では、Web ディレクトリの設計について説明する。3 章では、インターネットアーカイブ、及び、データの分類について述べる。4 章では、Web ディレクトリの利用法について論じる。5 章で、まとめと今後の課題について言及する。

2. Web ディレクトリの設計

Web ディレクトリとは、Web 上の大量のリンクを集め、それを分野ごとに分類したものであり、Yahoo! Directory2) や Google Directory3) などが代表的である。分類によりまとめられた各リンクリストはディレクトリと呼ばれ、ディレクトリ間の関係は一般に階層構造によって表現される。Web ディレクトリの利用者は、目的の情報にアクセスするために階層をたどればよく、一種の文書検索ナビゲーションとして機能する 4)。あるリンク集合に対して、最適なディレクトリ構造を定めることは難しいが、その設計においては、少なくとも利用者の目的や Web ファイルの分布を考慮することが重要である。

本研究では、名古屋大学を対象とした Web ディレクトリを構築するために、学内の Web サイトを分析し、ディレクトリ構造を設計した。ディレクトリ構造を決めるにあたり、

- **ディレクトリ階層の深さ**：階層が深くなり過ぎると、目的のページに到達するための経路が長くなり、検索効率が低下する。
- **ディレクトリの種類**：ディレクトリの数が少ないとディレクトリに配置されるリンク数が膨大になり、一方、多すぎると配置されるディレクトリ数が膨大になり、検索効率が悪くなる。
- **ディレクトリの名前**：大学に精通していない利用者を想定し、利用目的を考慮した名前とする。

などを定めた。

設計した Web ディレクトリの一部を図 1 に示す。今回の設計では、ディレクトリ階層の深さを 3 とし、また、１つのディレクトリ内に配置可能なディレクトリ数を９とした。また、ディレクトリの第１階層は、抽象的な分類であることを考慮し、利用者の目的を表現する名称を与えた。図 1 のようなディレクトリとして実現することにより、学内の組織やサイトの構成に精通していない利用者が、目的とする文書にアクセスすることも容易になると予想される。

図 1. 名古屋大学 Web ディレクトリの設計（一部）

3. Web ディレクトリの構築

名古屋大学ドメインにある Web ファイルを収集し、Web ディレクトリへの分類を実施した。

3.1 インターネットアーカイブ実験

名古屋大学内の Web ファイルの収集を試みた 1)。これは、Web ハーベスティング実証実験と称し、名

古屋大学情報連携基盤センター学術情報開発専門委員会情報流通ワーキンググループの活動として実施している。実験は、学内デジタル文書の収集・保存・活用の可能性を検証することを目的としており、本研究における Web ディレクトリ開発の基盤となる。

収集対象は名古屋大学ドメイン(nagoya-u.ac.jp)内にある Web サイトにあり、かつ、名古屋大学のトップページ [5]から直接的、もしくは、間接的に到達可能な Web ページとした。対象とするファイルのタイプについては、Javascript や cgi 等の動的な生成の側面が強いファイルを除き、特別な制限は設けず広く収集した。収集には、Web ファイルを自動収集するソフトウェアロボットを使用した。

実験により、423 サイト内の 642,196 ファイルを収集した。サイズにして 46.9GB に相当する。収集した Web ファイル数のタイプ別の内訳を図 2 に示す。ファイルタイプは拡張子により分類している。上位 4 タイプ(html, jpg, gif, htm)だけで全体の 85%以上に上っており、HTML ファイル及び画像ファイル(jpg, gif)が多数占めている。収集した Web ファイル規模のタイプ別の内訳を図 3 に示す。画像ファイル(jpg, gif)に加え、ドキュメントファイル(pdf, ps)が上位を占めている。これらはいずれもアーカイブの対象として相応しいファイルである。

図 2. Web ファイル数のタイプ別内訳

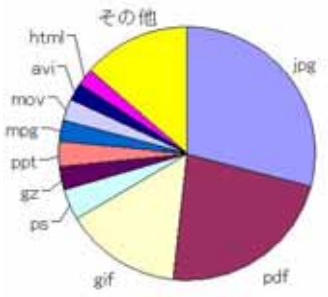

図 3. Web ファイル規模のタイプ別内訳

3.2 Web ファイルの分類

収集した Web ファイルを Web ディレクトリ上に分類した。分類は、作業者がファイルの内容を参照しながら、適切なディレクトリを選択することにより実施した。分類先のディレクトリ数に制限を設けず、必要な場合には複数のディレクトリに分類することも認めるとともに、Web ディレクトリの階層における分岐節点に位置するディレクトリへも分類を行った。今回の作業では分類対象を、HTML ファイル、PDF ファイル、動画ファイルとした。

なお、作業にあたり、ファイル分類ツールを作成し、作業の効率化を図った。作成したツールのスナップショットを図 4 に示す。

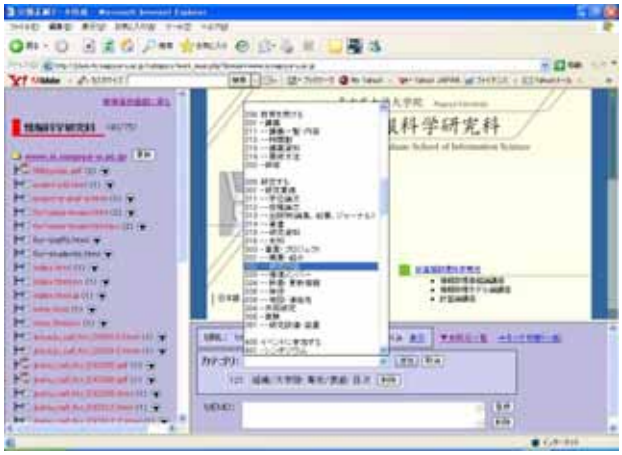

図 4. Web ファイル分類ツール

3.3 名古屋大学 Web ディレクトリ

前節までの設計及び分類に基づき、名古屋大学 Web ディレクトリを構築した。Web ディレクトリを図 5 に示す。現在、178 ディレクトリから構成されており、ファイルあたりの分類先は、1.06 ディレクトリである。

分類には、11243 ファイルを使用した。これは、収集した分類対象ファイルの 4.0%に相当する。実用の観点からみると必ずしも十分な規模を有しているとはいえないものの、たとえば、メタデータ・データベース共同構築事業 [6]において開発された国立情報学研究所の大学 Web サイト資源検索 [7]は比較対象の 1 つであるが、そこに登録されている名古屋大学の学術情報の収録件数は 1198 件であり、規模の優位性は高い。

開発した Web ディレクトリの機能の一つとして、Web ドメイン選択による検索対象の絞込み操作を導入した。これは、Web ディレクトリをたどる過程で、選択されたディレクトリ内のリンクを部局や施

設などの組織を指定することにより、表示されるリンクを絞り込む機能である。内容による分類が Web ページ群の縦方向の分割であるとすると、組織の選択は横方向の絞り込みを意味しており、情報アクセスのための多様な手段を提供している。Web ディレクトリにおける組織の選択画面を図 6 に示す。

Web ディレクトリは、引き続き、高機能化、大規模化を目指して開発を進めており、今後は、実利用も視野に入れた多面的評価を実施することを予定している。

図 5. 名古屋大学 Web ディレクトリ

図 6. 名古屋大学 Web ディレクトリの絞り込み機能

4. Web ディレクトリの利用

4.1 Web ディレクトリの自動構築

Web ディレクトリが実用システムとして機能するために、該当する Web ページの網羅性が重要となる。本研究での Web ディレクトリの構築は手作業によるものであり、システムの大規模化、及び、日常的なファイルの追加、更新、削除への対処には限界がある 8)。

この問題に対して著者らは、Web ディレクトリ構築を自動化するための方式について検討を進めている 9)10)。これまでに使用した Web ディレクトリ上のデータは、Web 文書を分類するための学習データとして使用できる。分類アルゴリズムとして、機械学習に基づく手法がいくつか提案されており、それらを階層的な分類に応用することは有用な方法の一つである。

4.2 知的情報検索インタフェースの開発

インターネットに限らず、これまでに様々な検索技術が考案されているが、検索精度、検索効率の双方を高めるために、適切に検索対象を絞り込むことは有力な方法である。開発した Web ディレクトリでは Web ページが属する分野が指定されているため、検索キーワードだけでなく、対象とする分野も併せて入力することにより、高い品質を備えた検索の遂行が可能となる。

5. まとめ

本論文では、名古屋大学ドメインを対象とした Web ディレクトリの設計と構築について述べた。名古屋大学 Web ディレクトリは、178 ディレクトリが 3 階層を形成しており、現在までに、11243 ファイルの分類が完了している。部局、施設等の指定による絞り込み機能が搭載されており、効率的なアクセスが可能である。

名古屋大学 Web ディレクトリは、

http://plum.itc.nagoya-u.ac.jp/

において実験的に公開している。今後は、データ規模、及び、インタフェースの充実をはかり、実用システムとして運用することを目指し、引き続き開発を進める予定である。

謝辞

有益なご議論を頂いた名古屋大学情報連携基盤センター学術情報開発専門委員会情報流通ワーキンググループメンバーの皆様に感謝します。Web ページの分類作業を担当いただいたアノテータの皆様に感謝します。本研究は、一部、名古屋大学附属図書館研究開発室、及び、情報連携基盤センターの教育研究基盤経費により実施しました。

参考文献：

1) 松原茂樹、鈴木祐介：Web サイト上の学術情報

をアーカイブする, LIBST Newsletter, No. 5, pp. 5-6 (2004).
2) Yahoo! Directory, http://dir.yahoo.com/
3) Google Directory, http://directory.google.com/
4) 谷津哲平, 新納浩幸, 佐々木稔：Web ディレクトリを用いた検索ナビゲーション, 言語処理学会第 11 回年次大会発表論文集, pp.1022-1025 (2005).
5) 名古屋大学ホームページ, http://www.nagoya-u.ac.jp/index.html
6) 伊藤真理、杉田茂樹：メタデータ・データベースの構築、逸村、竹内（編）：変わりゆく大学図書館、勁草書房, pp. 67-85 (2005).
7) 国立情報学研究所大学 Web サイト資源検索, http://ju.nii.ac.jp/
8) C. Santamaria, J. Gonzalo, and F. Verdejo: Automatic Association of Web Directories with Word Senses, Computational Linguistics, Vol. 29, No. 3, pp. 485-602 (2003).
9) 鈴木祐介, 松原茂樹, 吉川正俊：ハイパーリンクを用いた Web 文書の自動分類, 言語処理学会第 11 回年次大会発表論文集, pp.61-64 (2005).
10) 鈴木祐介、松原茂樹、吉川正俊：アンカーテキストを用いた Web ディレクトリの構築, 電子情報通信学会技術研究報告, Vol. 105, No. 203 (2005).