

ディレクトリサービスを利用した 分散型文献管理システムの構築について

副島健一* 古川善吾** 最所圭三† 荒木啓二郎*

* 九州大学工学部情報工学科 ** 九州大学情報処理教育センター

† 九州大学中央計数施設

ネットワークの普及、発達に伴い、分散する資源を効率よく管理することが重要になってきている。この問題を解決するために、データの分散管理を支援するツールが開発されている。これらのツールでは、特定の管理者が情報を管理するので、利用者自身が情報を管理できるものは少ない。利用者が保持している情報として、利用者の個人情報や文献などがあり、これらの情報は利用価値のあるものである。

本研究は、個人情報や文献情報を利用者自身が管理し他の利用者に提供するためのシステムを構築し、システム構築や利用時の課題を明確にし、解決することを目的とする。

Design and Implementation of Distributed Document Retrieval System Using Directory Service

Ken'ichi Soejima* Zengo Furukawa** Keizo Saisho† Keiji Araki*

* Dept. of Computer Science and Comm. Eng., Kyushu University

* Educational Center for Information Processing, Kyushu University

† University Computation Center, Kyushu University

Computer networks are widely used. It is important to effectively manage distributed resources on the networks. Some tools have been developed for supporting resource management. Since those tools are usually meant for administrators who control the resources, end users cannot control them. However, end users may have their own information which would be useful if it were made public. Therefore, resource management tools should be made available also to end users.

For this purpose, the authors developed a new resource management system in which end users can control their own information and can make it available to other users.

This paper discusses key issues in the design and use of such a system, and gives an overview of the prototype system developed.

1. はじめに

近年、計算機の普及やネットワーク技術の進歩によりローカルエリアネットワークやキャンパスネットワークが各組織で構築されてきた。さらにそれらのネットワークを互いに接続して、広域ネットワークが実現されている。わが国でも WIDE や JAIN、TISN などの広域ネットワークが形成され、そのうちのいくつかは海外のネットワークと接続している。

このようなネットワークの発展に伴い重要な問題となってきたことの1つに、ネットワーク上に分散した資源をいかに効率よく管理するかということがある。この対策として、現在さまざまなシステムが実現され利用されている。しかしながら、これらのシステムの多くは、管理は管理者が行うものであり、利用者はそれを利用するだけで、容易に利用者が情報を提供することはできなかった。

各利用者が提供できる情報として、個人情報や、自分が書いた文献などがある。現在でも、個人情報は whois サービスで、文献情報は文献データベースで利用できる。しかしながら、whois サービスは、登録や変更が容易でない、サーバの位置を知らなければならないなどの欠点がある。また、多くの文献データベースでは、使用に制限がある、登録されるまでに時間がかかる、自分で管理することができないなどの欠点がある。

個人情報を文献と同時に管理して提供することにより、興味ある文献を見つけた際に、その著者が書いた他の文献を利用することができ、連絡先を知ることでもできるようになる。

このような状況で、“利用者自身が自分の個人情報や文献を管理し、それを互いに公開することのできるシステムが欲しい”という要求が出て来るのは、自然なことである。

本研究は、個人情報や文献情報を利用者自身が管理し他の利用者に提供するためのシステムを構築し、システム構築や利用時の課題を明確にし、解決することを目的とする。

2. ディレクトリサービスについて

ディレクトリとは人名録や住所録のことであり、このディレクトリに含まれている情報を提供するサービスをディレクトリサービスと呼ぶ。しかしながら、最近はこれだけではなく、与えられたキーに関する情報を提供するサービスを、広い意味でのディレクトリサービスと呼んでいる(以降、単に「ディレクトリサービス」という場合、広い意味でのディレクトリサービスを示す)。ディレクトリサービスでは特に、個人や計算機などの実世界の個々の対象物(オブジェクト)に関する情報を取り扱いの単位とし、情報管理の単位とする。また、取り扱う情報が、広く分散している場合が多いため、情報の管理も分散して行うことが多い。

ディレクトリサービスの例として、NIS (Network Information Service) や BIND (Berkeley Internet Name Domain) などがある。しかし、これらは広域分散環境での使用に適していない、特定のデータ型しか取り扱うことができないなどの欠点がある。

他にディレクトリサービスを規格化したものとして、OSI ディレクトリサービスがある。これは X.500 シリーズや ISO 9594 シリーズで規定されているディレクトリサービスである。この OSI ディレクトリサービスは、広域分散環境で使用することを考慮して設計されており、多くの種類のデータ型を取り扱うことができる。

我々は、この OSI ディレクトリサービスを利用して、各利用者の個人情報および文献を管理し、提供するためのシステムを構築する。

3. ディレクトリサービスを利用した文献 管理システム

3.1 概要

現在、ネットワークを通じて多くの文献データベースを利用することができるけれども、それらの多くは以下のような欠点を持つ。

- 論文を書いてから登録されるまでに時間がかかる
- ユーザ自身が自分の文献情報を管理することができない
- 登録される文献に全国大会や研究会以上などという制限がある
- 文献情報は検索できるけれども文献本体は別に探さなければならない
- 公開する範囲を限定することができない

本システムは、各個人がそれぞれ自分の書いた文献の情報を管理するというものであり、これらの欠点を解消することができる。また、OSIディレクトリサービスを利用して実現しているため、個人情報も同時に管理できる。これにより、興味のある文献を見つけた際にその人が書いた別の文献を参照したり連絡先を調べたりすることができる。

ネットワークを通じて利用できる文献として、RFCやCCITTの勧告書などがあり、これらも同時に登録する。ただし、このようなアクセスコントロールをほとんど要しない文献を管理するシステムとして、gopherやwaisなどの優れたものがあるため、それらを利用するのも1つの手段ではある。

本システムを実現するためには、以下の点を考慮する必要がある。

- 著作権、プライバシーの問題

- 個人単位の管理

- キーワードの抽出支援
- ユーザインターフェース
- マニュアルの整備
- セキュリティ機能
(アクセスコントロールなど)

- 検索機能の充実

- 高速な検索
- 高機能な検索

- UUCPサイトへのサービス

今回は、これらのサービスを実現するための実験基盤の構築およびキーワード抽出支援ツールの作成を行った。

3.2 実現

3.2.1 スキーマの設計

設計したスキーマの概要を以下に示す。今回はこのようなスキーマを定義したけれども、今後これがあると便利というものが出てくる可能性がある。これらは、実際に運用して利用者の希望を聞きながら改良する必要がある。

個人情報

氏名、連絡先電話番号、連絡先郵便番号、連絡先住所、自宅電話番号、自宅郵便番号、自宅住所、肩書、役職、メールアドレス、よく使用する計算機、ユーザID、スケジュール、顔写真のイメージ、誕生日、好きな飲物、所有資格、本籍、趣味、所属学会、最終更新日時、アクセスコントロールリスト

組織情報

組織名、ドメイン名、ネームサーバ(海外

向け、国内向け)、使用 IP アドレス、技術責任者、ニュースサーバ、メールサーバ

文献情報

著者、参考文献、キーワード、説明、使用言語、最終更新日時、著作権、文献の状態、アクセスコントロールリスト、ファイル名、ファイルフォーマット

ファイル情報

ファイル名、ファイルフォーマット、ファイルサイズ、使用コード、キーワード、説明、ファイル本体、アクセスコントロールリスト

ここで、アクセスコントロールリストというのは、そのエントリに関するアクセス制限を記述したものである。

3.2.2 使用するソフトウェア

本システムを使用するためには、当然のことながらソフトウェアが必要である。このソフトウェアに求められる条件は、以下の通りである。

- 管理者が自由にスキーマを定義できる
- ソースを無料で入手できる
再配布が可能である
- 分散管理できる
- アクセスコントロールできる
- 特定のキーワードにより検索できる
- ユーザ自身が自分の情報を管理できる
- 文献を取り扱える
- 日本語を取り扱える

これを始めから作成することも考えられるけれども、上述の条件をある程度みたすものに QUIPU というソフトウェアパッケージがある。QUIPU は OSI ディレクトリサービスを実現したものの 1 つである。本システムでは、この QUIPU を使用することにする。ただし、上述の条件で QUIPU に欠けているものとして、文献の取り扱いおよび日本語の取り扱いがある。これらについては、QUIPU を拡張することで実現する。

3.2.3 QUIPU について

QUIPU は、ISODE という既存の TCP/IP ベースのネットワーク上で OSI を学習するために作成されたソフトウェアパッケージに含まれている。ISODE はフリーソフトウェアで、ソースコードが公開されている。ISODE は C 言語で書かれており、BSD 系 (4.2,4.3) および System V 系 (R2,R3) の UNIX 上で使用できる。平成 5 年 2 月現在の最新バージョンは 8.0 である。QUIPU は OSI ディレクトリサービスの規格に対して、ユーザ認証やアクセスコントロールが拡張されている。

QUIPU には dish、fred、pod、sd、de などの、多くのクライアントプログラムが付属している。また、日本語の取り扱いについては、QUIPU 自身が多国籍語を意識して設計されているため、日本語化は容易である。

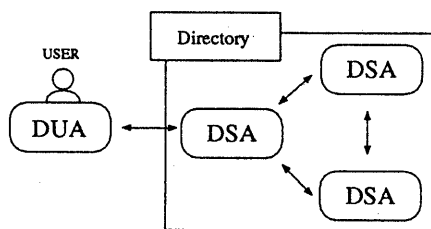
QUIPU は、ルーティング情報などのように動的に変化する情報を扱う応用や、一時的にでも情報の一貫性がとれていないと致命的になるような応用での使用は考えられていない。この点については、個人情報や文献情報はそれほど動的に変化することはないため、支障はないと考えている。

● 動作原理

表 1: 現在動作しているサーバ

マシン名 (IP アドレス)	DSA 名	管理している組織
nic.karrn.ad.jp(192.50.15.19)	Gladiolus	ルート、日本、KARRN
kyu-cs.csce.kyushu-u.ac.jp(133.5.19.4)	Apricot	九大、九大情報、 九大大型計算機センター
mikan.taurus.cse.kyutech.ac.jp(131.206.37.10)	Violet	九工大、九工大情報
reiko.keisu.kyushu-u.ac.jp(133.5.10.11)	Wisteria	九大中央計数施設
rhea.csce.kyushu-u.ac.jp(133.5.16.18)	Hydrangea	九大情報材料分室

QUIPU は、サーバである DSA(Directory System Agent) とクライアントである DUA (Directory User Agent) からなる。



ユーザは DUA を用いて、適当な DSA に要求を出す。その DSA が必要なデータを持っていない場合に DSA の間で行われる通信の形態として、OSI ディレクトリサービスでは、以下の 3 通りがあげられている。

- referral
DUA から問い合わせを受けた DSA が、DUA に他の DSA を紹介する。
- chaining
DUA から問い合わせを受けた DSA が、他の DSA に問い合わせに行く。それでも不十分なら、またその DSA が別の DSA に問い合わせに行く、という動作を繰り返す。
- multicasting
DUA から問い合わせを受けた DSA

が、他の複数の DSA にそのデータを持っていないかどうかを問い合わせる。

これらの方法のうち、QUIPU では referral と chaining をサポートしている。

3.2.4 文献の管理 / 転送方法

文献を管理、転送する方法として、以下の 2 つの方法がある。

- 他の属性と同様に管理し転送する
- 別に管理して ftp などのファイル転送専用のプロトコルを使用して転送する

後者の方法では、アクセスコントロールを行うことが困難であるため、前者の方法を選択した。

3.2.5 実験基盤について

実験基盤として、九州地域研究ネットワーク KARRN(Kyushu Area Regional Research Network)^[3] 上にプロトタイプを構築した。現在のところ、表 1. の組織でサーバが立ち上がっている。

これらのサーバは互いに 10Mbps の Ethernet または 128Kbps の専用回線で接続されており、図 1. のような名前空間を形成している。

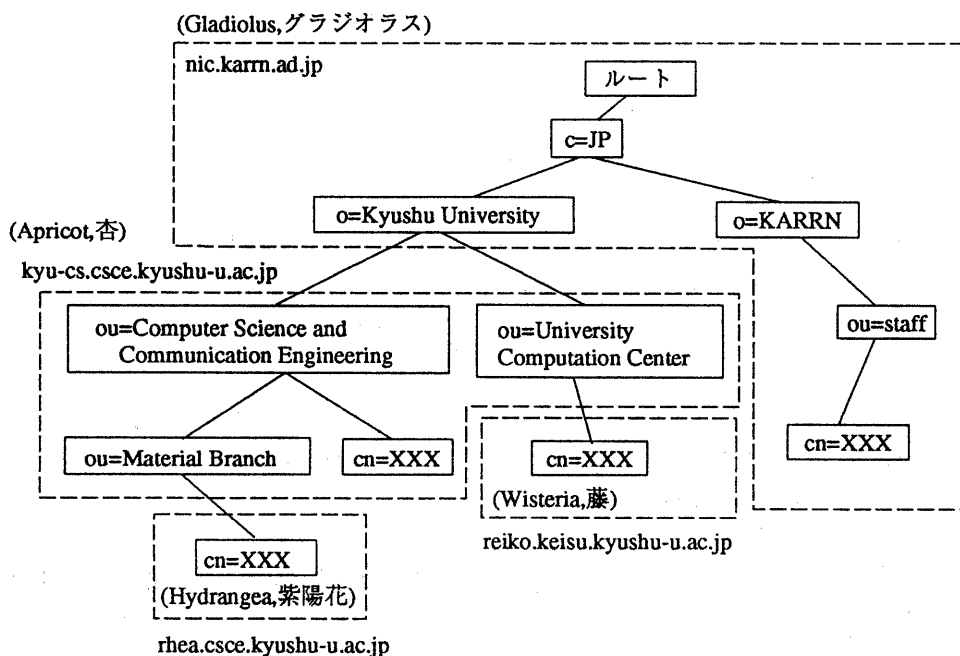


図 1: サーバの構成する名前空間

3.2.6 キーワード抽出支援

ユーザ自身が管理を行うため、文献の登録時のキーワード付与作業もユーザ自身が行う。しかしながら、各ユーザがばらばらに文献からキーワードを抽出すると、人によりかなりばらつきが出る。そこで、図 2. に示すように、自動的にキーワード候補を抽出し、それを参考にしながらユーザがキーワードを決めていく方法を採用した。この方法をとることにより、人によるばらつきをある程度なくすることができる。

本ツールがキーワード抽出支援のために行う作業は、以下の 2 つである。

- キーワード候補辞書に登録されている単語を文書中から抽出する
- 残りの文章から漢字 / カタカナ / アルファベット / 数字列を抽出し、不要なものの削

除、重要度評価、整列を行い、出力する

我々は、本ツールを C 言語を用いて作成した。不用語辞書は Wnn に付属の辞書から名詞のみを取り出して使用するけれども、キーワード候補辞書は、始めは空にしておき、ユーザがキーワードとして選んだものをキーワード候補辞書に加えていくことにする。そのため、始めは辞書に登録されている単語が不十分であっても、辞書が充実するに従い、抽出精度が向上する。

本ツールでは、キーワードを単に抽出するだけでなく、その文献におけるキーワードの重要度も共に登録する。すなわち、キーワードが XX である、というのではなく、キーワードが XX で重要度が 3 であるとする。これをそのまま登録することにより、検索時に、キーワード XX が重要度 3 以上である文献を探す、ということが可能になる。

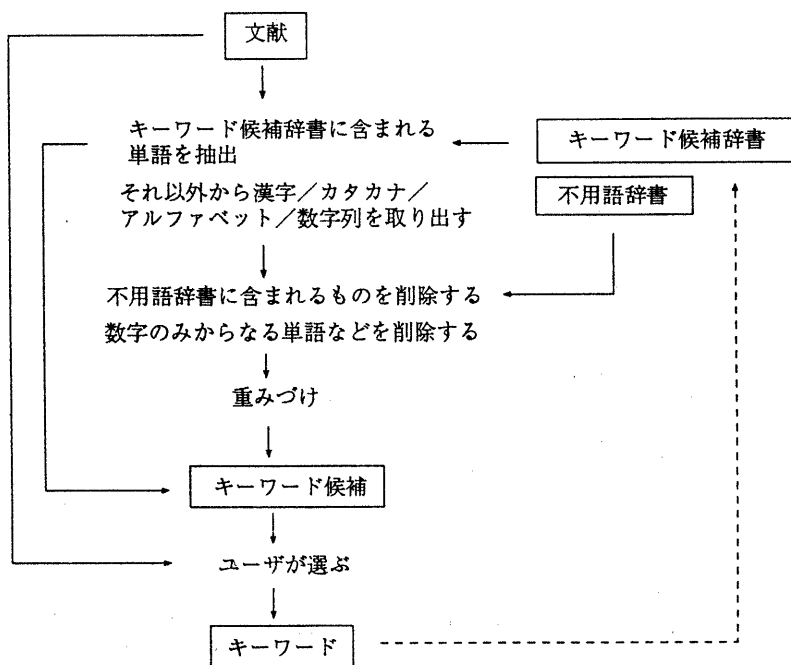


図 2: キーワード抽出作業の流れ

3.2.7 whois サービス

個人情報を提供するサービスとして、whois サービスが広く利用されている。この whois サービスは、サービスを提供するサーバが情報を保持し、それを提供するだけの単純なものであり、サーバ間での情報交換をするということはいわれない。そのため、どのサーバがどのような情報を提供しているかをあらかじめ知っておく必要がある、データの登録や更新の度にサーバが保持している情報を変更しなければならないなどの欠点がある。

本サービスでは個人情報を管理、提供しており、分散して保持された情報は、それを意識することなく利用することができる。また whois サービスとは異なり、登録、更新を手元で行うことができる。

そこで、本サービスで提供している個人情報

を、whois サービスと同じインターフェースで利用できるツールを作成した。

要求がある度に検索を行うのでは時間がかかるため、インデックスを用いて検索時間を短くする。

4. 評価

4.1 システムサイズ

本システムを使用するには、最低限、表 2. に示すディスク容量が必要である。コンパイラは Sun OS 4.1.2 に付属の cc を使用した。また、単位は Kbytes である。表 2. で、tex ファイルを $\text{\LaTeX}$ にかけたものが dvi ファイルで、それを PostScript プリンタの出力形式に変換したものが PostScript ファイルである。

表 2. でクライアントのみの場合で約 2Mbytes、サーバを含めて約 2~3Mbytes のディスク容量が必要であることが分かる。ただし、ログを取

表 2: 必要なディスク容量

	内容	Sparc Station 2	Sun 3
システム	DSA(設定ファイルを含む)	1,600	1,300
	DUA(dish のみ、設定ファイルを含む)	1,200	900
	設定ファイル(共通分)、マニュアル	100	
登録データ (1 人分)	個人情報	2	
	顔写真 (100x100 dot,8plane)	10	
	文献情報	0.5	
	文献 (概要)	2	
	文献 (本体 (B4,6 枚程度),tex)	22	
	文献 (本体,dvi)	26	
	文献 (本体,PostScript)	550	

(Kbytes)

る場合などには、これ以上のディスク容量が必要である。また、X-window system 上で動作する DUA が必要な場合は、さらに 1~2Mbytes が必要である。また、100 人がそれぞれ個人データ (顔写真を含む) と文献データ (文献本体を含む) を 50 個 (tex 形式、B4 で 6 枚程度) 登録したとすると、データだけで約 110Mbytes のディスク容量が必要である。ただし、組織内のすべてのデータを 1 台の計算機で管理する必要はないため、登録するデータが増えてくると組織内の部署ごとに別々の計算機でサーバを立ちあげ、分割することができる。

メモリ使用量は、DUA の 1 つである dish を使用した場合、最低 1.5Mbytes 必要である。ただし、データを読み込むためサイズが大きくなり、B4 で 6 枚程度の文献を 2~3 個読み込むと 3Mbytes 程になる。また、DSA については、起動時に読み込むデータの量によって異なるが、3~5Mbytes 程である。

4.2 キーワード抽出支援ツール

今回作成したキーワード抽出支援ツールを、約 25,000 字の文章を用いて評価した。また、評

表 3: 評価結果 (単位:%)

重要度	再現率	適合率
5	100	50
4	18	13
3	58	26
2	71	25
1	80	5
重みが 0 より大	57	42

価のために、事前に人手によりキーワードを 5 段階に分けて抽出しておいた。結果を表 3. に示す。ここで、再現率、適合率を式 (1)、(2) のように定義する。ここで、重要度 n の再現率とは、人手で抽出したキーワードで重要度が n 以上の単語と、本ツールで抽出したキーワードで重要度が n 以上の単語から計算した再現率である。

キーワード候補辞書は空のものを使用した。また、不用語辞書は Wnn Version 4.1 の基本語辞書から名詞を抽出して使用 (単語数 18,678 個、平均語長 5.8 字) した。実行時間は Sparc Sta-

$$\text{再現率} = \frac{\text{人手で抽出したキーワードのうち、本ツールで抽出したキーワードの数}}{\text{人手で抽出したキーワードの数}} \quad \dots \quad (1)$$

$$\text{適合率} = \frac{\text{人手で抽出したキーワードのうち、本ツールで抽出したキーワードの数}}{\text{本ツールで抽出したキーワードの数}} \quad \dots \quad (2)$$

tion 2 で約 10 秒であった。

再現率、適合率ともによい値とは言えないが、キーワード候補辞書が充実すると、これらの値は向上ことが期待できる。さらによい精度を望むならば、重みづけの条件を増やしたり形態素解析を行うなどの手段が必要である。

4.3 分散管理システムとしての評価

システムの柔軟性 システムの一部を変更する(スキーマの変更やデータ型の変更など)際に、全てのサーバ、クライアントで変更を行う必要があった。ただし、スキーマを変更した場合は、他のサーバに反映させる、あるいは他のサーバから変更ができる、などの機能が必要である。

名前空間 全体が1つの名前空間から構成されているため、ユーザは遅延の問題を除いて分散を意識しなくてよい。

検索機能 現在はローカルエリアネットワーク上で動作しているため、検索速度に問題はないけれども、さらに低速の回線を介して使用すると、検索に時間がかかる可能性がある。

5. おわりに

一般に、X.500 は計算機の負荷が重い処理である。本システムも例外ではなく計算機の負荷が軽いものではないけれども、計算機やネットワーク技術の発達によって、問題にならなくな

ることが期待できる。そのために、資源の管理を効率良く行うことが重要である。

本システムは、実験基盤を構築したのみである。十分なサービスを提供できる環境にいたっていない。今後は、さまざまな実験を行い、サービスの充実、拡大を目指す必要がある。

謝辞

本研究は WIDE プロジェクト ISODE WG の方々の御協力により検討したものである。また、九州地域研究ネットワーク KARRN の方々に感謝する。

参考文献

- [1] "The Directory - Overview of Concepts, Models and Services", CCITT Recommendation X.500, Dec., 1988.
- [2] C.J.Robbins, S.E.Kille, "QUIPU", The ISO Development Environment User's Manual, Vol.5, Jul., 1991.
- [3] 梅田政信: "九州地域研究ネットワーク KARRN の構築", 東京大学大型計算機センター研究論文集「地域ネットワークの課題」, pp.67-74, Jul., 1992.
- [4] 吉田、片山、松山、砂原: "OSI ディレクトリサービスの実験", 情報処理学会 マルチメディア通信と分散処理研究会, 1991.

- [5] 副島、木實、最所、古川、荒木:“九州地区ディレクトリサービスについて”, 情報処理学会九州支部研究会,1992.
- [6] 副島、古川、最所、荒木 :“九州地区ディレクトリサービスのための実験基盤の構築について”, 電気関係学会九州支部連合大会論文集, Oct.,1992.