ヒューマンインタフェース 77－2
（1998．3．13）

WWWドキュメント検索における ドメイン名クラスタリングの利用

島村 栄　高野 元
NEC C&Cメディア研究所

近年のWWWにおける提供情報の爆発的増加に伴い，必要なドキュメントを発見するためにはWWWディレクトリサービスが必要不可欠となっている．現在，多くのディレクトリサービスが提供するキーワード検索機能では，検索結果の表示がシステムの計算する重要度順のURLのリストである．このため，検索結果のリストが非常に長くなった場合，その中からユーザが求める情報を発見することが困難であった．筆者らはこの問題を解決するために，検索結果の候補URLのインターネットドメインから組織の名称を取り出し，検索結果を組織名称によってクラスタリングし階層化して木構造の形式でユーザに提示するシステムを開発した．

A Domain Clustering for WWW Document Search

Hisashi Shimamura, Hajime Takano

C&C Media Research Laboratories, NEC Corporation

Because of the recent explosive increase in the number of WWW documents, directory services are indispensable in finding needed documents. In the keyword search function of most directory services, search results are displayed as a URL list ordered by importance calculated by the system. It is difficult to find useful documents from the list when it is very long. To solve this problem, the authors have developed a new WWW search system that clusters the documents in the search result by the organization name, which is derived from its URL domain name. The system displays the clusters in a hierarchical tree view form.

1. はじめに

近年の World Wide Web (WWW)の急激な広がりに伴い，大量で多種多様な情報が WWW 上で公開されるようになった．この大量の WWW ドキュメントの中から，一般ユーザが必要な情報を WWW 上で効率よく発見することは非常に困難になっている．

現在，WWW 上での情報発見を容易にするために多くのディレクトリサービスが公開されている．このディレクトリサービスを用いてキーワード検索を行うことにより WWW 上に存在する膨大なドキュメントの中から求めるものを選び出すことができる．このようなディレクトリサービスとして，NETPLAZA (http://netplaza.biglobe.ne.jp/), AltaVista (http://www.altavista.digital.com/), Infoseek (http://japan.infoseek.com/), goo (http://www.goo.ne.jp/), Yahoo! (http://www.yahoo.co.jp/)等があげられる．

しかし，検索の対象となるドキュメントの数が増えるに連れ，単純なキーワード入力のみの検索では検索結果としてヒットするドキュメントの数が膨大になりがちである．1回の検索に対して数千～数万個の検索結果が返されることも珍しくない．そこで我々は，膨大になりがちなディレクトリサービスの検索結果を整理してユーザに提示し，より効率的に必要な情報を発見するための手段を提供するシステムを開発した．これは，検索結果としてヒットしたドキュメントの Uniform Resource Locator (URL)ドメインから組織名や部門名を抽出し，それに基づいて検索結果を分類，整理して提示するシステムである．これによってユーザは，たとえ大量の検索結果がヒットしても，自分の必要とする情報に近い組織によって提供されるドキュメントからチェックしていくことができ，また，必要ないと感じたドキュメントを提供する組織内の検索結果をチェックしなくてすむ．

我々は，このシステムを WWW ディレクトリサービスの一部として実装した．以下では，2節で従来のディレクトリサービスの持つ問題について整理し，3節で今回実装したドメインクラスタリング検索システムについて説明する．4節ではこのシステムに関する考察を行い，5節でまとめをおこなう．

2. 従来の WWW 検索システムの問題

WWW 上の膨大な情報からユーザが必要なものを探す手段として WWW ディレクトリサービスがある．このようなサービスとして現在，NETPLAZA, Yahoo!, goo, AltaVista, Infoseek 等が利用されている．このようなディレクトリサービスの多くは，あらかじめ収集した WWW ドキュメントの検索手段としてキーワード検索機能を提供している．検索結果は，各ドキュメント中の検索キーワードの出現頻度や見出しなどの出現場所に応じて重要度が計算され，重要度の高い順にユーザに提示される．現状の多くのディレクトリサービスでは，検索結果から必要な情報のみを絞り込むために以下のような条件を指定することができる．

- キーワード
- 複数キーワードを組み合わせた論理式
- 更新期間
- ドメイン，言語
- ドキュメントに含まれるメディア
  (例：音声，映像)
- ドキュメントに含まれる Plug-in
  (例：Java，Active-X)

これらの機能を利用することによってユーザはディレクトリサーバで収集管理されたドキュメントから自分の求める情報を発見することができる．しかし，実際には以下のような問題が生じがちである．

- **適切な検索条件を考えるのは困難．**
  検索結果が多すぎる場合，その中から候補を絞り込むためには追加キーワードなどを組み合わせて論理式を構成したり，ドキュメントの更新期間の指定など複雑な検索条件を入力しなけれ

ばならない．しかしこのような検索条件を入れることで，何も検索にヒットしなくなったり，逆に候補ドキュメントの数がほとんど変わらずまったく絞り込めない場合がある．検索結果を適当に絞り込む検索条件を考えるのは非常に難しい．

■**検索結果は一つ一つチェックしなければならない**

検索結果はリストの形式でユーザに示される．ユーザは必要なドキュメントを発見するまでリストの順に一つ一つ候補をチェックしていかなければならない．このため，すでにチェック済みのドキュメントの親や子ドキュメントや，類似したドキュメントを繰り返しチェックすることが多い．

■**検索結果リストの出力順序が必ずしも便利でない．**

検索結果ドキュメントは重要度順のリストでユーザに提示される．この重要度は一般にWWWドキュメント内でのキーワードの出現頻度，出現場所から機械的に計算される．これはユーザにとって一種のブラックボックスであり，必ずしもユーザのニーズと合致しない．このため，非常に多くのドキュメントをチェックした後に必要なドキュメントが現れることがある．

3. クラスタリング検索システム

この節では，本研究で開発したドメインクラスタリング検索システムについて説明する．

3.1. 木構造インタフェース

前節で挙げた問題に対する解決方法として，以下のような点に考慮した検索システムを開発した．

1.インタラクティブな検索インタフェース
2.検索結果に対するランダムアクセス
3.明確なルールによるクラスタリング

この検索システムの検索結果表示には木構造のビューを持つインタフェースを用い，明確なクラスタリングルールとしてURLドメインによるクラスタリングを採用した．

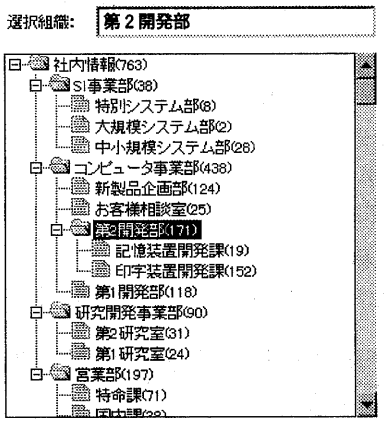

図 1: 木構造インタフェース

図 1は木構造インタフェースの例である．このインタフェースはキーワード検索の結果を示していて，URL ドメインが表す階層的な組織構造で結果ドキュメントをクラスタリングしている．フォルダアイコンが下位クラスタの集合としてのクラスタを示しており，ファイルアイコンは末端クラスタを示している．それぞれのクラスタは一つのURL ドメインに対応した組織をあらわしているが，URL ドメインに現れない組織名を下位組織の集合からなる上位クラスタとして登録ことができる．

各アイコンの後ろの文字列がそのクラスタの示す組織名で，その後ろの数字はその組織によって提供される検索結果の件数である．フォルダアイコンをダブルクリックすることで下位構造を展開，縮退表示することができ，ファイルアイコンをダブルクリックすると，そのアイコンの示す組織内の候補ドキュメントを確認することができる．

このユーザインタフェースは以下のような特徴を持っている．

■インタフェース上でのマウスクリックだけで検索結果の絞り込みができる．
■さまざまな詳細度で検索結果全体を眺めることができる．

■どこからでも好きな部分から検索結果をチェックすることができる．

3.2. システム概要

このシステムは大きく，検索インタフェースとクラスタリングサーバに分けることができる．以下ではそれぞれの部分の構成について説明する．

3.2.1. 検索インタフェース

図 2は検索インタフェースの外観である．

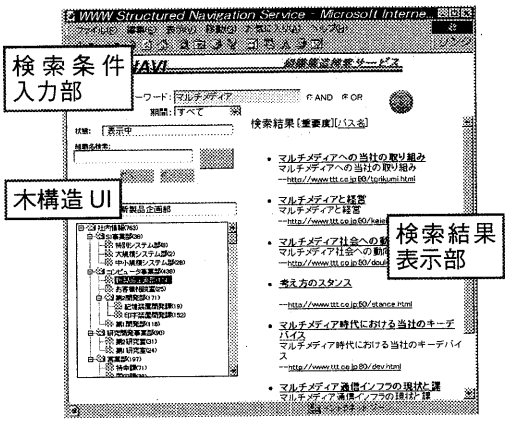

図 2：検索インタフェースの外観

画面上部が検索キーワード入力部，左下が検索結果をドメインによるクラスタリングによって表示する木構造ビュー，右下が木構造ビュー内で指定された組織内の検索結果ドキュメントの URL，タイトルなどを表示する部分である．

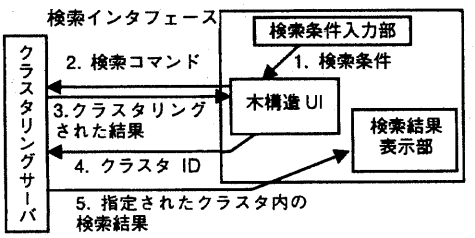

図 3：検索インタフェースの構成

図 3は検索インタフェースの構成とデータフローを示している．一般的な検索操作において，検索インタフェースは以下のように動作する．

1. 検索条件入力部は木構造ユーザインタフェース(UI)へ検索条件を渡す．
2. 木構造 UI は与えられた検索条件に基づいてクラスタリングサーバに対して検索コマンドを発行する．
3. クラスタリングサーバはクラスタリングされた検索結果を返す．木構造 UI はこの階層化された検索結果を表示する
4. ユーザが木構造 UI で指定したクラスタ識別子(ID) と，検索条件をクラスタリングサーバに渡す．
5. 検索結果表示部はクラスタリングサーバから返された指定クラスタ内の検索結果ドキュメントの情報を表示する．

3.2.2. クラスタリングサーバ

クラスタリングサーバはキーワードによる検索結果を URL のドメイン名をもとにクラスタリングし，そのドメイン名から現実の組織の名称を求める．現実の組織名はユーザにとって非常にわかりやすいクラスタリングキーであるといえる．

クラスタリングサーバは以下の 2 種類のデータベースを持つ

■ディレクトリサーバデータベース
あらかじめ収集された WWW 上の多くのドキュメントに関する情報を蓄積したデータベース．

■組織名データベース
URL ドメイン名と現実の組織名の組を格納したデータベース．例えば，"nec.co.jp"とそれに対応する組織名「日本電気株式会社」の組が格納される．

図 4はクラスタリングサーバの構成とデータフローを示している．クラスタリングサーバの動作は以下の手順で行われる．

1. 検索インタフェースはキーワード検索部に検索コマンドを渡す．
2. キーワード検索部はディレクトリサーバデータベースを検索コマンドに従って検索する．得られた検索結果ドキュメントを URL ドメイン名の順にソートして検索結果クラスタリング部へ渡す．

3. 検索結果クラスタリング部は，検索結果に現れる URL ドメイン名を現実の組織名へ変換し，同じ組織名の URL のドキュメントをクラスタリングする．この結果を検索インタフェースへ返す．
4. 検索インタフェースは木構造 UI で指定されたクラスタ ID と検索条件を渡す．
5. クラスタリングサーバは指定されたクラスタ内の検索結果ドキュメント情報を検索 UI へ返す．

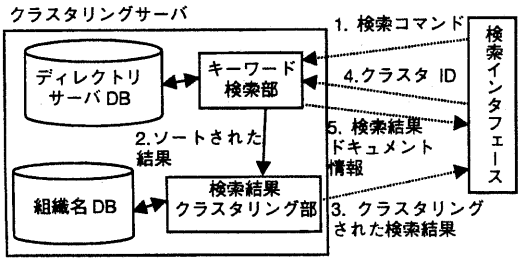

図 4：クラスタリングサーバの構成

3.3. WINGNAVI における実装

このクラスタリング検索システムは，NEC のイントラネット内での実験ディレクトリサービス WINGNAVI [10]上で運用されている．実装上の特徴は以下の通りである．

検索インタフェース：
- 検索条件入力部は HTML と Java Script で記述
- 木構造インタフェースは Java アプレットとして実装
- 検索条件入力部と木構造インタフェース間のやりとりには LiveConnect [4]を利用
- 検索結果ドキュメント表示部は HTML で記述

クラスタリングサーバ：
- クラスタリングサーバでは Netscape サーバを使用
- ディレクトリサーバデータベースは Oracle で実現
- 組織名データベースは GNU dbm で実現
- キーワード検索部と検索結果クラスタリング部は NSAPI を用いて Netscape サーバアプリケーションとして実装

4. 考察

4.1. 評価

この節では，クラスタリング検索システムによる検索実験の結果について述べる．実験は NEC イントラネット内の実験ディレクトリサービス WINGNAVI で収集された約 32,000 のドキュメントに対して検索をする形式で行った．表 1 は階層化クラスタリングされた検索結果の第 2 階層，第 3 階層に存在するクラスタ数と個々のクラスタにどの程度のドキュメントが含まれているかを示している．表の各行は左端の列のキーワードによる検索結果を示しており，各列の意味は以下の通りである．
- キーワード：検索キーワード
- 検索ヒット数：検索結果に含まれる総ドキュメント数
- クラスタ数：第 2，3 階層それぞれに含まれるクラスタ（組織）数．
- 最大ドキュメント数：第 2，3 階層の各クラスタ内に含まれる検索結果ドキュメント数の最大値

キーワード	検索ヒット数	クラスタ数		最大ドキュメント数 (documents/cluster)		平均ドキュメント数 (documents/cluster)	
		第2階層	第3階層	第2階層	第3階層	第2階層	第3階層
データベース	495	28	69	123	84	17.7	7.2
携帯	211	17	43	36	26	12.4	4.9
マルチメディア	656	32	76	117	56	20.5	8.6
ネットワーク	1796	35	97	312	146	51.3	18.5
NT	787	28	71	138	81	28.1	11.1
検索	1182	35	93	280	112	33.8	12.7
平均	854.5	29.2	74.8	167.7	84.2	27.3	10.5

表 1：イントラネット内の検索結果

■平均ドキュメント数：第 2，3 階層の各クラスタ内に含まれる検索結果ドキュメント数の平均値

この結果は，ユーザがもし第 2 階層までの組織名の知識があれば，855 個の検索結果ドキュメントを平均で 27 個に絞り込むことができることを示している．この数はユーザが一つ一つチェックすることが比較的容易な数だといえる．

また，検索の結果，第 2 階層には平均 30 個のクラスタがあるため，一つのクラスタをマウスクリックするのみで検索結果ドキュメントを平均で全体の 3.4% (1/30)へ絞り込むことができることになる．さらに第 3 階層までの知識があれば，一つのクラスタを指定することで全体の 1.3% (1/75)へ絞り込むことができる．

4.2. 今後の課題

上記のように，このクラスタリング検索システムによってユーザの検索を支援することができるが，今後の課題として以下のものがある．

■検索結果ドキュメントがそれほど多くない場合，クラスタのクリックがかえって煩わしい場合がある．

■組織について知識がほとんどないユーザにとってメリットが小さい．

■インターネットへ応用する場合，組織名データベースの維持管理が難しい．

5. おわりに

これまでのディレクトリサービスにはいくつかの問題があったため，求めるドキュメントを発見することは非常に困難であった．この問題を解決するため，検索結果ドキュメントをクラスタリングし，そのクラスタを階層化された木構造インタフェースで視角化する検索システムを開発した．このクラスタリングには URL のドメイン名を組織名に変換したものを利用した．このシステムには以下のような利点がある．

■GUI 上の簡単な操作で検索結果ドキュメントを絞り込むことができる．

■組織に関する知識によって効果的に必要なドキュメントを発見できる．

■階層化された木構造インタフェースで検索結果を概観することができる．

このシステムをイントラネット内で運用した結果，検索支援の有効性が確認できた．

参考文献

[1] Achick, P.G. Vaithyamathan, S., "Exploiting Clustering and Prases for Context-Based Information Retrieval", SIGIR'97 pp.314-323, Philadelphia PA, USA, 1997.

[2] Chang, C., Hsu, C., "Customizable Multi-Engine Search Tool with Clustering", Sixth World Wide Web Conference, pp.257-264, California, USA, 1997.

[3] "THE JAVA DEVELOPERS KIT 1.0.2", http://www.javasoft.com/products/jdk/1.0.2

[4] "The LiveConnect/Plug-in Developer's Guide", Netscape Communications Corporation, http://home.netscape.com/eng/mozilla/3.0/handbook/plugins/

[5] Lamping, J., Rao, R., "Laying out and Visualizing Large Trees Using a Hyperbolic Space", Proceedings of UIST'94, pp. 13-14, 1994.

[6] Maarek, Y. S., Shaul, I.Z.B., "Automatically Organizing Bookmarks per Contents" http://www5conf.inria.fr/fich_html/papers/P37/Overview.html, Fifth World Wide Web Conference, Paris, France, 1996.

[7] Nowell, L. T., France, R. K., Hix, D, Heath, L. S., Fox, E. A., "Visualizing Search Results: Some Alternatives To Query-Document Similarity", SIGIR'96, pp. 67-75, Zurich, Switzerland, 1996.

[8] "The Netscape Server API", Netscape Communications Corporation, http://home.netscape.com/newsref/std/server_api.html

[9] Robertson, G. S. Mackinlay, J. D., "ConeTrees: Animated 3D Visualizations of Hierarchical Information", Proceedings of CHI'91, pp. 189-194, 1991.

[10] Takano, H., Kubo, N., Shimamura, H., Matsuura, H., "A Directory Service Architecture for the World-Wide Web", http://www5conf.inria.fr/fich_html/posters.html, Fifth World Wide Web Conference, Paris, France, 1996.