グループウェア 33−13
(1999. 10. 22)

イントラネット向け文書ディレクトリ管理システム

鵜飼孝典
ugai@flab.fujitsu.co.jp
富士通研究所ドキュメント処理研究部

概要
インターネットに置かれた有用な情報を階層化されたカテゴリに分類して整理して提供する文書ディレクトリは，有用な情報を探すためのツールとしても，情報共有システムのひとつとしても非常に有効なシステムである.こういった文書ディレクトリを維持するには多大な労力が必要であるが,イントラネットでは,小人数で維持,管理しなければならないことが多く,省力化が必須である.本稿では，文書ディレクトリの管理に使われる既存の方法,技術をまとめ,イントラネットで文書ディレクトリを維持管理するためにより重要になる要件をまとめて,最後にわれわれが開発中の文書ディレクトリ管理システムについて報告する.

Browseable Directory Management System for Intranet

Takanori Ugai
Fujitsu Laboratories Limited, Document Processing Laboratory

Abstract
Browseable document directory is a system, which categorize lots of information in the Internet and Intranet as a graph structure. It is an important tool as not only information retrieving tool but also information sharing tool in Intranet. It is time-consuming, expensive, and error-prone to build and maintain a document directory. The directory is maintained by only a few staffs especially in Intranet and the maintenance cost is critical. In this report, we summarize existing technologies and tools to maintain directory and requirements for maintenance a directory in Intranet. Finally we report our browseable directory management system for Intranet.

1. はじめに

インターネットでは，近年一般のユーザが自由にホームページを開設することが容易になった一方で，本当に有用な情報が何処にあるのかを探すことが難しくなった．インターネットに置かれた有用な情報を階層化されたカテゴリに分類して整理して提供する Yahoo[4] のような文書ディレクトリは，有用な情報を探すためのツールとしても，情報共有システムのひとつとしても非常に有効なシステムである．しかしこういった文書ディレクトリを維持するには多大な労力が必要であり，例えばYahooでは，数百人もの「サーファ」と呼ばれる専門家がディレクトリへの登録申請の内容をチェック，分類，登録を行ない，古くなった登録内容を更新する作業を日夜行っている．インターネットでは多大な労力を払っても広告収入などのビジネスモデルが成立するが，イントラネットでは，小人数で維持，管理しなければならないことが多く，省力化が必須である．

本稿では，イントラネットで文書ディレクトリの維持管理を行なうために必要な機能を重視したシステムを提案する．第2章で文書ディレクトリのサービスとしての特徴をまとめ，第3章で文書ディレクトリの管理に使われる方法，技術をまとめる．その上で第4章でイントラネットで文書ディレクトリを維持管理するためにより重要になる要件をまとめ，最後にわれわれが開発中の文書ディレクトリ管理システムについて報告する．

2. 文書ディレクトリ

インターネットで必要な情報を見つけ出すためのサービスには，インターネットに置かれた有用な情報を階層化されたカテゴリに分類して整理して提供する文書ディレクトリ以外にもキーワードを入力して,そのキーワードに関連する URL や文書を提示する全文検索サービスがあり，多くのサイトで全文検索サービスを提供するようになってきている．全文検索サービスの特徴として，検索をするときにあらかじめ必要なキーワードを知らなければならない．これは，特に情報の山に何が埋まっているかを知らずに検索をはじめるときには非常に困難なことがある．一方，文書ディレクトリには利用者から見た場合次のような特徴がある．

- クリックするだけで目的の文書に到達したり，ナビゲーションできる簡便なインタフェースをもつ．
- 質の良い文書のみを登録されている．
- 静的なタクソノミやシソーラスではなく，ダイナミックに階層を更新しされ文書の交叉分類も許されている．
- ページはカテゴリ毎に整理されていて，まとめた範囲を代表するタイトルとしてカテゴリとして名前がつけられているので，情報を見つけ出すときのヒントになる．
- 各 URL に関する簡単な説明を検索対象とする検索サービスを提供している．

といった特徴があり，何か面白い物はないかなと比較的あいまいな要求を持って情報を検索する時や，なるべく良質な情報だけ見たいと言う時に使われることが多い．そのためディレクトリには次のようなことが期待されている．

- ディレクトリやカテゴリに対して，正確にそのカテゴリに格納されている情報を反映した名前がつけられていること．
- 適切な情報が各カテゴリに格納されていること．
- ディレクトリの階層や構造が適切であり，十分広い範囲がカバーできていること．
- 文書は適切な場所に置かれていれば，必ずしもディレクトリの一番下の階層ではなく，また，ただひとつのカテゴリの中であるとも限らないこと．
- 一定の矛盾の無い視点からのカテゴリが構成されていて，局所的にカテゴリが作成される視点が変わったりしていないこと．

- 内容は常に更新され，ディレクトリの構造やカテゴリの名前も時代に沿ったものであること

しかし，先にも述べたようにこういった文書ディレクトリを維持するには多大な労力が必要である．例えばYahooでは数百人もの専門家が,申請されたホームページの内容をチェックし，分類，登録，更新する作業を日夜行っている．

3. 文書ディレクトリの管理技術

本章では文書ディレクトリの管理に使われる既存の方法，技術について述べる．

3.1. オープンディレクトリ

先に述べたように，文書ディレクトリを維持するには多大な労力が必要である．そこでオープンソースの考え方と同じようにボランティアユーザの手によって，文書ディレクトリの維持管理を行うプロジェクトが考えられ，開始された．

そのうちの最初のひとつである Open Directory Project[2]は Netscape, Lycos, HotBot によって 1998 年 7 月に開始された．1999 年 9 月 19 日現在 946,756 のサイト，15,448 の編集者が登録され，カテゴリは 152,814 に登り，毎日 1000 ほどのサイトが新規に登録されている．Open Directory Project には 10 人ほどの専任編集者がおり，それ以外の編集者はすべてボランティアである．ボランティアは自分が編集したいカテゴリに編集者として応募し，承認されると応募したカテゴリから下位のカテゴリの編集を行うことができる．編集者は,主にカテゴリに格納する適当な URL を探してきて登録することをおこない，編集者以外から登録依頼の内容をチェックし，分類，登録，更新をおこなう．スーパーエディタと呼ばれる全権を持つ専任編集者は，カテゴリのボランティア編集者間のいざこざに対する調停役も行う．Open Directory Project では編集者の作業を軽減するために，ディレクトリに登録されている URL を自動的にチェックするシステムや URL の登録を簡単化するために，タイトルや内容を HTML ファイルの情報を自動的に参照する機能などが開発されている．このプロジェクトで作成されたディレクトリのカテゴリの構造，登録されたサイトのデータはすべて公開され，自由に利用することができる．Open Directory Project で最近問題になっている点は

- 編集者が足りず，チェック，分類，登録を待つ登録依頼がさばききれないこと．
- 編集者の興味の偏りによって URL の登録が公平に行われないことがあること
- 編集者間のコミュニケーションの不足によりディレクトリ全体としての編集方針を一定に保つことができず，局所的な編集方針で編集が行なわれること．
- サイトの紹介文が簡単過ぎるなどの理由で不適切になる場合があること．

などである．

Infoseek がはじめた GO Guides[3]は，Open Directory Project と同様，誰でも参加でき，ディレクトリに紹介したいサイトを提案できる．Open Directory Project と異なる点は，GO Guides のコミュニティで承認されてはじめてそのサイトがディレクトリに掲載される点である．編集者が紹介したサイトを他の会員が評価する過程を経ることで,「スパム」をなくし一定の品質を維持できることを期待したものである．編集者の興味の偏りによって URL の登録が公平に行われないことやサイトの紹介文が簡単過ぎるなどの理由で不適切になる場合があることなどは防ぐことが期待できるが,編集者が足りず，登録依頼をさばききれないという問題点については，Open Directory Project に比べて不利といえる．登録したサイトが掲載された場合にはポイントが与えられランクアップする．ランクアップすると,他の参加者が提案したサイトの承認を審議したり,新しいディレクトリを作成できるなどの特典がある．GO Guides のサイト紹介では，特別なプロフィールを作成し，編集者の電子メールアドレスなどを掲載することもできる．それにより，電子メールでサイト紹介に対する意見を受け取ったり，編集者名でサイト紹介のレベルを判断するといったことができる．

3.2. 文書ディレクトリ管理ソフトウェア

文書ディレクトリを管理するためのシステムで公開されているものに GNU License で配布されている Senga の Catalog[6]がある. Catalog は MySQL[8]をバックエンドのデータベースシステムとして利用する Perl で作られたプログラムで文書ディレクトリの作成, 維持, 管理, 表示を行うためのものであり,すべての操作が Web ブラウザを通して行うことができる. XML の形式のひとつである RDF 形式[9]をサポートし Open Directory Project で配布しているデータを利用することができる.

3.3. ディレクトリの自動化,省力化システム

本節では, こうした文書ディレクトリの構築や運用を, 自動化, あるいは半自動化しようとする技術についてまとめる.

従来の文書自動分類に関する研究では,

- クラスタリング: カテゴリの階層の構築と, 文書のカテゴリへのマッピングの両方を自動的に行なう.
- クラシフィケーション, カテゴライゼーション: あらかじめ与えられたカテゴリ階層に対して文書のカテゴリへのマッピングを自動的に行なう.

という二系統が古くから研究されており, 精度という点では人手によって維持されているものには及ばないというのが現状である.

文書ディレクトリの運用を, テキストマイニング技術でソースとなる文書を分析してクラスタリングとクラシフィケーション, カテゴライゼーションを行うという研究[1]も発表されている.

1999 年になって, ディレクトリの自動化, 省力化に関する商品の発表が相ついでいる. 1999 年 4 月 Semio 社が Semio Taxonomy を発表した. Semio Taxonomy ではトップカテゴリを人手で作成するとトップ以下のレベルのカテゴリは自動的に作成される. Semio 社は使われている技術の詳細については公開していないがソースとなる文書を分析してクラスタリングとクラシフィケーション, カテゴライゼーションを行う.

1999 年 6 月 Inktomi は, ポータルサイトが自社の検索サービスに簡単にディレクトリサービスを付加できるというサービスを発表した. "Directory Engine"と呼ばれるこのサービスは, あらかじめ作成した知識に基づいて目的にかなった文書を検索, 整理し, ユーザーが項目別に検索できるようなディレクトリに収める. ディレクトリに加えてショッピングモールなどをサービスに加えることもできる.

4. 文書ディレクトリ管理システムに要求される要件

4.1. 共通要件

本節では, インターネット, イントラネット共通に文書ディレクトリを管理するシステムに必要とされる機能を整理する. 文書ディレクトリの作成, 維持, 管理, 表示を行うという基本的な要件に加えて, あると便利なものとして次のようなものが考えられる.

編集支援として

1. Web インターフェイスによるカテゴリの編集ができること.
2. UNIX のコマンドを知らなくても編集できること.
3. HTML の文法を知らなくても編集できること.
4. URL をひとつひとつ追加する他に,追加したい URL を複数まとめて追加できること.
5. URL を追加するとき html の title を自動的に取り出してくれること
6. 同じ URL が他のカテゴリーで登録されているかどうかをチェックしてくれること.

編集者による更新作業支援として

7. 更新されたカテゴリーや URL の一覧を What's New として作ってくれること.

8. Dead link などの無効な URL を自動的に検出して編集者にそれを伝えること.

システム管理の支援として

9. 編集者が編集履歴を閲覧できること.
10. 複数の編集者による編集をサポートするために編集者毎に編集範囲を制限できること
11. 一般ユーザからの URL 追加, 変更を受け付けるインターフェイスを持ち, そのために必要な機能を持つこと
12. マルチバイトキャラクタが扱えること
13. 多国語対応されていること

4.2. イントラネットで特徴的な要件

イントラネットでの文書ディレクトリは, その品質を維持するために人手で維持することが必要である. 一方, 業務の一環として維持されているため, Open Directory Project のようにディレクトリの編集をボランティアに頼ることは難しい. ディレクトリを維持するための本来の業務に注力し, それ以外に関し, できる限り省力化できるシステムが必要とされている.

前節で述べた用件以外に, イントラネットでのディレクトリシステムでさらに必要となるのは次の要件である.

- 編集時, 検索時に表示する情報の表示範囲の制御ができること.
- 情報の登録手続きを実装 (configuration) できること.

インターネットで提供されるディレクトリには一般に誰にでも参照可能なサイトが登録される. 一方イントラネットでは, 情報に対するアクセス制御が重要になる場合があり, アクセス制限される URL も登録対象になりうる. そこで, URL 毎にまた編集者毎に編集できるかどうかという制御を行う必要が生じる. またカテゴリ毎にも同様にアクセス制御が必要となる場合がある. これを実現するために認証 (Authentication) サーバや承認 (Authorization)サーバとの連携をサポートする必要がある.

また同様にセキュリティ上の理由から情報の公開にあらかじめ決められた手続きを必要とすることがある. この機能を備えることで Go Guides のコミュニティで承認されてはじめてそのサイトがディレクトリに掲載されるという仕組みも容易に実装可能となる.

5. 開発中のディレクトリ運用システム

われわれが開発中のシステムは適切なカテゴリを作成し, 適切な(有用な) URL を探して来て, いれるカテゴリーを決め, ディレクトリで表示する適切な名前と説明を編集し, ディレクトリの更新を続けるというディレクトリの維持における本来の業務に注力し, それ以外に関し,できる限り省力化できることを目標としていて, 次のような機能が実装されている.

前述の要件のうち 1 から 4 を満たすために

- Web インターフェイスによるカテゴリのメンテナンスが可能になっている. そのため, サーバとなる OS のコマンドを知らなくても編集でき, HTML の文法を知らなくても編集できる.
- さらに表示画面はテンプレート機能を利用してデータベースに格納されているデータから自動的に作成されるため HTML file を直接編集するのではないので編集誤りが減る.
- また, 各カテゴリの画面は統一性を保つことが可能であり, テンプレートは柔軟にカスタマイズすることができる.

要件 5 を満たすために

- URL 追加時にタイトルを HTML ファイルから自動的に取り出し, HTML ファイルを解析し, 説明文を自動的に作り出す.

要件 6 を満たすために

- URL の追加時には同じ URL が他のカテゴリーで登録されているかどうかをチェックし, スパム行為の防止を行ったり, 同じ URL に関する情報は同期して更新できる.

要件 7 を満たすために

- 更新されたカテゴリーや URL の一覧を What's New として自動的に作る機能を持つ

要件8を満たすために

- ディレクトリに格納されている Dead link をなどの無効な URL を定期的に自動的に検出し編集者にそれを伝える.

要件9,10 を満たし,複数人でのディレクトリの編集をサポートするために

- 編集者毎に編集範囲を制限する機能と編集者が編集履歴を閲覧する機能をもつ

要件11を満たすために

- 編集者以外の一般ユーザからの URL 追加, 変更を受け付ける機能をもつ

さらにイントラネット向けとして,

- カテゴリ毎, 登録される URL 毎に閲覧できる編集者, ユーザを制限することが可能である.
- URL の登録時の認証手続きを実装 (configuration) することができる.

要件としてあげたもの以外に次のような機能を備えている.

- ディレクトリに登録されている URL のタイトルと説明の検索機能をもつ
- データベースから直接動的に画面をつくり出す他に, データベースからファイルシステムにダンプをつくり出す機能を備えることで, 大規模なディレクトリをあまり能力の無いシステムで運用することが可能.
- クリック回数による広告収入というビジネスモデルをサポートするために各 URL をユーザーがクリックした回数を数える機能をもつ
- ディレクトリのデータを複数のサーバで分割して保持することが可能.

6. 終わりに

本稿では, 文書ディレクトリのサービスとしての特徴をまとめ, 文書ディレクトリの管理に使われている既存の技術をまとめるた. その上で文書ディレクトリを維持管理するためにより重要になる要件をまとめ,イントラネットで文書ディレクトリを維持管理するために特に必要な機能について分析し, 最後にわれわれが開発中の文書ディレクトリ管理システムについて紹介した. イントラネットでの文書ディレクトリは,自動的な, あるいは半自動的なディレクトリ構築技術や管理技術を併用することはあっても, その品質を維持するために人手で維持することが必要である. 一方, 業務の一環として維持されているため, Open Directory Project のようにディレクトリの編集をボランティアに頼ることは難しい. 適切なカテゴリを作成し, 適切な(有用な) URL を探して来て, いれるカテゴリーを決め, ディレクトリで表示する適切な名前と説明を編集し, ディレクトリの更新を続けるというディレクトリの維持を行なうための本来の業務に注力し, それ以外に関し, できる限り省力化できるシステムが必要とされている. 本稿でイントラネットで特に必要とされている機能としてあげたものも, インターネットでサービスを提供するディレクトリの管理にも必要とされるものである. 将来的には, 文書処理技術を発展させた自動的なディレクトリ管理システムを併用することになると考えられる.

参考文献

1. Hiroshi Tsuda, WIND: Hyper Keyword Index as a Web Document Directory, IJCAI99 Text Mining Workshop
2. Open Directory Project : http:://www.dmoz.org/
3. Go Guides http://guides.go.com/
4. Semio Automatic Taxonomy Building http://www.semio.com/products/tax.htm
5. Yahoo http://www.yahoo.com/
6. Catalog, Senga http://www.senga.org/
7. Inktomi によるディレクトリ構築サービス http://www.inktomi.com/products/portal/directory/
8. MySQL http://www.mysql.org/
9. Resource Definition Format http://www.w3c.org/RDF/