3X−4

グループウェア(Groupmax)におけるインデックスサーバ (Bibliotheca2 Web Search) の適用

亀田正美† 高杉正勝‡ 若松隆司† 星幸雄†
(株)日立製作所 ソフトウェア開発本部† (株)日立西部ソフトウェア‡

1．はじめに

グループウェア（以降 Groupmax）におけるインデックスサーバ（以降 Bibliotheca2 Web Search）の適用について報告する。Bibliotheca2 Web Search は、イントラネット/インターネット上の WWW サーバから収集した情報をもとにインデックスを作成し、検索サービスを提供するシステムであるが、日立のグループウェアである Groupmax サーバからも情報収集し、検索サービスを提供することが可能である。また、収集した文書を構造化文書として扱い、構造を指定した検索が可能となっている。

2．検索システムの概要

検索システムの概要を図1に示す。

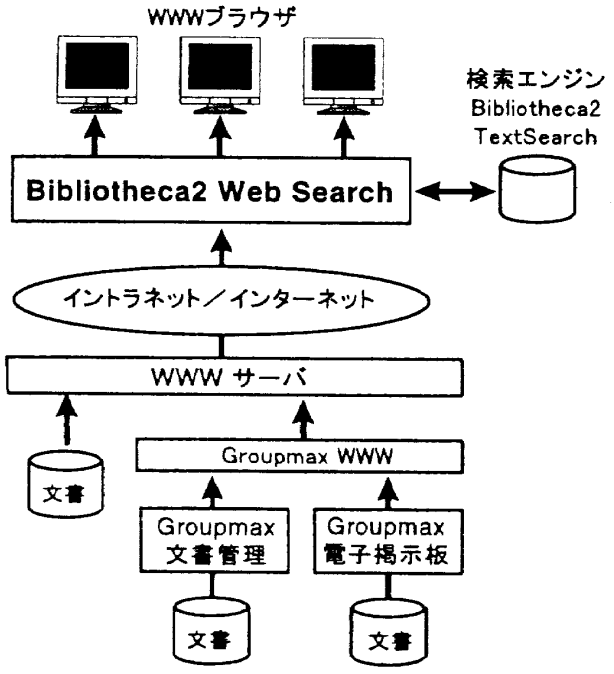

図1 検索システムの概要

Bibliotheca2 Web Search は、通常は、WWW インターフェースを使用して、WWW サーバ上の文書を収集する。文書にリンクが張られている場合、自動的にリンクを辿り文書を収集する。こうした機能を持つプログラムは一般的に"ロボット"などと呼ばれており、Bibliotheca2 Web Search もロボットの一種である。

Bibliotheca2 Web Search のロボットとしての特徴は、WWW サーバだけでなく、Groupmax サーバ（文書管理・電子掲示板）の文書・記事も収集できることである。この場合、WWW インターフェースのほかに、Groupmax サーバの WWW インターフェースを提供する Groupmax WWW を利用する。

WWW サーバ・Groupmax サーバから収集した文書は、構造化文書対応全文検索エンジン Bibliotheca2 TextSearch に登録される。

3．Groupmax 情報収集方式

図2に Groupmax 情報収集処理を示す。ここでは、Groumpmax サーバ情報のうち、文書管理サーバからの情報収集について述べる。

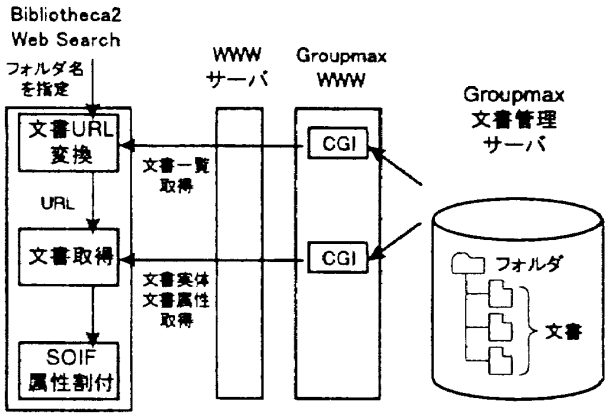

図2 Groupmax 情報収集方式

† Masami KAMEDA, Takashi WAKAMATSU, Yukio HOSHI
‡ Masakatsu TAKASUGI
† Software Development Center, Hitachi, Ltd.
‡ Hitachi Seibu Software Ltd.

Groupmax 文書管理サーバから文書を収集する場合、ユーザは、収集したい文書が格納されている"フォルダ"と呼ばれる入れ物の名称を指定する。フォルダ名を与えられたBibliotheca2 Web Searchは、Groupmax WWWが提供するCGIプログラムを利用して、そのフォルダに含まれる文書一覧をGroupmax文書管理サーバから取得する。文書一覧には、文書を示すURLが文書毎に記述されている。次に、各々の文書について、文書を示すURLをもとにして文書実体及び文書の属性一覧を取得する。

4．SOIF属性値作成処理

SOIF (Summary Object Interchange Format)とは、ロボットが収集したデータをロボット間で送受信できるようにするために、コロラド大学のHarvestプロジェクトで開発されたフォーマットで、"Title""Author"などの属性と、それに対応する値で構成されている。

Bibliotheca2 Web Searchでは、SOIFを利用することにより、収集したデータをロボット間で送受信できるだけでなく、文書の構造を指定して検索できるようにしている。

図3にSOIF属性値作成処理を示す。

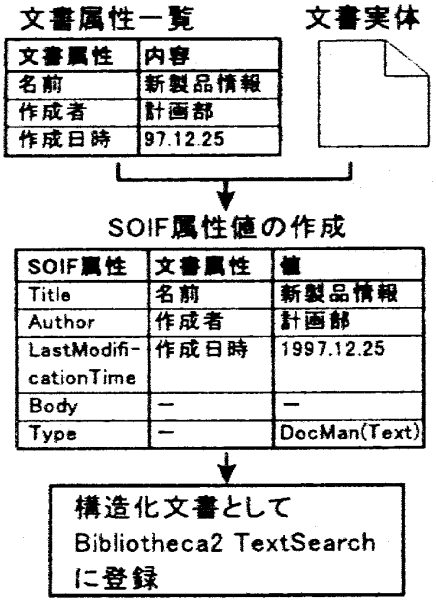

図3　SOIF属性値作成処理

SOIF属性値作成処理では、情報収集処理で取得した文書属性と文書実体からSOIF属性値を作成する。文書属性の"名前"をSOIFの"Title"に、"作成者"を"Author"に、"作成日時"を"Last Modification Time"に、文書の実体をBodyに割り当てる。また、Groupmaxの文書であること及び文書種別を示す識別子として"DocMan(文書種別)"を"Type"に設定する。このように、Bibliotheca2 Web Searchでは、収集した文書とその属性を、属性が並列に並んでいるI階層の構造化文書（SGML）として扱い、構造化文書に対応した全文検索エンジンBibliotheca2 TextSearchに登録する。

WWWサーバから収集したHTML文書の場合は、HTML文書のタグを解析し、TITLEタグの内容を"Title"に、ADDRESSタグの内容を"Author"になどと割り当てる。また、"Type"には"HTML"が設定される。

検索は、SOIF属性を指定して検索できるため、例えば、"Type=DocMan"という検索条件を組み合わせて、"Groupmaxサーバだけを検索対象にする"といった検索が可能である。

5．まとめ

Bibliotheca2 Web SearchのGroupmaxへの適用にあたって、以下の機能を開発した。

（1）Groupmax WWWを利用したGroupmaxサーバからの文書収集機能
（2）文書属性情報をもとにSOIF属性値を作成する機能
（3）SOIF属性を持つ文書を構造化文書として検索エンジンに登録する機能

上記機能の開発により、WWWサーバ・Groupmaxサーバからの文書収集の自動化及びサーバを意識せずに構造を指定した一括全文検索を可能にした。

6．参考文献

[1]菅谷他：「n-gram型大規模全文検索方式の開発 —インクリメンタル型n-gramインデクス方式—」，情報処理学会第53回全国大会 5T-2
[2]D. Hardy, M. Schwartz, and D. Wessels, Harvest User's Manual -- Version 1.4