8Y－07　　教育コンテンツ向け分類キーワード生成方法の提案

梅村　武久*1*2、 安田　孝美*1*3

通信・放送機構*1　NTT西日本 ITアプリケーションセンタ*2　名古屋大学　情報文化学部*3

1. はじめに

今後、教育シーンでのインターネット利用は必要不可欠なものとなることが予想される。しかし、限られた授業時間内に欲しい情報の発見や入手が困難であるといった問題がある。

この問題を解決するために、筆者らは情報の検索結果をカテゴリ分類体系に整理して表示することにより、検索結果の絞込みを支援するWWW検索システムを、通信・放送機構岡崎公共システム開発リサーチセンタに構築し、WWW情報の分類体系化を試みた[1][2]。本方式で、情報を教育向けにカテゴライズする図1の自動分類定義ファイル（以下、定義ファイル）を作成し、定義ファイルに用いる分類キーワードの追加・削除・重み付け等を繰り返し行うことで、分類精度が増し実用にも耐えうる成果を得た。

```
/学校の授業::
/学校の授業/国語:国語,国語,国語,絵文字,民話,語源,読み方,片仮名,昔話,語句,季語・・
/学校の授業/音楽:音楽,リズム,曲,メロディー,作曲,笛,ピアノ,えんそう,音階,楽譜・・
/学校の授業/保健体育:体育,体育,体育,ボール,水泳,体力,なわとび,スイミング,陸上・・
/学校の授業/小学校::
/学校の授業/小学校/算数:算数,算数,算数,さいころ,さいころ,偶数,奇数,暗算,三角形・・,
/学校の授業/小学校/理科:理科,理科,自由研究,自由研究,ふしぎ,空気,太陽,花粉,摩擦・・
        ⇩
/お遊び情報/料理・お菓子:レシピ,レシピ,キッチン,キッチン,献立,おやつ,おやつ・・
/行政:行政,行政,行政,自治体,自治体,官公庁,官公庁,県庁,県庁,市役所,市役所,区役所・・
/行政/郵政省:郵政省,郵政省,郵政省,はがき,年賀,小包,小包,郵便,郵便,放送,放送・・
```

図1　自動分類定義ファイル

2. 研究の目的

教育シーンでの情報検索においては、WWW情報に限らず、今後の学内ネットワークの普及によりイントラ型データベースへの検索も利用度が増す。また、学校におけるIT技術者は少なく、システムの運用・管理にかかる人手は極力最小限に押さえるシステムが要求される。

本稿では、これらの状況を考慮し、1）SQLやOracleに代表される汎用のデータベースサーバに対して定義ファイルを適用させる方法、2）TF－IDF法を応用した定義ファイルの分類キーワード自動生成方法について報告する。

3. 汎用データベースサーバへの適用

本研究において、WWW情報分類にキーワードによる手法を選定したのは、WWW情報以外にもイントラネット環境で用いられている教育関連のデータベースについても適用可能となるなど、幅広く応用できることがある。よって、分類するためのキーワード群は、図1のように汎用的なCSV形式のテキストファイルを使用している。このファイルを用いて、実際に岡崎市の小中学校ネットワークで運用している教育素材DB（Oracle,SQL サーバ）に対して、カテゴリ分類するシステムを構築した。

3．1　システム概要

本システムは、通信・放送機構岡崎公共システム開発リサーチセンタの「ネットワークアーキテクチャに関する研究開発」の一環として構築しているものであり、その概念図を図2に示す。

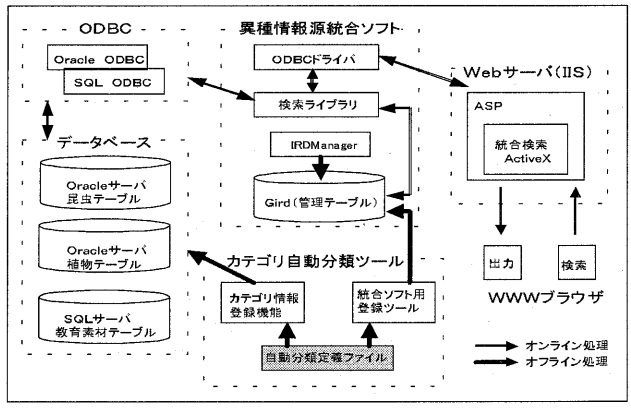

図2　教育素材 DB カテゴリ分類検索システム

このシステムは、WindowsNT上のアプリケーションサーバとして構築され、教育素材 DB を先述した定義ファイルを使って検索結果を分類表示するシステムである。主な機能を以下に挙げる。

a. データベース部

Proposal of method of generating classification key word for Education contents

Takehisa Umemura(NTTWEST IT Application Center),

Takami Yasuda(Nagoya University)

OracleとSQLのデータベースのうち、昆虫・植物・各種教育素材の静止画データ約16，300個を検索の対象としている。

b. 異種情報源統合ソフトウェア部

ODBC(Open Database Connectivity)経由で異なるデータベースに対して統合検索を可能としている。

c. カテゴリ自動分類ツール部

今回の研究において、新たに機能追加した部分であり、作成済みの定義ファイルを用いて、検索結果をカテゴリ分類する。尚、概要を次項で紹介する。

d. Web サーバ部

画面表示機能として、ASP(Active Server Page)を用いて、検索結果をフォルダ階層型に分類表示する。

3．2 教育向けカテゴリ自動分類方法

カテゴリ自動分類ツール部の作成にあたっては、以下の手法を用いている。

①定義ファイルのカテゴリ毎にカテゴリ ID を付与
②既存 DB にカテゴリ ID 用のフィールドを追加
③定義ファイルのキーワード群を用いて、既存 DB の情報内容フィールド(画像説明文)を全文検索
④ヒットした頻度に応じて、既存 DB のカテゴリ ID 用フィールドにカテゴリ ID をセット
⑤定義ファイルのカテゴリ名と ID を統合ソフトに登録

このように、既存DBに対してカテゴリIDフィールドを追加し事前に登録しておくことで、検索結果のカテゴリ分類表示を可能としており、現在分類精度のデータを収集中である。

4．分類キーワード生成方法

WWW情報の分類手法としては、URL名による分類やハイパーリンクの共起による分類[3]など様々な手法が考えられる。また、情報検索分野ではキーワードの抽出方法としてTF－IDF法などを用いた研究[4]も進められている。本研究では、TF－IDF法を応用し、定義ファイルの分類キーワードを自動生成する方法の提案をすると共に、利用者の環境を教育分野に特化することでカテゴリ項目を整理し、教育分野向けカテゴリ分類体系を確立することを目指している。

これまで、定義ファイルの分類キーワード抽出は、単語の出現頻度を基に人手に頼る方法を用いてきたが[1]、現在はTF－IDF法(図3)を応用した自動抽出を目指している。以下、その概要について述べる。

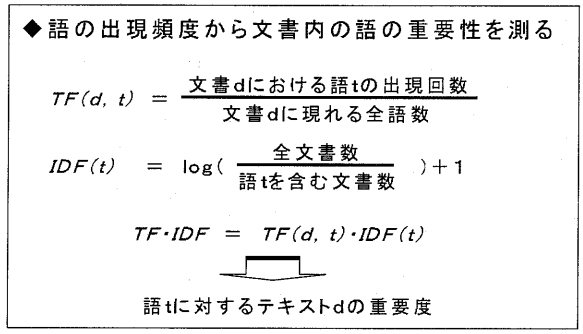

図3 TF－IDF法

4．1 分類キーワードの自動生成

IDF(逆文書頻度)値を文書に対する特徴度とし、ある一つのカテゴリ文書に対する特徴度が小さく、かつ全カテゴリ文書に対する特徴度が大きいキーワードを分類キーワードとして抽出する。また、その特徴度の大小に応じて重み付けを行う。初めに、極少数のWebページを収集し人手を使って分類する必要はあるが、その後は自動(プログラミング)化することが可能となる。

現在、WWW検索システム上で動作するプログラムを作成中であり、人手を中心とした既に作成済みの定義ファイルのキーワードと、自動生成したキーワードを比較し手法の改善を行い、その妥当性を実証していくこととする。

5．おわりに

教育分野を対象にイントラ型データベースサーバへの適用方法とWWW情報の分類キーワード生成方法について述べた。本稿では、その方法について述べるに留まったが、現在岡崎市の小中学校ネットワーク環境でフィールド実験中であり、今後実証データを収集し、本手法の有効性について検討していきたい。

参考文献

[1]梅村武久,安田孝美:教育分野における WWW 情報の分類体系化,第60回情処全大,pp4-433～434,(2000).

[2]梅村武久,安田孝美:教育分野における WWW 情報の分類体系化とその実証,第61回情処全大,pp4-347～348,(2000).

[3]大久保雅且,杉崎正之,田中一男:ﾘﾝｸの共起関係を用いた Web ﾍﾟｰｼﾞ分類法式の検討,第 59 回情処全大,pp3-81～82,(1999).

[4]中川 裕志:言語メディア論,http:// www.r.dl.itc.u-tokyo.ac.jp /~nakagawa/lmedia/skelton/skelton.html