教育コンテンツ向け分類キーワード生成方法の提案とその実証

2S－6

梅村　武久*1*2、安田　孝美*1*3

通信・放送機構*1　NTT西日本 名古屋支店*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教育向け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(図1)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、実用にも耐えうるWWW情報の分類体系化の基盤を確立した。また、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利用度も増すため、WWW情報の分類体系化を、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適用させる方法を提案した[1]。

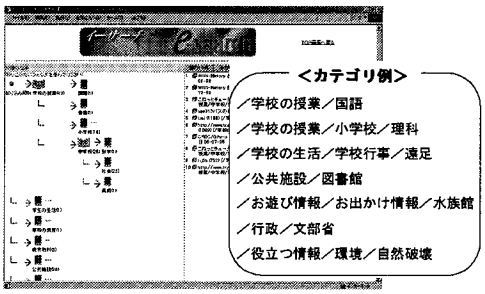

図1　WWW検索システム画面

2. 研究の目的

教育シーンでのインターネット・イントラネットの利用が増すにつれ、システムを運用管理するIT技術者が必要となり、その稼動も増大していくなどの問題が発生する。よって、運用管理にかかる人手を最小限に押さえるシステムが要求される。

構築したWWW検索システムは、ロボット型(自動収集)であり、カテゴリ分類表示も分類キーワード(自動分類定義ファイル)により自動化されているが、分類キーワードの作成や情報更新に伴うメンテナンスについては、単語頻度などを基に人手に頼る点が多くある。そこで、本稿ではTF－IDF法(図2)を応用した分類キーワードの自動生成方法とその実証結果について報告する。

◆語の出現頻度から文書内の語の重要性を測る

$$TF(d,t) = \frac{\text{文書dにおける語tの出現回数}}{\text{文書dに現れる全語数}}$$

$$IDF(t) = \log\left(\frac{\text{全文書数}}{\text{語tを含む文書数}}\right) + 1$$

$$TF\cdot IDF = TF(d,t)\cdot IDF(t)$$

語tに対するテキストdの重要度

図2　TF－IDF法

3. 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[2]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[3]も進められている。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、これらの様々な手法と組み合わせることが可能となるなど、幅広く応用できることがある。また、分類キーワードを格納した定義ファイルについても、図3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。

```
/学校の授業:
/学校の授業/国語:国語,国語,国語,絵文字,民話,諺源,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3　自動分類定義ファイル

次項にて、本定義ファイルの分類キーワードをTF－IDF法を応用して自動抽出する方法について述べる。

Proposal of method of generating classification key word for Education contents and its Proof

Takehisa Umemura(NTTWEST Nagoya branch),

Takami Yasuda(Nagoya University)

3．1 分類キーワードの自動生成処理フロー

IDF(inverse document frequency:逆文書頻度)値を各単語の文書に対する特徴度とし、ある一つのカテゴリ内の文書に対する特徴度aと全カテゴリ内の文書に対する特徴度bからカテゴリ適合度(特徴度b－特徴度a)を算出、カテゴリ適合度の大きな単語(キーワード)を分類キーワードとして抽出する。すなわち、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単語を分類キーワードとして抽出する。また、カテゴリ適合度にTF(term frequency:単語頻度)値を乗算することでキーワードの重要度を算出し、その大小に応じて分類キーワードの重み付けを行う。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。図4に、自動生成の処理フローを示す。

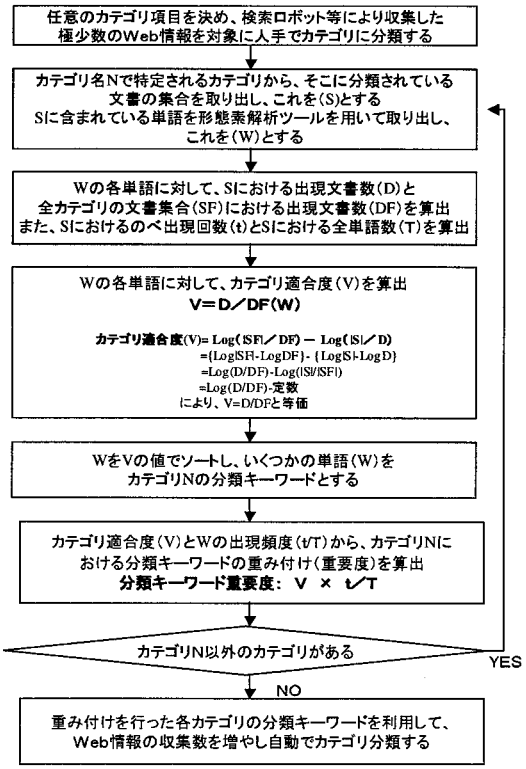

図4 分類キーワード自動生成処理フロー

3．2 実証結果

図4の処理フローが動作するプログラムを、実験運用中のWWW検索システム上に作成し、実証実験を行った。ある一つのカテゴリ(学校の授業／音楽)に分類したWeb情報25ページを対象に、各単語のカテゴリ適合度及び重要度を算出した結果の一部(上位25単語)を表1に示す。

カテゴリ毎でばらつきはあるものの、表1のようにほぼ妥当な分類キーワードを抽出することができた。抽出した各カテゴリの分類キーワードを用いて約2万ページのWeb情報をカテゴリ分類した結果、全126カテゴリ中ランダムに選択した10カテゴリにおける分類精度は約70%であり、まだサンプル数は少ないがこれまで人手を中心に作成してきた分類キーワードの分類精度とほぼ同様の結果が得られた。

表1 カテゴリ適合度、重要度の算出結果(抜粋)

No	カテゴリ	カテゴリ適合度	単語頻度>＝5	ページ数／該当カテゴリ	ページ数／全カテゴリ	重要度 評価値×(単語頻度／該当カテゴリ単語
1	/学校の授業/音楽					
2	楽器	0.272727273	49	9	33	5.923597679
3	伴奏	0.777777778	9	7	9	3.102836879
4	mayu	1	7	1	1	3.102836879
5	tramp	1	7	1	1	3.102836879
6	土笛	1	6	1	1	2.659574468
7	演奏	0.144927536	35	10	69	2.312673451
8	ピアノ	0.212121212	24	7	33	2.25660864
9	ホルン	1	5	2	2	2.216312057
10	音楽	0.070422535	66	20	284	2.050233745
11	soream	0.5	8	1	2	1.773049645
12	曲	0.191489362	18	9	47	1.358085579
13	バイオリン	0.5	6	3	6	1.329787234
14	タンゴ	0.25	12	1	4	1.329787234
15	鍵盤	0.333333333	8	2	6	1.182033097
16	太郎	0.166666667	15	1	6	1.108156028
17	西城	0.333333333	6	1	3	0.886524823
18	打ち込み	0.333333333	6	1	3	0.886524823
19	作曲	0.222222222	9	4	18	0.886524823
20	合唱	0.137931034	14	4	29	0.855955001
21	オーケストラ	0.363636364	5	4	11	0.805931657
22	オペラ	0.142857143	12	2	14	0.759878419
23	アンサンブル	0.333333333	5	4	12	0.738770686
24	音	0.038866359	33	8	217	0.539268556
25	オルガン	0.222222222	5	2	9	0.49251379

4．おわりに

教育分野を対象にWWW情報の分類キーワード生成方法とその実証結果について述べた。実証結果より本手法の有効性が把握できたので、今後は更にWeb情報の収集数を増やすと伴に、分類精度の向上を目指した手法の改善を行っていき教育向けカテゴリ分類体系を確立していきたい。

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,5M-06,(2000). 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,5S-05,(2000). 教育コンテンツ向け分類キーワード生成方法の提案,第62回情処全大,8Y-07,(2001).

[2]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).

[3]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