情報科学関係情報の自動分類を行うポータルの試験運用と評価

阿部 淳也† 出石 大志† 岸本 義晴†† 堀 幸雄††† 今井 慈郎†††
香川大学大学院工学研究科† 香川大学工学部†† 香川大学総合情報センター†††

1. はじめに

近年，アンテナと呼ばれるウェブサイト更新チェック用のソフトウェアやウェブサイトが登場している．しかし，これらのアンテナでは，基本的に新着文書のみを取り扱い，過去に収集した文書へのアクセスなどが提供されていない場合が多い．

また，ユーザがウェブページに特徴語を付与して管理する例としてソーシャルブックマークが挙げられる．このようなシステムの代表的な例として，株式会社はてなが提供しているサービスが挙げられる[1][2]．ウェブページをブックマークする際に文書に応じたタグの推薦を行い，タギングを容易にしている反面，ウェブページ自体の収集はユーザが手動で行う必要があり，時間を要し，誤操作も無視できない．

そこで，本研究では，複数のウェブページを対象に自動的に情報を収集・分類する文書管理ポータルを提案する．本稿では，システムの概要と設計を示す．また，試験運用を行った結果の評価についても言及する．

収集する文書群として，本研究では情報科学情報を取り扱っているウェブページを対象とする．主な理由は以下の通りである．

- 特徴的な専門用語が多く，分類精度に期待できる
- 情報系学科内において，多くの学生が興味を持ちやすく，学習意欲の向上が期待できる

2. 特徴語の自動分類

本システムでは，各記事に付与される特徴語として，記事中から自動抽出した名詞句によるキーワードとユーザが手動で付与した文字列によるタグの 2 種類を用いる．以下で，それぞれの特徴と付与方法について述べる．

Design and Brief Evaluation of Automatic Categorization System for Information Science
Junya Abe†, Hiroshi Izuishi†, Yoshiharu Kishimoto††, Yukio Hori†††, Yoshiro Imai†††
† Graduate School of Engineering, Kagawa University
†† School of Engineering, Kagawa University
††† Information Technology Center, Kagawa University

2-1. キーワード

ユーザが全ての記事に特徴語を付与することは作業量という観点から現実的ではない．そこで，記事中に出現する名詞句を抽出し，特徴語として付与する．これにより，ユーザの負担無しに全ての記事に特徴語を付与することができる．また，特徴語は TF-IDF 値を記事中における重要度として持ち，後述するクラスタ分析において特徴ベクトルの重みとして利用される．

2-2. タグ

前項で述べたキーワードは記事中の文章に依存しており，文章に含まれない語句を特徴語として付与することができない．そこで，キーワードの補助という形で，ユーザが手動で特徴語を付与する仕組みを提供する．これにより，記事に関する適切な特徴語を補うことができる．

2-3. 自動分類

特徴語を一方的に付与するだけでは特徴語の種類数が膨大になり，ユーザが目的の特徴語を見つけることが困難になることが考えられる．また，意味は同じだが表記の異なる異表記同義語の存在で特徴語検索の漏れが生じる可能性もある．

そこで，クラスタ分析を用いた特徴語の分類を行うことで上記の問題に対応する．クラスタ分析には，階層的手法よりも計算時間が短い非階層的手法の K-平均法を用いる[3]．

3. システム概要

3.1 システム構成

本システムは図 1 のような構成となる．

システムは，定期的に登録されたインターネット上のメディアから新規文書を収集しキーワードを自動抽出するキーワード抽出モジュール，キーワードのクラスタリングを行うキーワード分類モジュールを用いて情報の蓄積を行う．

ユーザは，蓄積された情報に対し，提供されるユーザインタフェースから，文書検索モジュールを介して，各種メディア群の中から，横断的に望みの文書を検索でき，情報を取得することができる．

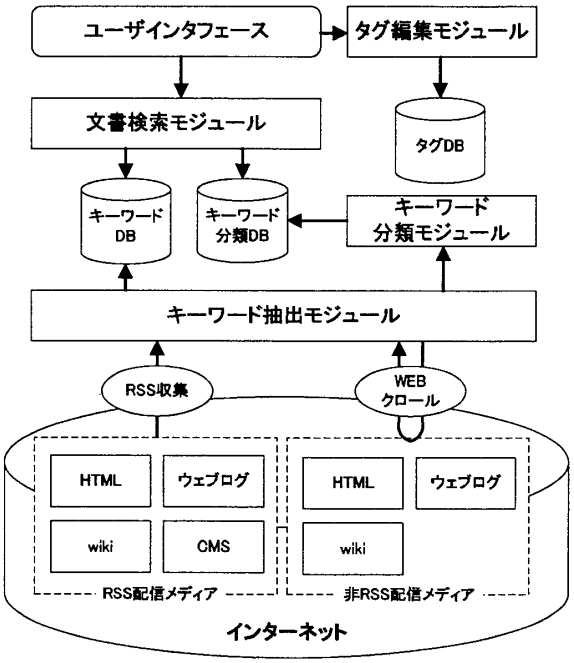

図 1　システム構成

3.2 キーワード抽出モジュール

文書の収集には主に各メディアが配信している RSS を利用する．RSS 配信に対応していない HTML 形式のメディアでは，WEB クロールを行うことで文書を収集する．文書に対して形態素 N-gram により名詞句のみをキーワードとして抽出する．本システムでは N=3 とした．各キーワードの TF-IDF 値の算出も本モジュールで行う．

3.3 キーワード分類モジュール

抽出されたキーワード群の重みを要素とした特徴ベクトルに基づき，単語文書行列を生成する．得られた単語文書行列にクラスタ分析を行うことでキーワードの分類を行う．

3.4 タグ編集モジュール

本モジュールでは，本システムで収集した各文書に対して，タグの追加・削除をユーザ自身で行うための機能を提供する．

3.5 文書検索モジュール

ユーザは，登録されている特徴語を基に文書の検索を行う．複数の特徴語を組み合わせることで多面的な観点からの検索を行うことができる．また，文書中の特徴語について，同じ分類に属するクラスタの代表語を検索結果に併記することで，関連する他の文書への誘導も可能とし，検索範囲や精度を向上させる．

4. 試験運用と評価

4.1 試験運用

運用開始直後からユーザが利用しやすくするために，事前にある程度文書を収集して特徴語を分類しておく必要がある．そこで，試験運用として，情報科学関係情報を扱っている 2 つ*のウェブサイトを対象として文章の収集を行った．文書の収集は収集期間中に不定期に行った．試験運用における文書の収集状況を表 1 に示す．

表 1　文書の収集状況

収集期間	2007 年 11 月 14 日 ～2007 年 12 月 4 日
文書数	478
キーワード種類数	8590

文書を特徴付ける精度は文書に付与されるキーワード数の数に依存する．付与されるキーワード数が少なければ，その文書への誘導性は下がり，結果的にユーザに提示されにくくなる．そこで，各文書に付与されたキーワード数について調査を行った．その結果を表 2 に示す．

表 2　全文書中のキーワード付与状況

付与数平均値	38.0293
付与数最頻値	33
付与数最小値	8
付与数最大値	116

4.2 評価

本システムのインタフェースでは，文書ごとに TF-IDF 値の高い順に 10 個のキーワードを併記して，ユーザに検索結果を提示している．現在，文書に付与されているキーワード数の平均値は，インタフェース中のキーワード提示数を大きく上回っており，十分利用しやすいものと考える．しかし，付与キーワード数が 10 を下回る文書が 4 件（約 1.15%）存在した．これらの文書はタグが付与される可能性が低い．こうした文書をより特徴付ける仕組みが必要である．

5. おわりに

本稿では，文書を自動で収集・分類する文書管理ポータルの提案と設計を行った．また，試験運用を行い，その結果について評価を行った．

今後は，実際に運用を行い，タグ付与状況や分類精度の評価を行う．また，ユーザの意見・要望に基づいたシステムの機能向上を目指す．

* IT 総合情報ポータル「ITmedia」Home(http://www.itmedia.co.jp/)，@IT - アットマーク・アイティ(http://www.atmarkit.co.jp/)

参考文献

[1] 株式会社はてな，“はてなアンテナ”，http://a.hatena.ne.jp/，(2008 年 1 月確認)．
[2] 株式会社はてな，“はてなブックマーク”，http://b.hatena.ne.jp/，(2008 年 1 月確認)．
[3] 新納浩幸，“R で学ぶクラスタ解析”，オーム社，(2007)．