教官公募情報のダイジェスト自動生成

4T−1

見舘 潔　佐藤理史
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

インターネットの普及により、様々な情報がインターネット上に発信されるようになった。誰もが自由に情報発信できるという特徴をもったインターネット上には、雑多な情報が様々な場所に存在するため、欲しい情報を素早くみつけだすことは容易でない。

これを支援する情報検索には、(1) サーチエンジンに代表されるような、情報が必要になった時に探索する方法と、(2) あらかじめ後の利用を想定して情報を編集しておき、これを利用する方法、の二つがある。

本稿では、後者の方法を採用し、情報を自動的に収集、編集し、ユーザに提供するシステムの実現について述べる。対象として、WWW 上に存在する大学などの教官公募ページを取り上げ、これらを自動的に探索・収集し、ダイジェスト形式に自動編集し、ユーザに素早く提供できるようなシステムを作成した。

2 システムの概要

作成したシステムは、(1) 収集、(2) 選別、(3) 情報抽出、(4) 検索、の 4 つのモジュールからなる。システム構成図を図 1に示す。

2.1 収集モジュール

WWW から、公募情報らしきページを取得するモジュールである。このモジュールは次の 3 つのプログラムからなる。

- Goo などのサーチエンジンを用いて「教官公募」「教員公募」「教官募集」「教員募集」などで検索し、得られた URL リストのページを取得するプログラム
- 公募対象機関（国公私立大学など）の公式 WWW ページのトップノードから 4 階層までのページを自動巡回して取得するプログラム
- 公募リンクデータベース中の公募リンクページ (公募ページへのリンクを持つページ) からリンク先ページを取得するプログラム

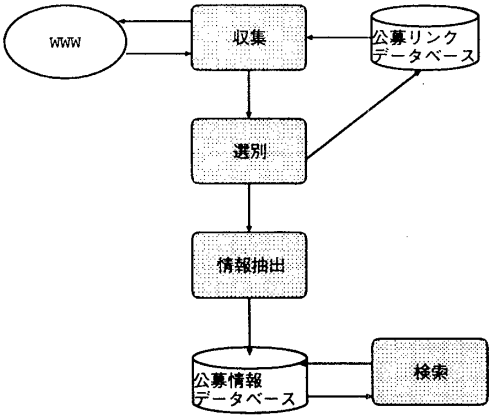

図 1: システム構成図

これらのプログラム (定期的に実行される) によって取得されたページは、選別モジュールへ渡される。

2.2 選別モジュール

収集モジュールにより取得されたページを、以下のアルゴリズムで、(1) 公募ページ、(2) 公募リンクページ、(3) その他、に分類する。

1. 公募ページ、公募リンクページに共通に使われるキーワード (「教官公募」など) が存在するかどうかを調べ、ない場合はその他と判定する。
2. 「職種」「職名」「締切日時」「提出先」「問い合わせ先」などの見出しが存在するかどうかを調べ、ある場合は公募ページと判定する。
3. アンカータグのなかに公募リンクページ特有の表現（「公募」など）が存在するかどうかを調べ、ある場合は公募リンクページと判定し、ない場合はその他と判定する。

公募ページの場合は、次の情報抽出モジュールへと進む。公募リンクページの場合は、その URL を公募リンクデータベースに追加更新する。

なお、本モジュールの判定精度は、90%程度である。

2.3 情報抽出モジュール

公募情報から、(1) 公募ページの URL、(2) 公募機関名、(3) 所属先、(4) 公募職種、(5) 分野・内容、(6) 応募締切日、(7) 公募機関の地域、の 7 つの情報を抽出しデータベース化する。抽出は以下の方法で行う。

Automated Editing of University Employment Bulletins.
Kiyoshi MITATE and Satoshi SATO.
School of Information Science, Japan Advanced Institute of Science and Technology.

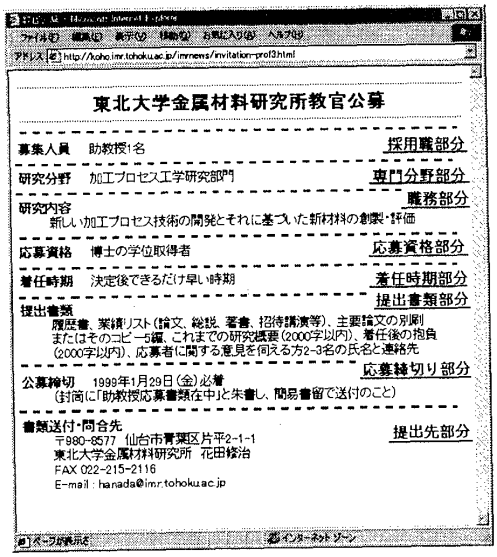

図 2: 見出しを用いていくつかの部分に分割の例

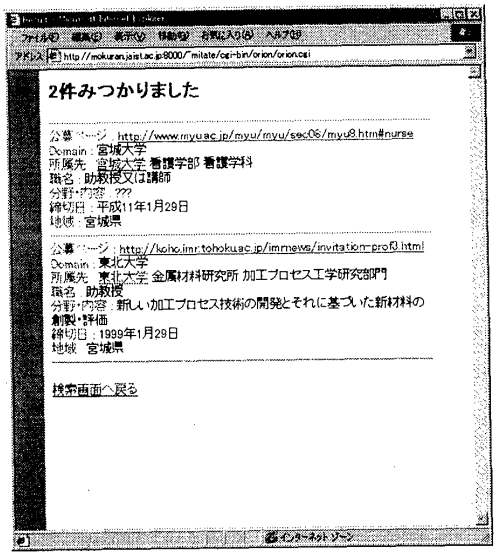

図 3: 検索結果

1. 見出しを用いて、ページをいくつかの部分に分割する。図 2に例を示す。この図の点線は、分割境界を表す。
2. それぞれの項目を抽出する [1]。
   (1) **公募ページの URL** : WWW からページを取得した時に、ファイルに添付しておいた URL を抽出する。
   (2) **公募機関名** : URL をもとに、JPNIC (Japan Network Information Center) から取得したドメイン対応表を用いて公募機関名を得る。
   (3) **所属先** : 所属、採用職、提出先、問合せ先の各部分と (2) で得られた公募機関名、TITLE タグの中身で総合的に決定する。
   (4) **公募職種** : 採用職部分から抽出する。
   (5) **分野・内容** : 専門分野、職務部分から抽出する。
   (6) **応募締切日** : 応募締切部分から抽出する。
   (7) **公募機関の地域** : 問い合せ先、提出先部分の郵便番号または、住所表記から決定する。

抽出モジュールの精度は各項目に差はあるが、おおむね 85-95%の正解率を得ている。

2.4 検索モジュール

ユーザの要求に応じてデータベースを検索し、その結果をダイジェスト形式で表示するモジュールである。「公募機関名」「公募機関の地域」「分野・内容」「公募職名」の 4 つの条件を指定することができる。東北地区の公募情報の検索結果を図 3に示す。この図の 2 件目は図 2の公募ページのダイジェストである。なお、ダイジェスト表示の所属先には、その大学のホームページへのハイパーリンクが付加される。

本システムは、現在 JAIST において試験運用しており WWW を通じてアクセスすることができる。[1]

3 おわりに

本稿では、WWW 上に存在する教官公募情報を自動的に収集して、ダイジェスト形式に自動編集し、ユーザに提供するシステムについて述べた。

WWW 上で教官公募情報を検索できるサービスに学術情報センターの「研究者公募情報」[2]がある。このサービスは現在、学術情報センターに掲載依頼があったもののみを人手により編集しているので、掲載依頼がない限り掲載されることはない。これに対して本システムは、WWW 上に存在する教官公募情報を自動的に見つけ出すことができる。今後、学術情報センターと連係を図り、本システムで得られた教官公募情報を学術情報センターに提供していく予定である。

参考文献

[1] 佐藤 円, 佐藤理史, 篠田陽一: 電子ニュースのダイジェスト自動生成, 情報処理学会論文紙, Vol.36, No.10, pp.2371-2379, 1995.

[1] http://mokuran.jaist.ac.jp:8000/~ mitate/orion/
[2] http://nacwww.nacsis.ac.jp/