修 士 論 文

教官公募情報のダイジェスト自動生成

指導教官　佐藤理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

710099 見舘 潔

1999 年 2 月 15 日

要 旨

ワールド ワイド ウェブ上に存在する教官公募情報を収集して、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を作成した。本システムは、(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構成される。収集モジュールは、サーチエンジンなどを用いてワールド ワイド ウェブから公募情報らしきページを取得する。選別モジュールは収集したページを、公募ページ、公募リンクページ、その他、に選別する。情報抽出モジュールは、公募ページから、公募ページの URL、公募機関名、所属先、公募職種、分野・内容、応募締切日、公募機関の地域、の７項目を抽出しデータベース化する。検索モジュールはユーザの検索要求に従っ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出力する。

目 次

第 1 章

序論

1.1 本研究の背景と目的

近年、インターネットの普及により電子的にアクセスできる情報量が爆発的に増加した。これに伴い、ユーザが欲しい情報を素早く見つけることが困難になり、これを支援するシステムが不可欠となっている。

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 [1][2]、の二つの方法がある。

サーチエンジンは、ユーザに「得たい情報に関するキーワード」を指定してもらい、そのキーワードを含むテキストを提示するシステムであり、多くのサーチエンジンが実用に供されている。しかし、サーチエンジンは以下の理由により万能とは言えない。

- ユーザ側から見れば、適切なキーワードの入力が必要である。
  キーワード検索システムは、一つの語、または、複数の語を論理演算子で統合したものを、検索質問として受け付ける。そのため、検索者は検索要求をこのような形式の質問に適切に変換する必要がある。

- システム側から見ればキーワードから分かることが少なすぎる。
  システムは、検索質問が適切なものであることを前提としているため、ユーザが知りたい情報とかけ離れた言葉が検索要求された場合はどうしようもない。

そこで、本研究では、(2) の方法を採用し、ワールドワイドウェブ上の大学などの教官公募情報を対象に、情報を自動的に収集、ダイジェスト形式に編集しユーザに提供するシ

ステムの作成を行った。この方法は、上記の問題に対して、あらかじめ情報をフィルタリングして保存しておくため、ユーザの適切なキーワードの入力を省くメリットがある。

教官公募情報をワールドワイドウェブを用いて見たいと考える人々は、全ての教官公募情報を網羅的に読むのでなく、自分の興味のある公募情報だけを取捨選択して読みたいと考える。このため教官公募情報がダイジェスト形式で提示され必要に応じて、詳細な情報が参照できるようにすることが望ましい。

しかし、学術情報センターの「研究者公募情報」[1] のように、公募情報を人手により編集してダイジェストとして提供するには、作成に時間がかかり、またコストがかかる。そこで、ワールドワイドウェブ上の教官公募情報を機械処理によりダイジェストを自動生成し、ユーザに提供する方法を検討する。これを実現することは、以下のようなメリットがある。

(1) 機械処理により、ダイジェスト作成、更新が短時間で大量にできる。

(2) ダイジェスト作成コストがほとんどかからない。

現在、紙の媒体による教官公募が主流であるが、今後さらにインターネットが普及することにより、紙の媒体と同時に各公募機関がワールドワイドウェブ上で教官公募情報を発信していくと思われる。そうなると、本システムのようなシステムの重要性がさらに高まり、上に示した機械処理によるダイジェスト自動生成のメリットがより大きくなるであろう。

1.2 本論文の構成

本論文ではまず、2章で作成した教官公募案内システムの概略を述べたのち、3章、4章、5章、6章で、本システムの４つの主要なモジュールとそれぞれの評価実験について述べる。7章で本研究のまとめと今後の課題について述べる。

[1] http://nacwww.nacsis.ac.jp

第 2 章

教官公募案内システムの概要

2.1 ワールドワイドウェブ上の公募情報とそのダイジェスト

大学などの教官公募情報は、これまで学会紙などの紙の媒体に掲載されることが大半であった。しかし、近年のインターネットの発展によりワールドワイドウェブのページに掲載されることも多くなりつつある。図 2.1に、ワールドワイドウェブ上の公募ページの例を示す。

このような公募ページは、各公募機関のワールドワイドウェブサーバ上に存在することが多い。そのため、ワールドワイドウェブ上の教官公募情報を見たいと考える人々は、学術情報センターの「研究者公募情報」[1]を見るか、あるいは goo[2]などのサーチエンジンを用いて公募情報を検索しなければならない。前者は、公募機関から掲載依頼があった公募情報のみを人手により編集し掲載している。そのため、公募機関から掲載依頼がなかったものは掲載されないことになる。後者は、「教官+公募」などのキーワードで検索を行うことになるが、その検索結果が膨大な量になることが多く、情報の取捨選択にかなり手間取る。

公募情報を得たいと考えている人々にとって、最も重要な項目は、「公募機関名」「公募分野」「公募職種」「公募締切日時」「公募機関の地域」であろう。これらをふまえて、ワールドワイドウェブ上の公募ページを収集し、公募ページから情報抽出し、ユーザにダイジェスト形式で提供すれば、ユーザによる、公募情報の取捨選択を効果的に支援することができると考えられる。

[1] http://nacwww.nacsis.ac.jp

[2] http://www.goo.ne.jp

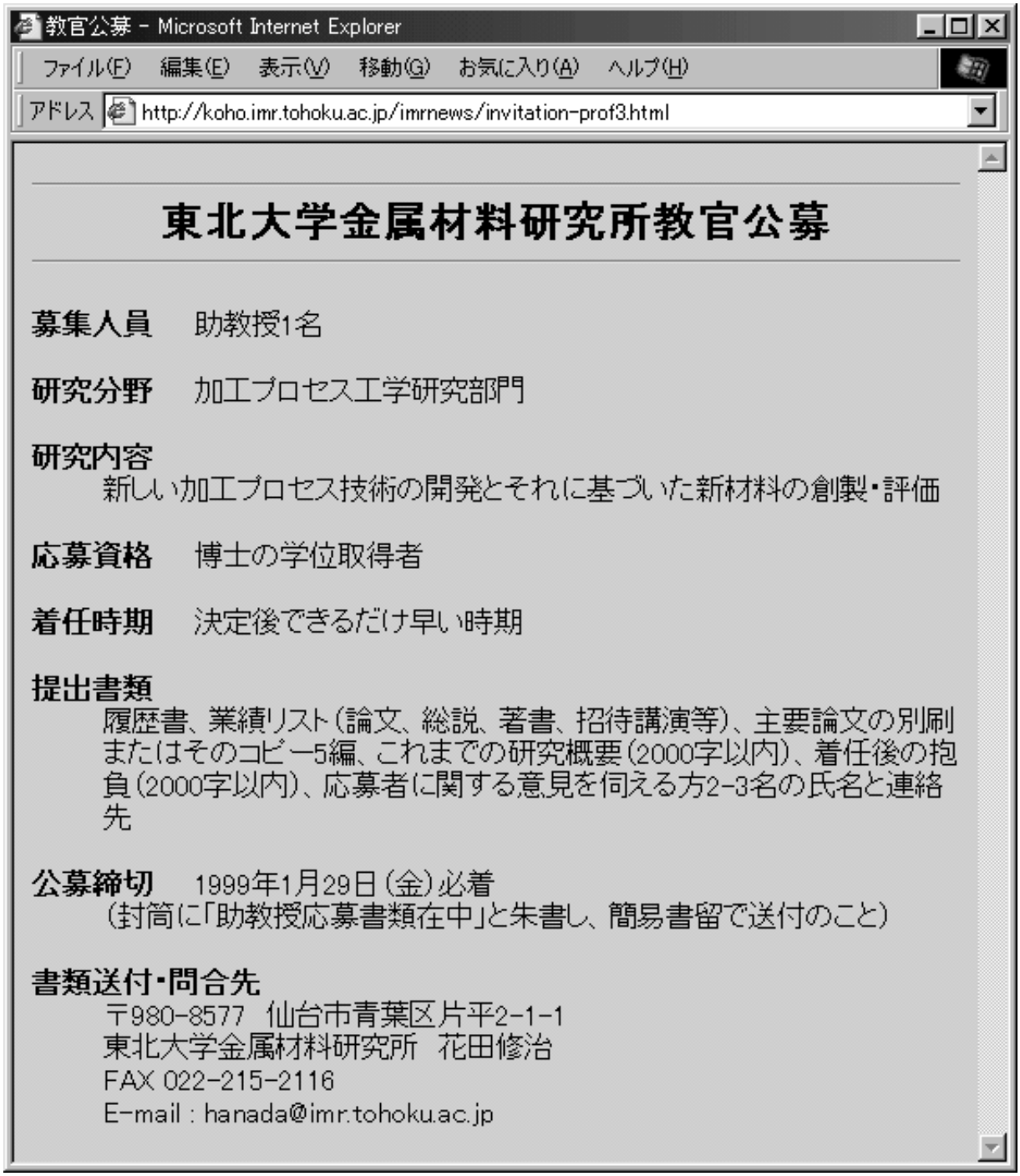

図 2.1: 公募ページの例 : http://koho.imr.tohoku.ac.jp/imrnews/invitation-prof3.html

2.2 システム概要

上記の考えに基づき、ワールドワイドウェブ上の教官公募情報のダイジェストを自動生成するシステムを作成した。その概要を図2.2に示す。本システムは、以下の４つのモジュールと２つのデータベースからなる。

(1) 収集モジュール
ワールドワイドウェブ上に存在する、公募ページらしきページを定期的に収集する。

(2) 選別モジュール
収集モジュールが取得してきたページを、公募ページ、公募リストページ[3]、その他のページ、に選別する。

(3) 情報抽出モジュール
公募ページから、「公募ページのURL」「公募機関名」「所属先」「公募職名」「分野・職務内容」「公募締切日時」「公募機関の地域」を抽出する。

(4) 公募情報データベース
情報抽出モジュールで抽出された情報を「公募ページのURL」「公募機関名」「所属１」「所属２」「所属３」「所属４」「公募職名」「分野・職務内容」「公募締切日時」「公募機関の地域」の１０フィールドからなるレコードとして格納したデータベース。

(5) 検索モジュール
ユーザの要求に応じて公募情報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。「公募機関名」「公募機関の地域」「分野・職務内容」「公募職名」の４つの条件を指定することができる。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。

(6) 公募リンクデータベース
公募リンクページと判定されたページのURLのデータベース。収集モジュールで利用される。

本システムは、プログラム言語Perl5[3]を用いて実装されており、その規模は、約3000ステップである。なお、本システムは、現在、JAISTにおいて試験運用しており、ワールドワイドウェブを通して、アクセスすることができる。[4]

[3]公募リストページとは、公募ページへのリンクをもつページのことである。

[4]http://mokuran.jaist.ac.jp:8000/~ mitate/orion/index.html

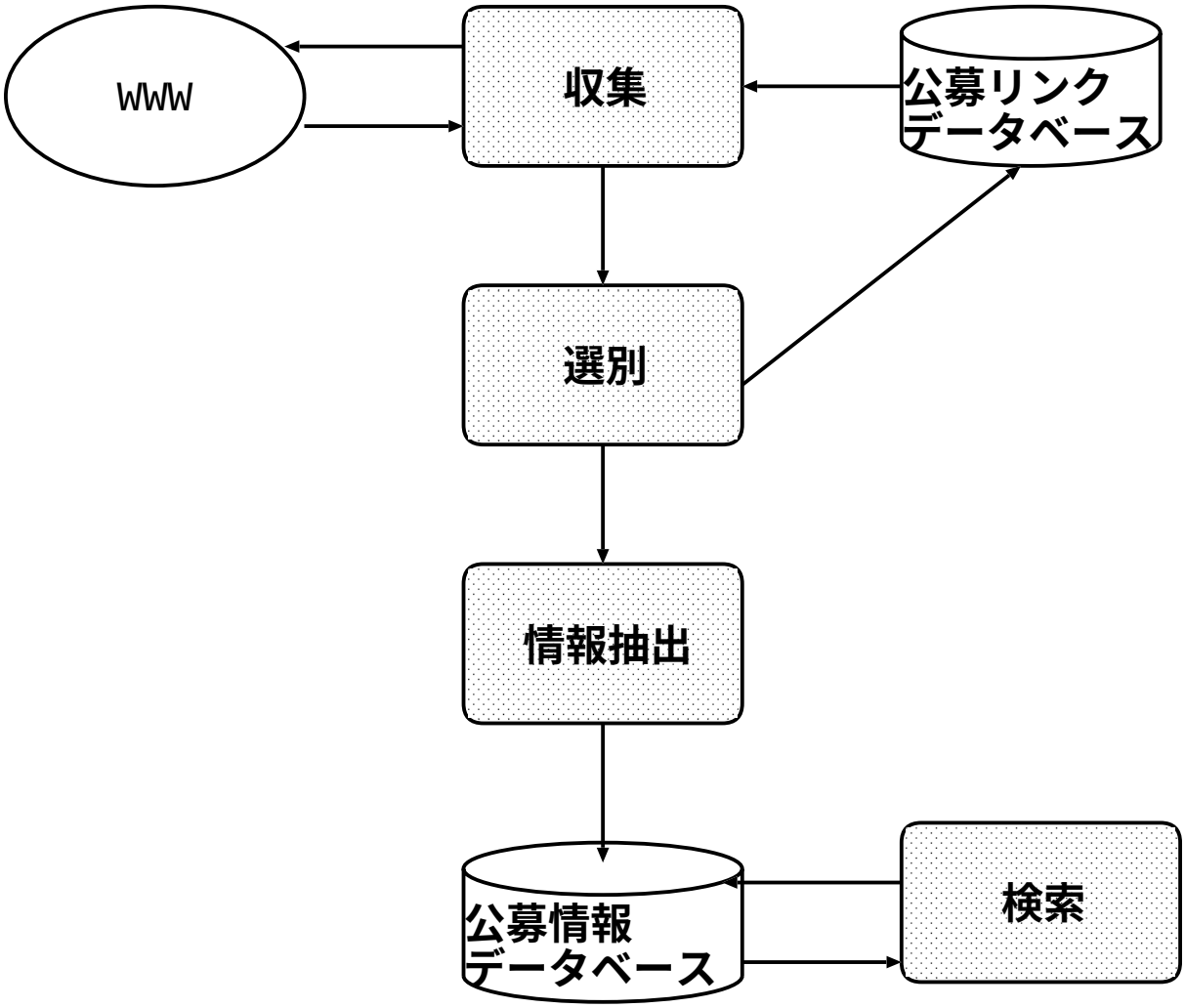

図 2.2: 教官公募情報ダイジェスト生成システムの概要

第 3 章

収集モジュール

本モジュールは、ワールド ワイド ウェブから教官公募ページらしきページを収集するモジュールである。(1) サーチエンジンを用いる方法、(2) 自動巡回ロボットを用いる方法、(3) 公募リンクデータベースを用いる方法、の３つの方法を実装している。

3.1 サーチエンジンを用いる方法

本方法では、教官公募ページらしきページを収集するために、サーチエンジン goo[1]を用いる。表3.1に示すようなキーワードとオプションで検索を行い、得られた URL のページを、WebRobot[4][5] を用いて取得する。

図 3.1に本方法の概要を示し、以下にそのアルゴリズムを示す。

(1) goo を用いたキーワード検索

goo を用いて、表 3.1に示したキーワードに対してそれぞれ検索を行い、検索結果のページからリンク先 URL を抽出する。

(2) URL のページを取得

(1) で得られた URL のそれぞれに対して、WebRobot を用いてそのページ (HTML ソース) を取得する。

現在、本方法は、月２回、定期的される。取得したページは選別モジュールへ渡される。

[1] http://www.goo.ne.jp/

表 3.1: goo への検索キーワードとオプション

キーワード	1. 教官+公募 2. 教官募集 3. 教官+公募 4. 教官公募 5. 教員+募集 6. 教員募集 7. 教員+公募 8. 教員公募 9. 研究員+募集 10. 研究員募集 11. 研究員+公募 12. 研究員公募
goo の検索オプション	WWW 日本語のサイトを対象 キーワードのすべての語を含む 最大結果表示 100 件 3 カ月以内に作成更新されたページを対象 日本のページ

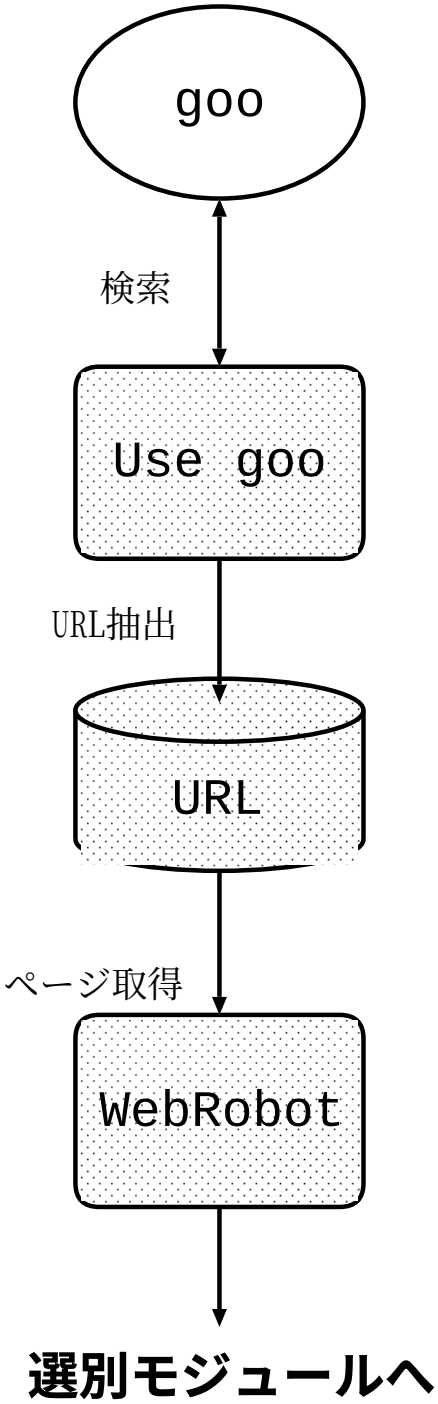

図 3.1: サーチエンジンを用いる方法

3.2 自動巡回ロボットを用いる方法

本方法では、公募対象機関[2]のホームページから出発し、最大４リンク先までリンクをたどりページを取得する。図3.2に本方法の概要を示し、以下にそのアルゴリズムを示す。

(1) 公募対象機関のホームページの URL データベースの作成

公募対象機関のホームページを自動巡回してページを取得するためには、まず公募対象機関のホームページの URL が必要となる。そこで、公募対象機関のホームページの URL データベースを以下の方法により作成した。

1. JPNIC (Japan Network Information Center) からドメイン対応表 [6] を取得する。ドメイン対応表の一部を図3.3に示す。
2. ドメイン対応表から、公募対象機関のドメイン (これを **domain** と表記する) を抽出する。例えば、北陸先端科学技術大学院大学なら"jaist.ac.jp" となる。
3. URL を http://www.**domain** と仮定し、この URL が存在するかどうか調べる。存在した場合は公募機関名とその URL を公募対象機関 URL データベースへ追加する。
4. URL を、http://**domain** と仮定し、この URL が存在するか調べる。存在した場合は公募機関名とその URL を公募対象機関 URL データベースへ追加する。
5. URL を、http://www3.**domain** と仮定し、この URL が存在するか調べる。存在した場合は公募機関名とその URL を公募対象機関 URL データベースへ追加する。存在しなければ、ホームページはないと判断する。

なお、公募対象機関 URL データベースの作成は、２ヵ月毎に自動的に行われる。

(2) ページ収集

作成した公募対象機関 URL データベースと WebRobot を用いて、一日に２０機関を対象に各機関のホームページから出発し、最大４リンク先までリンクをたどり、ページを取得する。取得されたページは選別モジュールへ渡される。

[2] 公募対象機関は、学術情報センターの「研究者公募情報」に準拠する。

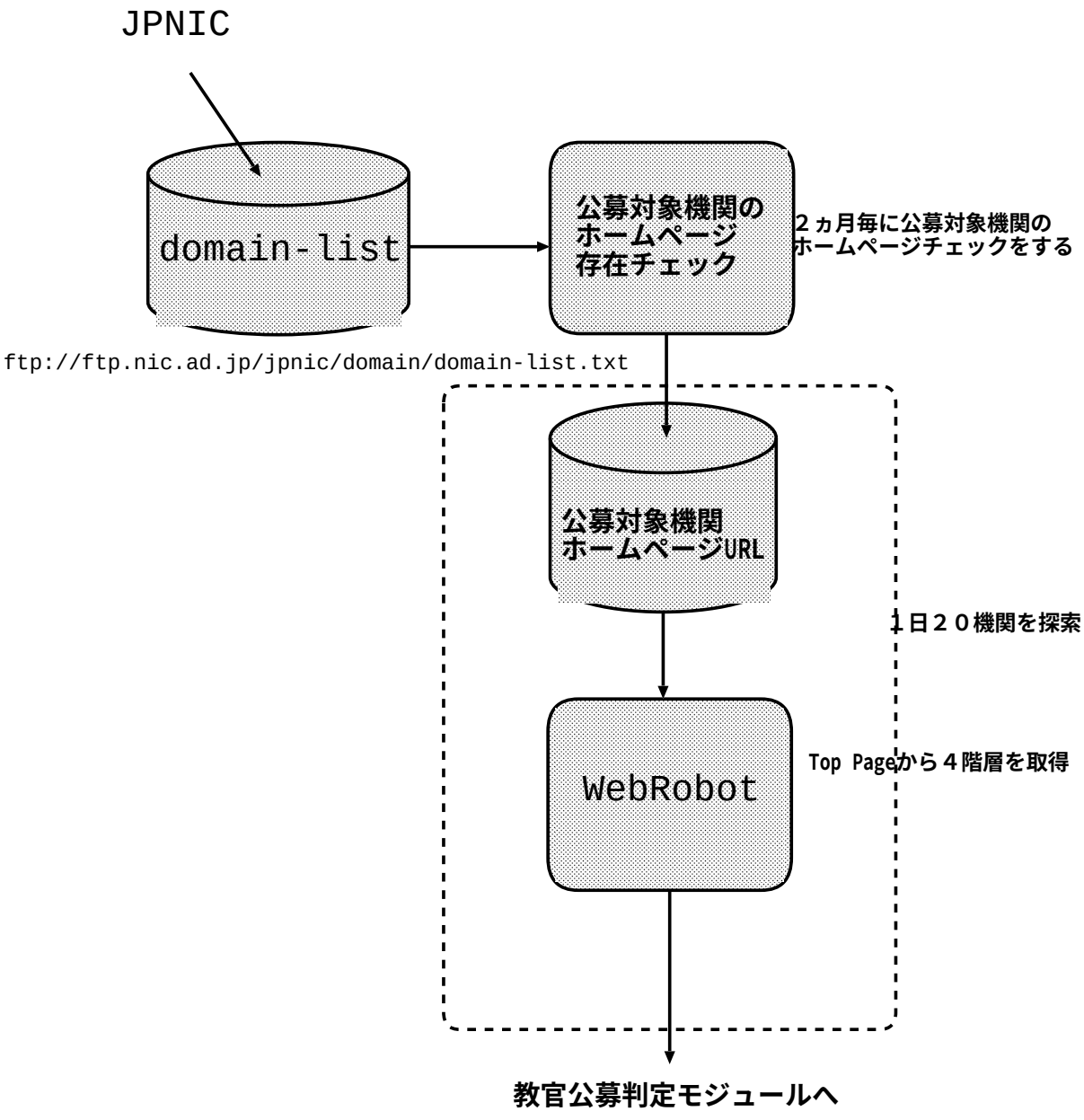

図 3.2: 自動巡回ロボットを用いる方法

```
Registered Domains in JP (Jan 10 1999): 59252
Connected  Domains in JP (Jan 10 1999): 54546
(Domains in parentheses are not connected.)

------ JP domains:    1 (0)
 KEK                  # 高エネルギー物理学研究所

------ AD domains:  207 (2)
 AAA                  # アーキテック・アンド・アーツ
 ADMIRAL              # アドミラルシステム
 ADMIX-NET            # ADMIX
 AGSIS                # あさひ銀総合システム株式会社

～ 中略 ～

 JAIST                # 北陸先端科学技術大学院大学
 JASRA                # 大原スキーテクノカレッジ
 JBC                  # 新潟情報ビジネス専門学校
 JC-21INT             # 東北電子計算機専門学校
 JCFL                 # 学校法人 文際学園日本外国語専門学校
 JCMW                 # 日本医療福祉専門学校
 JEWELRY              # 学校法人 水野学園
...
```

図 3.3: ドメイン対応表の例

3.3　公募リンクデータベースを用いる方法

ある機関（大学など）が公開しているウェブページのなかには、その機関内の、公募情報ページへのリンクがリスト形式で書かれているものがある。そのようなページの例を図3.4に示す。以下では、このようなページを公募リンクページと呼ぶ。公募リンクページは、その機関が新たな公募を行った時、書き換えられる。そのため、公募リンクページのURL をデータベース化[3]しておき、この URL を定期的に取得することにより、新しい公募ページを効率良く取得することが可能となる。

本方法は、公募リンクページからたどれるページを定期的 (月 2 回) に収集する。図 3.5 に本方法の概要を、以下にそのアルゴリズムを示す。

(1) 公募リンクデータベースから URL を取り出す

(2) URL のページとそこから辿れるページを１つづつ取得して選別モジュールへ渡す

[3] この公募リンクデータベースに、４章で述べる選別モジュールによって「公募リンクページ」と判定されたページの URL が順次追加される。

図 3.4: 公募リンクページの例 : http://www.edu.wakayama-u.ac.jp/gakubu/koubo.html

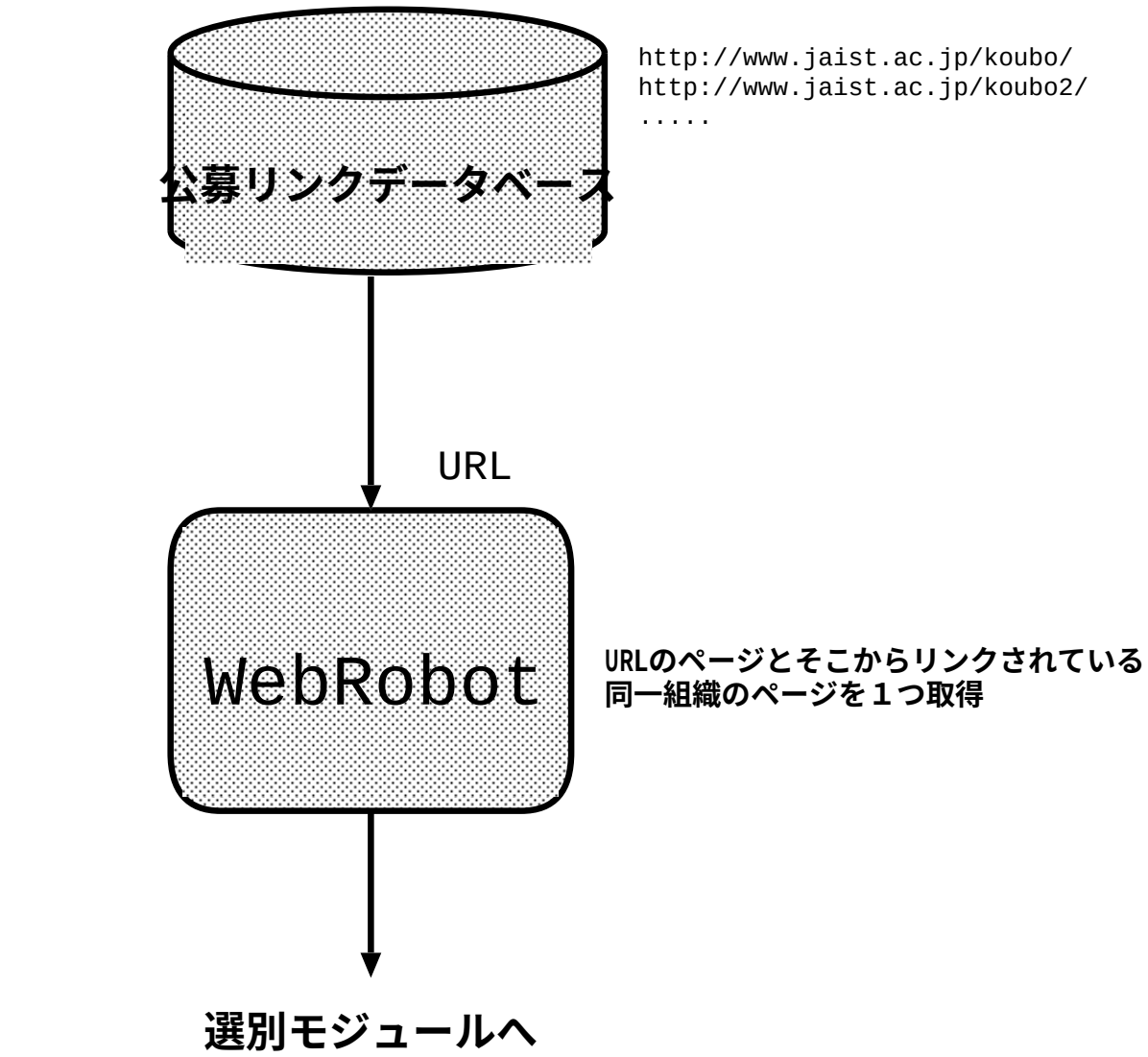

図 3.5: 公募リンクデータベースを用いる方法

第 4 章

選別モジュール

選別モジュールは、収集モジュールにより取得されたページを、(1) 公募ページ、(2) 公募リンクページ、(3) その他、に選別する。

4.1 公募ページ、公募リンクページの言語表現パターン

収集モジュールが取得したページを、(1) 公募ページ、(2) 公募リンクページ、(3) その他、に選別する方法を検討するために、公募ページ１０８ページ、公募リンクページ１８ページ、その他のページ２４ページの合計１５０ページを収集して調査を行った。この調査の結果、以下のものが手がかりとなりそうなことが明らかになった。

(1) 見出し
公募ページには「職種」「職名」「締切日時」「提出先」「問い合わせ先」などを表す見出しが存在する。その例を図 4.1に示す。

(2) 公募ページ、公募リンクページ共通に使われるキーワード
「教官公募」「教員公募」「教官募集」「教員募集」「公募案内」「公募」「募集」といったキーワードが両者に存在する。また、公募リンクページには、アンカータグの中に「公募」「募集」「教授」「助教授」「助手」「講師」といったキーワードが存在する。

以上を踏まえて、以下に示すようなアルゴリズムを利用することとした。

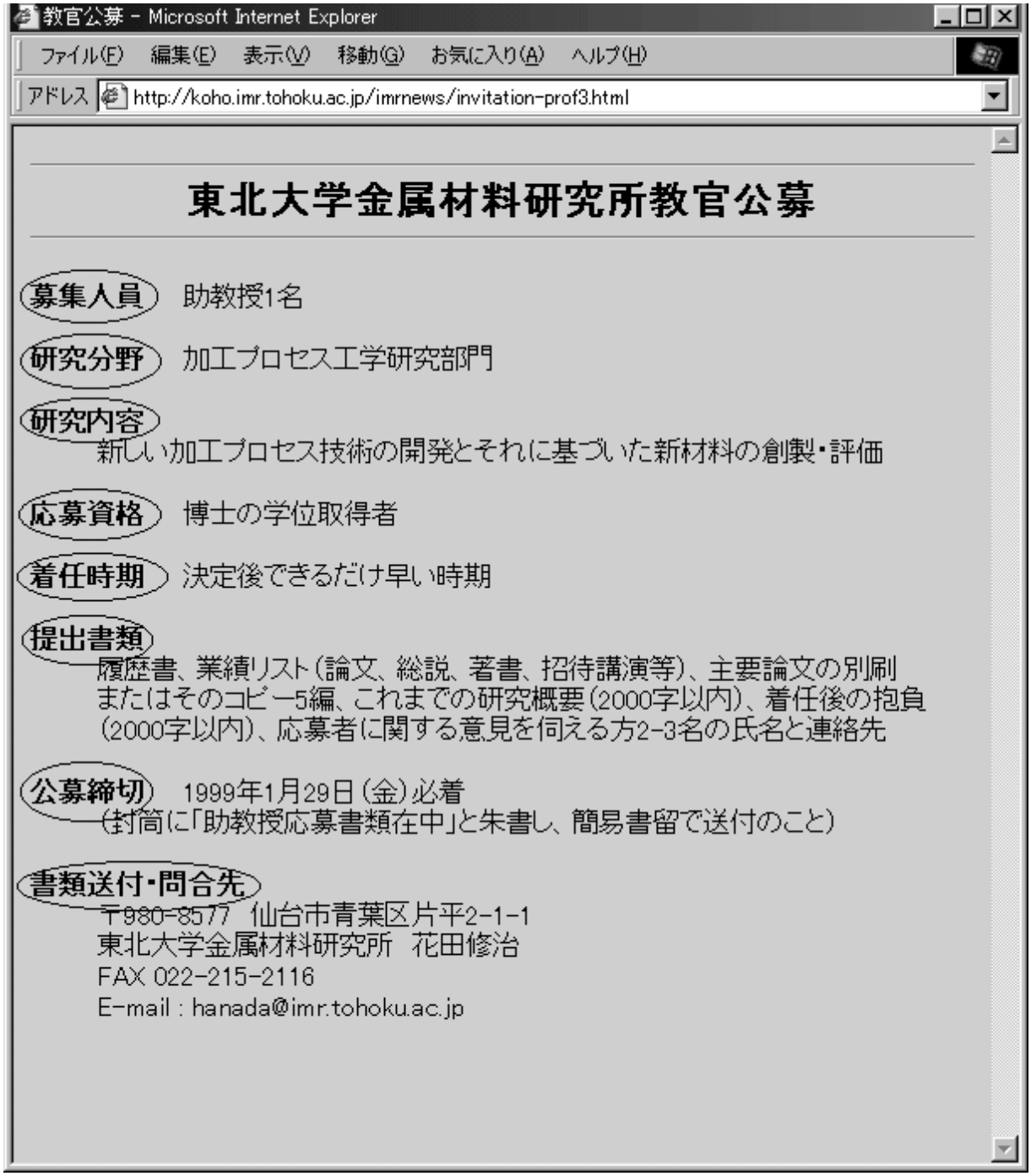

図 4.1: 公募ページの見出しの例 : 丸で囲った所が見出しの箇所

表 4.1: 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 | 教員) 公募
(教官 | 教員) 募集
公募案内
公募
募集

4.2 ページ選別アルゴリズム

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、に選別する。

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」など表4.1参照) 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」「締切日時」「提出先」「問い合わせ先」などの見出し語や職名 (表4.2、4.3、4.4参照) 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」など表4.5参照）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する。

4.3 選別モジュールの評価実験

選別モジュールの評価実験を行った。ここでは評価対象として選別モジュールが、公募ページ、公募一覧ページ、その他のページ、と選別したページを、ランダムに１００ページつつ取り出した合計３００ページを用いた。実験結果を表4.6に示す。

公募ページ、その他のページの選別には満足いく結果が得られた。しかし、公募リンクページの選別は、63 %と不本意な結果になってしまった。この原因は、システム作成に用いた公募リンクページのサンプルが１８ページと少なかったため、公募リンクページ特

表 4.2: 見出し語のパターン : 職種、職名

職種	公募人員，専門分野および応募資格 採用職名 採用人員 公募人員 (及び | および)?(職種)? 募集人員 人員 採用予定職 (名)?(及び | および) 人員 公募職種 (及び | および) 人員 任用予定職 (名)?(及び | および) 人員 職名と人員 職名 採用職 職種 (予定)?職位 職名 採用予定官職
職名	教授 助教授 助手 講師研究員

表 4.3: 見出し語のパターン : 締切日時

締切日時	(提出)?(締 | 〆) め?切 (り)?日? 応募 (締 | 〆)(め)?切 (り)?(日)? 募集 (締 | 〆) 切り?(日)? 公募 (締 | 〆) 切 (り)?(日)? 公募期限 提出期限 書類提出期限 応募期限 募集期間 応募期間 公募期間

表 4.4: 見出し語のパターン : 提出先・問合せ先

提出先・問合せ先	提出先 提出書類先 書類提出先 書類送付先 書類送付 応募書類提出先 提出先と問 (い)?合 ((わ)?せ)?先 提出・問 (い)?合 ((わ)?せ)?先 送付先 (及び | および) 問 (い)?合 ((わ)?せ)?先 提出先 (及び | および) 問 (い)?合 ((わ)?せ)?先 (問 (い)?合 ((わ)?せ)?先 (及び | および))?資料請求先 問 (い)?合 ((わ)?せ)?先 応募書類の提出先 宛先 (及び | および) 問 (い)?合 ((わ)?せ)?先 送付先 (／ |/)?照会先

表 4.5: 公募リンクページ特有のパターン

アンカータグの中のパターン	(教官 | 教員) 公募 (教官 | 教員) 募集 (教官 | 教員). * 募集 助手 助教授 教授 研究員 公募中 募集要項 詳細
文章中のパターン	(教官 | 教員) 公募 (教官 | 教員) 募集 公募中 (教官 | 教員) 公募掲載 (教官 | 教員) 募集掲載

表 4.6: 選別モジュール実験結果

	判定成功	判定失敗	計
公募ページ	97 (97%)	3 (0%)	100
公募一覧ページ	63 (63%)	37 (37%)	100
その他	100 (100%)	0 (0%)	100

有のパターンがうまく特定できなかったことが大きいと思われる。このパターンを強化すれば精度は向上すると考えられる。

第 5 章

情報抽出モジュール

公募情報を求める人々にとって、重要な情報は、「どこの機関が」「どんな分野の」「どんな職種の」公募をしているか、また、「応募締切日」はいつか、という情報である。さらに、「公募機関の地域」も分かればなおよい。そこで、情報抽出モジュールでは、公募ページから、(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・職務内容、(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。

5.1 公募ページの特徴

公募ページから、重要情報を抽出する方法を検討するために、まず、ワールドワイドウェブ上に存在する公募ページを人手により収集し、調査を行った。この調査の結果、以下の二つが重要情報抽出の際に、手がかりとなりそうなことが明らかになった。

(1) 見出し語

公募ページは、重要な情報を読者が見つけやすいように、見出し語がつけられていることが多い。見出し語の種類には、「採用職名」「所属先」「専門分野」「公募職種」「応募資格」「着任時期」「提出書類」「応募締切」「提出先」「問い合わせ先」「その他」などに大きく分類される。公募ページの見出し語の例を図 5.1に示す。このことより、見出し語を用いて、公募ページをいくつかの部分に分割し、これをうまく利用することによって、情報を抽出すべき場所をかなり限定できる。

(2) 抽出する情報に特有な言語表現パターン

抽出すべき情報には、それぞれ特有な言語表現が存在する。所属先には、「学部」「学

科」などの所属先に関する言葉が含まれる。また、日付の表記は、基本的に「X 年 X 月 X 日」のバリエーションである。これらの言語表現をパターン化したもの（言語表現パターン）とのパターンマッチングによって、多くの場合、抽出する情報を特定することができる。

5.2 公募情報抽出アルゴリズム

情報抽出モジュールの概要を図 5.2に、アルゴリズムを以下に示す。

(1) １ページ１公募、１ページ複数公募の判定

公募ページには、１ページに１公募情報が書かれているものと、１ページに複数公募情報が書かれているもの、が存在する。

以下の処理で、１ページ１公募情報を仮定するために、１ページ複数公募情報のページはそのままでは、うまく抽出できない。そこで、１ページ複数公募情報である場合には、そのページから１つずつ公募情報を切り分けて以下の処理を行う。

公募情報を切り分けるために、１ページに複数の公募情報があるページを調査したところ、次のパターンに当てはまるもので切り分ければ良いことが分かった。

(a) NAME タグ
　複数公募ページの場合、ページのトップに各公募情報へのリンクが $< a\ name >$ タグにより張られていることが多い。

(b) 「-」「=」などが連続したもの
　公募ページを分ける時に使われることがある。

以上のパターンを用いて、(a)、(b) のパターンが存在した時に、分割し、(2) へ渡す。

(2) 見出し語を用いて、公募ページをいくつかに分割する

見出し語を用いて公募ページを、「採用職」「所属先」「分野」「職務」「応募資格」「着任時期」「提出書類」「応募締切」「提出先」「問合せ先」「その他」の部分に分割する。表 5.1, 5.2, 5.3, 5.4に分割のために用いる見出し語のパターンを示す。図 5.3 に分割例を示す。この図の点線は、分割境界を表す。なお、公募ページの$< TITLE >$ タグの中の文字列と、あらかじめ付加しておいた公募ページの URL もあわせて抽出しておく。

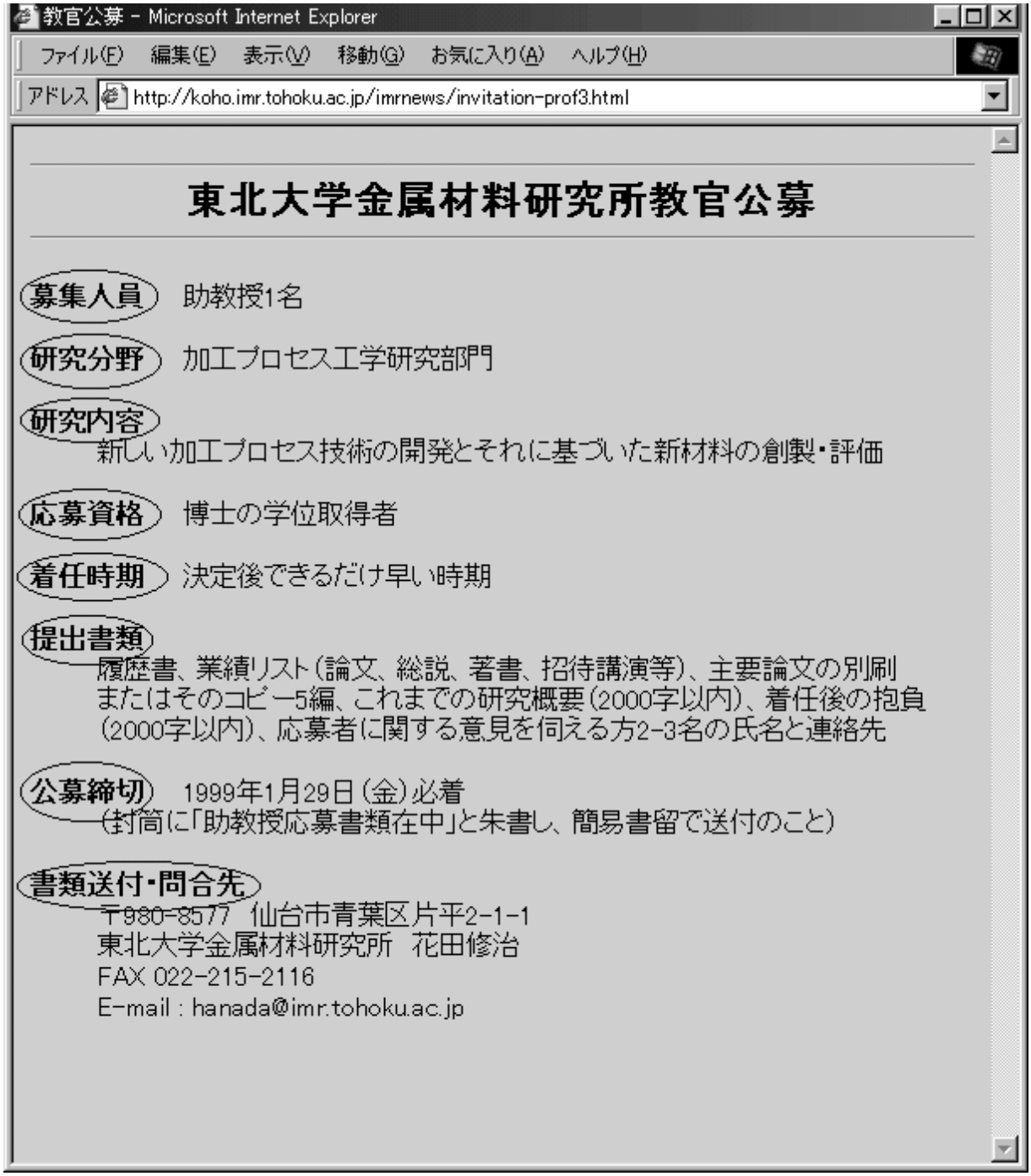

図 5.1: 公募ページの見出しの例 : 丸で囲った所が見出しの箇所

表 5.1: 見出し語のパターン : 採用職, 所属先

採用職	採用予定職名 募集する教官の職名及び人数 採用職 (名 | 種)((及び | および) 人員)? 採用 (予定職 (及び | および) 人員) 公募人員 ((及び | および) 職種)? 公募職種 ((及び | および) 人員)? 公募学科 (及び | および) 人員 職名 (と人員 | ・人員 |(及び | および) 人員)? 職種・人員 任用予定職名 (及び | および) 人員 所属・職名・人員 (職位 ・)?募集人員 任用職 職種
所属先	所属学科 (及び | および) 分野 所属学科 (所属)?講座 (名 | 等 | 等名)? 所属・職名・人員 研究分野 ((及び | および) 研究内容)? 所属 採用学部・学科 学部学科

表 5.2: 見出し語のパターン : 分野, 職務内容, 応募資格

分野	専攻 (科目 | 分野 | 領域) 専門 (分野 | 領域 | 研究 (分野 | 領域)) 研究 (職務) 内容 研究分野 教育・研究内容 募集分野と担当科目 公募対象 教育研究分野 分野
職務内容	担当授業 (科目)? 担当 (予定)?科目 担当予定の (授業科目等 | 講義) 担当する講義等 担当 (分野 | 内容 | 教育内容) 主要担当科目 主要担当授業科目 業務内容 職務内容 ((及び | および) 概要)? 研究 ((職務))?内容 教育・研究内容 職務
応募資格	応募資格 (等)? 応募年齢 資格 年齢

表 5.3: 見出し語のパターン : 着任時期, 提出書類, 応募締切, その他

着任時期	採用 (予定)?年月 (日)? 採用予定 (期)?日 着任 (時期 | 期日) 任用 (予定日 | 時期 | 予定時期) 就任 (時期 | 希望時期) 採用時期
提出書類	(提出—必要) 書類
応募締切	(募集)?(締 | 〆)(め)?切日 応募 ((締 | 〆)(め)?切 | 期限)(日)? 公募 (締 | 〆)(め)?切 (日)? 公募期限 (書類)?提出期限 応募期間 (締 | 〆) 切)
その他	その (他 | た) 選考 (方法)? 任用予定 採用 (予定期間 | 人数) 任期 待遇 捕捉 備考 応募条件 面接 目的

表 5.4: 見出し語のパターン : 提出先, 問合せ先

提出先	応募先 提出先 (と問 (い)?合 (わせ | せ)?先)? 提出 ((及び | および) 問 (い)?合 (わせ | せ)?先 問 (い)?合 (わせ | せ)?先 書類先 書類送付 (先)? 書類提出先 書類の送付 (先)?(及び | および) 問 (い)?合 (わせ | せ)?先 応募書類 (の)?(提出先 | 送付先) 送付先 ((及び | および) 問 (い)?合 (わせ | せ)?先)? 資料請求先 宛先 (及び | および) 問 (い)?合 (わせ | せ)?先 問 (い)?合 (わせ | せ)?先 (、| ・) 書類提出先 応募方法問い合わせ、及び応募書類送付先 送り先
問合せ先	問 (い)?合 (わせ | せ)?(先)?((及び | および)(資料請求先 | 書類送付先))? 問 (い)?合 (わせ | せ)?(先)?(、| ・) 書類 (提出 | 送付) 先 提出先と問 (い)?合 (わせ | せ)?先 提出 (及び | および) 問 (い)?合 (わせ | せ)?先 提出・問 (い)?合 (わせ | せ)?先 資料請求先 宛先 (及び | および) 問 (い)?合 (わせ | せ)?先 本件の照会先 本件に関する問 (い)?合 (わせ | せ)?先 照会先 応募方法問い合わせ 問合わせ並びに書類提出先

(3) 分割部分と URL、$< TITLE >$タグの中身を用いて情報を抽出

「URL」「公募機関名」「所属１」「所属２」「所属３」「所属４」「職種」「分野・研究内容」「応募締切日」「地域」の情報を以下の方法で抽出し公募情報データベースへ追加する。

1. **公募ページの URL の抽出**
   公募ページを取得した時に、あらかじめ付加しておいた URL を抽出する。
2. **公募機関名の抽出**
   JPNIC (Japan Network Information Center) のドメイン対応表を用いて、抽出したURL から公募機関名を特定する。例えば、URL が http://www.jaist.ac.jp/koubo.html であったなら、ドメイン対応表の”jaist.ac.jp” に対応する「北陸先端科学技術大学院大学」が公募機関名となる。
3. **採用職名を抽出**
   採用職部分から採用職名を抽出する。以下の正規表現パターンにマッチする文字列を抽出する。

   ```
   (学長 | 副学長 | 教授 | 助教授 | 助手 |(専任)?講師 | 研究員 |(又は | または)|
   、|, )+
   ```

   もし、上記の方法で採用職が得られない場合は、ページ全体を職種のパターンでスキャンして採用職名を抽出する。
4. **公募締切日を抽出**
   応募締切部分から日時のパターンを用いて、締切日時を抽出する。
5. **専門分野・職務内容を抽出**
   現在、分野の特定はしておらず、分野、職務部分をそのまま、抽出する。
6. **公募機関の地域の特定**
   提出先、問合せ先部分を用いて公募機関の地域の特定を行う。まず、提出先、問合せ先から郵便番号情報が存在するかどうか調べ、存在すれば郵便番号から都道府県を特定する。存在しなければ、都市名を抽出し、都市名から都道府県を特定する。
7. **所属先の抽出**
   採用職、所属、提出先、問い合わせ先の各部分と $< TITLE >$タグの中身、公募機関名、から抽出する。このとき、所属は「公募機関名」「大学の学部に相

表 5.5: 所属情報存在調査パターン

(学部 | 研究科 | 学群 | 言語文化部 | 教養部 | センター | 施設 | 研究所 | 資料編纂所 | 図書館 | 分館 | 資料館 | 学科 | 教室 | 専攻 | 学類 | 課程 | 部門 | 天文台 | 実験所 | 観測所 | 試験所 | 病院 | 分院 | 学校 | 植物園 | 高等学校 | 小学校 | 中学校 | 幼稚園 | 牧場 | 農場 | 工場 | 博物館 | 演習林 | 診療所 | 研究船 | 練習船 | 系 | 分室 | 開発資料部 | 領域 | 講座)

表 5.6: 所属区切りパターン

所属 1	(大学大学院 | 大学院大学 | 大学校 | 大学院 | 大学 | 高等専門学校)
所属 2	(学部第.1, 2 部 | 学部 | 研究科 | 学群 | 言語文化部 | 教養部 | 研究所 | 資料編纂所 | 分館)
所属 3	(センター | 施設 | 図書館 | 資料館 | 学科 | 教室 | 専攻 | 学類 | 課程 | 部門 | 天文台 | 実験所 | 観測所 | 試験所 | 病院 | 分院 | 小学校 | 中学校 | 高等学校 | 付属.∗?学校 | 植物園 | 幼稚園 | 共同利用施設 | 牧場 | 農場 | 工場 | 博物館 | 演習林 | 診療所 | 第 1, 2 部 | 実習所 | 実験室 | 研究船.∗?丸 | 練習船.∗?丸 | 系 | 分館 | 分室 | 開発資料部 | 別科)
所属 4	(講座 | 実験場 | 分野 | 領域)

当するもの」「大学の学科に相当するもの」「大学の講座に相当するもの」４つの階層に分類する [8]。

まず、採用職、所属、提出先、問い合わせ先の各部分、$< TITLE >$タグの中身の各々から、以下の方法で所属情報を抽出する。

(a) 採用職部分に表 5.5に示すパターンが存在するかどうか調べる。ない場合は、所属情報がないので抽出終了。

(b) 余分な記号や、文字を削除する。

(c) 表 5.6の所属 1-4 にマッチするパターンで文を区切り、所属情報を抽出する。

次に、各々の部分から抽出した情報をマージする。

例えば、図 5.3に示したページから所属先の抽出を行うと、提出先部分から、「東北大学金属材料研究所」、専門分野（所属）から、「加工プロセス工学研究部門」が抽出される。これらを、マージし、最終的に得られる所属先は「東北大学金属材料研究所加工プロセス工学研究部門」となり、これを４つの階層に分類すると「東北大学」「金属材料研究所」「加工プロセス工学研究部門」「???」(この部分は存在しない) となる。

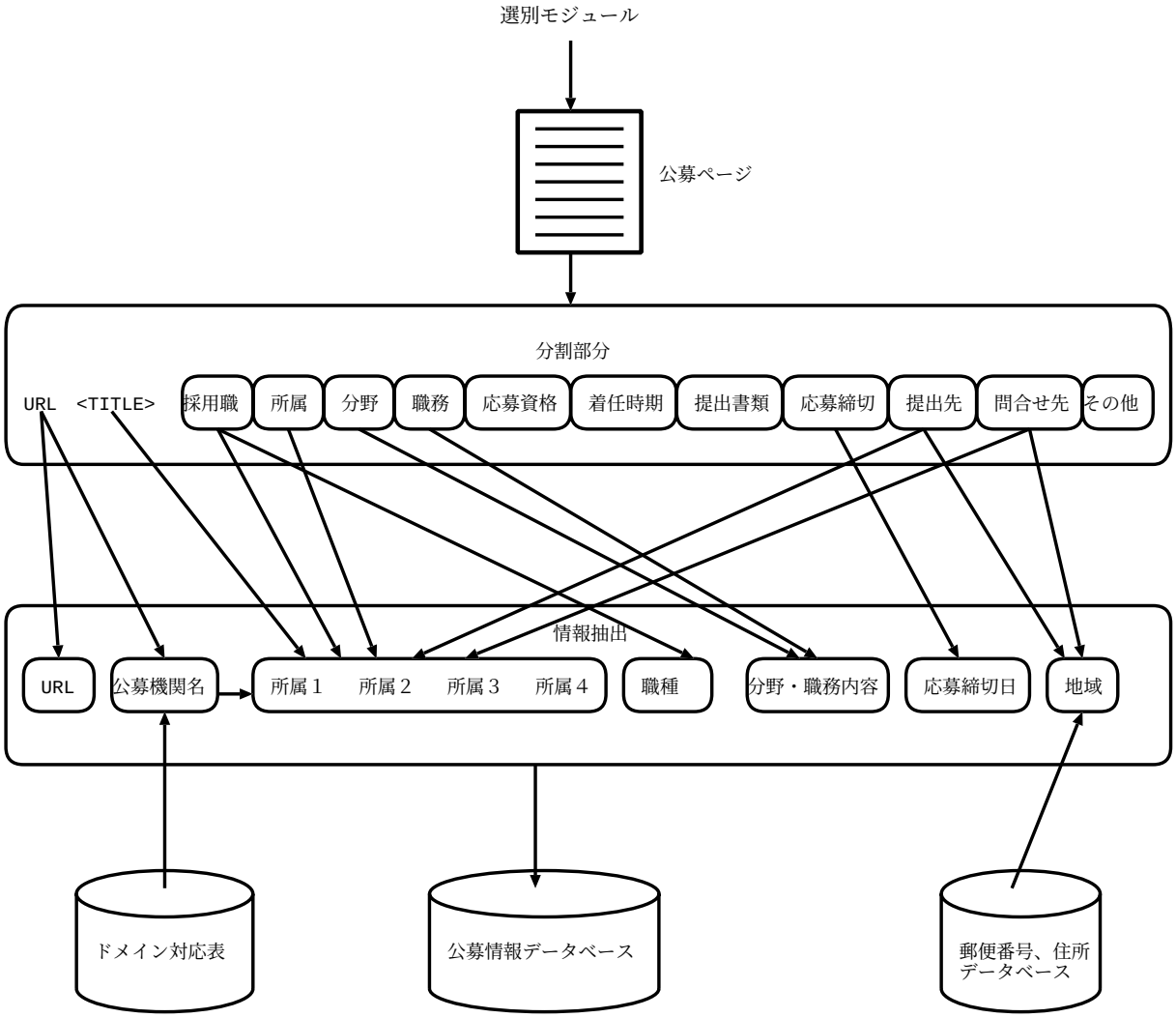

図 5.2: 情報抽出モジュールの概要

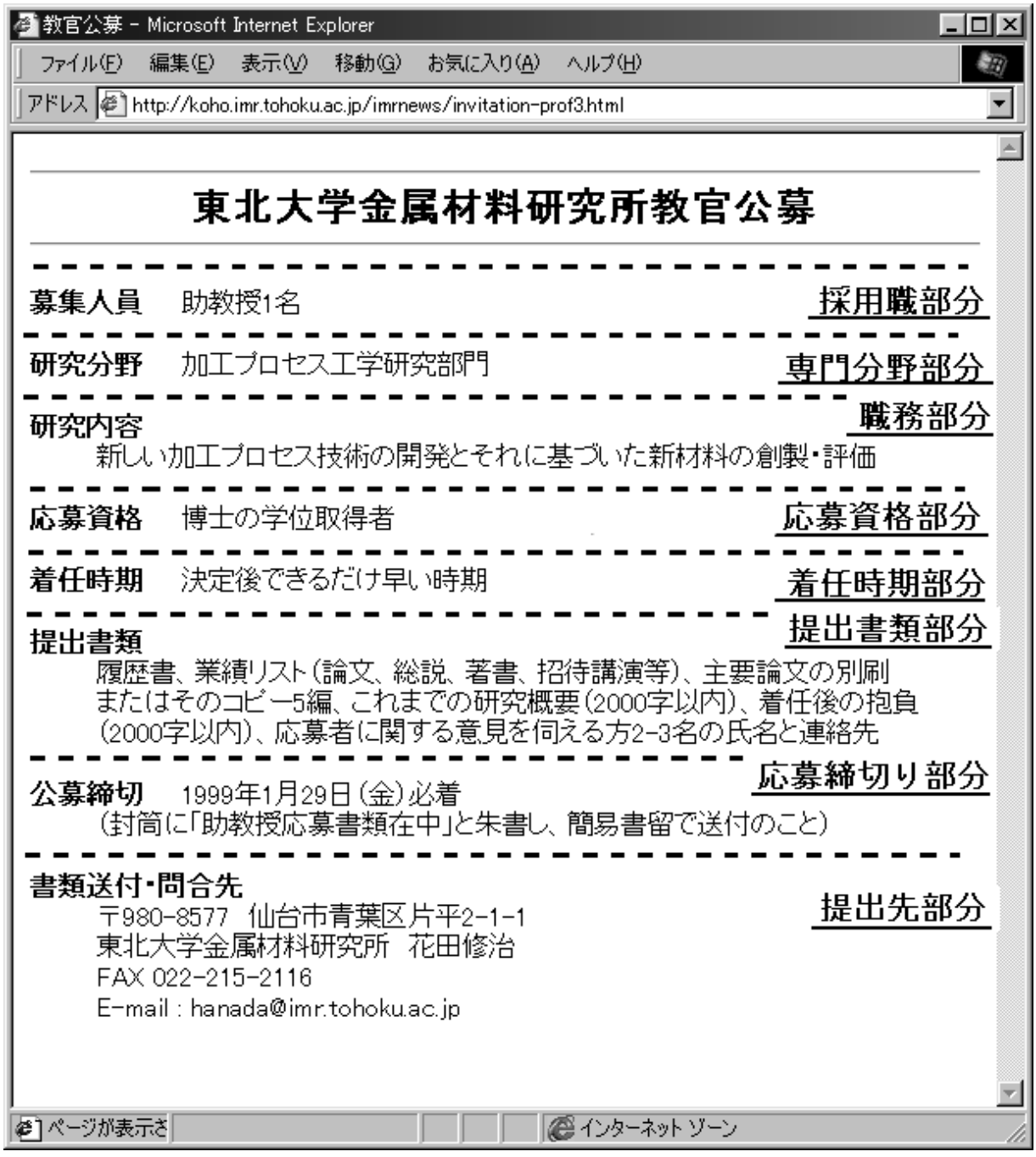

図 5.3: 見出し語を用いていくつかの部分に分割の例

5.3　公募情報抽出実験

情報抽出モジュールの評価実験を行った。対象ページは平成１１年１月５日現在、システムがワールドワイドウェブから収集した 213 公募ページである。表 5.7に実験結果を示す。

この表の評価基準は以下の通りである。

- **所属先**

  ○：所属情報が完全に抽出できた

  △：所属情報が一部欠落していたり、まちがっていたりして抽出した

  ×：所属情報が全く抽出できなかった

- **職種**

  ○：職名情報が完全に抽出できた

  △：職名情報が一部欠落していたり、まちがっていたりして抽出した

  ×：職名情報が全く抽出できなかった

- **分野・職務内容**

  ○：人間が読んで、専門分野や職務内容がわかる

  △：人間が読んで、専門分野や職務内容が推測できる

  ×：人間が読んで、専門分野や職務内容がわからない

- **公募締切日**

  ○：抽出できた

  ×：抽出できなかった

- **公募機関の地域**

  ○：地域情報が得られた

  ×：地域情報が得られなかった、または間違っていた

表 5.7に示すように、情報抽出モジュールの精度は各項目に差はあるが、おおむね 80-95 %の正解率を得られた。また、△までを正解とするならば、正解率は 85-95 %となる。

表 5.7: 公募情報抽出実験結果

	○	△	×	正解率 ○/213	正解率 (○+△)/213
所属	173	36	4	81 %	98 %
職種	197	3	13	92 %	94 %
分野	171	11	32	80 %	85 %
締切日	192		21	90 %	
地域	202		11	95 %	

第 6 章

検索モジュール

本モジュールでは、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つの条件で検索することができる。図 6.1に検索インターフェースの画面を示す。

- 公募機関名
  ここに、入れられた文字列は、公募情報データベースの公募機関名のフィールドと照合される。したがって、短期大学のみの公募情報を調べたい時などには、「短期大学」と入力すれば、短期大学の公募の一覧がダイジェスト形式で表示される。その例を図 6.2に示す。

- 公募機関の地域
  公募機関の地域を特定することができる。地域は、「北海道地区」「東北地区」「関東甲信越地区」「東京地区」「北陸地区」「東海地区」「近畿地区」「中国・四国地区」「九州・沖縄地区」から選べる。なお、この地区分けは学術情報センターの「研究者公募情報」のディレクトリ検索に準拠している。

- 分野・内容
  ここに、入れられた文字列は、公募情報データベースの分野・内容フィールド内の文字列と照合される。したがって、「情報工学」の分野の公募を探したい場合はここに、「情報工学」などと入れればよい。但し、分野・内容の抽出精度はそれほど高くない (80 %〜) ので、適切に検索できない場合もある。

- 公募職名
  公募職名を特定して検索できる。職名は「学長」「副学長」「教授」「助教授」「講師」

「助手」「研究員」から選ぶことができる。

システムは以下の処理を行う。

(1) ユーザの検索入力に対して、システムはそれに合致する公募情報だけ公募情報データベースから選び出す。

(2) 次に、それぞれの公募締切日を調べ、公募締切日が過ぎているものを排除する。

(3) (2) で残ったそれぞれの公募情報のレコードを HTML に変換して出力する。このとき、所属先には、その公募機関のホームページのハイパーリンクを付加する。

本システムは現在、学術情報センターの「研究者公募情報」の情報もあらかじめ取得して、データベース化してあるので、本システムがワールドワイドウェブから収集した公募情報と学術情報センターが提供する公募情報を同時に検索することが可能である。

図 6.3に公募機関の地域を「北海道地区」職名を「教授」で検索した結果を示す。

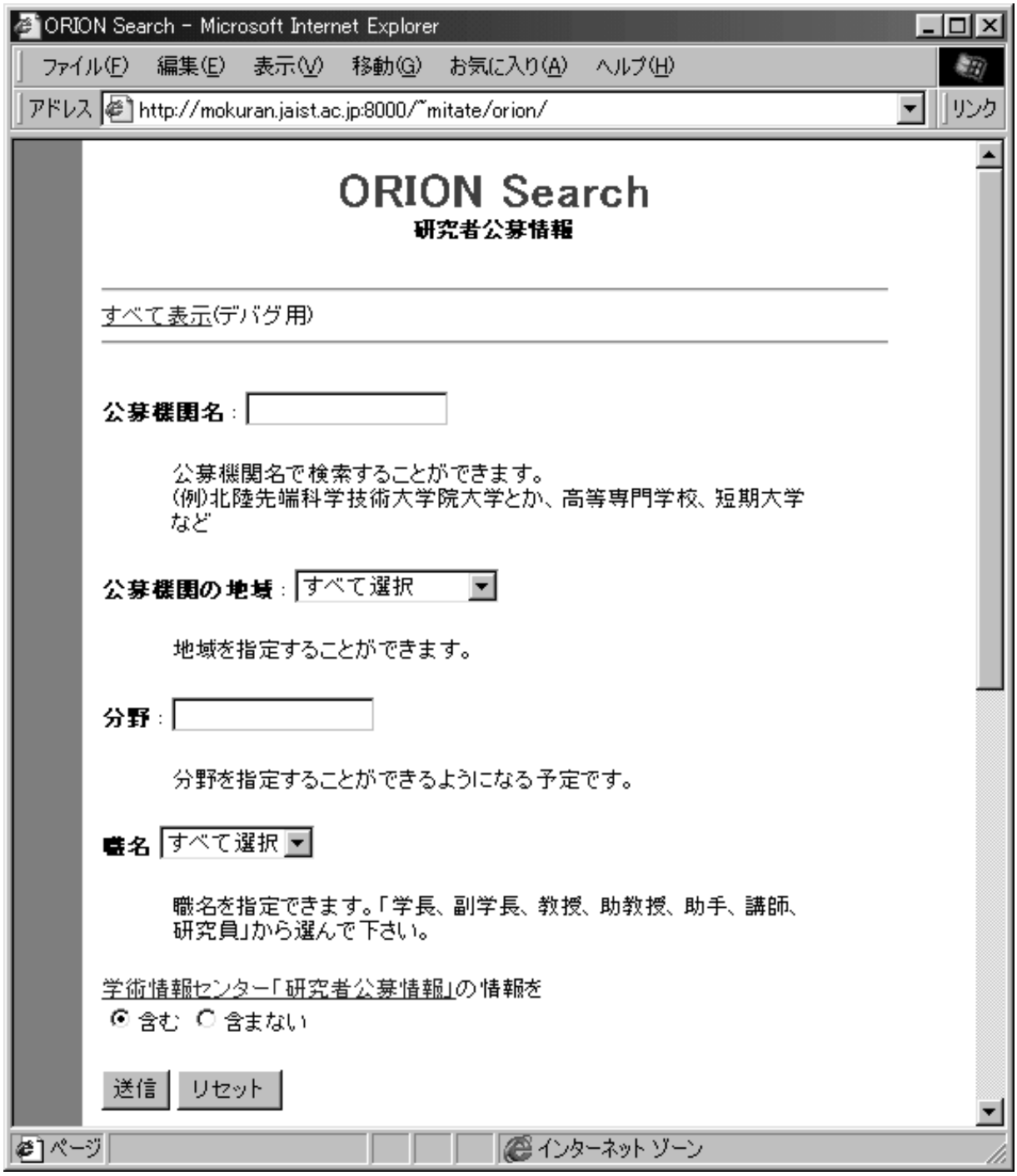

図 6.1: 検索インターフェイス

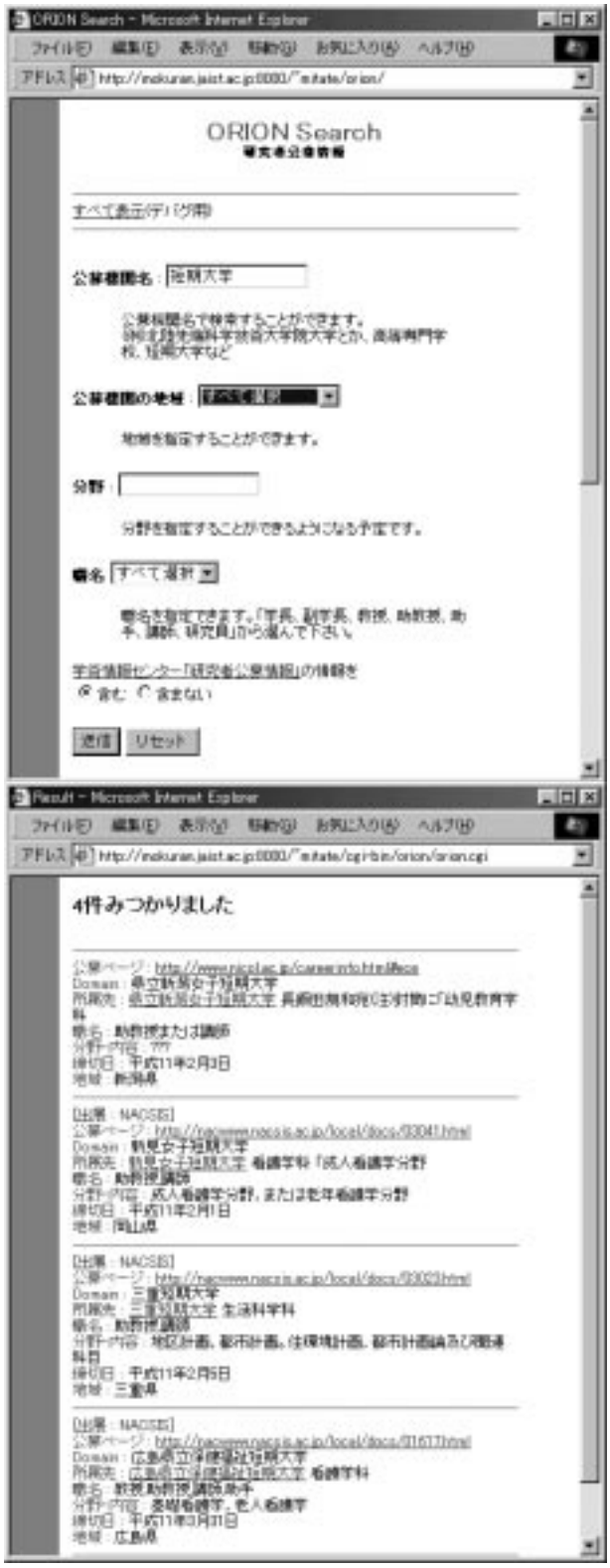

図 6.2: 短期大学の検索とその結果

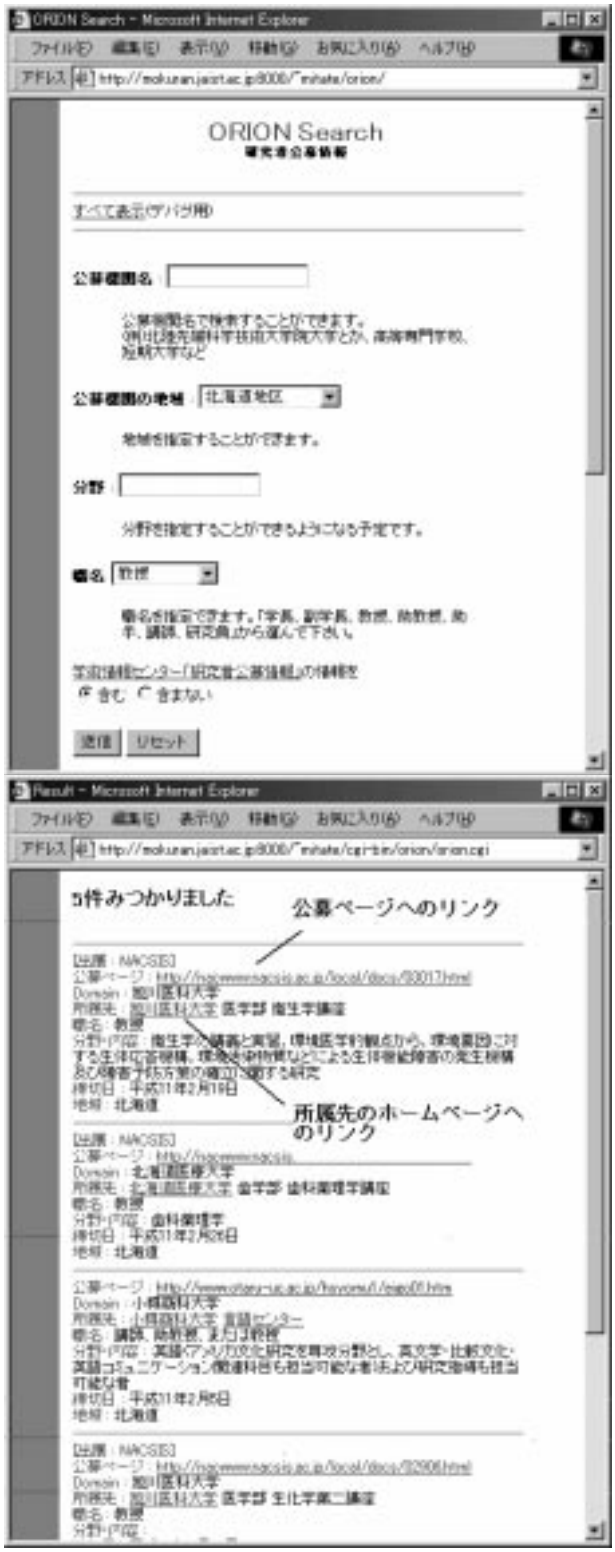

図 6.3: 検索例：地域「北海道地区」、職名「教授」

第 7 章

結論

本稿では、ワールドワイドウェブ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。

本システムは、収集、選別、情報抽出、検索、の４つのモジュールと、公募リンクデータベース、公募情報データベース、の２つのデータベースから構成される。収集モジュールでは、ワールドワイドウェブから教官公募ページらしきページを収集し、それらのページを選別モジュールで、教官公募ページ、公募リンクページ、その他のページに選別する。公募リンクページの URL は公募リンクデータベースへ保存される。情報抽出モジュールは、教官公募ページから教官公募情報の主要な情報を抽出し、公募情報データベースへ保存する。検索モジュールでは、ユーザの要求に合致する公募情報を公募情報データベースから検索し、ダイジェスト形式で表示する。

本システムの中心的部分は、収集モジュールと情報抽出モジュールである。収集モジュールは、教官公募ページらしきページを収集するために、サーチエンジンを用いる方法、自動巡回ロボットを用いる方法、公募リンクデータベースを用いる方法、の３つの方法を実装した。情報抽出モジュールでは、教官公募ページによくみられる見出し語を用いてページをいくつかの部分に分割し、分割された各部分から、抽出する情報に特有な言語表現パターンを用いて情報を抽出する。抽出する情報は、所属先、公募職名、分野・内容、公募締切日時、公募機関の地域である。分野・内容の情報は、書き手により表現がさまざまであり、かつ書かれている文章から専門分野を特定することが現段階では困難なため、80-85 %と精度はそれほど高くない。これ以外の情報は、90 %以上の抽出精度を得られた。

ワールドワイドウェブ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」がある。このサービスは現在、学術情報センターに公募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、本システムでは、ワールドワイドウェブ上に存在する教官公募ページを自動的に見つけ出すことができる。また現在、紙の媒体による教官公募案内が主流であるが、今後さらにインターネットが普及することにより、紙の媒体と同時に各公募機関がワールドワイドウェブ上で公募情報を発信していくと思われる。そうなると、本システムの重要性がさらに高まると考えられる。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。

謝辞

本研究を進めるにあたり、終始熱心な御指導を賜わりました佐藤理史助教授に心から感謝致します。

さらに、常日頃より議論を重ね、研究に関して良きアドバイスを下さった知識工学講座の皆様に心より感謝の意を表したいと思います。

最後に、２年にわたる JAIST での生活を支えてくれた家族、そして友人に感謝致します。

見舘 潔

1999 年 2 月 15 日

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.

[2] 佐藤理史, 佐藤 円 : ネットニュースグループ fj.wanted のダイジェスト自動生成, 自然言語処理, Vol.3, No.2, pp.19-32, 1996.

[3] Larry Wall, Tom Christiansen, Randal L. Schwartz 共著, 近藤 嘉雪 訳, プログラミング Perl 改定版, オライリー・ジャパン, 1997.

[4] Clinton Wong 著, 法林 浩之 監訳, 須田 隆久 訳, Web クライアントプログラミング, オライリー・ジャパン, 1997.

[5] $The Web Robots Pages$, http://info.webcrawler.com/mak/projects/robots/robots.html

[6] ドメイン対応表, ¡ftp://ftp.nic.ad.jp/jpnic/domain/domain-list.txt¿

[7] 長尾 真, 黒橋 禎夫, 佐藤 理史, 池原 悟, 中野 洋, 言語情報処理, 岩波書店, 1998.

[8] 全国大学職員録, 廣潤社, 1992.