CiNiiを利用したエキスパートサーチシステム

増田　浩司†　　太田　　学††

† 岡山大学工学部　〒 700–8530　岡山県岡山市北区津島中 3-1-1
†† 岡山大学大学院自然科学研究科　〒 700–8530　岡山県岡山市北区津島中 3-1-1
E-mail: †masuda@de.cs.okayama-u.ac.jp, ††ohta@de.cs.okayama-u.ac.jp

あらまし　本稿では, 入力された専門用語に関係する分野に詳しい研究者を検索して, その人物のプロファイルを自動生成するシステムを提案する. ユーザがまず専門用語を入力すると, 提案システムがそれに関連する学術論文を学術文献データベース CiNii で検索して, その著者をエキスパートとして抽出する. そしてその著者名と専門用語を用いて Web からその著者の関連情報を抽出し, 経歴や関連人物, 研究事例からなるプロファイルを自動生成する. 提案システムを利用すれば, ユーザはその研究分野のエキスパートとそのプロファイルを容易に把握できる.

キーワード　エキスパートサーチ, 人物情報抽出

An expert search system using CiNii

Kouji MASUDA† and Manabu OHTA††

† Faculty of Engineering, Okayama University
3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama 700–8530 Japan
†† Graduate School of Natural Science and Technology, Okayama University
3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama 700–8530 Japan
E-mail: †masuda@de.cs.okayama-u.ac.jp, ††ohta@de.cs.okayama-u.ac.jp

Abstract　This paper proposes a system to search domain experts who are familiar with inputted technical terms and to generate their profiles automatically. The proposed system searches CiNii for research papers relevant to the inputted technical terms to extract their authors as the domain experts. Then, it searches the Web for information relevant to the experts to generate their profiles using the technical term and the extracted authors' names. The profile consists of an expert's carrer, related people, research topics. With the use of the proposed system, users can easily find domain experts and know their profiles.

Key words　Expert Search, Extraction of Person Information

1. はじめに

近年, 大学・研究機関の研究者と行政, 企業等の産学官連携が盛んである. そこでは, 異なる分野の専門家や研究者同士の共同研究が頻繁に行われている.

しかし, このような連携や協力を始めるには, その専門分野にどのような専門家がいるのかをまず把握する必要がある. そこで本研究では, 学術論文データベース CiNii を利用し, ある専門用語からその分野の専門家と思われる人物の氏名を検索し, その人物のプロファイルを自動生成するシステムを提案する.

提案システムではまず, ユーザが専門用語を入力すると, システムがその専門用語に関連する学術論文を CiNii で検索し, その著者をエキスパートとして抽出しリスト化して表示する. ユーザは表示された人物リストからプロファイルを生成する 1 名を選択する. システムはユーザが選択したエキスパートの氏名と専門用語を用いて Web からも関連情報を抽出し, そのエキスパートの経歴や関連する人物, 研究事例などから構成されるプロファイルを自動生成して表示する.

本稿では 2 章で関連研究, 3 章で提案システムについて説明する. 4 章でプロトタイプシステム, 5 章でシステムの評価実験について述べ, 6 章でまとめる.

2. 関連研究

2.1 人物検索エンジン SPYSEE [1]

SPYSEE [1] は, オーマ株式会社が運営する人物検索サイトである. このサイトでは, 数十万人の人物情報をウェブから自動抽出し, 各人物の人物像や人物間の相関図, 関係のある人物などを表示している. ユーザは任意のクエリで人物を検索するこ

とで各人物のページにたどり着き, その人物情報を得ることができる.

本研究は専門用語から人物を検索するが, SPYSEE は人物名を入力として, その人物の情報を出力するサービスである. この点が提案システムと大きく異なる. また, SPYSEE は同姓同名の人物の識別に関しては, 現在その分類方法の開発を進めている段階である. 提案システムでは, 入力される専門用語を利用して人物情報を獲得するので, この同姓同名問題にもある程度対処できる.

2.2 研究者逆引きデータベース

キーワードを入力し，そのキーワードに関連した研究者を出力するデータベースとして, 橋本ら [2] の提案した研究者逆引きデータベースシステムがある. 彼らの提案したシステムは, 特許と研究者をキーワードにより関連づけ，キーワードによる特許検索結果から関連研究者を出力するものである. 具体的には, 事前に, 研究者の学術論文から研究者の研究を代表するキーワードを抽出し, そのキーワードを多く含む特許を研究者と関連度が高い特許であると見なして, 研究者と特許の関連度を計算しておく. そして, キーワードによる特許検索を行い, 検索結果の上位の特許と関連度が高い研究者をランキングし出力するというものである.

彼らは実際に, 東京工業大学に所属していた 125 名の研究者と 2004 年から 2007 年の公開特許公報との関連度を計算し, 約 20 万件の特許に対して研究者を関連づけたデータベースを構築している.

2.3 Web からのキーワード抽出

人物情報抽出では, 森ら [3] が Web から人物に関するキーワードを抽出する方法を提案した. 森らの手法はまず, 検索エンジンを利用して人物の検索を行い, 検索結果の上位の Web ページを得る. そして, 獲得した Web ページに対して HTML タグの除去および形態素解析を行い, Termex(注1) を用いて用語を抽出する.

次に Web ページから抽出した用語と人物名の共起の強さを Jaccard 係数を用いて求める. すなわち, 人物名 n を含む Web ページ集合を N, 用語 w を含む Web ページ集合を W として, n と w の単独のヒット数をそれぞれ |N|, |W|, n と w の AND 検索のヒット数を $|N \cap W|$ とみなす. このとき, Jaccard 係数 J(n,w) は次のように計算できる.

$$J(n,w) = \frac{|N \cap W|}{|N| + |W| - |N \cap W|} \quad (1)$$

森らはこの Jaccard 係数の大きい用語を人物に関するキーワードとして自動抽出した.

2.4 学術文献データベース CiNii [4]

CiNii [4] は, 国立情報学研究所が運営する学術文献のデータベースで, 正式名称は NII 論文情報ナビゲータである. 2008 年 1 月の時点で, 271 の学協会から許諾を得て, 紙媒体の学協会誌約 1000 タイトル（紀要も含む）に掲載された約 280 万件の論文本文を PDF として蓄積している.

論文を検索する際には, 論文情報全体をキーワードで検索する簡易検索の他, 検索条件を細かく設定できる詳細検索がある. 例えば詳細検索では, 著者名や著者所属などを指定することでより絞り込んだ検索を行うことができる. また, 発表年の新旧や論文の被引用件数などを指定して, 検索結果の表示順序を変更することも可能である.

3. 提案システム

本章では, 提案システムの具体的な処理について述べる. 3.1 節では提案システムの処理の概要を述べ, 3.2 節でエキスパートの氏名の獲得方法について述べる. 続く 3.3 節で Web から人物情報を抽出する方法について述べ, 3.4 節ではプロファイルの生成方法について説明する.

3.1 処理の概要

本システムは, 専門用語を入力として, その分野におけるエキスパートの経歴や関連する人物, 研究事例からなるプロファイルを自動生成して出力するものである. 本システムは, 大きく分けて 3 段階の処理を行う. 図 1 にプロファイルを生成するまでの簡単な処理の流れを示す.

1 段階目はエキスパートの氏名の獲得である. 入力である専門用語を受け取り, CiNii を利用してエキスパートの氏名を抽出する. 詳しくは, 3.2 節で説明する.

ユーザが抽出されたエキスパートのリストから 1 名を選択すると, 2 段階目の処理である人物情報の抽出を行う. ここでは, ユーザが選択したエキスパートの関連情報を Web から抽出して収集する. 詳しくは, 3.3 節で説明する.

3 段階目はプロファイルの生成である. 収集した人物情報を基に CiNii を利用してエキスパートのプロファイルを生成し, 表示する. 詳しい方法は 3.4 節で説明する.

3.2 エキスパートの氏名の獲得

本研究では, 学術論文を発表している人物を, その論文の中で用いられている専門用語に関する分野のエキスパートであると仮定する. この仮定により, 専門用語と人物名を結びつける.

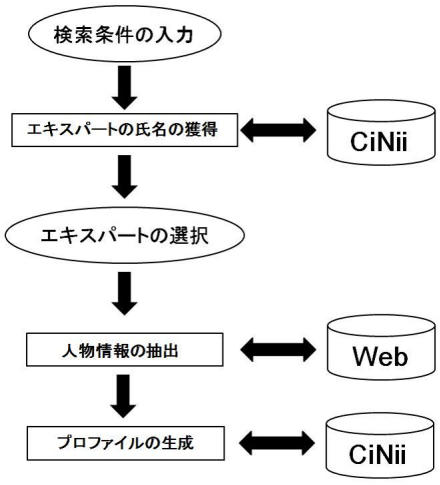

図 1 提案システムの処理の流れ

(注1)：東京大学中川研究室，横浜国立大学森研究室で開発された用語抽出システム, http://gensen.dl.itc.u-tokyo.ac.jp/win.html

まず, システムはユーザから入力として専門用語を受け取る. 次に, その専門用語に関連する論文を学術文献データベース CiNii で検索し, ユーザが指定した件数分の論文から著者情報を得る. そして, 獲得した著者情報を出現頻度に基づいてソートし, 上位 10 名の氏名を表示する.

3.3 人物情報の抽出

本システムでは, 森らの手法 [3] を利用して Web から人物情報を抽出する.

まず, CiNii から獲得したエキスパートの氏名と入力された専門用語を検索キーワードとして Yahoo Web 検索 API [5] による AND 検索を行い, 上位 100 件のサマリを得る.

次に, 獲得したサマリを形態素解析(注2) し, 用語抽出システム Termex による用語の抽出を行った後, Jaccard 係数を計算する.

本研究では, |N| をエキスパートの氏名と入力された専門用語の AND 検索のヒット数, |W| を用語のヒット数とし, $|N \cap W|$ はエキスパートの氏名と入力された専門用語と用語の AND 検索のヒット数とする.

計算した Jaccard 係数の値が大きい用語 w をその人物に関連する用語として抽出する.

3.4 プロファイルの生成

本システムが出力するエキスパートのプロファイルは経歴, 関連人物, 研究事例からなる.

まず, 人物情報として Web から抽出した用語を, 形態素解析の結果と語尾から, 「人物」,「組織」,「学歴」,「役職」,「その他」に分類する. そして分類した用語を利用してプロファイルを生成する.

3.4.1 用語の分類

抽出した用語の中で形態素解析の結果が人名とされた用語は「人物」に分類する. 形態素解析の結果, 先頭が組織とされた用語はその用語の語尾を手掛かりとして「組織」,「学歴」,「役職」のいずれかに分類する. 例えば語尾が “卒業” や “修了” であればその用語は「学歴」に分類し, “教授” や “長” などであれば「役職」に分類する. このとき, 「学歴」や「役職」に分類された用語から, “卒業” や “教授” などの語尾を取り除き新しく用語を作る. このようにして作られた用語は「組織」に分類する. 語尾から「学歴」や「役職」に分類できなかった用語は「組織」に分類する.

抽出した用語のうち,「人物」,「組織」,「学歴」,「役職」のどれにも分類されなかった用語は「その他」に分類する. 「その他」に分類された用語は, 研究事例の生成で用いる論文を検索する際に利用する. 詳しくは 3.4.4 項の研究事例の生成で説明する. このとき, 論文情報から共著者が獲得できればその共著者の氏名を「人物」に分類する. プロファイルの生成を行うエキスパートの所属が獲得できれば, その所属していた組織名を「組織」に分類する.

3.4.2 経歴の生成

経歴の内容は「組織」,「学歴」,「役職」に分類された用語から作成する.

「組織」に分類された用語についてはエキスパートがその組織に所属していた期間を推定する. 具体的には, CiNii でエキスパートの氏名と組織をそれぞれ著者と著者所属として指定し, 発表年が最も新しい論文と最も古い論文を検索する. この 2 つの論文の発表年の間をエキスパートがその組織に所属していた期間と推定する.

また, 経歴の時系列に関連人物を反映させる. そのため, 関連人物についても所属していた組織と所属時の年代を推定している. 関連人物の情報を生成する詳しい方法は 3.4.3 項で説明する. 本研究では, エキスパートの経歴の時系列の中に, この関連人物の情報を埋め込む. 具体的には, 関連人物の所属組織と所属時の年代から関連人物を分類してまとめる. そして, 関連人物の所属時の年代を手掛かりにして分類し, 関連人物をエキスパートが組織に所属していた期間にあてはめる. 関連人物を含む経歴のプロファイルは 4.3 節で説明する.

3.4.3 関連人物の生成

関連人物の内容は「人物」に分類された用語を用いて作成する. このとき, Web から抽出した人物名か, 論文の共著者の氏名かに注目し, その人物とエキスパートの関連度の強さの目安を示す. すなわちその強さを大中小の 3 段階で表す. Web から抽出され, かつ論文の共著者である人物の関連度は大とする. Web からは抽出できなかったが論文の共著者である人物の関連度は中とする. 論文の共著者ではなく Web からしか抽出できなかった人物の関連度は小とする.

各関連人物についても所属する組織と所属時の年代を推定する. 論文の共著者として獲得した人物はその論文から所属組織を得て, 論文の発表年をその組織に所属していた年代とする. Web からしか抽出できなかった人物は次の手順で所属を推定する. まず, CiNii からその人物が著したと思われる論文を検索する. 具体的には, 検索するフリーワードに専門用語を入力し, 著者の項目をその人物の氏名で指定して検索する. 検索した論文からその人物の所属した組織名を獲得できれば, その論文の出版年を所属していた年と推定する. 論文から所属した組織を得られない場合や論文が検索できなかった場合は, エキスパートの経歴として獲得した「組織」の中で最も共起するものを所属する組織と推定する. 具体的には, 所属組織を推定する人物の氏名と「組織」に分類された用語で Yahoo Web 検索 API [5] による AND 検索を行い, 「組織」の用語の中で最も検索結果数が多かった組織を所属していた組織と推定する. 所属していた年代は, その組織にエキスパートが所属していた期間のちょうど真ん中とする. 例えば, ある関連人物の組織がこの共起による推定で岡山大学工学部となり, そこにエキスパートが所属していた期間が 2000 年から 2004 年までだった場合, その関連人物は 2002 年に岡山大学工学部に所属していたと推定する. 「組織」の用語の中に共起する組織が無ければ, 所属不明とする.

このようにして推定した関連人物の所属組織と所属年はエキスパートの経歴の生成に利用する.

3.4.4 研究事例の生成

研究事例はエキスパートの論文から作成する. 「その他」に

(注2): 形態素解析ツールとして茶筅を利用した.

分類する用語は研究分野に関連する用語と判断し, CiNii で論文を検索する. 検索するフリーワードに「その他」に分類された用語を入力し, 著者にエキスパートを指定して検索する. 検索された論文を利用して研究事例の内容を作成する. その論文から共著者を得ることができれば, その共著者を関連人物に追加する. エキスパートの所属に関する情報を得たときは, その所属を経歴の内容を作成するための「組織」の用語に追加する.

3.5 画像の検索

プロファイルに表示するためのエキスパートの画像を Yahoo 画像検索 API [5] を利用して検索する. 具体的には, クエリをエキスパートの人物名として画像を検索し取得する.

4. プロトタイプシステム

本章では実装したプロトタイプシステムについて, その実行画面を順を追いながら説明する.

図 2 はシステムの実行画面の遷移を示している. システムは次の手順で実行される.

(1) システムを起動し画面に従って検索条件を入力する. 詳しくは 4.1 節で説明する.

(2) システムは検索条件に基づいて検索を開始し, 検索結果としてエキスパートのリストを表示する. ユーザはこのリストからプロファイルを閲覧したい人物を 1 名選択する. 詳しくは 4.2 節で説明する.

(3) システムは選択されたエキスパートのプロファイルを自動生成し表示する. ユーザは生成されたプロファイルの経歴, 関連人物, 研究事例の表示を切り替えながら閲覧することができる. このとき, 表示された関連人物を 1 名選択するとその人物のプロファイルを新たに生成し, 閲覧することができる. 詳しくは 4.3 節で説明する.

(4) プロファイルを生成した後, 人物名を変えずに専門用語のみを変更しプロファイルの生成をやり直すことも可能である. 詳しくは 4.4 節で説明する.

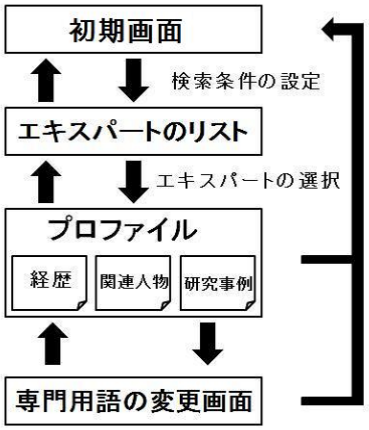

図 2 実行画面の遷移

図 3 初期画面

4.1 システムの起動

プロトタイプシステムを起動すると, 図 3 が初期画面として表示される.

この画面で検索条件は次のように設定する. テキストボックスに検索したい専門用語などを入力する. 「着目点」では検索する論文を発表年の新旧, または被引用件数のいずれの観点でランク付けするかを選択する. 「検索論文数」は検索する論文の数であり, 200 件, 600 件, 1000 件の中から選択する.

このようにして検索条件を設定した後,「検索」ボタンを押すことで, 検索を開始する.

4.2 エキスパートのリスト

システムは検索結果をエキスパートのリストとして表示する. 検索結果の例を図 4 に示す. この図 4 は入力語を “CP 対称性の破れ”,「着目点」を “新”,「検索論文数」を “200 件” という条件で検索した結果である. この実行例では “CP 対称性の破れ” という用語に関する論文が 39 件検索できたのでこれらの論文からエキスパートの情報を獲得する.

図 4 では, 著者情報を検索された論文数に基づいてソートし, 上位 10 名を表示している. この実行例では, 木村恵一氏が論文数 7, 高村明氏が 7, 横枕英和氏が 5 となっている.

ユーザはこの中から興味を持った 1 名を選択すると, その人物のプロファイルを閲覧できる.

システムはユーザが選択した人物の情報の収集を開始し, 収集した情報を基にその人物の「経歴」や「関連人物」, 「研究事例」からなるプロファイルを生成する (図 5, 図 6, 図 7). 図 5, 図 6, 図 7 は益川敏英氏について生成されたプロファイルで, それぞれ「経歴」,「関連人物」,「研究事例」を表している.

4.3 プロファイルの閲覧

ユーザは「経歴」,「関連人物」,「研究事例」の各ページの画面を切替ながらプロファイルを閲覧できる. 経歴のページの見方と操作については 4.3.1 項で, 関連人物のページについては 4.3.2 項で, 研究事例のページの見方については 4.3.3 項で説明する.

4.3.1 経　　歴

図 5 のように, ユーザが生成されたプロファイル画面の上部にある「経歴」の項目を選択すると, 選択した人物の経歴に関する生成結果が表示される.

生成したプロファイルの初期画面として, 経歴の生成結果が表示される.

図 4　検 索 結 果

経歴の画面の上部には「学歴」が表示される．学歴が抽出されなかった場合，この項目は表示されない．図 5 から，益川敏英氏の学歴が “名古屋大学大学院理学研究科博士課程修了” や “向陽高等学校卒業” などであることが分かる．

その下の「職歴」については，論文の発表年から推定した組織に所属していた年に基づいて，所属した組織を年表形式で表示している．例えば，この図 5 からは益川敏英氏が 1971 年から 1976 年まで “京大理” に所属していたことが分かる．また，2006 年以降は “京都産業大学理学部” に所属しており，2010 年からは “名古屋大学” に所属していることが分かる．

「役職」に分類された用語の内，「職歴」の組織と関連するものはその組織の下に「(役職)」として列挙する．表示された組織と関連を持たない役職があれば，「その他の役職」として「職歴」の年表の下に列挙する．図 5 からは，益川敏英氏が京都産業大学理学部に 2006 年以降所属しており，そこで “京都産業大学理学部教授” という役職に就いていることを読み取ることができる．また，「その他の役職」から益川敏英氏は “日本学術会議会員” を務めていたことも分かる．

また，年表中の「[関連人物]」という項目を選択するとその組織に所属していた当時の関連する人物を表示する．関連人物は組織ごとに分類して表示され，また各人物名の右側の◎や○の色でエキスパートとの関連度の大きさを表す．◎は関連度が大で Web から抽出され，かつ論文の共著者である人物を示す．○は関連度が中で Web からは抽出できなかったが論文の共著者である人物である．○は関連度が小で論文の共著者ではなく Web からしか抽出できなかった人物であることを意味している．例えば，2006 年から 2010 年にかけて “京都産業大学理学部” に所属していた当時の「関連人物」として “名古屋大学” に所属していた小林誠氏や “京都産業大学理学部” に所属していた

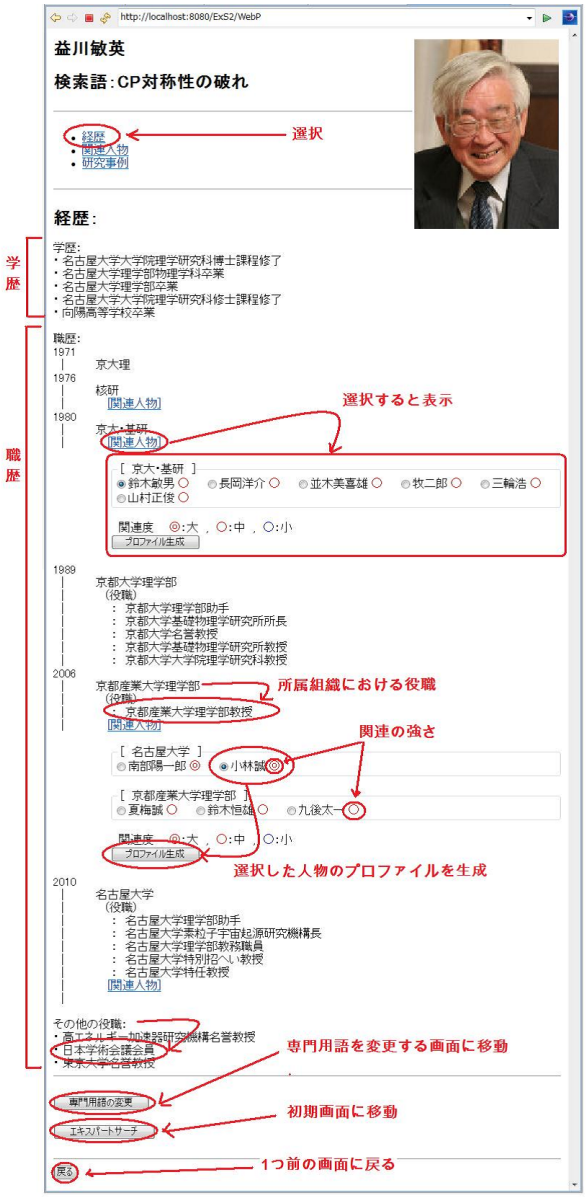

図 5　プロファイル (経歴)

九後太一氏らが表示されている．ユーザは益川敏英氏が “京都産業大学理学部” に所属していた当時，小林誠氏や九後太一氏らと関わりがあったことを知ることができる．また，小林誠氏の関連度は◎であることから，小林誠氏は益川敏英氏と関係の深い人物だと読み取ることができる．

ユーザは関連人物として表示された人物から 1 名を選択してその人物のプロファイルを生成して閲覧することもできる．例えば，“京都産業大学理学部” 所属当時の「関連人物」の小林誠氏を選択して，小林誠氏のプロファイルを閲覧することができる．

プロファイルの画面の下にある「戻る」で 1 つ前の画面に戻り，「エキスパートサーチ」で最初の検索画面に移動する．「専門用語の変更」を選択すると，移動先で専門用語を変更してプロファイルの生成をやり直すことができる．詳しくは 4.4 節で説明する．

図 6　プロファイル (関連人物)

4.3.2　関 連 人 物

図 5 で表示したプロファイルの上部の「関連人物」という項目をユーザが選択すると画面は図 6 のような関連人物のリストに切替わる．

図 6 は，Web から抽出した人物名と論文の共著者の氏名を表示している．これらの人物名を関連の強さから ◎，○，○ の順に表示している．各人物名の下に関連度と所属を表示する．さらに，◎，○ に該当する人物は共著の論文を 1 件表示する．人物名の下にある「論文」を選択すると共著の論文のタイトルとその書誌情報を表示する．

「経歴」のページに表示した「関連人物」のリストと同様に，ユーザは表示した関連人物のリストから 1 名を選択することで，その人物のプロファイルを閲覧することができる．

4.3.3　研 究 事 例

図 7 のように，プロファイルの上部の「研究事例」の項目を選択するとエキスパートの研究事例集に切替わる．

ここではエキスパートが執筆した論文の情報を研究事例として表示する．

画面上部に「研究関連ワード 一覧」を表示し，その下に各「研究関連ワード」に関連する論文のタイトルと共著者，書誌情報などを表示している．

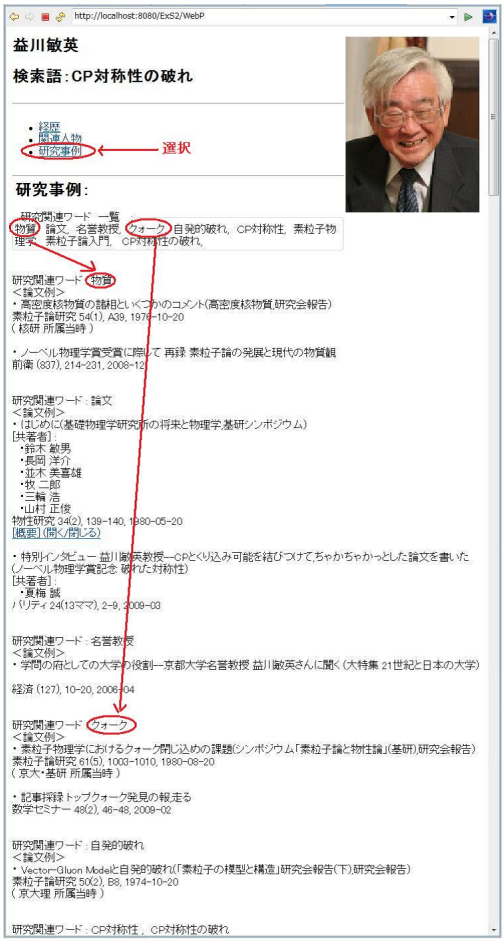

図 7　プロファイル (研究事例)

図 8　専門用語の変更

例えば，“物質” という研究関連ワードに関連する “高密度核物質の諸相といくつかのコメント (高密度核物質研究会報告)” という論文があることがわかる．また，その論文の発表された年に益川敏英氏は “核研” に所属していたことなどもここから読み取ることができる．

4.4　専門用語の変更

図 5 などのプロファイルの表示画面の下にある「専門用語の変更」というボタンを押すと，画面は図 8 のようになる．

ここでは，人物名を変更せずに検索を開始する際に入力した専門用語のみを変更して，プロファイルの生成をやり直すこと

ができる. 例えば, 図 5, 図 6, 図 7 のプロファイルは, “CP 対称性の破れ” という専門用語に結び付けられた益川敏英氏のプロファイルだが, 専門用語の “CP 対称性の破れ” だけを変更してプロファイルを再生成できる.

例えば, 専門用語の “CP 対称性の破れ” を “超弦理論” と変更してプロファイルを再生成することで, “超弦理論” の研究に携わった当時の関連人物や研究内容を知ることができる.

5. 評価実験

本章では, 提案したエキスパートサーチシステムに対する評価実験について述べる. 5.1 節では, 岡山大学工学部情報工学科に所属する教員のプロファイル生成結果について評価する. 5.2 節では, 提案システムで生成したプロファイルを SPYSEE [1] の検索結果と比較して定性的に分析する.

5.1 表示内容の精度

2011 年 1 月の時点で岡山大学工学部情報工学科に所属する助教以上の教員 19 名についてプロファイルを生成し, 「経歴」の表示内容である「学歴」,「組織」,「役職」の再現率と適合率, F 値を算出した. 正解データは Web などを通じて独自に収集した. なお, 関連人物の表示内容については関連の範囲が曖昧なので本実験では評価しない. また, 研究事例も正解データが膨大で再現率の算出が困難であるので評価しない. 研究事例の表示内容はエキスパートの氏名と専門用語から検索した論文に基づいて作成したもので, 同姓同名の人物による誤りが見られた 1 名を除いて全て正しく, 結果として適合率は高い値となった. 同姓同名の人物による誤りについては 5.1.2 項で詳しく説明する.

正誤の判断をする際に次の点に注意した. まず, 所属年は評価の対象にとせず, 組織の名前の正誤のみを判断する. また, “九州大学大学院システム情報科学研究科” と “九州大学大学院システム情報科学研究院” のような語尾の僅かな違いは区別せずに同じものとみなした. 同様に “岡山大学大学院自然科学研究科” と “岡山大学自然科学研究科” のような例も区別せず, “准教授” と “助教授” や “助教” と “助手” も同じものとみなした.

エキスパートのプロファイルの「組織」,「学歴」,「役職」について再現率, 適合率, F 値を算出した結果を表 1 に示す.

表 1　プロファイル要素の検索効率

	再現率	適合率	F 値
組織	0.568	0.902	0.697
学歴	0.222	0.909	0.357
役職	0.472	0.833	0.602
計	0.596	0.871	0.598

表 2　誤表示の原因の内訳

	計 (個)
形態素解析	1
CiNii	1
サマリ	7
同姓同名	2

5.1.1 プロファイル要素の検索効率

表 1 では再現率の低さが目立つ. 特に学歴と役職の再現率は著しく低い. これには次のような原因があった. まず学歴や役職は, 同大学卒業や同大学助手のように省略した形で Web 上に記されることある点が挙げられる. 本システムではこのような表記には対応しておらず, 同大学卒業や同大学助手のような用語は “同大学” という名称の大学の学歴や役職として抽出してしまう. 2 つ目は検索の際に入力した専門用語と大学卒業時の研究分野が一致しないことがある点である. この場合, 卒業した大学と専門用語の関連が弱く, ○○大学卒業という学歴に分類される用語の Jaccard 係数は低い値となり, 正しい学歴なのに抽出されない場合がある. 3 つ目は Web ページ上の学歴や役職に関する情報がサマリに表示されないことがある点である. 本システムはこのサマリから用語の抽出を行うため, Web ページに学歴や役職が載っていても, 検索結果のサマリに含まれなければ抽出できない. 4 つ目は学歴の記述された方によっては学歴の用語が抽出できないことである. 例えば, “1980 年に岡山大学を卒業” というように記述されている場合, これを形態素解析し Termex を実行すると “卒業” と “岡山大学” の 2 つの用語が出力される. この場合, “岡山大学卒業” という 1 つの用語は獲得できず学歴として表示できない. これらの理由により学歴と役職の再現率は低い値となったと考える. 特に 3 つ目や 4 つ目に挙げた原因の解決は喫緊の課題と考えている.

また, 役職や学歴が抽出できなければ, これらの用語から組織を得ることができず, 組織の再現率を下げる原因ともなる. 他にも, 在学時の組織を抽出できても在学時の論文が CiNii で得られない場合, 論文の出版年から在籍した期間を推定できない. その結果, 在学時の学校名が表示できず組織の再現率を下げることがある.

一方, 適合率は組織, 役職, 学歴のどの項目も比較的高いことが分かる. ユーザに正しい情報を提示するという観点からみると, 森ら [3] の手法を基に本研究で提案した人物情報の抽出手法と CiNii を利用したプロファイル生成は適切だといえる.

5.1.2 誤表示の原因

本実験で生成した 19 名のプロファイルの中には合計で 11 個の誤りが含まれていた.

そこで, 誤表示の原因を分類したものを表 2 に示す. なお, 表 2 では組織, 学歴, 役職の誤表示をまとめて分類している.

原因の 1 つ目は形態素解析の問題である. 形態素解析の結果, 本来は単なる専門用語が「組織」や「人名」に分類されることがある. これにより, 「経歴」のページに専門用語が誤って表示される. 例えば, ある人物のプロファイルの「その他の役職」に “オイラー回帰長” という専門用語が表示された. この “オイラー回帰長” という語を形態素解析すると “オイラー” の部分が組織名と分類される. さらに語尾の “長” という文字から「役職」に分類し, プロファイルの「経歴」に役職として表示された.

2 つ目の原因は CiNii の論文情報の誤りである. CiNii の論文情報で著者の所属する組織の名称に誤りがあり, それが誤ったまま表示されることがあった. 具体的には, ある人物の「職歴」に “岡山大学工学部情報科学科” と表示された. 正しくは “岡山大学工学部情報工学科” であるが, CiNii で検索した論文に “岡山大学工学部情報科学科” と誤った名称が登録されていたため, 本システムが誤った名称のまま抽出し表示した.

3 つ目はサマリから用語を抽出する際の誤りである. Web 検索結果のサマリから用語を抽出する際, エキスパート以外の人物が所属する組織などを抽出することがある. その際, Jaccard

係数による判別で明らかな誤りは除去できるが, エキスパートと関連が強い人物の情報であった場合は除去できないことがある. 例えば, 図 5 の益川敏英氏のプロファイルの「その他の役職」に “高エネルギー加速器研究機構名誉教授” とあるがこの表示は誤りで, これは小林誠氏の役職である. 益川敏英氏と小林誠氏は関連が強いためシステムが誤抽出した.

4 つ目は同姓同名による誤りである. 本実験ではプロファイルを生成する際に専門分野の用語を利用することで, 同姓同名の問題は 19 名中 18 名については発生しなかった. しかし, 1 名のプロファイルに同姓同名の人物の情報が混在していた. これは, ある企業に所属する同姓同名の人物が類似した分野の研究をしており, この人物の所属した 2 つの組織を情報工学科の教員の組織として誤って抽出した結果である. このように, 専門分野の近い同姓同名の人物の識別は本システムでも解決できていない.

5.2 SPYSEE との比較

人物情報を出力する検索エンジンとして, 人物検索サイト SPYSEE [1] がある. このサイトは, 人物名を入力とする点が本システムと異なるが, プロファイルを表示するという点は類似している. そこで, SPYSEE と本システムを比較して, 本システムの有効性を分析した.

まず, SPYSEE では著名な人物であれば経歴を表示するページも存在するが, 大半の人物は人物像という項目に文章で経歴が記述されている (図 9). しかし, 本システムは経歴を年表形式で表示しておりユーザが理解し易いと思われる.

次に, SPYSEE は事前に作成している人物のプロファイルを表示するため, 検索時間が短いという利点がある一方で, 作成されていない人物の情報は表示できないという問題がある. 本システムはユーザが人物を選択してから情報抽出を開始するので, プロファイルが生成されるまで待たねばならないが, 人物の情報が表示されない可能性は低い. 実際に SPYSEE で, 5.1 節の評価実験の対象とした岡山大学工学部情報工学科に所属する教員 19 名を検索したところ, 表示内容の正誤は別にして, 結果が表示された人物は 14 名であった. これに対して, 本システムはプロファイルの生成自体は 19 名全員について成功している. ただし, SPYSEE はエキスパート専用の検索エンジンではなく芸能人なども検索できるシステムであるため, 単純には比較できない.

また, SPYSEE には同姓同名の問題がある. 本システムでは, 専門用語を人物情報を抽出する際に利用することで, 同姓同名による誤表示をある程度減らすことができる. 実際に SPYSEE では岡山大学工学部情報工学科の教員で検索結果が返された 14 名のうち, 人物像の記述が適当だと思われる人物は 10 名であった. 一方, 本システムは 5.1.2 項で述べたように同姓同名でかつ専門分野も類似する 1 名を除き, プロファイルの生成に成功した 19 名の中の 18 名には同姓同名の人物の情報が混在しなかった.

一方, SPYSEE の優れた点は相関図という項目で関連する人物のネットワークを表示していることである. これにより関連人物と別の関連人物の繋がりなどをユーザに分かり易く表示している. また, 各関連人物の写真や簡単な情報も表示しており機能的で, これらは本研究でも参考にすべき点だといえる.

図 9 SPYSEE の動作例

6. ま と め

本稿では, 専門用語を入力するとその分野に詳しいエキスパートを検索して, その人物のプロファイルを自動生成するシステムを提案した. さらに提案システムのプロトタイプを実装し, 専門用語を入力してその分野のエキスパートの経歴や関連人物, 研究事例からなるプロファイルを自動生成した. 評価実験では, 経歴のプロファイルにおける表示内容の再現率と適合率, F 値を算出したほか, 誤表示の原因を分析した. また, 生成したエキスパートのプロファイルを SPYSEE の検索結果と比較して, 本システムの有効性について考察した.

今後の課題としては, エキスパートの氏名を獲得する際に, CiNii のみならず他のデータベースなど多くの情報源を活用することが挙げられる. 提案システムでは, 経歴の年表は論文情報から取得するので, 組織に所属していた時期の論文が CiNii に存在しない場合, プロファイルに反映されない. よって, 経歴の年表を補完する上でも CiNii 以外の情報源は有用と考えられる.

文 献

[1] あのひと検索 スパイシー SPYSEE
http://spysee.jp/
[2] 橋本泰一, 乾孝司, 内海和夫, 石川正道, “研究者逆引きデータベースシステムの構築”, 人工知能学会全国大会, 2009.
[3] 森純一郎, 松尾豊, 石塚満, “Web からの人物に関するキーワード抽出”, 人工知能学会論文誌 , AI 20, 337-345, 2005-11-01.
[4] CiNii NII 論文情報ナビゲータ,
http://ci.nii.ac.jp/
[5] Yahoo!デベロッパーネットワーク,
http://developer.yahoo.co.jp/