

Title	ワールドワイドウェブにおける人物検索の実現
Author(s)	山本, あゆみ
Citation	
Issue Date	2000-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1322
Rights	
Description	Supervisor:佐藤 理史, 情報科学研究科, 修士

ワールドワイドウェブにおける人物検索の実現

山本 あゆみ

北陸先端科学技術大学院大学 情報科学研究科

2000 年 2 月 15 日

キーワード： ワールドワイドウェブ, 人物情報の自動抽出, 表解析, 情報抽出, 検索エンジン

ワールドワイドウェブ上には, 様々な情報が膨大に存在する. このような情報源から求める情報を探し出すために, ロボット型検索エンジンと呼ばれる汎用の検索システムが開発されてきた. この検索システムは, 「ロボット」または「スパイダー」と呼ばれるプログラムによって, 大量のウェブページを自動収集し, それらページに対するキーワード検索を提供する. そのため, 多くの場合, ある検索質問（クエリ）に対して, 大量の検索結果（URL）が得られ, それらの大部分が, 求める情報とはあまり関係がないものであることが多い. したがって, 利用者が効率良く情報収集するのは, 困難である.

この問題を解決する 1 つの方法として, 検索対象のカテゴリを限定した検索システムが考えられる. 検索対象のカテゴリを限定することにより, 異なるカテゴリを排除することが可能となるとともに, 主要な情報がそのカテゴリによって定まるため, 整理された情報を利用者に提供することが可能となる.

本研究では, 検索対象を人物とし, 人名からその人物の主要情報を提供する人物検索システムを実現した. 本システムにより, 利用者は, 効率良く人物情報を収集することができる. 本システムは, ユーザインタフェース, 人物情報データベース, 次に示す 2 つの人物情報収集方法を実装したモジュールから構成される.

(1) 職業別人名録を利用して情報収集する方法

ウェブ上には、ある職業に属する人物の主要情報をまとめて整理した職業別人名録が多数存在し、その多くは、表形式で記述されている。このような職業別人名録は、人名がわからなくても、職名がわかれば、収集可能である。本研究では、職名からその職業の人名録を見つけ出し表解析を適用することで、効率的に多くの人物に関する主要情報を収集する方法を提案する。まず、職名を入力とし、その職業の人名録があるページを検索エンジンとリンク情報を利用して収集する。次に、収集したページのうち表形式の人名録が存在するページから属性-属性値で表現された人物情報を抽出する。この方法は、オフラインで実行できるため、あらかじめ収集した情報をデータベース化し、検索時に、このデータベースを参照して利用者に情報提供するアプローチをとることができる。

(2) 人名からその人物の主要情報を収集する方法

人名録を利用する方法は、効率良く人物情報を収集することができるが、人名録に掲載されている人物の情報しか収集できないという問題がある。この問題を解決するために、人名からその人物のテキストで記述された主要情報を収集する方法を提案する。テキスト形式の人物情報は、人名の見出しと、その人物を紹介したテキストからなることが多く、そのレイアウトにいくつかのパターンが見られる。このレイアウトパターンを利用して情報抽出を行う。まず、人名からその人物の主要情報がテキストで記述されているページを検索エンジンを利用して収集する。次に、収集したページに対して、空行、インデントなどのレイアウト情報を示すタグを挿入し、これらのタグを利用して、人名の見出しを見つける。最後に、この人名の見出しの本文となる領域を判定し、この本文に「生まれ」など人物紹介でよく使用されるキーワードが存在した場合、人物情報として抽出する。この方法による収集は、人名が与えられないと実行できないため、オンラインで実行する。

本システムの有効性を検討するため、これら2つの方法それぞれに対して実験を行った。まず、(1)に示した方法を用いて、次の2つの実験を行った。第1の実験では、職名を「政治家」とし、国会議員、および、地方議員、の情報収集を行う。第2の実験では、職名を「著述家」とし、文学関係の受賞作品をもつ人物の情報収集を行う。結果を以下に示す。

- ・「政治家」の場合は、2245 人の情報を収集した。ただし、「著述家」の場合は、833 人しか情報収集しなかった。
- ・「政治家」の場合は、すべての人物に対して適切な属性-属性値対応の情報を収集した。ただし、「著述家」の場合は、68 人に対して、1 つだけ不適切な対応の情報が存在していた。

これらの結果より、人物情報の内容がよいものを収集できることが明らかになった。ただし、職業によって、あまり多くの人物を収集できない。より多くの情報を収集するために、職業別人名録の収集に改良が必要である。

次に、(2)に示した方法を用いて、参議院議員 50 名と、著述家 128 名、の名前を入力として実験を行った。結果は、以下の通りである。

- ・実験に用いた 178 名のうち 161 名 (90%) から有用な情報を収集していた。
- ・収集した人物情報は、全部で 691 件あり、そのうち正解は 549 件 (79%) であった。

これらの結果より、ウェブ上にプロフィールが存在する人物の名前が入力された場合、有益な情報を収集できることが明らかになった。ただし、参議院議員のように、ウェブ上にプロフィールが多く存在しない人物に対しては、有用な情報源が数 URL しかないという状況になることが多い。より多くの情報を提供するためには、ウェブ上にすでに存在しているプロフィールを抽出するだけでは不十分である。