

環境情報ポータル構築とその情報収集方法に関する研究

名古屋大学 学生員 ○石山 隆弘
名古屋大学 フェロー会員 伊藤 義人

1. 目的

インターネットは最新の情報技術とサービスを吸引して、今もっとも注目を集めているメディアである。しかしその情報量は膨大で 2002 年 5 月現在での総ネットワーク資源は 30PetaByte にもなる。この膨大な情報の中には教育や学習などの学術目的に利用できるものが含まれるが、あまりの情報の多さにその所在が不明確になってしまっている問題がある。本研究ではこの問題を解決するために、ある特定の分野の情報を収集するための枠組みとして情報収集を半自動化した新しいタイプのポータルサイトの構築を行う。この枠組みのケーススタディとして、情報収集する対象を環境情報に絞り、ポータルサイトの情報収集精度と利便性の評価を行う。

2. ポータルサイトの設計と実装

本システムによる情報収集は以下の手順を踏んで行われる。1) Web ページの取得、2) 取得した Web ページが環境情報であるか否かの判断、3) 環境情報と判断されたページの全文情報の蓄積。この 3 つのステップは全て自動化されているため情報収集に費やす労力は極小です。1) については図 1 に示すように、ページ内のリンクを次々と辿る自動収集ロボットを開発した。このスタートページは環境情報を含むものを事前に集めておき、リンクを辿る途中で環境情報以外のページが含まれた場合は、そこで環境情報との関連性が断たれたものと考え以後の収集を中止する。ここで や が環境情報であるか否かを機械的に判断するために、訓練データ（ある特定のジャンルについてのみ記述された良質な文章から成る単語の出現パターンデータ）と Web ページの単語出現パターンを比較する手法を採用した。訓練データのパターンと Web ページのパターンが酷似すれば環境情報を含んでいると判断した。訓練データには環境大事典（長瀬産業株式会社）と土木用語辞典（土木学会）を利用した。この判定精度は全体の 70-80% の Web ページに対して正当な結果が得られるものができた。

ポータルシステムをより便利なものにするためにポータルのパーソナライズ機能を準備した。パーソナライズすることにより、個人の細かいニーズに対応し、より良質な情報を提供することが可能になる。具体的な内容としては、個人の興味に対応したニュースの生成（ユーザ登録時にキーワードを入力することで興味対象を把握するものとした）、類似ページ収集（任意の Web ページや文章を与えることで、それらと類似する内容を含むページを収集する）などがある。今回のシステム

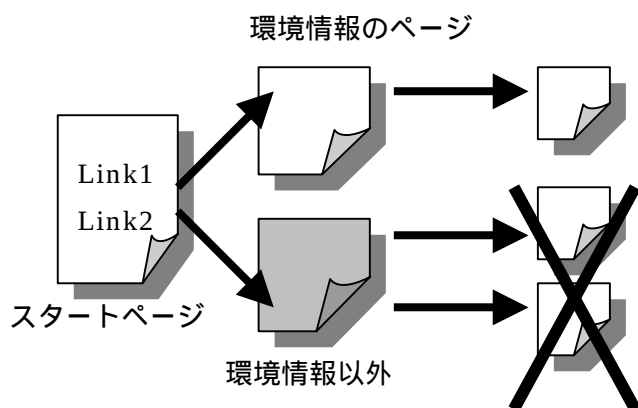


図 1 自動収集ロボット

はほぼ全ての処理を自動化しているため、情報の質に問題が残る。それを改善するために、システムが自動収集した結果に対してユーザが役に立つページであったかを評価し、その評価を蓄積して学習していく機能を導入した。この機能の導入によりシステムが多くの人に使い込まれることにより、より良質な情報が蓄積されていくことが期待される。

また、これら基本システムのプログラムソースは全て公開されており、誰にでも自由に手を加えることができるため、この基本システムをベースにして環境情報以外の情報収集に利用し、自由に改良を加えて

キーワード 環境情報，ポータル，情報検索，自然言語処理

連絡先 〒160-0004 名古屋市千種区不老町 名古屋大学工学研究科土木工学専攻

T E L 052-789-3733

Linux のような発展を遂げることも期待できる．公開した基本システムはフリーで利用できる PHP(Hypertext Preprocessor)と MySQL で構築されており，サーバさえ準備すればノーコストで導入することが可能である．

3．ポータルサイトのサービス

本研究で構築したポータルサイト（Top ページを図 2 に示す）では以下のようなサービスを提供するものとした．(1)Topics の自動生成，(2)カテゴリ別情報提供，(3)メタ情報検索，(4)スタートページの追加，(5)類似ページ収集リクエスト，(6)収集結果の閲覧と評価，(7)収集事例の検索．これらの機能に対してユーザの使い勝手を考えたインターフェースや補助機能（膨大な量のページが収集されたときに，その全てを見ずともいくつかの必要なページをチェックすることで，チェックを入れたページと類似するページのみに絞り込む機能など）を追加した．なお(5)-(7)の機能を利用するには，より良質な情報を効果的に提供するために，事前にユーザ登録が必要となっている．



図 2 ポータルサイト Top ページ

4．利用を通しての実証実験

訓練データに環境大事典と土木用語辞典を利用して実際の利用に耐えうる情報が収集できるかを調査した．環境大事典利用し，任意の環境情報を含んだページと類似するページを収集させる実験においては，大局的な環境情報を含むページが多く集まってしまう，より具体性のある情報を含んだページ（地球温暖化など）の情報も含まれるものの，全体からみるとそれは数%にしかならないことがわかった．しかし，本当に必要とされていたページは収集結果の中からさらに絞り込みを行うことでうまく抽出できる．この他に，論文を与えて類似するものを収集させたところ，直接論文内容に繋がるものは少なかったが，これから論文に関連する情報を収集するための足がかり的なものを収集することができた．

土木用語辞典を利用して同様の試みを行ったが，環境大事典に比べ内容判定が厳しく収集数が少ない（環境情報でないと判定されるページが多い）結果となった．そのため，収集対象を環境情報から土木情報に変更し，訓練データに沿った内容のものを対象としたが，やはり収集数が少なく内容判定アルゴリズムの見直しが必要であることがわかった．しかし，収集された少数のページはどれも良質かつ，自動的にカテゴライズされたカテゴリが正当なものであるものが多く，カテゴリ分けの精度面は問題ない．つまり環境情報（もしくは土木情報）であるか否かの判定ラインを緩くすることによって良い結果が得られる．

5．結論

本研究では，学術情報を効率良く収集するための枠組みとして，新しい運営形式を取るポータルサイトの考案を行い，そのケーススタディとして環境情報ポータルサイトの構築を試みた．その結果，以下のような結論が得られた．

- (1)環境情報ポータルサイトの構築に必要な情報収集の半自動化を行い，従来より少ない労力で学術情報の収集を行うシステムの基本システム部分を製作，公開した．
- (2)環境大事典と土木用語大辞典で構成される訓練データとのマッチング処理による内容判定方法を導入し，学術的な情報のみを網羅的に収集するための枠組みを構築した．

今後の課題として，訓練データの構成や文章内容判断手法の見直しによる類似性調査の精度向上や，ポータルで提供されるサービスのさらなる充実を行っていく必要がある．

参考文献

- ・ Peter Lyman, Hal R. Varian(2000)：How much information.
- ・ 塚本浩司，颯々野学(2001)：AdaBoost と能動学習を用いたテキスト分類，自然言語処理 No.146，pp13～20．