2Q-2

# ブラウザの履歴を用いて個人の嗜好を反映させる自動情報提供システム

細野 哲史 † 小高 知宏 † 黒岩 丈介 † 白井 治彦 ††
† 福井大学大学院工学研究科 †† 福井大学工学部

## 1 はじめに

インターネットの普及に伴い, 一般の多くの人からの情報発信が盛んになった [1]. 利用者は興味のある新しい情報などを入手するために様々なジャンルのサイトにアクセスして情報収集を行う.

本研究では自動で情報収集を行い, 利用者に対して必要であると思われる情報のみを提供するシステムの構築をすることを目的した. このシステムを自動情報提供システムとする. 自動情報提供システムとは利用者の嗜好などから情報を収集し提供するシステムである. このシステムによって利用者は効率的な情報収集が行えると考えられる.

## 2 システムの設計

### 2.1 システムの全体の構成

自動情報提供システムは図 1 のように Web 上から利用者に興味あると考えられる記事情報をのみを提供する. 図 1 では Web 上に A,B,C,D という情報があり, 利用者にとって重要な情報が A,B のみであったとする. この場合, ブラウザなどで情報収集を行うと, 右の利用者のように誤って A,B 以外に必要ない C の情報を閲覧してしまう可能性がある. 本研究では A,B のみの情報を提供しようとするシステムの作成を行う.

Web 上の情報が利用者に興味あるものかを弁別するためにブラウザの履歴情報を扱う. 履歴情報には閲覧したページの情報や, 検索に使用したキーワードなどの情報が含まれている. この中からよく行くページ, またキーワードなどの情報から利用者の趣味・嗜好を分析する. システムはその分析結果から重要と考えられる記事のみを利用者に提供する.

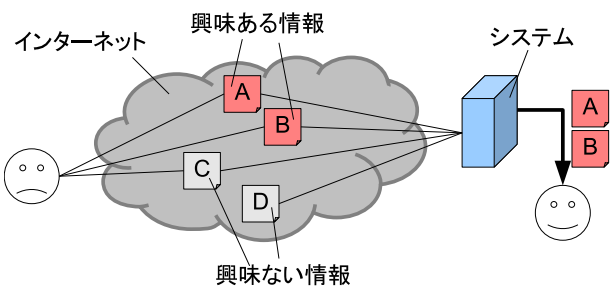


図 1: 一般的な情報収集方法との比較

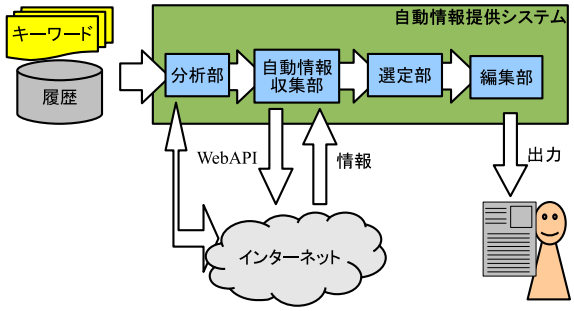


図 2: 自動情報提供システム

このシステムではブラウザなどで見ることのできる Web ベースのデータを扱う. 対象とする種類のデータとして, ニュースサイトの情報を扱う.

このような自動情報提供システムを構築するために図 2 のようなシステムを考える. このシステムには必要であると考えられる以下の機能を実装したものである.

- 個人の趣味嗜好を分析する機能
- 自動で情報を収集する機能
- 利用者に提供する情報を選定する機能
- 情報を閲覧しやすくレイアウトする機能

システムの全体の流れとしては, 最初に個人の嗜好を分析したデータを得る. その次に分析したデータから関連する情報をインターネット上から取得する. 次に, 取得した多くのデータの中から, 提供すべき情報を選定する. 最後に選定された情報を利用者にとって閲覧しやすい状態に出力する. 以下図 2 の各処理部について説明する.

**分析部** : 利用者がどのような分野に興味を持っているのかを調べる. 履歴やお気に入りなどの個人情報から興味ある分野を調べる.

**自動情報収集部** : 分析部の結果から探索する URL を決める. その URL に対し, 情報を URL, 文章などに分類して保存する.

**選定部** : 利用者にとって自動情報収集部で入手した情報が必要なものかどうか情報を取捨選択する. 情報に含まれる単語やタグによって識別を行う. 一定数まで情報を絞れた時点で処理を終了する.

**編集部**： 選定された情報のうち, 利用者にとっての情報の価値を付ける. 優先順位によって記事の順番や場所を決めてまとめる.

## 3 自動情報収集部の実装

自動情報収集部を開発するに当たり, 開発言語として C♯ を用いた. これは C♯ がネットワーク系のライブラリが充実しているからである. また, 開発環境として VisualStudio を用いた.

HTML の分析では HTMLParser[2] を使用する.HTML Parser は HTML 内の URL,IMG を抽出することができる. 今回は公開されている HTMLParser を改良し, 本文などの文章も抽出できるように改良した物を用いた.

## 4 記事の抽出方法

### 4.1 対象とするデータ

自動情報収集部において動作対象とするサイトを指定する. 今回は 3 つのサイトを対象とした. それぞれのサイトをサイト A, サイト B, サイト C とする.

サイト A は幅広いジャンルを扱うニュースサイト, サイト B はゲーム情報などのニュースサイト, サイト C は国際ニュースなどを扱うサイトである. これらのサイトの共通点として, サイトのトップがタイトルと画像のレイアウトで構成されていることが挙げられる.

### 4.2 ノイズ除去の対策

HTML によって URL,IMG, 本文を抽出したデータからノイズと思われるデータを除去する. このとき, ノイズの対象とするのは IMG の種類, および本文の内容,URL の内容である.

IMG の種類については jpeg 画像,png 画像,gif 画像のみを対象とする. 次に本文の内容については短すぎるもの, 年月日を用いたものを除去する.URL の内容については, ブラックリストを作成して通販サイトへのリンクなどを除去する.

### 4.3 記事の抽出

除去した後のデータを見たとき,URL,IMG, 本文が連続した順序不同のパターンとして出現することがある. これを記事データに必要なデータとして保存する. このとき, 本文に相当するデータは記事におけるタイトルに相当し,URL は記事の詳細へのリンクを表す.

## 5 自動情報収集部の実験

先に示した手法が記事の抽出方法として有効かどうかを調べるために検査を行う必要がある. 記事データを URL,IMG, 本文のセットで 1 つのデータとする. 検査データとしてサイト A, サイト B, サイト C を用いる. 評価に際し, 包括度と正確度という 2 つの評価指標を用いる. 両方の指標の定義は以下のものとする.

**包括度**： 抽出すべき記事数に対して, どの程度記事を抽出できているかを表す指標である. 数値が高いほど情報の抜け漏れ度合いが少ないことを表す.

**正確度**： 総データ数に対してどの程度正確に抽出できているかを表す指標である. 数値が高いほどノイズの影響の度合いが低いことを表す.

| サイト | 包括度 | 正確度 |
|---|---|---|
| サイト A | 100%(30/30) | 93.75%(30/32) |
| サイト B | 100%(18/18) | 81.81%(18/22) |
| サイト C | 100%(12/12) | 31.57%(12/38) |

表 1: 自動情報収集部の評価

実験を行った結果を表 1 に示す. 表 1 の括弧内の数字は包括度の場合, 取得した正しい記事データ数/取得すべき記事データ数を表し, 正確度の場合, 取得した正しい記事データ数/取得した全記事データ数を表す.

## 6 考察とまとめ

表 1 より, サイト A,B,C において包括度が 100 %となっている. このことから, 自動情報収集部で用いた手法で記事を抽出できることがわかった.

また, 正確度がサイトによって約 30%と低くなっている. このことから, サイトによってはノイズの影響が顕著に現れることがわかった. 対策としてノイズの影響が疑われるデータの検出機能などが必要であり, サイトごとやパターンによって対応する必要があると考えられる.

しかし, 過度にノイズ除去を行った場合, 正しい記事データまで除去してしまうことがある. そのため包括度が下がることを防ぐように対策をとる必要があると考える.

## 参考文献

[1] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕. blog ページの自動収集と監視に基づくテキストマイニング. 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004.

[2] Jeff Heaton. Parsing html in microsoft c#. http://www.jeffheaton.com.