

RSS アクセスログを用いたユーザ行動分析手法の提案

Suggestion of the user action analysis technique using a RSS access log

不破 拓[†] 和田 雄次[‡]
Taku Fuwa[†] Yuji Wada[‡]

1 はじめに

インターネット普及・高速化により、Web サイトが増加し、ネットワークを流れる情報量が膨大になっている状況において、Web サイトは広告・マーケティングなどの情報発信手段として重要な位置を占めている。インターネットを利用しているユーザがアクセスした Web サイトの情報が求めている情報と比べて不足もしくは異なるものであれば、ユーザは直ちに別の Web サイトに情報源を求めるか、情報が見つからないものとしてあきらめるだろう。故にユーザが効率よく情報を収集できる支援が注目される。

しかし、これまでの研究から、Web 上に存在する Web ページの記述方法は様々であり、広告や意味のないノイズページとよばれる存在があるため、適切に前処理するのが難しい。

そこで本研究では Web サイトの見出しや要約などのメタデータを構造化して記述する RSS にアクセスしたログを用いれば、広告ページなどのノイズを容易に除去でき、有効ページの割合がふえると考える。また、RSS を構成する要素を用いた重み付けで特徴語を抽出すれば、そのページの特徴をより現した特徴語を抽出できる可能性が期待できる。

よって RSS を用いたアクセス履歴に注目した、ユーザの利用傾向取り出す分析手法を発表する。

2 分析手法

本研究で使用するアクセスログは Apache 互換形式のものを利用する。ログの収集方法としてユーザが RSS リーダ経由で Web サイトにアクセスしたログをプロキシサーバ経由で収集する。図 1 に示している図は本手法の処理の流れを表す。本手法の構成要素である前処理部、セッション抽出部、ページ解析部、履歴分析部について記述する。

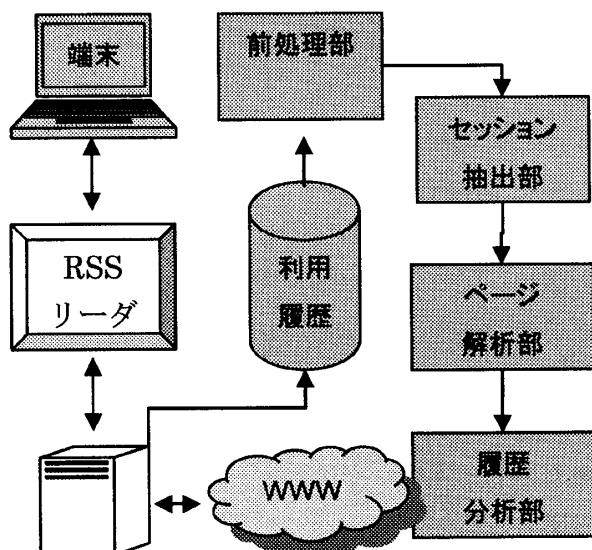


図1 構成概要図

2.1 前処理部

ここではアクセスログに含まれるノイズの除去を行う。ノイズページはポップアップするページなどが挙げられる。それらは形態素解析を行った場合、文章がないページを多く形態素解析を行うため、分析結果がノイズによって精度が下がってしまう。

収集したアクセスログを以下の条件で除去する。URL の拡張子が “rdf”・“html”・“htm” であること、HTTP のステータスコードが “OK” もしくは “NotModified” であること、HTTP コ

[†] 東京電機大学大学院 情報環境学研究科 TDU[‡] 東京電機大学 情報環境学部 TDU

マンドが”GET”のものであることである。

”rdf”からタイトルとリンクの一覧を抽出し、抽出したものを RSS アクセスリストと呼ぶ。RSS アクセスリストとアクセスログをつき合わせて、アクセスリストにないリンクを持つログを除去する。同時に “rdf” の拡張子のログも除去する。URL が一致したものを、図 2 のように IP アドレス、時間、URL を抽出したものを作成する。

192.168.1.1 [12/Jun/2006:16:54:09 +0900] http://www3.asahi.com/rss/index.rdf
192.168.1.1 [12/Jun/2006:16:54:35 +0900] http://www.asahi.com/TKY200606160127.html
192.168.1.1 [12/Jun/2006:16:54:35 +0900] http://adnet.asahi.com/js.ng/asahi.htm
192.168.1.1 [12/Jun/2006:16:54:36 +0900] http://ad.doubleclick.net/ad/N2830.BS?



図 2 前処理済

2.2 セッション抽出部

ただ単に特徴語を抽出して分析するのではなく、ユーザのアクセスごとのセッションに注目して分析を行えば、より高い精度の結果ができると考える。例えば特定のユーザが他のユーザより大量のアクセスした場合、その特定のユーザの興味・嗜好に結果が引っ張られてしまう。そこで前処理が終わったログからアクセスした時間の前後が一定の係数以上時間が離れたものを別のセッションとする。セッションを抽出したものを前述のログにセッション情報を付加して出力する。

2.3 ページ解析部

精錬されたログから Web サイトの HTML ファイルを取得する。収集した HTML ファイルの文字コードを変換した後に、タグの除去を行う。RSS アクセスリストからアクセスログのリンクに対応するタイトルを形態素解析し名詞を抽出する。タイトルの名詞を抽出した時点では、名詞同士の関係は等価である。そこでページを代表する名詞を抽出するために、アクセスした URL の記事の中の出現する名詞の出現率を用いて重み付けをし、そのページの特徴語を抽出する。図 3 にその流れを示す。

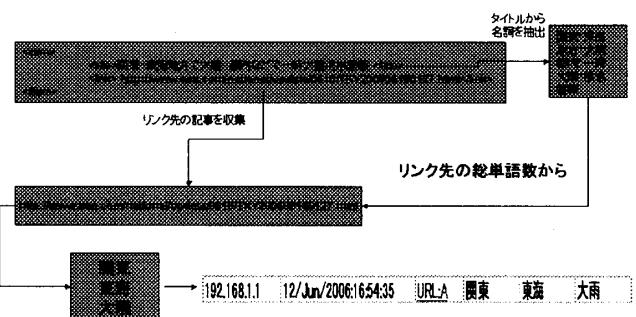


図 3 ページ解析

2.4 履歴分析部

履歴分析部ではページごとの名詞の部分に注目して、セッション全体のページの記事中の単語の出現率によって特徴語を抽出し分析を行う。

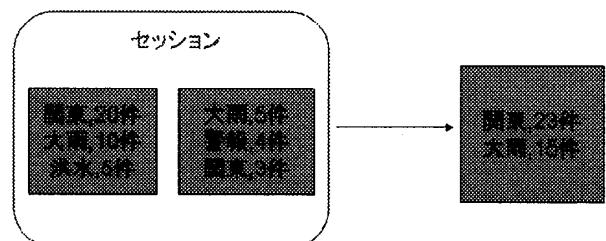


図 4 履歴分析

3 実験予定

本手法の実験には、数人の被験者にクライアント型 RSS リーダーを利用して情報収集したデータを用いる予定である。取得したログから、本稿で述べた手法で分析した結果とセッションを用いない特徴語を抽出して分析した結果等を比較して、ユーザの利用動向を把握する上でどちらの精度が高いか検証する。また、データを収集する際に被験者に IP とアクセスした時間を別途控えてもらい、抽出したセッションの妥当性の検証も行う。

参考文献

- [1]森本 和伸 他, "MineBlog:興味発見を支援する Blog 生地推薦システム" 情報処理学会論文誌, 2006
- [2]戸川聰 他, "WAVISABI:Web 閲覧特性に基づく管理者支援のための利用動向可視化システム", 情報処理学会論文誌, 2005
- [3]浦 勇亮 他, "Web アクセスログのクラスタリングによる問い合わせ拡張に関する検討:i タウンページ上での実験", DEWS2002