

4V-2 アクセスログデータに基づくデータマイニングの適用*

加藤 久慶[†] 西山 裕之[†] 溝口 文雄[†]

東京理科大学 理工学部 経営工学科[‡]

1 はじめに

WWW(World Wide Web)は、ユーザが気軽に利用でき、有益な情報をそこから得ることができる。そのため、そのユーザ数は年々増加し、社会的関心も非常に高まっている。このようなWWWにおいて、情報提供者やWebサーバー管理者は、たくさんの人に閲覧してもらうために、趣向をこらしたサイト構築を行なっている。この時、もし、Webのアクセスログからユーザの閲覧パターンを知ることができれば、彼らにとって効果的なサイトを構築するための1つの指標とすることができる。しかし、実際には、Webアクセスログの件数は膨大であるため、手動によるユーザの閲覧パターンを抽出することは困難である。

このような問題に対して、アクセスログ全体に対してデータマイニング技術を適用してユーザの閲覧パターンを分析する研究[1]がある。しかし、本研究では、アクセスログを共通カテゴリーでグループに分け、そして、その後に、データマイニング技術を適用し、Webサーバー管理者にとって有益と考えられる情報を抽出することを目的とする。

2 アプローチ

本研究では、膨大なWebアクセスログから、各ドメインに対する閲覧パターンを抽出し、各々に対して興味の傾向を見つける。

具体的には、ac(教育機関)、co(企業)などのドメインをそれぞれのグループとして考え、各々に対して頻度の高い閲覧ページの組合せを抽出する。そして、Webサーバー管理者にとって有用な規則を見つけ出す。

この時、頻度の高い閲覧ページの組合せを抽出するために、本研究では、データマイニングの手法の1つで、高速に組合せを抽出することにおいて成果をあげているAprioriアルゴリズム[2]を利用する。そのアルゴリズムを以下に示す。

Apriori アルゴリズム

Aprioriアルゴリズム[2]は、相関ルールの抽出技法の1つで、トランザクションデータベースの中から事象間の関連性を抽出することができる。このトランザクションデータベースとは、例えば、顧客IDと商品購買との関係を表したデータベースである。そのトランザクションデータベースから頻度の高いアイテム集合を生成し、その候補となる集合から規則を見つけ出す。具体的には、以下に示す処理を行なう。

1. 候補集合の生成：事象間の組み合わせを見つける。
2. ラージアイテム集合の生成：支持度による候補集合の枝刈りを行なう。
3. 1と2を繰り返す行う。

この時、ラージアイテム集合とは、規則の候補となる集合であり、導出される規則は、この集合から生成される。また、支持度とは、集合において、必ず満たさなければならない頻度である。

3 Web アクセスログ分析

ここでは、Webアクセスログの形式やそのログデータをどのように分析していくかについて述べる。

3.1 Web アクセスログの形式

Webアクセスログとは、IPアドレス、URLにリクエストがあった日時、そして、アクセスされたページに対してどのような要求があったかを示す文字列から成る。以下には、ウェブアクセスログの例を示す。

```
sutnproxy.ed.noda.sut.ac.jp - - [16/Nov/1999:20:24:01+0900] "GET /minilogo.jpg HTTP/1.0" 200 9921
```

これは、HTTP1.0のプロトコルに従って、sutnproxy.ed.noda.sut.ac.jpから要求があり、その日時は1999年11月16日20時24分01秒であることを意味する。

本研究では、そのアドレスから得られるac(教育機関)やco(企業)などのドメインと要求ページに注目した分析を行なう。

*Application to Access Log Data using Data Mining Technique

[†]Hisayoshi Kato, Hiroyuki Nishiyama, Fumio Mizoguchi

[‡]Department of Industrial Administration, Faculty of Sci. and Tech., Science University of Tokyo

3.2 対象データ

本研究では、表1に示すデータについて分析を行なう。データは、東京理科大学情報メディアセンターのウェブに対するアクセスログである。そして、件数は、そのサイトに対するアクセス件数である。

対象サイト	東京理科大学情報メディアセンター
URL	http://www.imc.sut.ac.jp/
期間	1999年6月～2000年10月
ファイルサイズ	8,180,000byte
件数	412,588件
平均件数/日	300～500件

表 1: 分析対象

3.3 分析方法

分析の流れを図1に示す。まず、Web アクセスログを入力とし、指定した日時における co(企業) や go(政府機関) などのドメインによって分類されたトランザクションデータベースを作成する。そして、そのトランザクションデータベースを入力とし、Apriori アルゴリズムにより頻度の高い閲覧ページの組合せを求める。結果として、次のような出力が得られる。

(/A.html) → (/B.html) ∧ (/C.html) (支持度 3%, 確信度 60%)

上記で示す閲覧パターンは、あるドメインに属し、A.htmlを閲覧するユーザの3%は、B.htmlとC.htmlも同時に閲覧し、その時の確からしさが60%であったことを意味する。

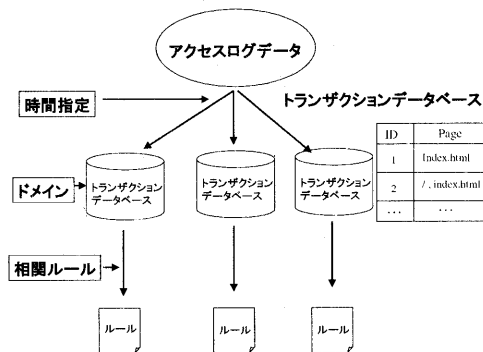


図 1: 分析の流れ

4 分析結果

アクセスログのうち、アクセスドメイン数が多い上位3ドメイン ac(教育機関), co(企業), go(政府機関) について得られた規則について述べる。

• ac(教育機関) ドメイン

次のような興味深い閲覧規則が得られた。

(/Guide/abstract.html) → (/) ∧ (/navi.html) ∧ (/contents.html) (支持度 3.0%, 確信度 92.0%)

この規則は、論文関連のページ (/Guide/abstract.html) 閲覧するユーザは、このように、メインページ ('/navi.html', '/', '/contents.html') も一緒に見るというものである。このように、このドメインでは、論文やイベント情報のものに関係するページの閲覧規則が数多く得られた。

• co(企業) ドメイン

閲覧パターンとして次に示す規則が得られた。

(/ITCourse/it.html) → (/) ∧ (/navi.html) (支持度 4%, 確信度 80%)

これは、IT 関連ページ (/ITCourse/it.html) とメインページ (/navi.html) を同時に閲覧する規則である。この閲覧パターンが示すように、企業ドメインは、企業と関係するページの閲覧規則が中心であった。

• go(政府機関) ドメイン

このドメインでは、「/Paper/Report/Jconf.htm(支持度 3%)」というような単一ページへの閲覧規則が数多く得られた。

5 考察

4章で示すように、得られた規則は、ac(教育機関) ならば論文関係の閲覧を示す規則が中心で、co(企業) ならば企業情報と関係のあるページ閲覧が中心であった。つまり、ドメインと閲覧対象は関連性があることがわかる。

6 まとめ

本研究では、データマイニング技術の1手法である Apriori アルゴリズムをウェブアクセスログ分析に適用し、各ドメインに対する閲覧特徴の抽出を行なった。そして、ドメインと閲覧対象には関連性があることを示した。これにより、Web サーバー管理者は、この情報を1つの指標として Web 構築を行なうことが可能である。

参考文献

- [1] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu "Mining Access Patterns Efficiently from Web Logs", Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000.
- [2] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules", In Proc. of the 11th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994