Case Study: Web Usage Mining

Mathias Goller

September 30, 2007

1 High Tech Inc.

High Tech Inc. is one of the leading companies in machinery tools industry. In the 1970s, electric motors were its core product. In late 1970s, High Tech Inc. was concerned with growing competition by companies from overseas. Therefore, High Tech Inc. diversified its product lines. Three new lines of products emerged, namely engines for trains, electric machine tools, and miniaturised electromechanic machines.

The new diversification strategy started a new era of success in the company's history. High Tech Inc. began to spread all over the world. It set up new factories and regional headquarters and invested in new product lines. Now, High Tech Inc. is a global player producing a vast set of high tech machinery goods such as assembly robots and generators.

High Tech Inc. sells its products to other companies which embed them in their own products or install them into their factories.

Most of High Tech Inc.'s products require support during set-up and maintenance. Usually, a company buying a product employs technicians and engineers to do the set-up process on their own.

Technicians and engineers can access documentation, tutorials, and hints to several technical issues of any product on the company's web site. As High Tech Inc. offers highly sophisticated products and many of them, there are quite a lot of documents and web pages on the company's web site.

2 Improving Accessibility of Relevant Information

As High Tech Inc. has steadily introduced new products, it failed to keep information systems small and simple. For reducing time-to-market it omitted several review processes during setting up the support web site. Therefore, many customers complain about support as relevant information is found after a hard time of searching—if found at all.

As a consequence, High Tech Inc. installed a project team to analyse and improve the corporate's support web site.

In the first meeting of project *Improving Accessibility of Relevant Information* (short: IARI) the most important project team members stated their opinion concerning the project.

Project Leader Users report that they are unable to find the information they need. They complain about ill-linked documents. Each pages gives only a bit of information. If it is linked, it is linked to pretty abstract documents. For instance, many users complain that documents with specific information about a product often link to management summaries and pages that describe how to buy that product—although, these users are customers that already possess that product. Hence, we need to re-structure our support web site according to the needs and interests of our customers.

Thus, finding out what our customers are interested in is our major concern in project IARI. I suppose to use a tool that performs web usage mining on our web servers. By doing so, we can

determine frequent click patterns, i.e. we know how our users search for information. We would gain the opportunity to re-structure our web site according to the way our users access our site.

Server Administrator We store all data of server accesses on one of our web servers. Therefore, we can trace each page request of a user. Unfortunately, these data is stored decentralised as each server has a log file on its own. There is a cluster of web servers in each region in which we operate a regional headquarter. Although all servers use the common log format, which is a W3C standard protocol, it will be a hard task to integrate these data.

Web Content Manager Analysing pages a user has clicked is insufficient for determining what a user has intended to look at. We regularly run a program on the web logs that shows us the most important pages of each web site. The top-level pages of our products are the most-accessed pages. The results are pretty much the same, regardless in which country we do this test.

Therefore, I suggest to analyse the key words of the web site's search engine instead. Users enter key word in our search tool. By that way they tell us what they are looking for.

3 Your Task: Consider Data Mining Techniques and Give Suggestions

Propose your suggestions about if and how data mining techniques might support the success of this project in a short email to the project leader. Consider at least of one the questions below.

1. Which data mining technique is adequate?
2. What sources of data are potentially beneficial?
3. Are there pre-processing tasks necessary? Which ones?
4. Is there all information given to make a decision? If not, what could one do to gather that missing piece of information?
5. Does the web content manager justifiably criticise analysing log files?
6. Is the web content manager's alternative a better option compared to analysing log files of page accesses?

Feel free to use other sources of information about web usage mining but be careful in choosing quality of chosen sources. All external information must be correctly marked and referenced.

Case Study: Web Usage Mining - High Tech Inc.

Mario Hofer
0255509

Abstract

This paper gives a rough overview on how data mining techniques can support the success of the IARI project. All conclusions drawn in this document are based on the description of the case study. Therefore the paper only focuses on the topic of Web Usage Mining. Furthermore the recommended courses of action may suffer from incomplete information.

1 Introduction

The recommended courses of action, described in the following sections are based on the statements given by the most important project team members. The main aim of the project IARI is to improve the support web site of High Tech Inc., since customers claim that information and documentations on products are, if at all, very hard to find. Therefore the analysis starts with a description on potential data sources. After that there will be a section on pre-processing to show, why pre-processing is necessary and which actions can be taken. The main section is a description of adequate data mining techniques. This section also includes comments on missing information. The conclusion sums up the main results.

2 Data Sources

In order to analyse frequent click patterns etc. one has to collect data on the web traffic. In case of High Tech Inc. the server administrator offers the logs of the companies webservers, which trace each page request of a user. As mentioned in the assignment, the data is stored decentralised at different reagional web servers. Web server logs as the single source for data may be incomplete because of "data gaps" described by [2] and [4]:

- Web server logs do not record cached page views.
- Web server logs do not record data which is transmitted with the POST method.
- Web server logs may fail to record data encrypted in URI used for scripts like CGI.

The afore mentioned problems are also critical to the IARI project. Therefore it is adviseable to incorporate further data sources. Besides web server logs [4] suggests two further levels of data collection, namely Client Level Collection and Proxy Level Collection. Client Level Collection either uses remote agents like JavaScript and JavaApplets or modified web browser versions as a data source. Whereas Proxy Level Collection uses cached data as a source. Since there is no information on the server structure of High Tech Inc. a further analysis of Server Level Collection will be omitted. Furthermore the development of a new web browser is not feasible. Considering these aspects Client Level Collection seems to be the only further source of data.

In addition to Web Usage Mining, the web content manager of High Tech Inc. suggests to analyse the key words of the web site's search engine. Therefore a log of these key words is a further source of data. The applicability of this approach will be discussed later in this paper.

3 Pre-Processing

The previous section already shows that web logs are a unpurified source of data. This section gives more reasons why pre-processing is necessary and it also highlights different types of pre-processing.

3.1 Why Pre-Processing is necessary

[2], [3] and [4] claim that pre-processing of data is vital for the success of Web Usage Mining. [3] mentions that the analysis focuses on the behaviour of users, nevertheless web logs only store the IP which may not correspond to a single, specific user. Furthermore besides the problem of cached websites [2] and [3] say that web log data is contaminated because of spiders or robots visiting the web site[1].

[1] A more detailed discussion on the need of pre-processing as well as an overview on pre-processing methods in general would be beyond the scope of this paper and can be found in [1]

3.2 Aspects of Pre-Processing

In order to carry out pattern discovery [4] suggests to convert usage, content and structure information into adequate data abstractions.

- Usage Pre-processing: This aspect deals with the matching of a web log entry to a specific user. [4] considers this part as the most difficult task in Web Usage Mining. The identification of users can be supported heavily if the web site requires the users to authenticate themselves.
- Content Pre-processing: This step consists of converting data like texts, images or scripts into a form which is usefull to the Web Usage Mining process. Furthermore the content of the web site and the web log entries have to be merged together in order to get significant results.
- Structure Pre-Processing: The structure of a site, which consists of hypertextlinks between pages, can be obtained by using clustering or classification algorithms.

Considering the facts above it is advisable that High Tech Inc. integrates the web logs from the different servers, although the server administrator says that this will be a hard task. However if the website is presented in different languages it is only meaningful to merge regional web server logs.

4 Data Mining Techniques

This section gives a short overview on mining techniques applicable to Web Usage Mining[2], what they do and in which way they are useful to the IARI project[3].

Statistical Anlaysis

To get a grasp of the data High Tech Inc. should carry out basic descriptive statistical analysis. In order to do so one should gather dimensions like frequency, mean, variance, etc. on differnt variables, e.g. page views, viewing time and the length of a navigational path. Especially the viewing time of a page could be an indicator of the relevance of its content.

Association Rules

Since the project leader reports of customers claiming that the pages often only give bits of information, one could introduce association rules in order to discover which pages are viewed together in a single session. This could give hints on how to better organise the content of the web site. Compared to structure mining, this method has the advantage that the pages do not have to be connected directly through hyperlinks.

Cluster Analysis

Since clustering *"is the process of grouping a set of physical or absract objects into classes of similar objects"*[4], it is an adequate method to build clusters of documents for each product offered by High Tech Inc.. This allows improvement of the search engine, leading to shorter search paths of the users. This method could also be applied to analyse the keywords entered into the search engine. However the implementation might be difficult, because the clustering algorithm has to work with some kind of "semantic similarity". Although the web content manager's criticism on the log files is justifyable because of the problems mentiond in the sections on data sources and preprocessing, his solution still has, among other things, one drawback: Analysing the key words gives hints which words are used for searching. Nevertheless one has to match the correct sites to the indentified words.

Classification

Classification provides algorithms to allocate data items to specific, predefined classes. This technique could be used to build several web page categories, e.g. management summaries, marketing pages, tutorials, hints, technical documentations, etc.

4.1 Missing Information

The information provided in the assignment is not sufficient to give detailed recommendations. For example it would be interesting to know whether the websites are offered in different languages. Furthermore to identify data sources it would be useful to have detailed information about the IT-infrastructure at High Tech Inc.. In addition, one should also run scripts to gather information on dead links within the web site. A script to determine dead-pages (pages which are not linked at all in the system) would be useful aswell. Those two metrics would provide insight into the general applicability of Web Usage Mining. As [1] describes, the number of links to a page is an indicator the *authority* of a page. Dead-pages will not receive a high authority with this method, altough they might be very important.

5 Conclusion and Open Questions

This paper has shown different ways how Web Usage Mining can be applied to the IARI project and which steps have to be taken. What this paper has

[2] Taken from [4]

[3] [4] also describes *sequential patterns* and *dependency modeling* as techniques for Web Usage Mining. However the domains of these methods are rather marketing and online retailing. Therefore a description of them is left out.

[4] see [1] p. 335

not analysed are further, alternative approaches to improve the utility of the website. For example High Tech Inc. could incorporate metadata, like the relevance of a site, in their system. The company could also introduce some kind of reporting system, to enable users to report dead- or senseless links.

References

[1] J. Han and M. Kamber, *Data Mining - Concept and Techniques,* Morgan Kaufmann, 2001.

[2] E. Rahm, "Web Mining," *Datenbank-Spektrum*, Vol. 2, pp. 75-76, 2002.

[3] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation," *Comm. ACM*, Vol. 43 (8), pp. 127-134, 2000.

[4] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, Vol. 1 (2), pp. 12-23, 2000.