Design and Evaluation of Query Prefetching Prediction Model for Effective Web Information

MD. ALAMGEER

Bioinformatician & Software Engineer; Head, Department of Bioinformatics, Singhania University, Rajasthan, India; and http://www.gensolution.org

Abstract: With the explosive growth of WWW applications on the Internet searching on the WWW is became an everyday task for billions of people around the world. People search for content (text pages, images, sound, video, e-shops and many more) on every possible work of life, and rely on search engines to help them locate their points of interest from amongst the billions of resources on the Internet. Search engines accept the queries and are faced with two challenging tasks, which relate to two well accepted criteria in the field of information retrieval:

1. A recall task, of finding resources relevant to the query in the Web.
2. A precision task, in which the relevant resources are ranked according to the relevance to the query, in order to be presented to the user in the best possible order.

Perhaps the most search operations done on WWW is querying for relevant text pages based on content analysis, this approach worked reasonably well when the web was smaller and when sites were designed without authors' knowledge of search engines' ranking techniques. This degradation has prompted at finding additional means for finding & ranking information. Then link-structure of WWW has been acknowledged as a rich source of information in web retrieval problems.

The most widely used approach for this purpose is the query pre-fetching using pattern discovery process of Web usage mining that entails many techniques like Markov model, association rules and clustering. This thesis presents a formal model of QUery PREFetching PreDiction (QUPRED), which is capable of effectively answering queries in almost blink of the eye. Where we implement various techniques to predict the future page to be accessed by the web user thoroughly analyzing users previous browsing patterns & queries. First we combine low-order Markov model and association rules. Markov model analysis is performed on the data sets. The data sets are clustered and Markov model analysis is performed. The outcome of the integration is better accuracy with less state space complexity than higher order Markov model. The lastly, combining all three techniques together: clustering, association rules and low-order Markov model. The integration model is applied, which achieves better Web page access prediction accuracy, less state space complexity and less number of rules generated than the previous two models.

An empirical analysis of the characteristics of user behavior in real world data warehouse environment and performance measurements using simulated traces with data from real world application demonstrates the usefulness of our model.

Keywords: Query prediction, prefetching, dynamic view, prediction algorithm, multidimensional data modeling, web caching, OLAP, Markov Model, IPM
1. OVERVIEW

Piece of work gives an insight into the background of data mining, Web data mining and Web usage mining in particular and their importance in Web access prediction. It describes the various Web Caching and Web-Prefetching techniques respectively. Also gives a comprehensive overview of the topic area of this thesis (i.e., prediction of user behavior).

It contains a formal specification of the encountered challenges that have to be addressed by a prediction environment. As the presented framework is independent of its application to online tools, this work is also valuable in prediction methods for other application areas (for example, predictive www proxy caches). The theoretical background of online database systems, namely the data model and the according query formalisms also defined here. The proposed formalisms constituting the formal foundation for the thesis’ core are derived after comparing the most important scientific approaches in this area.

It also examines efficient prefetching policies depending on the behaviour patterns of users, architecture of the Search engine’s index Structure and Query Processing Models. It introduces the three Web usage mining tools: Markov model, Association rules and clustering emphasizing on their importance in Web access prediction and presents a model that integrates Markov model with association rules.

2. OBJECTIVE

The main objective behind this research work is the correlation between web usage mining and web personalisation. Information is more and more becoming the driving factor of today’s economy and social life. This demand for interactive ad-hoc analysis of both structured and unstructured data has led to the development of a large variety of interactive information systems based on different paradigms. The World Wide Web (WWW) using the hypertext paradigm is certainly the most well-known system for semi-structured qualitative data.

With this Web data mining gains its importance to focus on the Web pages link structure, their content and their usage. Web usage mining concentrates on tools and techniques (i.e preprocessing, pattern discovery and analysis) used to predict users’ navigational paths by discovering their Web access patterns and querying styles. Knowledge about this pattern can be used for improving the performance. Automatically adapting the user interface to the user’s current tasks will certainly improve productivity of the overall system.

Therefore, it is extremely important to understand the kind & way the web user interact with the online system (issuing of query) and always be one step ahead of them when it comes to predicting future accessed pages. Pattern discovery achieves this by extracting useful knowledge and patterns applying different tools
and techniques (association rules, clustering and Markov models etc.). Each of these pattern discovery techniques has its own strengths and weaknesses. Discovered patterns of accessed Web pages helps predict the future page to be accessed by the user.

3. WEB PRE-FETCHING TECHNIQUES

3.1. Web Data Mining

Data mining involves the study of data-driven techniques to discover and model hidden patterns in large volumes of raw data. Web structure mining involves the techniques used to study the web pages schema of a collection of hyper-links. Web usage mining on the other hand, involves the analysis and discovery of user access patterns from web servers in order to better serve the users’ needs. Figure 1 summarizes the processes involved in each of the web data mining phases.

![Web Data Mining Architecture](image-url)
With the advancement in computer technologies, there has been a major need to apply data mining techniques to electronic data as the amount of information stored is increasing at a very high pace.

3.2. Web Content Mining
Web content mining involves mining web data contents. It focuses on various techniques that assist in searching the web for documents whose content meets a certain goal. Those documents, once found, are used to build a knowledge base.

The emphasis here is on analyzing the Internet hypertext material. The Internet data that is available in digital form has to be prepared for analysis. For instance, [1] focused on resource recovery on the web. Moreover, several intelligent search agents, information filtering/categorization and personalized web agents have been developed for information retrieval and for organisation of structured and semi-structured information on the web [2].

3.3. Web Usage Mining
Web usage mining involves the automatic discovery and analysis of patterns in data as a result of the user's interactions with one or more web sites. It focuses on tools and techniques used to study and understand the users’ navigation preferences and behavior by discovering their web access patterns. These techniques are effective means that help e-commerce businesses improve their Web sites in an efficient manner [3].

3.4. Preprocessing
Before starting any mining technique, web data has to be cleaned and preprocessed. Preprocessing prepares data for the pattern discovery stage. It transforms web log files into web transaction data that can be processed by data mining tasks. Web data could take many forms. The primary data sources are the server log files that include web server access logs and application server logs. The most important and the most easily accessed data is the web server log report that keeps track of every single user access to the server. In general, the log entries include information like date, time, client IP, URL of the source, name of the script or file requested and the server status [1]. There are three types of preprocessing: usage preprocessing, content preprocessing and structure preprocessing [4].

3.5. Pattern Discovery & Analysis
During this stage, algorithms are run on the data and patterns are extracted from it. Pattern discovery involves the employment of sophisticated techniques from artificial intelligence, data mining techniques, psychology and information theory in order to extract knowledge from collected and preprocessed data. Some of the
most widely used pattern discovery approaches are statistical analysis, association rule mining, clustering, classification, sequential patterns and dependency modeling [4]. Statistical analysis techniques are the most common tools used to extract knowledge about Web site users. These tools could provide user information like the most frequently accessed pages, average time of viewing a certain page, average time the user spends browsing a certain site etc. However, knowledge extracted using statistical analysis could be very useful for improving the system performance and for providing support for marketing purposes especially for e-commerce applications [2]. Association rule mining refers to the sets of pages that are accessed together in a single server session. It is useful in providing personalized web content to users. Another type of clustering is page clusters which discover pages that have related content.

3.6. Pattern Analysis

Not all discovered patterns are useful and this step aims at identifying the patterns which represent new and potentially useful knowledge. Pattern analysis involves filtering out the unneeded patterns or rules discovered through the pattern discovery phase. The most common pattern analysis technique is the use of query language like SQL. Another technique could be the usage of online analytical processing (OLAP) tools [5]. Figure 2 summarizes the Web usage mining architecture.

![Figure 2: Web Usage Mining Architecture](image-url)
4. MODELING OF WEB CACHING

As the popularity and size of Internet continue to grow, the scalable design of a network system becomes more and more important. Caching is a mature technology that has been widely adopted in many computer areas such as operating systems and databases. Most of the Information Retrieval (IR) underlying principles have great impact on web caching. It is also effective to reduce network traffic, latency as well as server loads. Therefore, intensive research has been conducted in this area. Here we aim to sit a predictive caching replacement algorithms based on web access characteristics. Further, we also extend our prediction model to pre-fetch documents in order to improve the network performance.

![Figure 3: The Architecture of Web Caching](image)

The main thrust of this chapter is to study the design of replacement algorithms of web caching. First of all, we generalize a model of web caching which is widely used in the literature. Our further discussion will be based on it. Secondly, we characterize the web access and point out the key issues to build a good replacement algorithm. We also review the Caching related previous work in this chapter. Figure 3 represents a replacement algorithm which is the heart in the generic caching model.

4.1. A Generic Model

In a web caching system, no matter it is proxy caching or transparent caching, its kernel usually consists of a Cache Manager, some Metadata, a Page Replacement Algorithm A, and the Cache Memory (there all storage of web objects). A generic model can be mapped in Figure 4. The Cache Manager is a central controller; it coordinates all the operations among other components. Metadata stores indexing
information of cached objects for fast access. It also keeps additional information, such as last access time, access frequency, of web objects for the sake of replacement decision. Compared with the size of a cache, Metadata is usually negligible.

![Figure 4: A Generic Web Caching Model](image)

Page Replacement Algorithm or Replacement Policy is the heart of a caching system, it takes necessary metadata input to decide whether to cache a new object or not and if storage space is limited, which object(s) should be evicted in order to make room for the new comer. The Cache Memory is referred as the physical repository where web objects are kept. A caching system sits between clients and remote servers, interacting with them. This can be illustrated by the marked steps in Figure 6.2:

1. A client sends a request for a webpage $P$ to Caching Manager;
2. Caching Manager then searches $P$ in Metadata to see if it is in Cache Memory or not. If found, go to step 3; otherwise, go to step 4;
3. Caching Manager loads $P$ into memory go to 10;
4. Caching Manager requests $P$ from the remote server;
5. $P$ is returned by the remote server;
6. Caching Manager consults with Page Replacement Algorithm $A$;
7. $A$ gathers necessary information about cached web objects
8. $A$ makes a decision on whether to admit $P$ into the cache or not. If $P$ is admitted, $A$ also picks out a list of object(s) to be kicked out from cache if space is not enough to hold $P$;
9. If the decision is to bypass $P$, (i.e $P$ won’t be deposited into the cache), go to 10 Otherwise, Cache Manager replaces those listed objects with $P$;
10. Cache Manager updates metadata to synchronize it with the cache status; Cache Manager responses the client with $P$. 
This model and its variations have been frequently used explicitly or implicitly in the web caching context. Extensive research has been done to the design of better Page Replacement Algorithm A[5, 7] data structure of Cache [7] and optimization the Metadata [8].

5. CONCLUSION
We modeled and formally described the structure of a framework that uses patterns in user behavior to dynamically predict future user behavior. The framework consist of different communicating processes. It can be used to compare our approach with other approach thus satisfying our general objective and achieving comparability of related work.

We provided a comprehensive formal description of multidimensional model which constitutes a combination of previously proposed data model descriptions considering Static part online system and dynamic online user behaviour. Query transformations are also shown differently from algebraic formulation of a query. This is a distinctive feature of our approach remedying shortcomings of existing approaches in this area.

Based on our pattern model, we presented an algorithm that can predict the next interaction from current session context. To this end thesis discussed such prediction techniques for WWW data access, which can be applied to predict the query behavior for users of online systems. Our theoretical analysis of the problem has shown, that the online area exhibits certain properties (mainly the number of different queries and the query sub assumption problem) that makes prediction as well as the application of prediction results significantly more complicated. On the other hand, limitations of online domain compared to database systems in general (semantically rich multidimensional data model containing classification semantics, graphical query formulation and thus a limited class of queries with locality properties) can be exploited such that prediction still makes sense.

ACKNOWLEDGEMENTS
The author thanks Dr. S.I. Ahson and Dr. Khursheid Haider for discussions and encouragements. He gratefully acknowledges the partial support of faculty members of School of Computer Science and School of Research, Singhania University. A research work is never solely the result of the efforts of a single person but is a mirror of the scientific and social environment of the author. Therefore, I would like to take the opportunity to express my gratitude to all the people who have directly or indirectly contributed to this work.
REFERENCES


