Optimization of adaptive websites from an AI perspective for improved user experience.

Deepak Potsangbam

17946429

A report submitted for

300598 Master Project 2

in partial fulfilment of the requirements for the degree of

Master of Information and communication Technology

Supervisor: Professor Yi Zhou

School of Computing, Engineering and Mathematics
Western Sydney University

May, 20xx

ABSTRACT

Web development is a constantly evolving concept and acts as a vast information source. This information is able to users in the form of webpages and through web browsing users are able to access their required information. However, website design and orientation are significant correlation factors related to users preferences and needs. Being able to adapt a website to meet individual user needs and preferences has been challenging the concept of website development. This report examines the capabilities of introducing Artificial Intelligence (AI) to website development. Through the analysis of previous research it is identified that the Hill Climbing search algorithm is a suitable means to achieve such a result that can bring about an adaptive website. The development of website designs is explored through the use of a structural design. Through the identification of specific web metrics, a standard for web design can be developed to aid in optimization. The Hill climbing search algorithm is used on dummy data to obtain a feasible solution to using AI to improve the users experience. There is potential to achieve website optimization to meet user preferences and improve their overall with self-optimizing and adaptive websites.

ACKNOWLEDGMENTS

I would like to thank and express my gratitude towards Professor Yi Zhou, who has been a motivating factor for this research and me. He has led, guided, taught and supported me at every instance I needed on this research. Without his enthusiasm, and efforts, delivering this research would have been in big doubts. His immense knowledge and interest has been a driving force behind this research. I am fortunate to have him as a mentor and a supervisor.

Also I would like to thank and appreciate Dr Jianhua Yang for his understanding and support shown towards me for delivering this research. His encouragement and support has made this research a success.

My sincere appreciation towards Western Sydney University for the opportunity to explore and understand what research is all about and its effects to the society and community.

And I would like to express my gratitude towards my colleague and research mate, Anisa Dean for the continuous support and belief in me to be able to deliver this research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

CHAPTER I: INTRODUCTION

In the last 5 years, there has been lot of changes in the technical world, which includes the web design as well. Our initial understanding or assumption of the web has been redefined in the last few years. From a 640 pixels wide screen that started in the 1990s to 1024 pixels in early 2000s, there saw a bigger change in the last few years, which saw the entry of the small screens in the web world. These screens have different orientations and varying sizes. Our assumption of the size of the web did not work any longer. The last website itself is already 20 years old. The difference that we have from the first website and first screen to what we have now is enormous. This led to the quest in the field of web designing where web design can have intelligence and understand the environment it exist in and evolve with the surroundings.

Reiterating from Master project 1, creating a web design that suits everyone’s needs and desires is quite a difficult task. With every individual having different sets of needs and desires, it often leads to having a design that is preferable to a group of people and not to others. Even a single person changes their likings in course of time. Such changes are hard to capture and reflect on a web design. As the project initially started as a research to find a small solution to address these differences and issues, the research has undergone various learning’s and understanding of the web and its design. Developing a solution, still is an ambitious task, but understanding the process of web design, itself has lead to various realizations.

This report is a continuation of the research carried out in MP1 and it provides various analysis feedbacks on the development of such a solutions. It also carries out a detail research on how the web designing works from a structural design point of view. As evolving of websites necessary includes evolving the structure primarily, it is a mandatory aspect to understand the underlying concept of the structural design. The report finds how the structure design is related to the websites popularity or worth. It furthers analyses the web metrics that forms a standard for web designs and its applications. It is important to understand how artificial intelligence will impact web design and will developing a truly self-evolving, self-optimizing become a reality.

CHAPTER II: Literature Review

Adaptive websites as already describes as part of master project 1, is a complex task and it is considered complex mainly because any such adaptation are based on user's needs and likings. Each individual has its own likes and needs and it also varies with time. It becomes a hard task to design a website that suffice and satisfy each and everyone's needs. Also many of the existing websites are out-dated with older technologies used. It no longer serves the user's needs. Also there are instances where a website is created for a specific function but ends up being used for another. It is easier to say websites are just mere fossil cast in HTML. There has although been various development in this field from self-evolving chess playing computer to adaptive personal assistant. Main concern with adaptive website is treated to its feasibility (Alkan et.al., p. 57) . Can subjective relations be automated? Will it result in chaos rather than better results? It has been a topic of concerns from various researchers and technical individuals. Generating such an adaptive website is not a simple task, as stated by Perkowitz and Etzioni, "*The goal of creating self improving websites requires various breakthrough in artificial intelligence*". . (Perkowitz et al. 2000, p. 255)

Self-improving websites generally involves automatic customization or improvement and some amount of prediction of the user's work. Armstrong in his work defined that it is possible of predicting the links a user will follow based on their interest. Certain application have been developed that can perform this task. Also Kautz and Pollack have defined *Goal recognition* as a method of identifying what a user is trying to achieve from a set of actions available. Also Lesh and Etzioni have worked on this problem in a domain independent framework depicting user's action as operators. Taking into consideration, that users behave rationally, they deduce from a users precondition and post condition of an action to reason why a user is doing vs to what should have been done.

Various projects also exist that has worked on this field. AVANTI projects [Fink et al.,1996] (see http://zeus.gmd.de/projects/avanti.html) are mainly concerned with dynamic customization of the websites based on users requirements. Also there has been studies related to meta information by STRUDEL website management system who provides meta information by putting the contents in a framework. Interface design quantification has been describe well by Raskin(Raskin, 2000). Nakayama also has developed a technique of finding the gap between a web designers expectation and the actual users's access (Alkan et.al., p. 57).

Adaptive website are also closely related to data mining where data mining plays a main part in the derivation of the adaptive process. With web mining being used in the web domain from around 1995, it has expanded its usage to web structure mining, web content mining and web usage mining. Spiliopoulou and Faulstich (1998), Wu et al. (1998), Zaiane et al. (1998), Shahabi et al. (1998) have done extensive work on web usage mining. An example of the work is WebSIFT.

There are various other studies and research carried out as well which highlights the usefulness of automating web design. This research is also focused towards achieving and evaluating if adaptive website adds logical values to the process and how does works. This study discusses various issues and topics related to adaptive website. Although the name may sound very straight forward, it requires a deeper understanding of the domain to actually understand what adaptive website is about.

CHAPTER III: Understanding Web

Adaptive Websites

Adaptive websites are the websites that evolves the organisation and presentation of the website based on certain user's access pattern and online behaviour. Various researches have been carried out in analysing and understanding user's access patterns and behaviours on the web. A lot of data mining methods has been developed such as the association rules, clustering and pattern sequences (Alkan et.al., p. 57). Each of these approach has its downfall, such as creating clusters leads to multiple interface or accommodating all navigation patterns often confuses the users on the path to return to some information they previously seek. Realizing an adaptive websites and its optimization required much more analysis and understanding of the web, the web designing process, and the structure and layouts. . (Perkowitz et al. 2000, p. 255)

Need to Develop Adaptive Website

It has been fairly understood the need to develop adaptive websites. As discussed earlier, the needs and requirements differ from one person to another. Relating to websites we see two major issues of all that exists.

- In a structure of links in hypermedia, users may become unable to navigate properly and lost with the information overload. This is typically known as lost-in-hyperspace problem.

- Websites is a mode of displaying static content and information, which is viewed by diverse set of users. This is quite an issue for people who are already aware of the information and may be too difficult for those who do not have the needed background.

> To some it may just be uninteresting. This is often termed as one-size-fits-all issue of the non-adaptive sites.

This led to the need of developing technologies that understand the needs of the users and render effective mode of channeling the contents and layouts to improve the user experience. The current websites and development process is non-dynamic. (Perkowitz et al. 1998, p.729) With time the structure and layouts becomes outdated and it involves a huge cost to redo the structure and layouts with changing time, technology and requirements. Also the parameters through which users' experience is calculated needs to be regularly monitored and accessed. This is one driving reason behind the move towards adaptive websites. It may also be noted that with rise in big data, most of the website that exists has information overload, and this leads to tougher and confusing navigation often leaving users lost in the process. One important task is also to consider the navigation to be easier for the users keeping intact the quality and fulfillment of the experience as well (Perkowitz et al. 1998, p.729). Considering all these factors, adaptive websites seems to be a good alternative for the current existing HTML idiosyncratic blocks. An adaptive website theoretically should be able to recognize users and actions and identify changes along way. It should be capable of reasoning and plan for the upcoming unknown circumstances. It is important to understand why an adaptive system is developed for and for whom. In this context it becomes important to understand the kind of users that such a system must satisfy.

Stand alone users: are the individual users whose presence on the web is for personal needs and their activities are based on their personal interest. Based on this interest they receive the feedbacks (Perkowitz et al. 1998, p.729).

User groups: Group of people who share common thought or are like-minded. They get feedbacks based on the interest or activities of anyone in the group.

Website operators: Content analyst, investors, admins are another set of people that the system should satisfy as well. They analyses the web information on how the system is implemented and used.

Goals

Various goals need to be achieved to satisfy these diverse sets of people such as selection, usage analysis, personalization and feedbacks. Each of these goals can be described as below.

Selection: This related to the task of identifying of rules and criterions from larger sets. In case of an adaptive website, certain pages may be included or removed or filtered.

Usage Analysis: It is a difficult task to determine the effectiveness of a website in terms of performance. E.g. it is hard to identify if a website leaves confused or lost or if it is hard to navigate through the pages. Analyzing an adaptive website may give us information on how to improve the effectiveness of a website by providing relevant information or links to the user easily. Identifying feedbacks and suggestions for improvement or even the usage of a website may directly or indirectly assist in improving the user experience (Kilfoil et al. 2003, p. 111). A good indicator of user's interest is the amount of time a user spends on a page or how often they visit the page.

Personalization: An adaptive website needs to reflect an individual's needs, interest and knowledge. This method of identification and implementation is termed as personalization. Personalization can be achieved through various methods, filtering and profiling are some methods to name (Kilfoil et al. 2003, p. 111). The main task here is to bring about changes or alter a website based on a user’s needs with information collected prior or in real time.

Recommendations: When users are overloaded with information and they get lost with the number of possible web pages available, adaptive websites can provide hyperlinks termed recommendations with the most relevant information that the user requires (Kilfoil et al. 2003, p. 111). An example of recommendation if providing user with links of pages which the user is most likely to visit. This is known as path- shortening. This reduces user’s navigation and assist with the overall user experiences.

Challenges

Implementation of adaptive system has its own share of challenges and it directly affects the feasibility and performance of such systems. Few of the challenges are analyzed here

A) Information shortage

Although the web is a major source of information, for a proper implementation of adaptive systems, various data sets are required. It is often the case that such data sets are unavailable or insufficient. Datasets such as web structure, web content, user profiles and usage data are the important data that are required for this system. Such data is rare or the available data needs to be modified to extract the useful information.

Web structure

The physical layout of a website assist in understanding the user’s preferences and recommendations. Also in web, the factor why a structure or layout exists is not immediately available and it is a huge scale issue to try and retrieve such information. Also there are multiple problems associated with the changeability of the web structures.

Web Content

Content plays an important part in understanding the user’s interest and the relationships between the pages. But it suffers more from scalability, availability and changeability problem.

User Data

User data can provide a great deal of information about an individual’s interest and in determining similar set of groups who share same interest and behavior. Profiling is a technique of extracting such information from various sources. This data can be obtained from the client, server or indirect proxy methods and are categorized as demographic, behavioral, and attitudinal or click stream data which results in two categories. Data that defines individuals and data that defines groups of users and their interest. User profiles are based on the information obtained during the information access. Obtaining user data can be an issue and validating the data, be another issue. Users in general are not interested to provide such data or there are legal issues concerning privacy concerns. Indirect methods have to be implemented to obtain these data.

Website Usage Data

In this system, most important data is the user’s interaction with the website and the associated behaviors. It can be classified as page navigation, views, transaction and events. Of all the data, this is the major volume of data available from the web server’s log files. Although there is large volume of this data, it is less informative on its own. Website structure and individual users information are required to derive information out of these data.

B) Measurement

Analyzing the effectiveness of adaptive websites is quite a complex task. It is required to create Parameters and metrics to determine the effectiveness of the suggestions and recommendations applied. Although such metrics exist, to describe the web or the user's access, such metrics are not absolute.

C) Effect on user experience

When suggestions and recommendations are applied, it brings about changes to the website, mainly to the structure, content and links that are provided to the user. But it is important to consider the effect of such changes to the user's experience. From this it can be determined is a changes should be applied or rejected based on how the user experience will be affected. This rise to the problem of what recommendations or changes should be applied or and should not as each changes will not necessary be according to all users preference. It yet goes back to the lost-in-hyperspace problem. This is still a known issue with no known solutions yet found.

D) Semantic information Processing

Most of the data that is considered in adaptive websites are semantically related. But semantic information processing is still in its early phase and its effectiveness is in doubt. Such process works absolutely in theory but in real time, they are not tested yet.

E) Unavailability of negative feedbacks

It is a hard task to determine from user's access if a user's page visit needs to be considered as a positive or negative or a random experience. Even with backtracking, the basic idea is a user wills backtracks if it does not find a page or link where it expected to be. In this case it should be carefully noted if such a behavior will be considered as traversing a target or it is a normal behavior. Also it is worth mentioning that certain procedures such as k-nearest neighbor, Bayes theorem focuses only on the positive feedbacks.

F) Caching

Every website implements caching in order to improve the page browsing. This means that each page that a user visits is not requested each time. So a web log of the user will not necessary have the complete or correct log information.

Types of Adaptations

Adaptive websites relies all its functions in its ability to adapt or change according to the usage. Following discussion describes the types of changes that may be implemented.

A) Content

Adaptive websites main changes relates to the contents of the page based on model that has been derived from the user’s preferences. Contents are altered, removed or modified to suit the changes required. Changes in the contents are mainly carried out for achieving certain goals.

Optional information: Additional information may be inserted or removed based on user’s preferences.

Personalized recommendations: mainly related to e-commerce, offers or sales recommendations are provided depending on user’s interest.

Content substitution: depending on browser capabilities, contents may be substituted or added giving different look and feel.

B) Presentation

Along with the contents, the presentation is also changed to suits the user.

Page variants: An example is for a multi lingual website where different pages are stored with different language variants.

Natural language generation: This is an example of generating different text description for various users as implemented in online page translators.

C) Links

Changes in the navigation related to changing the links between the web pages allowing faster page access and search. It also assists in solving the lost-in-hyperspace problem. Various techniques exist for realizing navigational adaptations.

Direct links: Providing direct links to users in form of buttons which links to the best options as per the prediction from users' access.

Link sorting: depending on user's preferences, first all relevant pages are selected, and are further sorted based on the applicability. This is then visually displayed to the user as hyperlinks. Based on the relevance, the links are ordered accordingly. But this is hard to apply to indexes and content pages and it does not apply to non-contextual links.

Link annotations: Links are represented in various ways to define its applicability. This also assists in link removals and link hidings.

D) Structure

The adaptive website may as well change the structure of the website permanently rather than per request basis. Currently a human administrator makes any changes on the structure but the suggestions are provided by the system. Some of the indications are mentioned.

New Index pages: depending on the viewing pattern of certain group of users, system may suggest generating index page which defines the links and act as the central point of the user group.

Website Structure

Understanding website structure is a complicated task and it can be termed as a method of reverse engineering in order to determine the layout and URL patterns of a website automatically and how they are linked together to become the actual website. Understanding of the web at this level allows us to be able to grasp how the structure of a web can be changed to suit the needs. About a billion websites available on the web and each of them have a different structure and design. Often websites are designed with different layout for pages that has different functions. They are then interlinked using hyperlinks denoted by the URL strings through syntactic patterns.

Many of the websites uses HTML entities and each layout defines which of these entities are used and how they are allocated visually on loading of the page. As understood, a page layout is denoted by a DOM (Document object model) tree. A layout is considered as a set of pages that has similar DOM. Pages in a website are usually generated based on different layouts as per their functions. This means pages that are similar usually has similar functions, which makes it easy for a user to have a glance at a page and understand their functions.

URL patterns are sets of URLs that has syntactic similarities in their formats. In a simple example URL patterns can often be represented by regular expressions. Few examples of these are given. Here we have observed that a layout can usually have more than one URL patterns that are related. A Records selling website usually has one layout to display all its record lists. It then provides multiple query parameters to generate this list with varying conditions based on the parameters. Each different query parameter generates different URL pattern but the results are shown in the same layout.

From the layout and the URL patterns, we can build a graph which represents the website structure. We consider each layout as a node and if any interlink exist between the pages of the two nodes, those nodes are considered as linked nodes. Each hyperlinks defined the direction of the link. We also observe that if a hyperlink have more than one URL pattern, it can results to multiple links from one node to another. One of the examples taken from ASP.Net forums shows this relation.

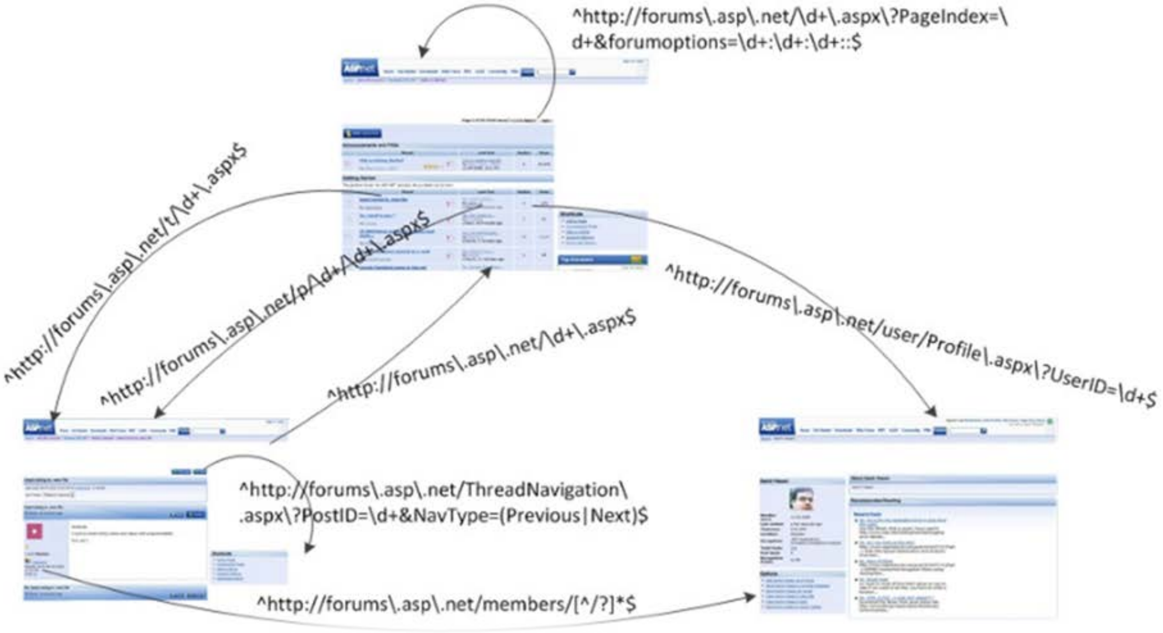
^http://forums\.asp\.net/\d+\.aspx\?PageIndex=\
d+&forumoptions=\d+:\d+:\d+::$
^http://forums\.asp\.net/t/\d+\.aspx$
^http://forums\.asp\.net/p/\d+/\d+\.aspx$
^http://forums\.asp\.net/\d+\.aspx$
^http://forums\.asp\.net/user/Profile\.aspx\?UserID=\d+$
^http://forums\.asp\.net/ThreadNavigation\
.aspx\?PostID=\d+&NavType=(Previous|Next)$
^http://forums\.asp\.net/members/[^/?]*$

Figure 1: Node Relations

It becomes an important factor to understand how a website's organizational structure can be read automatically. Pipelining is introduced as analysed for a possibility for the research purpose. Here few pages of a random website is sampled.

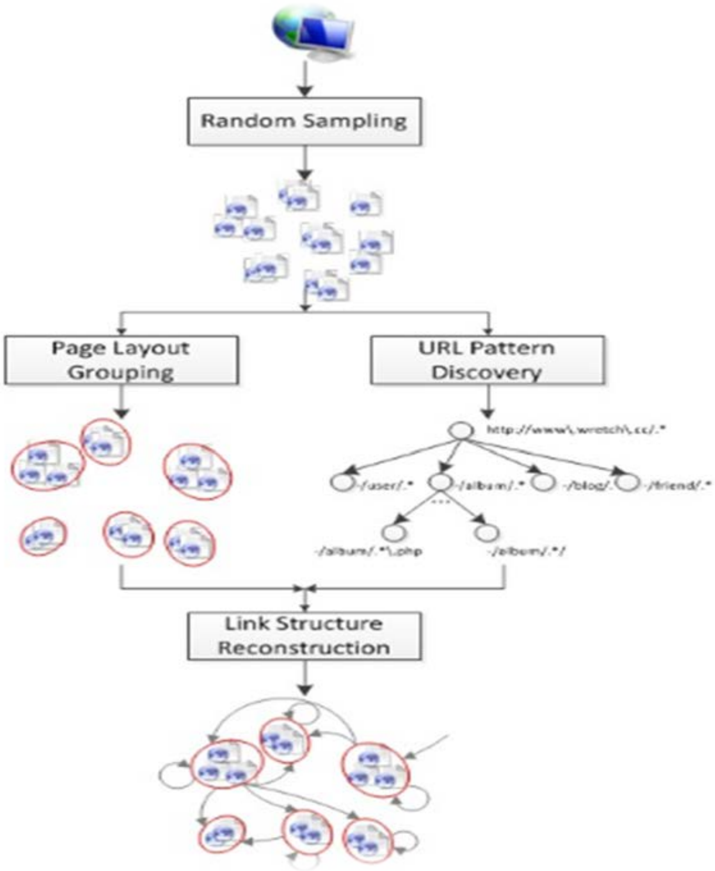

Figure 2: Pipelining

After this the layout and the URL patterns are analysed and with these we develop a link graph.

Random Sampling: By random sampling, its main purpose is to be able to provide a brief of the website by just considering few pages from the site. A lot depends on the sampling quality and the mining process. It is important to maintain these pages as diverse as possible and to attain this a breadth-first and depth-first procedure is applied which assist in retrieving pages from high depth levels with ease.

Page grouping by layout: Pages with similar HTML entities usually have the same layout and vice versa. In order to determine the layout of a webpage, the DOM path is taken into account. DOM path is defined as a path from a leaf node to the root of the tree. Each leaf node describes the entity and the path to the root from the leaf describes a rough location of the entity at page rendering.

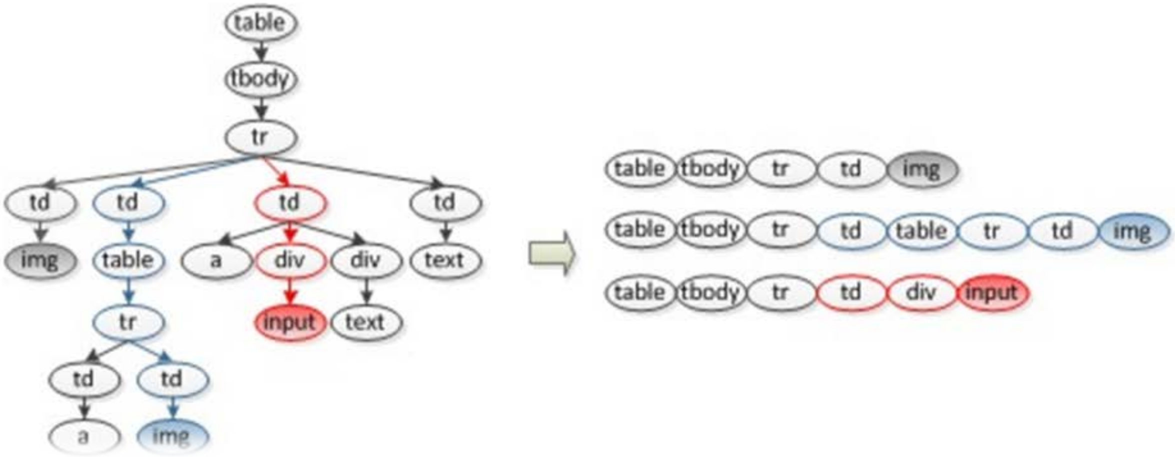

Figure 3: DOM Tree Relation

In a collection of HTML pages, each DOM path is exclusively separated to form a feature space. In this space each of the page is represented by a point and based on this, the layout similarity can be calculated. Similar pages are grouped together based on bottom up strategy. Each group forms a layout.

Analysing URL patterns: a URL has a definitive syntax as defined by W3C. Based on this syntax, a key- value pair can represent a URL.

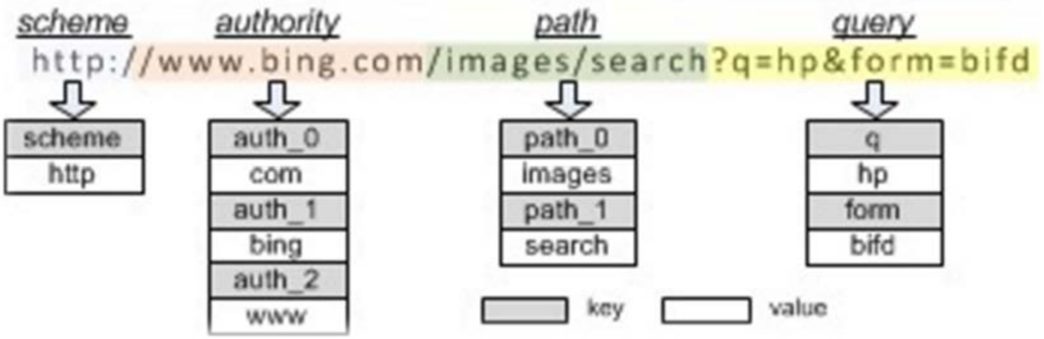

Figure 4: Key Pair

From this, it is observed that different URL entity (key) has different functions and have different roles as well in a website. Document types and function keys or URL entity has limited or few values and are specified in the pattern explicitly. Whereas keys or entities that represent parameters have a wide range of values. From this a pattern tree is generated to define a set of URLs using top-down iterative-split procedures.

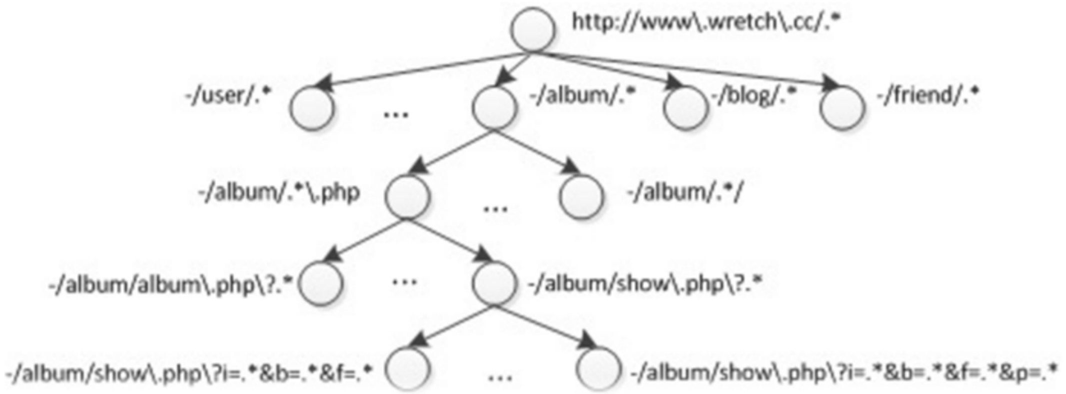

Figure 5: Pattern Tree

Link structure construction: After the layout and the URL pattern is generated it is an easy task to generate the link graph.

Web Standards and Metrics

Web has transformed dramatically over the last few years to serve users and their needs. Every website in one-way or the other satisfy some users or group of users. Differentiating between similar websites, which is good or better, is not an easy task. Website and the technologies has grown and evolved manyfold but there has been little progress made in this regards. Most system in this world has a reference point or guidelines. There are certain standards and metrics through which the performance, the structure and layout of a website can be analysed and decide if it is a good design and as per expectation. Currently anyone can build any website for any user and not conform to any standards as long as it satisfy the owner and the users. This makes it even tougher to understand the outcomes of the websites.

Talking about metrics, there are lots of best practices and well-known metrics suggested by many users and website owners or tech gurus on the Internet. It does definitely helps for a starter to understand but in spite of all this there is no uniform standards or metrics established. W3C or the main governing body of the web has released certain metrics and guidelines for accessibility of the dis-abled ones such as the WCAG 2.0, but there is no trace of any standard guidelines made for web development. Under such circumstances, the question arises as how do we decide if a website is good or not, how will an adaptive website algorithm decide if the website is working as expected or if it needs improvement. Having said these, it all comes down to the point of using the best practises that suits us the best. Although this is definitely as setback to not have such a reference point for comparison or a benchmark in this field. But it’s not as easy as having said it, the web is a very complex system and with the varying and diverse needs and requirements, it is a big challenge to develop a system that everyone can conform and adhere to. Few metrics we may consider as a part of understanding the adaptive websites are mentioned here.

Some of the common web metrics that we can find every web developer refers to be stated as.

- Visits – Number of transaction by an individual browsing one or more pages. This is the most fundamental method of comparing the traffic of websites. It is also often used as methods to compare between websites on their performances.

- Number of page per visits: This is the number of times a page is viewed during a single visit. This is a good metric to understand content usage.

- Conversion: This is directly linked on how many users returns and is satisfied with the experience they had with a particular site. This is an important aspect from a customer experience point of view.

- Bounce rate: This gives us the amount of users that just viewed a single page and left immediately. This defines the depth and quality of a visit

- Time on site: Amount of time spent on an individual page. This gives the extent of the visit.

- New visitor/returning visitor: This is an important metric that shows the customer loyalty and the sites attractiveness.

- Page load times: An important metric from user experience point of view. This defines how quickly the pages are linked and displayed.

These are few known metrics, although there are many available on the web and it is up to an individual to pick the best that suits most in the absence of standard guidelines. Also we might need to take into account few of the response time metrics such as HTML load time, full-page object load time.

Data Mining

1. Data Mining and Web Mining

For data transaction, retrieving of data and communication, World Wide Web has indeed become one of the most crucial mediums. Data mining is a way to extract a similar content or a similar pattern from different sets of data sources. Web mining is similar to data mining but it is typically done over web data. The useful information extracted over a web with similar pattern can be said as web mining. This extracted data in return can be used for various purposes like for studying operator's interests and also for observing users interests. It can also be used to improve the website structure along with enhancing services on the website. This extracted information can further be used to facilitate personalization and support adaptive websites. Web mining can be separated into three main classes, which are discussed below (Perkowitz et al. 1998, p.729).

1.1 Web Usage Mining

Web Usage Mining can define web pages usages. Proxy server, Web Server Data, Web Client Side Data can be used as different sources where the usage data can be collected. As the name suggests web server side data can be extracted from web servers and client server side data can be extracted from hosts that are accessing the website. The host may access this data through remotely agents who are implemented on JavaScript or Java (Hamilton et al. 2001, p. 129)

. Proxy server on the other hand log access files saved when a web page is requested from a host and response is given from server side.

Web Usage Mining puts tries to put more sense to the data. Different users on web may generate this data when they are over a session or following similar pattern or behavior. When a user is interacting with the web, the web usage mining tries to predict user's behavior from the past usage of the user. With this web usage mining also learns the user navigation pattern to predict similar navigation when a user logins again in a new session. This information can then be used for various purposes like business intelligence, personalization, and usage classification and also for adaptive websites (Hamilton et al. 2001, p. 129)

.

Data pre-processing, pattern analysis, and pattern discovery are main three phases of web usage mining (Hamilton et al. 2001, p. 129)

. Set of methods, techniques and set of algorithms, which are used to extract patterns from a web log file, can be defined as pattern discovery. Statistical analysis, classification, consecutive pattern and clustering and many other techniques are used for pattern discovery. In order to govern important and interesting patterns, after the discovery of patterns, they must be examined thoroughly. This examination of the pattern has many forms like visualization techniques, loading usage data and knowledge query mechanism (Lee et al. 2004, p. 131)

.

1.2 Web Content Mining

Mining information from a web page can be referred as web content mining. This content can be tables, graphs, data blocks, equation formulas, data records, texts, audio, graphics, video, audio, etc. How to improve the content on web page has always been one of the prime research areas that are considered very crucial research even now. This research is done to improve quality of the content and quality of search results, which helps in extracting interesting web page content (Lee et al. 2004, p. 131)

.

1.3 Web Structure Mining

The association of the content of the web where a proper format of the structure is defined can be called as Web Structure Mining. Web Structure Mining is thus defined by HTML commands from formatting within a page and also defined by hyperlinks between different pages (Lee et al. 2004, p. 131)

. It proves very crucial to keep an overview of the website by understanding the concept relation between contents and the structure of the website.

2. Web Usage Mining Techniques

As mentioned before different data mining techniques like clustering, classifications, organization rule, consecutive pattern mining, and statistical analysis techniques are descried as below (Hartigan et al. 1979, p. 102).

2.1 Clustering

When any population of events or running items with similar sets of features are partitioned a cluster is formed. Two main types of cluster discovered in web usage mining are pages cluster and usage cluster. These cluster are the prime clusters type since they are the most interesting clusters of web usage mining. Number of clustering techniques is applied to discover subsets of web pages in a log files that are required for connection to enhance the performance of already connected pages (Hartigan et al. 1979, p. 102).

2.2 Statistical Analysis

The process of implementing statistical methods on web log files which defines sessions and addresses user navigation aspects like session time and length of navigation path can be defined as Statistical Analysis. To predict some page or web document, which might be accessed, statistical analysis can be used to predict these outcomes.

2.3 Classification

When existing sets of events or transactions are divided into another predefined classes on the basis of some features, classification is said to be done (Mobasher et al. 1999, p. 21). This classification can be applied on group of different users following a similar navigation pattern in web usage mining. As a result a profile of users consisting of a similar class or pattern can be put together and new approaches for web page classification can be discovered. Index pages and content pages are two important categories of classification of web pages.

2.4 Organization Rule Mining

Attribute value that are occurring frequently than other values when put together in a given set of information can be called as Organization Rule Mining formally known as Association Rule Mining (Hartigan et al. 1979, p. 102). When certain pages are viewed together very often, Association Rule Mining can be applied in a web usage mining to show with in the same session by the same user which page is to be visited more frequently (Mobasher et al. 1999, p. 21). Similar pages can be identified and can be related through a similar navigation patterns can be used for web personalization, which is facilitated by Association Rule Mining.

2.5 Consecutive Pattern Mining

When a number of events or actions in a given set of time or other constraint are determined in a sequence, consecutive pattern mining is said to be done. Consecutive Pattern Mining is formally known as Sequential Pattern mining (Mobasher et al. 1999, p. 21). User’s future behaviour while visiting pages can be predicted in web sage mining through this consecutive pattern mining. Sequential Pattern Mining is also used by some Web Usage Mining or Analytical Tools such as SpeedTracer to get needed pattern which a user may require.

3. Data Pre-Processing

There are many prerequisites processes that need to be done before application of data mining techniques. This is required so that the data log files are made available for data mining. These web log files stores information like time logged in, session detail, date of request, URL (Uniform Resource Locator), etc. Data cleaning, data filtering, path completion, session identification and session formatting make the main tasks of data pre-processing.

3.1 Data Cleaning

Pre-processing task is initiated by data cleaning. In any log file analysis, the items that are not important are eliminated. This can be achieved by checking suffix of the URL. This also includes ruling out the failed server requests.

3.2 Path Completion

This mainly does the auto completion in page references that are missing as a result of local browsing caching like when one presses back the previous visited page can be re visited.

3.3 User Identification

Local cache, corporate firewall and proxy server makes identifying unique user a very difficult task. A reasonable assumption would be each different IP address could be taken as a unique user. Heuristic will assume that there is a new user with the same IP address is the request page is not directly reachable by the hyperlink for any of the pages. Other hypothesis can be made when the same user requests the same page multiple times. Although, in some scenario it becomes very difficult to identify the user. For example if two different user uses the same device, and same browser and obviously same IP address since the same device and when both the users are requesting the same sets of pages, it becomes very difficult to differentiate the users (Coenen et al. 2000, p. 78).

3.4 Session Identification

In one visit to the same website in a defined interval of time, when the same user visits the set of pages a session identification can be made. Timeout threshold can be used to identify user sessions. For example if time exceeds 30 minutes then it can be assumed that the user has started a new session. One other assumption can be made is the same user visits sets of pages within the time constraint then it can be considered as a single session (Coenen et al. 2000, p. 78).

3.5 Session Formatting

This is the final pre-processing procedure, which is used for formatting transactions or sessions. Since all the sessions and transaction after completion needs to delete the log files for security purpose session formatting is done in the end.

Web Usage Mining

An important aspect for adaptive websites where all the data will be derived from for the purpose of generating recommendations. Adding links and changing the layout of the pages usually achieve usability. Fu *et. Al*, 2001 discusses about organizing the pages in a websites for users to find the data or information they need with minimum efforts. [Perkowitz and Etzioni, 1998] defined effort as a function of the number of links traversed and the difficulty of finding these links in the webpage. When reorganization of a web page is required, access patterns data is foremost mostly relevant and hence extracted. Index and content pages are than generated depending on the Characteristics and access information. This continues for the whole website and finally a reorganization is developed. [Srikant and Yang, 2001] proposed and algorithm which will automatically locate pages in a webpage when user do not find them in the place they expected it to be. This is added as a new link based on the user's interest and patterns. [Spiliopoulou and Pohle, 2001] described improving a website performances using data mining Techniques.

CHAPTER IV: METHODOLOGY

As part of this research, the Hill climbing search algorithm is used to obtain a feasible solution to using AI to improve the users experience. The implementation of the Hill Climbing algorithm under the IntWEB framework to adjust accordingly and meet users requirements is considered as a ‘search problem’. The implementation of this framework to adapt websites suggests that as soon as all improvements are made to adapt the website IntWEB will stop (stopping condition).

For the arrangement and as discussed, the Hill Climbing Search has the capabilities to provide varying solutions. Hill Climbing is a method in AI that can be used to take care of issues with numerous conceivable arrangements and where every one of these arrangements makes up the entire search space.

The main components of a search problem on the IntWEB framework

- States of the website

Every state in the framework is the complete webpage. Every links are taken to be a part of the optimization to achieve the improved user experience. The initial configuration also known as the initial state of the website being adapted using the Hill Climbing algorithm includes no links (Perkowitz et al. 2000, p. 155). In addition, in order to adapt the website through the Hill Climbing technique, links are added one by one to meet the user’s preferences.

- Actions to adapt the website

The action under the Hill Climbing algorithm and the IntWEB framework is implemented as transformation functionality. With the initial state the first configuration transformation to adapt the website using the Hill Climbing algorithm is achieved through the *add_shortcut* action. As a transformation functionality, the add_shortcut action produces a new state (S') from an existing state (S). The new state is adapting by implementing links to the webpage of the website.

- Evaluation Functionality

According to the Hill Climbing algorithm of adapting websites, webpages that are accessed in a single session are link to one another. The evaluation function is a method used to determine the varying sessions that a particular state's link Rate. This can also be considered as the goodness of a state in relation to the particular search problem in account. As a measurement, the states that show a strong link configurations are assigned a higher score level. The evaluation function is also applied when a state's link configuration is developed overtime as the number of paths that are included in a state increase. (Perkowitz et al. 2000, p. 155)

- Stopping Condition

The stopping condition for the Hill Climbing algorithm is IntWEB will be halted as soon as no more improvements can be made to a website.

The main components of the IntWEB framework

- Pre-processor

The Pre-processor module of the IntWEB framework yields sessions from log file documentations. These log files have accessibility to varying users pages and sessions sets of pages that are acquired together, as yield. Firstly, this pre-processor module eliminates any irrelevant log data such as any access information that is not relevant so as to clean the log file. Secondly, the module performs a user identification so as to recognize the users IP address and their respective browser information on the cleaned file. Hereafter, in the third phase of the pre-processor module, performing a re-process of the cleaned log file and here with the user information sessions are identified.

In the IntWEB pre-process module, the output is the session list that is created for the user session identification and is used as the input for the evaluation functionality in the framework.

- Web Crawler

In This module states are developed. The web crawler component of IntWEB processes the websites webpages and develops states. The states are created as the Web crawler module has the ability to automatically extract a website topology.

- Web Adapter module

This is the most crucial module for adapting websites. It develops a list of possible adaptation links for the websites and webpages to meet users specifications. Here the output list presents several link suggestions that can be configured. Each link suggestion consist of 3 main components namely the link from, the link to and the number of sessions that the state link configuration has.

These values are depicted in the following format:

<Li_to, Li_from, session_Li>

The Web adapter modules main goal is to maximize the goodness of a state or the number of sessions that a state's link configuration has. It implements and runs the Hill Climb algorithm along the entire state space. Through the use of the evaluation function, the Web adapter uses the pre-processor model session list to enhance the number of sessions and the state configurations.

The Web adaptor consists of two main modules usage mining module and adaptor module.

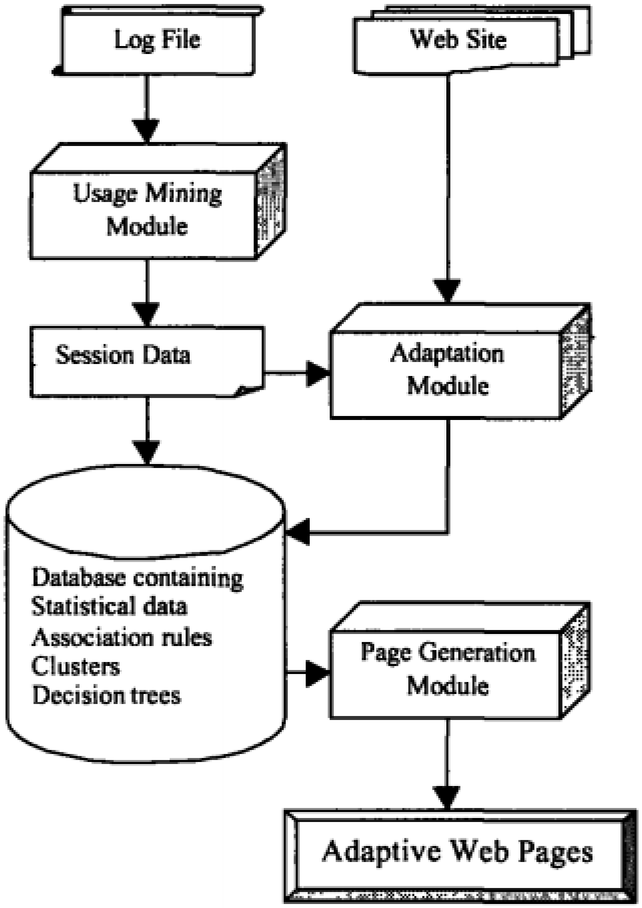

Figure 6: Sampling Model of a working Web Adapter

Usage mining module

All the log files are cleaned and identified before the algorithm can be applied. Usually web server logs consists information of all access data. This data can be analysed in multiple ways to extract the information required.

```
fcrawler.looksmart.com  -  -  [26/Apr/2000:00:00:12  -0400]  "GET  /contacts.html
HTTP/1.0" 200 4595 "-" "FAST-WebCrawler/2.1-pre2 (ashen@looksmart.net)"
fcrawler.looksmart.com  -  -  [26/Apr/2000:00:17:19  -0400]  "GET  /news/news.html
HTTP/1.0" 200 16716 "-" "FAST-WebCrawler/2.1-pre2 (ashen@looksmart.net)"
```

It consist of the IP address or domain name of the individual user, authentication, time stamps, host, HTTP request, URL, transaction code, and bytes transferred. This log also consists of many redundant data. Cleaning of these data removes the unrelated information such as for example image details are irrelevant.

Once the data is cleaned, sessions are identified. Using association rules and clustering algorithm, patterns and relationships are identified from these log files and categorized into meaningful sessions. IP address, page, time stamp are used to classify a single sessions. Single IP denotes an individual users, different IP means different users. But there may be instances different people utilizing same IP, browsing and software information are analyzed to further define the access.

The association rules are then identified to define relationships among these sessions. Session support for all elements above a threshold level minimum support are identified and the association rules are defined based on these.

Input is a unique set X of n unique URLs in the log file where

X ={url1, url2,……urln} and

Set Y of m user sessions where

Y = {Y1, Y2, …Ym}

For a set of URLs x $\subseteq$ X , it is stated as

Support = $= \frac{|\{y \subseteq \mathrm{Y}: x \subseteq y\}|}{|\mathrm{Y}|}$

The Algorithm

As per the Hill Climbing algorithm, AI can be implemented to find a range of varying possible solutions to a particular problem. The range of possible solutions together consists of the entire search space.

At the initial state, where there is no feasible solution to the problem, the action add_shortcut makes transition to the feasible states. Dependent on the evaluation function the closest to best state is chosen until the existing state improvement has been maximized.

Pseudo Code shown below for the Hill climbing search algorithm consists of two sections namely the GENERATE_NEIGHBOURS function and the evaluate function (EVAL(x)).

```
PresentNode = FirstNode;
Loop do
    M = GENERATE_NEIGHBOURS (PresentNode);
    nextRate = -INF;
    nextNode = NULL;
    for all x in L
        Rate = EVAL(x);
        if (Rate > nextRate)
            nextNode = x;
            nextRate = Rate;
    else if nextRate <=EVAL(PresentNode)
```

```
        return PresentNode;
    PresentNode = AfterNode;
```

The GENERATE_NEIGHBOURS function implements the *add_shortcut* actions to adapt the website. Here, a list of the potential states and link configurations between webpages are produced. When a possible link from page 1 to page 2 has been selected, it is added through the use of the evaluation function ((EVAL(x)). Hereafter, the next potential links for page 2 to other pages are generated. When this happens, the best possible link from page 2 to page 3 is taken into consideration. The process is then continued till there is a list of links for all webpages of the website. Once complete, the Hill Climb search algorithm is be rerun from the beginning from the initial state to determine whether there are any possibilities of a link that consists of an addition to an alternative link. During this algorithm run, page 1 is not used as the source link.

The evaluation function (EVAL(x)) measures the strong webpage link configuration. Below is the pseudo code for the evaluation function. This enables users to link directly. However this pseudo code shown below for the evaluation function is dependent on the number of pages that need to be transverse before leading to a specific webpage. For example, a link from page 1 to page 3 needs to be transverse page 2. Through this code, a link can be proposed between page 1 to page 3 directly. Each individual session is considered as a state. The path_length_threshold acts as a measurement for the evaluation function to only evaluate webpage link paths that meet the threshold.

```
Rate = 0;
for each $L_k$ <from_$L_k$, to_L > current state **S**
for each session **s** in session_Li
        if (( **s** contains path from_$L_k$ to to_$L_k$ ) and
          (path_length (from_$L_k$ to to_$L_K$ ) is **s** not exceeds
               path_length_threshold))
          Rate = Rate + path_length (from_$L_{k,}$ to_$L_k$ );
Return Rate;
```

(EVAL(x)) defines every session that a state has including the path that exist , the cover value is calculated and returned. For example, considering 3 webpages w1(2), w2(4) and w3(7), number in the bracket denoting the number of sessions which each page has. W1 is allocated as from which all possible states are generated which will have w1 as *from field* and wi as *to field* . (EVAL(x)) generates the rate value of all the existing states. From here with w2 as source all states without w1 as destination are calculated and this loop continues until all pages are completed.

Evaluation

This program has been evaluated using dummy data and not real time data hence it is tough to state the effectiveness of it currently. During this the path_length_threshold is fixed at a different value to see the affects on the results.

From the dummy data set, the server web log consist of the IP address, authentication credentials, timestamps, HTTP request, response, transaction size, URL and browser identifier. The sessions are extracted from these dummy log files. These logs are cleaned and removed of unnecessary and redundant data such as robot access, error logs, images files. There are about 20000 access approximately. After cleaning up we are left with about 5000 access. From these dataset about 110 sessions are identified. These sessions are categorized based on the number of request it has. In the figure X is the amount of sessions whereas Y represents the number of request per sessions.

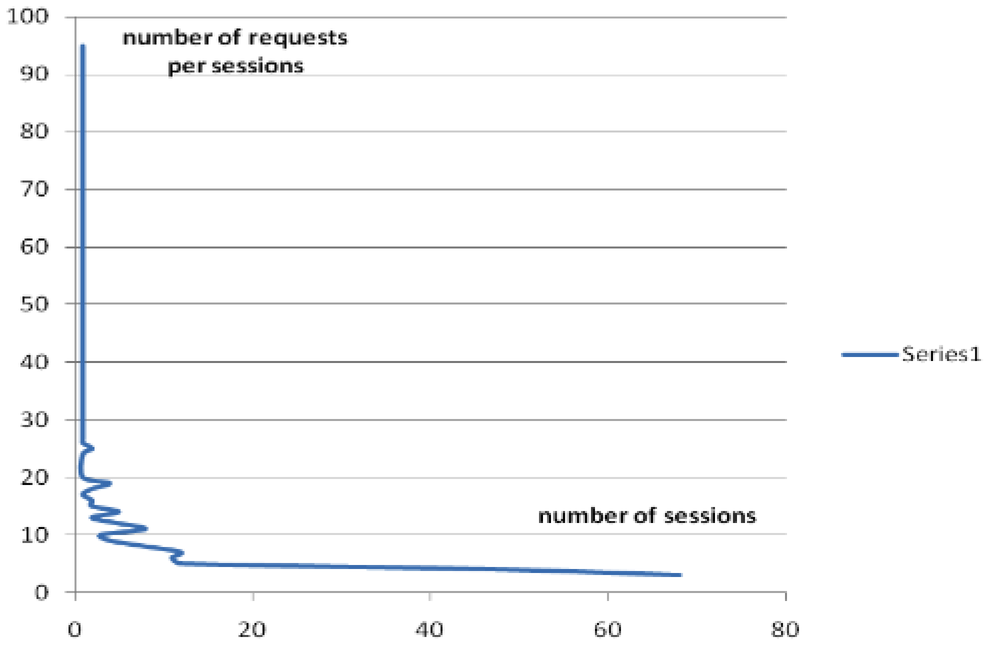

Figure 7: Request verses Sessions

From the graph it is evident that number of sessions with more than 20 requests are lesser than number of sessions that has less than 20 requests and it forms about 60% of the data.

CHAPTER V: RESULTS

Evaluation of the adapting list with different path_length_threshold values, we take the value for the threshold to be as 2, 3, 5, 7 and 9. As per the result, X-axis denotes pair of request where links needs to be generated. Y-axis denotes the amount of session this pair is present for or in other word it denotes the session_cover value. We observe that with increase in path_length_threshold, there is increase in session_cover also, as there is an increase in the possibility of the existence of the pairs. Other observation is length of adaptation list also increases with increase in the path_length_threshold. This is expected as the number of pairs linked is increased allowing longer sequence of request to exist.

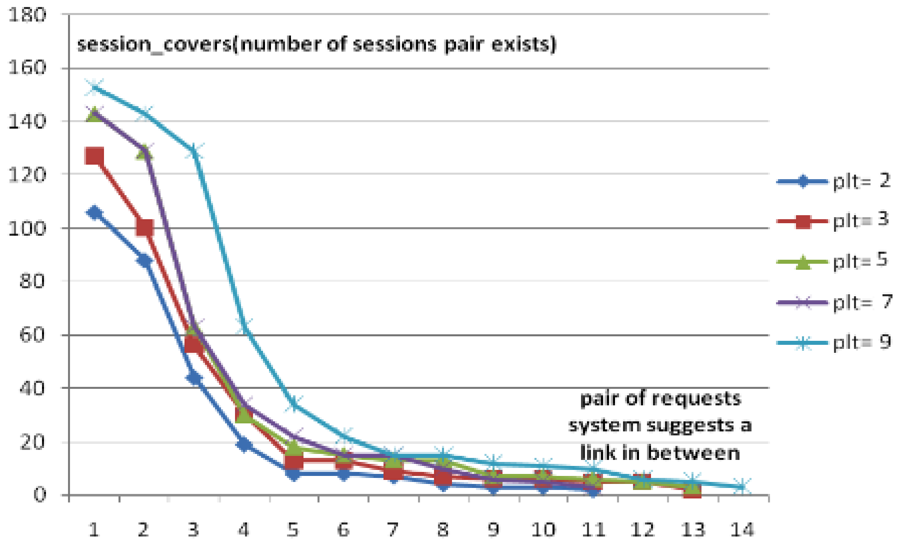

Figure 8: Cover value verses Path_length_threshold

The adaptation list for path_length_threshold = 7 is given. This produces about 8 adaptation lists as shown, which is mainly adding of links between pages for navigations. Upon analysis these adaptation are beneficial and logical. Although the recommendations are obtained, it is still a big task to determine if this will be implemented or rejected leaving it to the human administrator for now or to an algorithm that will weigh between each of the recommendations for the advantage and applies the changes based on these advantages over the other. At this point we can conclude that the system does produce feasible results, which make sense from a web's perspective, but it is still unclear if a fully automatic system that adapts and evolves without any human interaction can be developed. It is worth mentioning that in the evaluation , the website used does not contain any pages and all data s are generated as required. This data does not have many required information. Hence n the real time simulation, where more richer information is available, it is expected to provide more enhanced results.

FROM PAGE	TO PAGE	Rate
/Course/?sem=2000&course=ceng111&cedit=0	/Grades/grades_info.php	139
/Course/?sem=2000&course=ceng190&cedit=0	/Grades/grades_info.php	110
/Student/assessment.php	/Grades/grades_info.php	87
/Course/?sem=2000&course=ceng220&cedit=0	/Student/assessment.php?hid=1099	52
/Grades/grades_info.php	/Student/assessment.php?hid=2011	31
/Grades/grades_info.php	/Course/?sem=2000&course=ceng111&cedit=0	20
/Student/assessment.php?hid=1099	/Student/assessment.php?task_assessment=list&selector_homeworksassessment_course=ceng32	11
/Student/assessment.php?hid=2011	/Student/assessment.php?hid=1099	4

Figure 9: Sampling Adaptation Result

CHAPTER VII: CONCLUSION

Adaptive website is a very topic and a complex one. Adaptive websites with AI is even more complex where all aspects of AI come into play to achieve the outcome. This involves understanding human behaviours and actions. Also it has certain parts where predicting the outcomes becomes necessary to achieve a particular task. There are various questions concerning adaptive websites. This study is focussed to find an answer if such a system makes suggestions that are logically true and that leads to the improvement. Although in the vast domain of adaptive websites and AI, this study is a very small part that tries to understand the underlying concept behind the adaptive web and its functionalities. It also highlights the factors that play a role in such a system. Through the study, from the results it is observed that such a system as a standalone make logically correct improvements. But the implementation of this still requires far more research and developments. Achieving a truly self-evolving websites in a true sense is still a far go but with new developments and technologies coming up, it will definitely take us faster to the time when this becomes a reality.

REFERENCES

Alkan, O.K. and Senkul, P., IntWEB: An AI-Based Approach for Adaptive Web. In *Workshop chairs* (p. 57).

Raskin J. The human interface, first ed. Menlo, CA: Stratford Publishing, Inc.; 2000.

Perkowitz, M. and Etzioni, O., 2000. Towards adaptive web sites: Conceptual framework and case study. *Artificial intelligence*, *118*(1), pp.245-275

Perkowitz, M. and Etzioni, O., 1998, July. Adaptive web sites: Automatically synthesizing web pages. In *AAAI/IAAI* (pp. 727-732).

Perkowitz, M. and Etzioni, O., 2000. Adaptive web sites. *Communications of the ACM*, *43*(8), pp.152-158.

Kilfoil, M., Ghorbani, A., Xing, W., Lei, Z., Lu, J., Zhang, J. and Xu, X., 2003, May. Toward an adaptive web: The state of the art and science. In *Proceedings of Communication Network and Services Research (CNSR) 2003 Conference* (pp. 108-119).

Coenen, F., Swinnen, G., Vanhoof, K. and Wets, G., 2000, August. A framework for self adaptive websites: tactical versus strategic changes. In*Proc. of the Workshop on Webmining for E-commerce: Challenges and Opportunities (KDD'00)* (pp. 75-80).

Jacobs, N., 1999. Adaplix: Towards adaptive websites. In *In Proceedings of the Benelearn'99 workshop*. (Jacobs 1999)

Kosala, R. and Blockeel, H., 2000. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), pp.1-15. (Kosala et al. 2000, p. 10)

Mobasher, B., Cooley, R. and Srivastava, J., 1999. Creating adaptive web sites through usage-based clustering of URLs. In *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on*(pp. 19-25). IEEE. .

Hamilton, H.J., Wang, X. and Yao, Y.Y., 2001, May. WebAdaptor: Designing Adaptive Web Sites Using Data Mining Techniques. In *FLAIRS Conference* (pp. 128-132).

Mobasher, B., Jain, N., Han, E.H. and Srivastava, J., 1996. *Web mining: Pattern discovery from world wide web transactions*. Technical Report TR96-050, Department of Computer Science, University of Minnesota. (Mobasher et al. 1996)

Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), pp.100-108.

Lee, J.H. and Shiu, W.K., 2004. An adaptive website system to improve efficiency with web mining techniques. *Advanced Engineering Informatics*, *18*(3), pp.129-142.