QUALITY ANALYSES AND IMPROVEMENT FOR FUZZY CLUSTERING AND WEB PERSONALIZATION

AMIR KETATA

A THESIS

IN

THE DEPARTMENT

OF

COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

CONCORDIA UNIVERSITY

MONTREAL, QUEBEC, CANADA

MARCH 2009

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-63244-4
Our file Notre référence
ISBN: 978-0-494-63244-4

ABSTRACT

QUALITY ANALYSES AND IMPROVEMENT FOR FUZZY CLUSTERING AND WEB PERSONALIZATION

AMIR KETATA

Web mining researchers and practitioners keep on innovating and creating new technologies to help web site managers efficiently improve their offered web-based services and to facilitate information retrieval by web site users. The increasing amount of information and services offered through the Web coupled with the increase in web-based transactions calls for systems that can handle gigantic amount of usage information efficiently while providing good predictions or recommendations and personalization of web sites.

In this thesis we first focus on clustering to obtain usage model from weblog data and investigate ways to improve the clustering quality. We also consider applications and focus on generating predictions through collaborative filtering which matches behavior of a current user with that of past like-minded users. To provide dependable performance analysis and improve clustering quality, we study 4 fuzzy clustering algorithms and compare their effectiveness and efficiency in web prediction. Dependability aspects led us further to investigate objectivity of validity indices and choose a more objective index for assessing the relative performance of the clustering techniques. We also use appropriate statistical testing methods in our experiments to distinguish real differences from those that may be due to sampling or other errors. Our results reconfirm some of the claims made previously about these clustering and prediction techniques, while at the same time suggest the need to assess both cluster validation and prediction quality for a sound comparison of the clustering techniques.

To assess quality of aggregate usage profiles (UP), we devised a set of criteria which reflect the semantic characterization of UPs and help avoid resorting to subjective human judgment in assessment of UPs and clustering quality. We formulate each of these criteria as a computable measure for individual as well as for groups of UPs. We applied these criteria in the final phase of fuzzy clustering. The soundness and usability of the criteria have been confirmed through a user survey

ACKNOWLEDGMENTS

I am deeply indebted to my supervisors Dr. Sudhir P. Mudur and Dr. Nematollaah Shiri to whom I express my sincere gratitude and appreciation for their scientific guidance, and exemplary training throughout the course of this study. They have accepted me in their research team, introduced me to the world of web mining, shared with me their invaluable information and experiences, and tolerated my human imperfectness with inexhaustible patience and warm kindness. I appreciate the generous support they provided through their grants from Natural Sciences and Engineering Research Council (NSERC) of Canada and from SRT by Concordia University.

I am grateful to the Department of Computer Science and Software Engineering at Concordia University for the opportunity to work and use the lab facilities.

I am most particularly grateful to the anonymous respondents to my user study evaluation questionnaire, which forms an important part of this thesis.

I would like to thank all my friends and colleagues at Concordia and elsewhere, and in particular Abdallah Almaksoor, Rami Kassab, Mustapha El-Bouhssini, and Ahad Tahir for their friendship and moral support.

I cannot forget thanking my sisters Hanene and Hajer, and my brother Naoufel who has been my roommate, my friend and the most patient person I often disturbed and turned to in difficult times during the past two plus years.

Last and most importantly, my love and deep gratitude go to my parents Serra and Habib Ketata for their constant care, encouragement and continuous support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAD: Average Absolute Deviation

ARCA: Any Relational Clustering Algorithm

CARD: competitive agglomeration-based clustering

CF: Collaborative Filtering

FCM: Fuzzy C-means

FCMdd: Fuzzy C-Medoids

FH: Fuzzy Hybrid CF

LFCMdd: Linearized FCMdd

MAE: Mean Absolute Error

MB: Model-Based CF

MPC: Modified Partition Coefficient

PLSA: probabilistic latent semantic analysis

PC: Partition Coefficient

PD: Personality Diagnosis

PE: Partition Entropy

RFCMdd: Robust FCMdd

RFSC: Relational Fuzzy Subtractive Clustering

SHT: statistical hypothesis testing

UP: Aggregate Usage Profile

WAVP: weighted average visit per page

WPE: Windham Proportion Exponent

VI: Validity Index

XB: Xie-Beni Index

LIST OF SYMBOLS

e: convergence threshold for the clustering technique ARCA.

N_u: Number of pages in the website.

N: Number of sessions in the sample/population.

Url_i: the string defining the path from the root node to the i^{th} page.

V: The sessions matrix.

Vij: The occurrence of the j^{th} page in the i^{th} session.

μ_{ij}: The membership of the j^{th} session to the i^{th} cluster.

R_{ij}: The dissimilarity (distance) between the i^{th} and j^{th} sessions.

C (also C_n): The number of clusters in a clustering result.

m: is a weighting exponent (fuzzifier). Normally $1 \leq m \leq 2$.

$f_j^{\,a}$: The frequency of the j^{th} page in the a^{th} usage profile.

H_0: Null hypothesis: Saying there is no difference/correlation, etc. between observations.

H_a: Alternative hypothesis: the opposite of the null hypothesis.

s: For RFCMdd, a number of objects most likely to be real members (and not outliers), chosen based on how many objects one would like to disregard from clustering.

p: For LFCMdd, a number of objects that have highest membership values to clusters.

MIN_CARDINALITY: The cluster cardinality threshold, below which clusters are discarded from clustering results.

1 INTRODUCTION

The widespread and ever-expanding use of the world-wide-web by millions of viewers and website managers is resulting in a huge amount of data offered and fetched for different purposes. The many uses of the Web relate to many different domains, news, science, education, sports, industry, advertising, e-commerce, and others. Some of the information is of a general nature, such as news and scientific material, and are therefore posted on well designed websites, but with only casual changes made to the site structure. Providers of other types of information, especially those of e-commerce, compete to attract and satisfy customers, by offering their users the product or service they likely want. This need has led to the development of web personalization systems [Goldberg et al., 1992] [Resnick et al., 1994] [Shardanand and Maes, 1995]. Indeed, these systems aim at offering users specific information on particular products or services that match their needs, based on their navigational behavior. The development of personalization systems is based on capture or collection of suitable web data, their preprocessing, and their analysis for extraction of usage patterns [Eirinaki and Vazirgiannis, 2003] [Mobasher, 2007]. Collaborative filtering based on clustering of web usage data has been a successful web mining tool for efficient prediction of users' demands. Users are clustered into groups of similarly behaving individuals, each group being characterized through a devised aggregate usage profile (UP). Clustering techniques were often empirically evaluated by well-known validity indices, by manual analysis of the UPs produced, and by their effectiveness as models for collaborative filtering systems. Based on the similarity of the active user's profile to a specific cluster represented by its center or UP, the system predicts pages of items matching the active user's preferences. The performance of such systems depends on ability to handle large amount of data in real time.

1.1 Motivations

Appropriate personalization systems enhance the user experience with the website, often by enabling customization of the navigation path for individual users. In some cases, however, one may alter the navigation in a dissatisfying manner, and lead to opposing results, which would make the website owner question the usefulness of integrating personalization into the website. Similarly, in forming user categories, through web usage clustering, misplacing a user into the wrong cluster can lead to a wrong personalization and could affect user's experience with the website. Therefore, choosing the best personalization system and the appropriate model needs reliable evaluation criteria and procedures. It is essential to properly design and test the functioning of personalization systems and their underlying usage modeling techniques.

The assessment of relative performance of clustering techniques has been often based on the use of several types of indices [Corsini et al., 2005], [Rhee and Oh, 1996], [Zhang et al., 2008]. However these indices suffer from certain limitations. While some depend on clustering parameters such as cluster number, others do not take into consideration the ultimate objective of clustering, and only consider how well the clusters shape and separate. Yet, others are not valid for certain conditions. It is therefore important to assess the relative merits, complementarity and weaknesses of the criteria being used for clustering evaluation.

Experimenters usually resort to sampling instead of working on whole data sets, to reach conclusions. While such a procedure is efficient in terms of cost involved, care must be taken to assess and minimize sampling errors that are likely to occur. For better comparison of different procedures or algorithms, it is essential that detected differences between such procedures be assessed uniformly, using standard statistical procedures [Groebrner et al., 2006], [Black, 2008]. Although this has been often overlooked in sampled web data handling, application of statistical testing will allow more objective comparison, and adds more confidence in conclusions drawn from experiments.

Aggregate usage profiles (UP's), derived through clustering for use in prediction systems are often evaluated on the basis of human subjective judgment. It is essential to minimize such subjectivity to increase confidence and efficiency in such systems. Devising comprehensive evaluation methods that use quantified tools for UP assessment will enable a sound judgment and automation of UP characterization and their potential use for evaluating the relative performance of underlying clustering procedures.

1.2 Contributions

Research in the web usage mining area has increased tremendously in the past few years, explained by the need to deal with the increasing number of web data and internet users. More efficient and more scalable techniques are being developed to satisfy the needs of both information providers and users. Yet, more attention still needs to be given to objective evaluation criteria, consistency of results and statistically sound procedures of evaluation. Our main contributions are the following:

1.2.1 Dependable Performance Analysis of Fuzzy Clustering Techniques

Fuzzy clustering [Roubens, 1978], [Bezdek, 1981], [Corsini et al., 2005] is a popular method for modeling web usage data with the objective of client segmentation, usage prediction, recommendation for web personalization [Mobasher et al., 2002], website restructuring [Yan et al., 1996], page pre-fetching [Schetcher et al., 1998] and other applications. With the number of fuzzy clustering techniques increasing, potential users obviously would seek information as to which method would best suit their specific needs. It is important to adopt dependable performance analysis techniques. We propose that analysis should include – (i) use of least biased cluster validity indices, (ii) task-based performance to provide more ground truth by coupling with an application and (iii) statistical hypothesis testing to ensure that stated results from experiments are not by chance.

We conducted a comprehensive set of experiments within the framework of a dependable comparative study of 4 fuzzy clustering techniques. As mentioned above, we take three measures to ensure dependability on results. First, we assess cluster validity indices with regards to objectionable index characteristics and determine and use the indices least biased towards parameters of a specific clustering technique. Second, realizing that clustering accuracy in itself is not sufficient to assess the true value and usefulness of a clustering technique, we base our comparison on end-use quality of clustering, *i.e.* prediction quality. Finally, because results are usually derived from experimental web data that are not only noisy in nature, but also prone to sampling and experimental errors, we use relevant statistical testing to discriminate between significant and non-significant results and achieve good confidence in the derived conclusions.

1.2.2 Dependable Performance Analysis of Web Prediction Systems

The Fuzzy Hybrid (FH) filtering technique has been proposed by [Suryavanshi *et al.*, 2005 b] to combine the good accuracy of Memory-based collaborative filtering and the good scalability of Model-based (MB) collaborative filtering. Their research experiments have shown very promising results in favor of FH. In this study, we compare two collaborative filtering systems, FH and MB, for their prediction quality under different testing conditions. First, we perform a conventional comparison using prediction measures F1, MAE and R, applied on 10 samples of different sizes. We observe that while such measures generally give similar results, the difference in criteria values in favor of FH gets smaller as sample size increases. Next, we conduct experiments to investigate the performance consistency of the two prediction systems in well-designed samples of size 1000, such that they have different degrees of bias towards one of the two techniques. We find that FH outperforms MB in the majority of samples. Further confirmation and confidence in our conclusions is realized through statistical hypothesis testing.

1.2.3 Goodness Criteria for Web Usage Profiles

We propose several improvements in evaluating (aggregate) usage profiles. Usage profiles (UP's), or more generally cluster profiles, centroids, classification vectors, aggregate usage model [Mobasher, 1999], [Nasraoui et al., 1999 a], [Joshi et al., 1999], [Mobasher et al., 2002] have been frequently used to judge the final quality of clustering but, the evaluation process was usually based on human subjective judgments. To the best of our knowledge, UP quality has not been used before to decide on the best number of clusters, nor has it helped in choosing the best clustering produced by clustering techniques. Moreover, there was no systematic methodology to automatically compare usage profile sets produced by different clustering techniques in any clustering comparison study. Surely, usage profiles provide more ground truth to clustering evaluators than validity indices as they carry more pertinent information. They are also used as an input to certain clustering applications, such as personalization. In this work, we define and use 4 measurable attributes of usage profiles, and show its application is selecting the best set of clusters from fuzzy clustering of web data. The four criteria defined for usage profiles are: distinctness, coherence, strength and coverage. We validate the significance and effectiveness of the proposed criteria through actual user feedback on UP characterization based on the 4 criteria. Although our experiments are mainly based on web logs, such a methodology will suit virtually any "profilable" data domain, including basket data and item rating data.

The focus in this thesis is not on the mechanics of clustering and prediction algorithms, but rather on their use as procedural tools, or black boxes, for web mining. Therefore we do not provide full details of those algorithms or the technical reasons for their performance.

1.3 Thesis Outline

The rest of the thesis document is arranged into 5 major sections plus Appendices and Bibliography. Chapter 2 provides the background and review of related work, encompassing a presentation of web usage mining, selected fuzzy clustering techniques and prediction systems, statistical hypothesis testing, and a review of UP studies. In Chapter 3 we develop a dependable performance analysis of fuzzy clustering techniques. Chapter 4 provides a dependable performance comparison of the two prediction systems, FH and MB. In Chapter 5 we present our UP quality criteria together with appropriate measures and a UP quality based cluster filtering. In Chapter 6 we develop overall UP measures and use them to compare two clustering techniques. Conclusions and future work are presented in Chapter 7.

2 BACKGROUND AND RELATED WORK

Communication among people witnesses a real technological revolution, especially with the advent of new web technologies that allow people to communicate instantly and interactively as if they were physically in direct face-to-face contact. Most of people throughout the world nowadays use the computer to chat or talk over the internet, exchange stories, images, and music at their wish, and at very affordable prices. In fact, certain services offered through the web, such as information search or e-commerce cannot be provided as efficiently through direct contact. Providers of such services try their best to attract customers by offering them the services or products they would most likely acquire. Obviously, cost considerations make necessary the automation of service provision, so that the service providers or vendors can instantly respond to the simultaneous requests of large numbers (millions) of customers. The present challenge for these service providers is not only to offer potentially good items or services, but to anticipate customers' requests for, or interests in, such products or services. This task is getting even more challenging as the amount of information or products as well as the numbers of clients are getting larger and larger. Enormous efforts have been made to distinguish between different web user behaviors and to anticipate their types and needs. Developments in web mining and web personalization form a major part of these efforts. Web mining aims at discovering non-intuitive patterns of web user and web item relationships. Web personalization aims to automatically adjust the content of a website to fit the needs and interests of the different visitors, each treated differently. In order to determine such specific information, a variety of approaches are proposed by e-commerce researchers and specialists. These approaches can be divided in three major categories depending on the type of information filtering adopted: rule based filtering; content based filtering, and collaborative filtering. The choice of the proper filtering algorithm is constrained by the information provided.

- In rule based filtering, decision rules are designed offline. Then, the features of a visitor are normally supplied through explicit online questioning of each visitor. These features are then injected into the rule based system as facts, and conclusions are derived as to the changes needed to adapt the website design and content according to the visitors' answers. Such rule based systems are often highly dependent on the nature of the website, they lack reusability, and require extensive rule design.
- In content based filtering, usage-independent features of each website item are manually or automatically extracted. The features of the active user are overall features of the items in which the active visitor has expressed interest. The similarity between a website item and the active user can then be calculated based on the similarity of their features. The most similar items to the active user are filtered to assure better personalization. Surely, the extraction of all required features of each item is expensive.
- Collaborative filtering (CF) is the main filtering approach considered in this thesis, and is therefore described in more detail in the following section.

2.1 Collaborative Filtering and Web Usage Mining

Collaborative Filtering (CF) makes use of the information obtained from community to help individuals better perform their tasks. An original CF system, known as *memory-based CF*, takes the information provided by an active user and performs an extensive online search in the whole community for best match(es) to help estimating (or predicting) the needs or intentions of that user. Such type of CF suffers from several shortcomings, one of which is the scalability issue. In the web domain, indeed, CF systems have to deal with large or huge datasets. Online-learning performed by traditional CF systems become more and more time consuming as the size of dataset increases. In web navigation, users generally dislike long waits for page display, which results in loss of web site visitors.

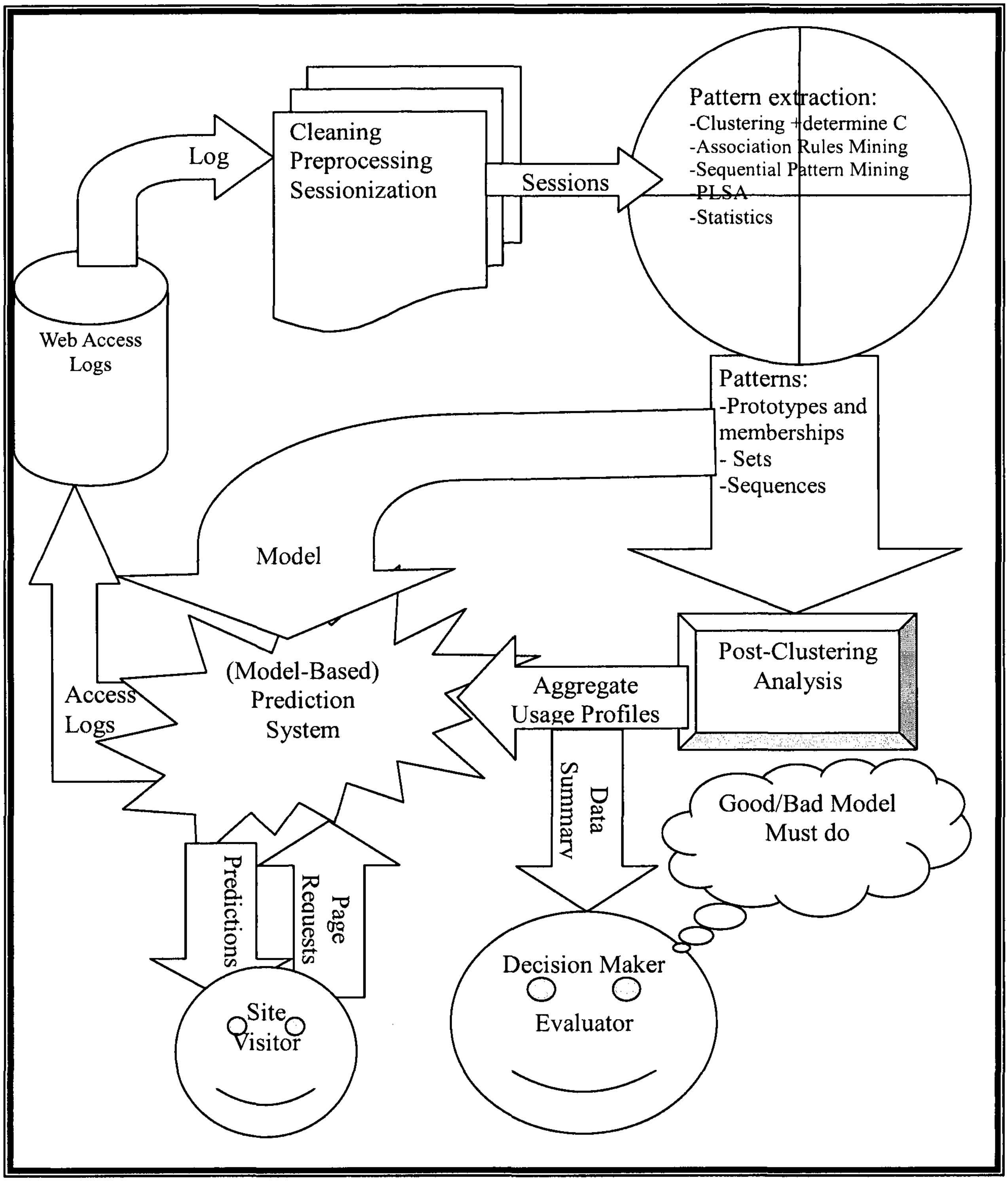


Fig. 1. Collaborative Filtering Life Cycle

This stimulated the emerging of a new brand of CF, named Model-Based CF, which makes use of Web usage mining to address such issues. Web usage mining is an effective tool for extracting meaningful knowledge out of the usage dataset, and delivering it as a summarized representation of the dataset. Traditionally, such knowledge can be used to make human decisions by visualizing it in the form of usage profiles or other representation. But also, model-based CF systems may deploy this knowledge in guiding the search that they perform. Such a process is divided for accuracy and simplicity into different stages.

We now describe the processing stages that we implement in our experiments in line with the general guidelines for model-based CF's, [Joshi et al., 1999], [Suryavanshi et al., 2005 a], [Cooley et al., 1999], [Nasraoui et al., 1999 a], [Nasraoui et al., 2002], [Suryavanshi et al., 2005 b]. This process is summarized in Figure 1.

2.1.1 Collection of Web Usage Data

Normally web usage data are log files recorded by web servers. Each log file includes among other information, the pages accessed, called pageviews, where pages are considered static, as assumed in the related literature. Following are the main information and descriptions of items normally provided in a pageview:

URL: the URI of requested file/object/item. They are in two main categories:

I. Complete file URI: showing the path and the name and extention of the file.

II. Folder URI: where the file is indicated by its folder. Expectedly the hidden name is '*index.html*'.

Date and **Time:** date and time of the pageview.

IP: IP of the user requesting the page.

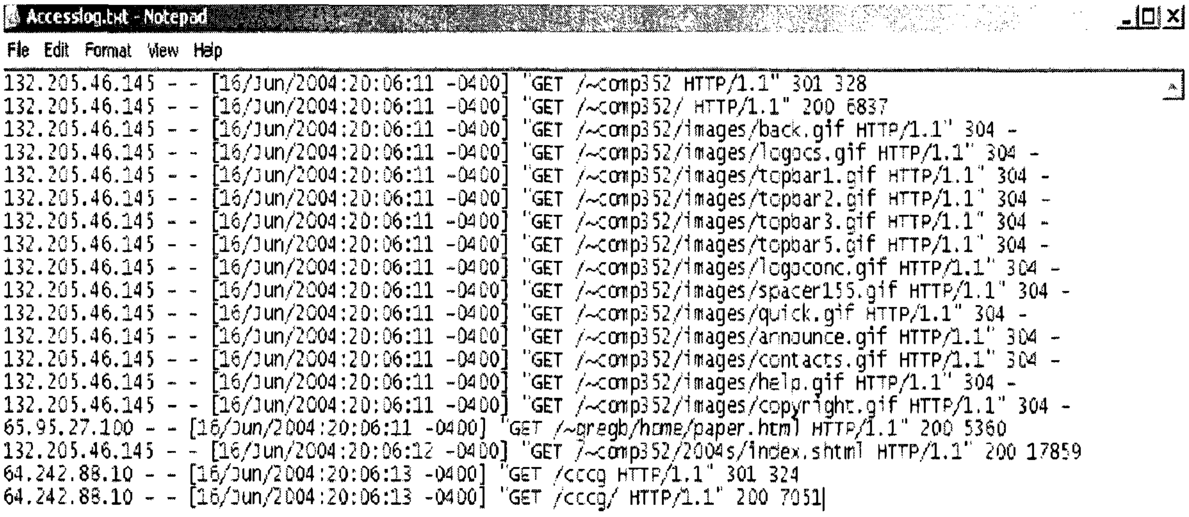
Additional, but less important information may include:

ID: The identification number of the pageview that uniquely identifies it from other pageviews.

Status: the server response code that indicates whether the request was successful.

Method: This is the fetching method used by the request. Potential values are: GET, POST, and HEAD. However, in a generally static website, most requests use the method GET.

Version: the protocol used in transfering web data, e.g. "HTTP/1.1".



```
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352 HTTP/1.1" 301 328
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/ HTTP/1.1" 200 6837
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/back.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/logocs.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/topbar1.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/topbar2.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/topbar3.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/topbar5.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/logoconc.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/spacer155.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/quick.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/announce.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/contacts.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/help.gif HTTP/1.1" 304 -
132.205.46.145 - - [16/Jun/2004:20:06:11 -0400] "GET /~comp352/images/copyright.gif HTTP/1.1" 304 -
65.95.27.100 - - [16/Jun/2004:20:06:11 -0400] "GET /~gregb/home/paper.html HTTP/1.1" 200 5360
132.205.46.145 - - [16/Jun/2004:20:06:12 -0400] "GET /~comp352/2004s/index.shtml HTTP/1.1" 200 17859
64.242.88.10 - - [16/Jun/2004:20:06:13 -0400] "GET /cccg HTTP/1.1" 301 324
64.242.88.10 - - [16/Jun/2004:20:06:13 -0400] "GET /cccg/ HTTP/1.1" 200 7051
```

Fig. 2. Sample web usage data

The input data used in our experiments in this thesis is web access log records from our CSE department. The number of distinct pages (and so the dimensionality) is 12,685. A sample of this dataset is presented in Figure 2, taken from [Suryavanshi, 2006]. For a period of 3 months, more than 500,000 pageviews were collected and processed.

2.1.2 Data Preprocessing (Cleaning)

Failed requests, web-crawler requests, non-page object requests, and root page (since most sessions include it) are removed from the data. This includes removing the images and other media files, the style sheet files, java script files and web application files. Further cleaning is performed after sessionization (the process of organization pageviews into sessions which is described in a later subsection). Cleaning may also include removing pages with very high and very low presence in order to reduce noise.

2.1.3 Pageview Similarity

In this work we adopt the URL similarity measure, based on page hierarchy structure, proposed by [Nasraoui et al., 1999 a] and adopted by e.g., [Suryavanshi et al., 2005 a], [Suryavanshi et al., 2005 b], [Azman and Ounis, 2004], and [Nasraoui et al., 2002]. The similarity between URLs url_i and url_j of two web pages is defined below by equation (1), where the range of i and j is from 1 to N_u, the total number of distinct URLs in the website:

$$SU_{ij} = \min(|url_i \cap url_j| / \max(1,\max(|url_i|,|url_j|) - 1),1) \qquad (1)$$

As url_i represents the path traversed from the root to the i^{th} page, and we consider it here as a set of the nodes in that path. So for example, if url_i = "/w/x/y/z", then we consider it as the set: {w, x, y, z} in this formula.

2.1.4 Sessionization

A *session* is viewed here as a set of pages accessed together by one visitor in an uninterrupted period of time. In our work, each session is a group of all pageviews from the same IP with successive pageviews accessed within a time period of 45 minutes. Consequently, we represent a session by a vector V_i of length N_u, where V_{ij} is the jth pageview and its value is 1 if the jth page is accessed in this session, and 0 otherwise. We use V as the representative array for all the sessions. Further cleaning is done by removing sessions with just one or two pageviews.

Additional useful information that can be obtained after identifying the sessions is the time spent on each pageview. For example, if t_1 is the access time of pageview V_{i1} and t_2 is the access time of the next pageview V_{i2}, then the duration of time spent on V_{i1} is t_2-t_1. This information may help estimate the interest by a visitor in a certain page. The representation of a session in this case needs to be modified to include the corresponding weights. We do not consider weights in our work.

From the data collected of the Computer Science Department, around 65,000 sessions were built.

2.1.5 Object Similarity

Our goal is to cluster sessions (the alternative would be "pages" in “page clustering”). The similarity between sessions *k* and *l* is defined as follows:

$$S_{kl} = \max(S\alpha_{kl}, S\beta_{kl}) \qquad (2)$$

This is the maximum of two quantities, one that takes URL similarity into account, and the other, called *cosine* [Miller et al., 2004] that does not. More explicitly,

$$S\alpha_{kl} = \sum_{i=1}^{N}\sum_{j=1}^{N} V_{ki} * V_{lj} * SU_{ij} \Bigg/ \left(\sum_{i=1}^{N} V_{ki} * \sum_{j=1}^{N} V_{lj} \right) \qquad (3)$$

$$S\beta_{kl} = \sum_{i=1}^{N} V_{ki} * V_{li} \Bigg/ \sqrt{\sum_{i=1}^{N} V_{ki} * \sum_{j=1}^{N} V_{lj}} \qquad (4)$$

where N is the number of sessions. The dissimilarity R_{kl} (distance) between sessions k and l is defined as:

$$R_{kl} = 1 - S_{kl} \qquad (5)$$

2.1.6 Usage Mining and Choice of Parameters

Various methods of mining are available to discover useful patterns for usage modeling and prediction, including association rules [Mobasher et al., 1999], sequential patterns, probabilistic latent semantic analysis (PLSA) [Jin et al., 2004] and clustering [Mobasher et al., 2002] [Lu et al., 2005] [Jin et al., 2004]. In this thesis, we focus on fuzzy clustering as a pattern discovery technique. In context of web personalization, the clustering may be classified into two approaches, described as follows.

2.1.6.1 Page/Item clustering

In the usage mining context, this clustering is performed by calculating the similarity between pages based on their similar usages. The more two pages are accessed together in same sessions, the more similar they are.

[Xu et al., 2006] proposed a recommender system based on web usage mining technique with Probabilistic Latent Semantic Analysis (PLSA). A main stage of the process was clustering of the website pages based on their access patterns. The similarity between pages was determined based on their common usage.

Other examples of item-based clustering include the well-known K-means [Ungar and

Foster, 1998], ROCK and agglomerative hierarchical clustering [Connor and Herlocker, 1999], and divisive hierarchical clustering [Kohrs and M'erialdo, 1999].

2.1.6.2 Session clustering

This approach aims to group sessions based on their similar pages, which we use in our work. Examples of these techniques include FCMdd (Fuzzy C Medoids), ARCA (Any Relational Clustering Algorithm), and RFSC (Relational Fuzzy Subtractive Clustering), which we will consider in our study. Other examples include K-means [Ungar and Foster, 1998], EM [Dempster et al., 1977], and divisive hierarchical clustering [Kohrs and M'erialdo, 1999].

2.1.7 Pattern Analysis and Evaluation

In this section we review existing methods for evaluating the cluster content. This allows us later to compare and contrast the criteria used to measure cluster quality with our proposed goodness criteria for usage profiles.

Evaluation of clustering results has been accomplished by a variety of approaches, classified as *external* or *internal*, depending on whether external input is required during evaluation or not.

Cluster validity index has been well studied as the main criterion in internal methods for web usage clustering [Wang and Zhang, 2007], [Perkowitz and Etzioni, 1998], [Bouguessa et al., 2006], [Rhee and Oh, 1996], [Zhang et al., 2008] emphasizing on intra and inter cluster distances, expressed as *compactness* and *separation* measures, respectively. A variety of crisp and fuzzy validity indices have been developed based on these two measures. A more detailed study of validity indices provided in Section 2.3.

Some researchers [Legany et al., 2006] have suggested categorizing clustering evaluation methods into three classes: internal, external, and *relative*. They consider both internal and external evaluations to have statistical basis, and consider validity indices as relative criteria.

Three external approaches for cluster evaluation have been reported [Crabtree et al., 2005], [Tonella et al., 2003]. They are the "Gold standard", the task oriented, and user evaluation. The "Gold standard" approach, also known as general ideal clustering, compares the candidate clustering to “ideal clustering” created and provided manually by an expert on the basis of the real structure and content of the data set. The two main criteria measured are *precision*, which measures how accurate the matching is between candidate clusters and ideal ones; and *recall* which measures the match relative to the ideal clusters. [Manning et al., 2008] used known measures reflecting the *recall* and *precision*, namely, F index, Purity, Mutual Information, and rand index. A new gold standard method was proposed by [Crabtree et al., 2005], called Quality and Coverage (QC4). QC4 is claimed to be valid across algorithms producing clusters with different granularity and hierarchy characteristics. It considers both *quality* and *coverage* as attributes of clusters which in turn are based on the notions of precision and recall. The four measurements defined are: average quality (AQ), weighted quality (WQ), average coverage (AC), and weighted coverage (WC). Traditional cluster evaluation measures such as F, Purity, and Entropy were considered but modified for clustering-specific characteristics to avoid biases. The Jaccard, Fowlkes-Mallows Index and Hubert's Γ Statistic [Gonzales, 2005] are other measures used for this purpose.

An example of task-oriented evaluation is measuring the quality of clustering results through applications such as web prediction systems [Ketata et al., 2009], [Mobasher et., 2002], [Bao et al., 2005]. Information protection is another example [Torra and Miyamoto, 2004] where information of clustered objects is masked by the general information content of the cluster (center).

User evaluation methods include those evaluations that involve implicit or explicit human judgment on the clustering [Zamir, 1999]. Well-known examples as described earlier are user evaluation of clustering-derived usage profiles [Mobasher, 1999], [Nasraoui et al., 1999 a], [Joshi et al., 1999], [Mobasher et al., 2002], which will be further described in Section 2.6.

2.1.8 Applications of Pattern Discovery: Prediction

Web mining techniques are deployed for in-depth analyses of web data with the objective of finding or unveiling potentially useful trends, correlations or patterns. These in turn are used for prediction as an end-product by users. Web prediction is used for different purposes, including page fetching [Schetcher et al., 1998], personalization and recommendations [Mobasher et al., 2002], and website restructuring [Yan et al., 1996].

[Herlocker et. al, 2004] reviewed various methods of collaborative filtering through which personalization systems are evaluated. They indicated that attention must be paid to the following aspects: user tasks, types of data sets and analyses, metrics related to accuracy and other traits, and user-based evaluation. Experimenting with a number of accuracy metrics, they were able to classify them into three distinct classes.

[Miller et al., 2004] addressed important limitations of recommender systems, related to portability, security, privacy, and lack of connectivity. They proposed a stand-alone architecture that works on palmtop computers, and is able to provide offline recommendations, store information locally, and encrypt data when synchronizing with the server. [Perkowitz and Etzioni, 1998] proposed a method to automatically generate an index page for each set of web pages based on identifying their co-occurrences.

In the next section we review the fuzzy clustering techniques selected for evaluation in this thesis.

2.2 Selected Fuzzy Clustering Techniques

Clustering is a statistical technique that divides a heterogeneous group of individuals (items, objects) into a number of subgroups or clusters of individuals such that individuals within a cluster are more similar to each other than individuals from other clusters [Halkidi et al., 2001]. The more distinct the groups are, the easier it is to decide into which cluster a new individual would fit. While different methods of clustering have been developed, some are called hard or crisp-clustering methods as they extract non-overlapping clusters, such that any individual belongs to one cluster only. In contrast, in fuzzy clustering, an individual may belong to the different clusters, with different degrees of membership, resulting in some cluster overlap. These methods are more appropriate to manipulate web data which bear an inherent fuzziness, typified for example by the varying rating of the same web page by the same user in the same period of time [Eirinaki and Vazirgiannis, 2003]. A crisp cluster is represented by the set of all objects belonging to that cluster, while a fuzzy cluster is represented by its prototype (cluster center or centroid, also medoid if it is a real object) and membership values of all objects in population to that cluster. Some of the fuzzy clustering algorithms used in the web domain include: Fuzzy C-Means (FCM) clustering [Bezdek, 1982], Fuzzy C Medoids (FCMdd) [Nasraoui et al., 2002], Any Relational Clustering Algorithm (ARCA) [Corsini et al., 2005], and Relational Fuzzy Substractive Clustering (RFSC) [Suryavanshi et al., 2005]. The word "relational" refers to the type of data handled through these methods. In clustering applications, the individuals or objects making up the original population are characterized for a number of attributes using assessment data, designated as feature or object data. Another type of data (like correlation) measures the pairwise relations among individuals or objects, and is referred to as relational data. Because of high dimensionality and correlation among web feature data, relational data are more suitable for clustering than object data [Nasraoui et al., 2000].

We now give additional details on the fuzzy clustering techniques that we will evaluate in experiments of this thesis. We remark that we will be dealing with these techniques almost as black boxes in our comparative study in Chapter 3.

2.2.1 Any Relational Clustering Algorithm (ARCA)

[Corsini et al., 2005] proposed a fuzzy relational technique based on the FCM algorithm and requiring no specific restriction on the relational matrix.

The new clustering technique, known as "Any Relational Clustering Algorithm" (ARCA) optimizes the following objective function where U is the membership matrix, and P the cluster center matrix:

$$J_m(U,P) = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}{}^{m} R_{p_i k}^{2} \tag{6}$$

where ∀ i, k: R_{Pik}, $u_{ik} \in [0,1]$.

In this formula, for each object x_k to be clustered, u_{ik} in [0,1] denotes its degree of membership to the cluster represented by prototype p_i, R^2_{ik} is the distance (dissimilarity) between ith cluster center(p_i) and x_k, and $m \geq 1.0$ is a weighting exponent (fuzzifier).

Given a specific number of clusters (C), (m), and initial membership values, ARCA runs iteratively, constrained by a maximum number of iterations, computes in each iteration the prototypes (cluster centers) then memberships using the formulae (7) and (8), and stops when there is no change higher than a certain threshold (e) in the membership matrix. This threshold is an input to ARCA defining when to stop the clustering process. The algorithm is summarized as follows:

```
The ARCA algorithm:
  1. Initialize membership matrix $u^0$. Set C and m.
  2. For l=1 to max_iterations
    2.1. Calculate the center vector $p_i$ of each i$^{th}$ cluster, Eq. (7).
     2.2. Calculate memberships of all objects to cluster i, Eq. (8).
     2.3. If the difference between $u^l$ and $u^{l-1}$ is less than "e": EXIT.
  EndFor
```

$$p_{i,k}^{(l)} = \frac{\sum_{j=1}^{N} u_{ij}^{m} R_{kj}}{\sum_{j=1}^{N} u_{ij}^{m}} \tag{7}$$

$$u_{i,k}^{(l+1)} = 1 / \sum_{j=1}^{C} \left(\frac{\delta_{ik}^{(l)}}{\delta_{jk}^{(l)}} \right)^{2/(m-1)} \tag{8}$$

$$\delta_{ik} = \delta(x_k, p_i) = \sqrt{\sum_{j=1}^{N} (R_{k,j} - p_{i,j})^2} \tag{9}$$

Our main motivations for choosing ARCA are that : (i) it is a version of FCM (Fuzzy C-means) method, which is most famous, with proven stability and high partition quality; (ii) it has been argued that ARCA is stable without any restriction on the square binary relation; (iii) it has shown good results on 4 benchmarks and 4 synthesized data sets; (iv) it outperformed two popular fuzzy clustering techniques, fuzzy non metric model [Roubens, 1978] and the assignment prototype [Windham, 1985] algorithm; (v) ARCA has a proven higher stability, scalability and convergence speed than non Euclidean relational FCM (NERFCM), one of the most reliable fuzzy clustering algorithms; and (vi) we wished to see how ARCA performs on web usage data and as a model for web prediction.

2.2.2 Fuzzy C Medoids (FCMdd)

This technique [Krishnapuram et al., 1999] aims at optimizing the function defined in Eq. (6). As suggested by [Krishnapuram et al., 1999], the formula shown in equation 10 is used for membership computation:

$$u_{ij} = \left(1/R_{P_i j}\right)^{1/(m-1)} \Bigg/ \sum_{k=1}^{C} \left(1/R_{P_k j}\right)^{1/(m-1)} \qquad (10)$$

FCMdd was chosen for our performance analysis for the following reasons: (i) it has wide applicability in the domain of web mining [Joshi et al., 1999], [Nasraoui et al., 2002]; and (ii) it has been proven to outperform several other mining techniques [Joshi et al., 1999], [Krishnapuram et al., 1999].

FCMdd chooses its best C by overspecifying an initial value for C, then after crispification, removes clusters with cardinality less than a threshold (we call it *MIN_Cardinality*). Crispification is simply performed by assigning each object to the closest cluster.

The original version FCMdd algorithm can be formally expressed as follows.

```
The FCMdd algorithm:
1. Initialize prototypes vector V. Set C and m.
2. For I=1 to max_iterations
  2.1 Calculate memberships of all objects to each cluster Eq.(10)
  2.2.Find the prototype $p_i$ of each $i^{th}$ cluster, such that the object k picked for the $i^{th}$ cluster minimizes the function:
$\sum_{j=1}^{n} \left( u^{m}{}_{ij} * R_{kj} \right)$
2.3. If there is no difference between previous and current V: EXIT.
EndFor
```

We used two versions of FCMdd: Linearized FCMdd (LFCMdd) which is more efficient and shown to be more accurate than basic FCMdd. This version, when choosing the new center candidate of a cluster in step 2 above, searches in a subset of size p < N of the total session set, for a subset of sessions that has the highest membership values to that cluster. The second version of FCMdd is designated as RFCMdd, and also called Fuzzy C-Trimmed Medoids by [Krishnapuram et al., 1999]. As reported in [Joshi et al., 1999], [Nasraoui et al., 2002], it is a robust version of the algorithm reported to produce a clustering less sensitive to noise than other FCMdd versions. It also chooses the *s* objects most likely to be real members (and not outliers), as prototype candidates.

2.2.3 Relational Fuzzy Subtractive Clustering (RFSC)

The Relational Fuzzy Subtractive Clustering (RFSC) technique has been recently developed [Suryavanshi et al., 2005 a] at the Department of Computer Science and Software Engineering at Concordia University. It is included in our study as it was claimed to be highly scalable to large web usage data. We have modified the original algorithm for choosing the best C. As in the case of FCMdd, we overspecify C, then prune out clusters for which the cardinality of crispified clusters is less than *MIN_Cardinality*. That indeed has given us better results.

The RFSC algorithm can be expressed as follows:

```
The RFSC Algorithm:
1. Calculate all objects potentials using Eq. (11)
2. For i=1 to C
  2.1.Choose the highest potential object as the center of the $i^{th}$ cluster, call it $p_i$;
  2.2 Calculate memberships of all objects to that cluster, Eq. (12)
  2.3 Subtract the potentials of each object j using Eq. (13)
EndFor
3. Crispify: Assign each object to its closest cluster.
4. Prune out low cardinality clusters.
```

where:

$$\mathrm{P}_i = \sum_{j=1}^{N} e^{-\alpha R_{ij}^2} \tag{11}$$

$$u_{ij} = e^{-\alpha R_{pi\,j}^2} \tag{12}$$

$$\mathrm{P}_j = \mathrm{P}_j - \mathrm{P}_{pi} * e^{-\alpha R_{pi\,j}^2} \tag{13}$$

and $\alpha = 4/\gamma^2$, with γ being the median value of dissimilarity across all pairs of objects, and R_{ij} the dissimilarity between objects x_i and x_j, and p_i as the cluster center (prototype) of the i^{th} cluster. e here represents the exponential function.

We now present an overview of well known cluster goodness metrics for fuzzy clustering used later for comparison in Chapter 3.

2.3 Validity Indices

2.3.1 Definitions and Examples

A validity index (VI) is a function that measures the goodness or validity of clustering. According to [Gath and Geva, 1989], a good clustering would have the following attributes: (i) evident separation of the clusters, (ii) minimal volumes of the clusters, and (iii) maximal number of data points positioned close to each cluster centroid.

A VI is used to assess how well the clustering separated the original set of items in well distinguished groups with a meaningful discrimination among the clusters that reflects the structure of the original data. Therefore, it can be used to compare the clustering performance of different clustering techniques. The VI can be also used to studying the relationship between the VI itself and clustering parameters in order to:

a. set the best values for such parameters, primarily the number of clusters (C), as used by [Rhee and Oh, 1996][Zhang et al., 2008][Kim et al., 2004 b]. This is referred to as the cluster validity problem [Zhang et al., 2008], or

b. to estimate the sensitivity of such VIs to clustering parameters, or

c. to estimate the sensitivity of the clustering performance to the parameters.

Some fuzzy indices VIs reflect the quality of fuzzy clustering. Other crisp indices are developed for crisp clustering, yet they may also be used for fuzzy clustering after crispification of the clusters.

Fuzzy Indices

A number of VIs belong to this category, but we will focus on those used in our studies. One of the most known of these is the Partition Coefficient (PC) introduced by [Bezdek, 1981] to reflect the extent of overlap among fuzzy clusters, defined as follows:

$$PC = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^{2} \qquad (14)$$

where μ_{ij} is the membership grade of the j^{th} object to the i^{th} cluster. Depending on membership grades, the PC values range between 1/C and 1, where C is the number of clusters. Large membership grades result in large PC, indicating low membership sharing among clusters, while small membership grades result in small PC and an increased overlap among clusters. Therefore, the larger the PC, the better is the clustering. However, this VI does not take into consideration the structure of the data (i.e., their dispersal or distribution) and therefore may not be a reliable indicator of clustering quality [Rhee and Oh, 1996], [Abonyi and Feil, 2007]. Other drawbacks attributed to this VI include: a monotonic negative relationship with C [Rhee and Oh, 1996], [Zhang et al., 2008], [Wang and Zhang, 2007], sensitivity to cluster overlap [Bouguessa et al., 2006], and failure to good estimate C as shown through several experiments in [Zhang et al., 2008].

In contrast to PC, the Xie-Beni (XB) index [Xie and Beni, 1991] considers, for clustering, both membership grades as well as the distance measure between objects.

XB is intuitively defined as the ratio of cluster compactness to separation:

XB = *Compactness / Separation*

With the numerator being equal to:

$$Compactness = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^{2} * R_{p_i j}^{2} \qquad 15)$$

where $R_{p_i j}$ is the dissimilarity between ith prototype and jth object.

Separation is defined as $min_{i \neq k}$ $R^{2}_{p_i p_k}$, $R_{p_i p_k}$ where is the dissimilarity between the ith and kth prototypes, i.e., the minimum of the distances between any pair of prototypes found. Small *Compactness* and large *Separation* indicate good clustering, as do small XB values. Therefore, smaller XB values are desired, reflecting a greater separation among clusters and/or their greater compactness.

Similar to Xie-Y index (discussed below), the XB index enforces the constraint:

$\sum_{i=1}^{C}\mu_{ij} = 1$. ARCA and both versions of FCMdd also impose this same constraint, but for improved noise detection purposes, RFSC does not. Hence, a normalized version of RFSC membership had to be calculated to enable the use of these indices.

XB is widely used for fuzzy clustering validation. However, new indices have been developed to avoid XB's detected weaknesses. [Bouguessa et al., 2006], [Rhee and Oh, 1996], [Zhang et al., 2008], and [Xie et al., 2002] mentioned a monotonic negative correlation of XB with C as C gets very large (close to N). Also [Kwon, 1998] mentioned XB's failure to validate the cluster centers and relative membership values for FCM algorithms when C gets large. It also failed to estimate C properly for certain datasets, even when the correct number of C is relatively small [Zhang et al., 2008], [Kwon, 1998].

'Xie-Y' is a VI proposed by [Xie et al., 2002], defined as follows:

$$\text{Xie-}Y = (C_{ns}/C) \times Compactness \times Separation \tag{16}$$

where C is the number of clusters, and C_{ns} is the number of non-singleton clusters (containing more than one object).

$$Compactness = C_{ns} \Big/ \sum_{i=1}^{Cns} \left[\left(\sum_{\sigma_{ij}} \mu_{ij}^{2} * R_{p_i j}^{2} \right) \Big/ \sum_{\sigma_{ij}} \mu_{ij}^{2} \right] \tag{17}$$

in which σ_{ij} is the predicate $\{X_j \in C_i,\ X_j \neq P_i\}$ where C_i is the i^{th} cluster, and P_i is its prototype (*i.e.,* its center).

$$Separation = \left(\sum_{i=1}^{C} \min_{1 \le j \le C,\, j \ne i} R_{p_i p_j} \Big/ C \right)^{2} \tag{18}$$

The larger the separation value, the better the clustering. The authors claimed that, unlike other indices, Xie-Y is a good indicator of the clustering quality even for large C.

Other validity indices include Partition Entropy (PE), Windham Proportion Exponent (WPE) [Windham, 1981], Modified Partition Coefficient (MPC) [Dave, 1996], Separation index (S) [Abonyi and Feil, 2007], Tang (T) index [Tang et al., 2005], P index

[Chen and Linkens, 2004], RFSC index [Suryavanshi et al., 2005 a], and FCMdd inter and intra cluster distances [Nasraoui et al., 2000].

Crisp indices

Crisp indices were primarily designed to evaluate the validity of crisp clustering. But they also may be used for fuzzy clustering after crispification.

Some of the well-known indices of this category include indices that are applicable to clickstream data such as Dunn index (DI) [Dunn, 1974] and, Davies-Bouldin index (DB) [Davies and Bouldin, 1979] [Legany et al., 2006].

More indices are presented in Appendix A.

2.3.2 Uses and Limitations

It is assumed that the performance of model-based prediction systems depends on how well the model captures the characteristic patterns of the target dataset. In using the model obtained via clustering, it is essential that the clustering operation is performed efficiently and correctly. There is a multitude of clustering techniques. This constitutes a challenge for practitioners and researchers to decide on the best clustering that would give a truthful picture of the variability or the pattern (structure) existing among the individuals or objects making up the whole dataset. First, there is a need to agree on what makes good a clustering algorithm, and devise a metric to measure such quality. This metric may be referred to validity index (VI), cluster validity index, cluster validation index, partition coefficient, etc. New names keep on appearing in the research arena with fine-tuning or brand new definitions of clustering quality. The second step is the application of the clustering algorithm to a dataset for an *a priori* chosen set of cluster numbers (C), the computation of VI for each clustering run, identifying the best VI value and taking the corresponding C value as the "right" number of clusters for that data. The increasing number of available VI's automatically leads to the question of "which one to choose." The relative performance of VI's may be assessed by subjecting the clustering results of a synthetic dataset with known structure (including C) to the various VIs being compared. This indeed has been done in several studies [Bouguessa et al., 2006], [Zhang et al., 2008], [Kwon, 1998]. The VI that leads to the identification of the correct C is the best VI for that clustering.

Thus many different VI's have been proposed each with its own definition, and characteristics, in turn require assessment of their performance/applicability. We will cite a number of studies dealing with the performance of various VI's. These are in addition to those we have presented in relation to partition coefficient (PC), XB index and Xie-Y index in the previous subsection. The emphasis in this section is on the uses, advantages and limitations of the indices. A list of reported validity indices, with related abbreviations and formulae provided in Appendix A.

[Gath and Geva, 1989] proposed a validity index, called F_{HV} that help estimate the best clustering parameters (mainly the number of expected clusters) for a newly-proposed clustering technique UFP-ONC. The experiments were conducted on a number of synthetic and real data (including Iris and Sleep EEG datasets). The hypervolume of this index was described as having an objective performance in comparison with others, such as DB index and PE (see Appendix A), as regards to cluster number and towards showing a clear extremum.

[Rhee and Oh, 1996] proposed another validity index I_G combining Compactness and Separation. They also proposed a new method to choose the best number of clusters (C) based on the large change in the index, rather than reaching a minimal or maximal value. The method does not depend on the clustering algorithm. They also reviewed several other indices including XB and PC that they compared with their approach. Their methodology was successfully tested on different datasets including butterfly data, Iris data, and some synthesis datasets.

[Kim et al., 2004 b] introduced a new validity index that measures the overlapping and separation between clusters. The index was compared and shown to be superior to indices PC, PE, XB, FS, K, SC, and others (see Appendix A for brief descriptions).

[Legany et al., 2006] reviewed several validity indices, namely Dunn [Dunn, 1974], DB [Davies and Bouldin, 1979], SD [Halkidi et al., 2000], S_Dbw [Halkidi and Vazirgiannis, 1996], RMSSDT [Sharma, 1996], and RS [Sharma, 1996], and compared experimentally their performance on FCM and K-means clustering techniques with "right" and "wrong" clustering configurations.

[Zhang et al., 2008] developed another validity index *W*, for validation of clusters recovered from fuzzyC-means clustering technique. The index is defined in terms of *variation* (the opposite to compactness) and *Separation.* Experiments results of comparing *W* to 9 previously known indices (PC, CE, MPC, FS, XB, K, F_{HV}, PBMF, and PCAES) showed superiority of *W* in terms of effectiveness, reliability and robustness in noisy environments.

Validity indices have been critically tested and reviewed for their effectiveness and consistent performance across a wide range of conditions and their independence vis a vis clustering input parameters, especially the number of clusters (C) and the fuzzifier *m*. Most comparative studies of fuzzy clustering techniques use cluster validity indices as a principal metric for evaluation. Yet, several of these indices have been judged inappropriate because of their dependence on clustering parameters, including the number of clusters [Bouguessa et al., 2006], [Oliveira and Pedrycz, 2007], [Xie et al., 2002], [Zhang et al., 2008], [Halkidi et al., 2002 b], [Gath and Geva, 1989].

A recent study [Hwang and Thill, 2007] compared fuzzy C-means clustering (FCM) with crisp clustering utilizing statistical hypothesis testing. The index used for comparison of cluster validity was the sum of squared errors (SSE). Our experience in this work has shown that SSE is highly dependent on C. [Corsini et al., 2005] compared ARCA to Non-Euclidean Relational FCM (NERFCM) [Hathaway and Bezdek, 1994] based on two validity indices: XB and Partition Coefficient, as applied on a number of widely different datasets. Critical reviews of these two indices were made in [Bouguessa et al., 2006] and [Zhang et al., 2008]. Studies of clustering algorithms RFSC [Suryavanshi et al., 2005 a], CARD [Nasraoui et al., 2000] and FCMdd [Krishnapuram et al., 1999] used specific validity indices for these techniques. Our experience with RFSC index shows that it is highly sensitive to C. [Pal and Bezdek, 1995] studied the influence of the input parameter fuzzifier *m* on important validity indices. [Kwon, 1998] indicated that the validity index K of the same author is not sensitive to *m*. [Wang and Zhang, 2007] carried out extensive experiments with various VI's in relation to fuzzy cluster validation for fuzzy C-means clustering. The validity indices included PC, PE, T, K, XB and F_{HV} (see Appendix A), and the comparison tested the ability of these indices to identify the correct C value for clustering of a variety of datasets.

Certain data structures presenting very atypical patterns may affect the performance of validity indices. Dunn index for example is known for its sensitivity to noise data. Robustness of the VI's has been investigated in noisy environments [Zhang et al., 2008], [Abonyi and Feil, 2007], or in situations of overlapping clusters [Bouguessa et al., 2006], [Kim et al., 2004 b], [Melegy et al., 2007].

In our experiments, we use 3 cluster validity indices and compare 4 fuzzy clustering techniques, for variable data set sizes and different cluster numbers.

2.4 Background on Statistical Hypothesis Testing for Web Mining

2.4.1 Statistical Hypothesis Testing and Significance Level

When experiments are conducted, the question often asked is whether the results are significant, e.g., they are not due to mere chance and specific to this data set. Statistical hypothesis testing (SHT) is conducted to shed light on such concerns. Such testing is based on probabilities and allows attaching a degree of confidence to the conclusions drawn from the experimental results [Groebrner et al., 2006], [Black, 2008].

In conducting experiments, we as experimenter have an idea in mind that would like to confirm or reject through experiments, for instance when comparing performance of two solution algorithms for the same problem, in terms of computation time, for instance. The idea is to test whether "a new algorithm" is different from a "traditional" one. For this, the experimenter formulates two possible situations, the first stipulating no difference between the two algorithms (this is called the null hypothesis, denoted as H_o), while the second may postulate a real difference between the two algorithms (this is called the alternative hypothesis, denoted as H_a). For H_a, the experimenter may be interested in any possible difference, *i.e.,* in any direction (and this is the case of a bilateral situation or test), or he/she may be interested in knowing whether the "new algorithm" is better or worse (time required is more or less) than the "traditional" one, which is the case of unilateral test.

The results of the experiments allow the experimenter to reach a verdict or conclusion, with a certain degree of confidence (usually 95% or 99%) indicating whether the null hypothesis H_0 can be rejected, in favor of H_a, for example to declare there is a real difference between the two algorithms in question (in case of a bilateral test). A degree of confidence of 95% corresponds to a significance level of 1- 0.95 = 0.05 or 5%. The significance level is usually denoted as α, and the confidence level as $100\ (1-\alpha)\%$.

2.4.2 Implementation of Statistical Hypothesis Testing (HST)

Briefly, the implementation of SHT procedure includes the following steps: (a) collecting sample or experiment data, (b) formulating the hypotheses, (c) plugging those data into a formula of a so-called "test statistic", (d) comparing the computed value of the test statistic to a theoretical value determined by the type of comparison or test being implemented, the sample size and the chosen significance level, and (e) based on that comparison, decide on the rejection (or non-rejection) of H_o. The components and dependencies of the SHT process are presented in more details in Appendix C.

The comparison of two (or more) techniques is generally based on the comparison of means, proportions, medians, variances or other statistical parameters of the compared processes.

2.4.3 Parametric and Non-Parametric Tests

In general, some assumptions are made with respect to the data generated by the processes (population or sample data of the characteristic being used to quantify the performance of the processes, say, the "time required to do the task"). The assumptions may include the normal distribution of the characteristic, and the homogeneity of variance of that characteristic for the techniques being compared. Such assumptions are required for SHT involving means, variances, proportions, correlation or regression coefficients or other statistical parameters. The related tests are described as parametric tests, in contrast to another class of tests, called non-parametric tests, usually based on ranks of data arranged in ascending or descending order. Non-parametric tests have the advantage of not requiring the usual assumptions of normal distribution or homogeneity of variances, but they are less powerful than parametric tests for, e.g., requiring larger differences to reach significance.

In the Web Mining domain, the evaluation of a technique is mainly based on empirical tests; which therefore requires the application of statistical testing procedures to reach acceptable conclusions.

In the domain of data mining, SHT was utilized for three main objectives. The first is to validate a mining technique by verifying a null hypothesis that says the technique provides no valid patterns [Halkidi et al., 2002 a]. The second is to evaluate the correlation between two variables [Jahanian et al., 2004]. In this context, Pearson correlation coefficient and Spearman rank correlation coefficient were normally deployed. The third is to estimate the significance of the difference between two techniques [Ketata et al., 2009], which is what we will be mainly using in this thesis. After conducting the comparative experiments, SHT is used to ensure that the observed difference in the results is not due to mere chance, i.e., to sampling errors, but in fact it is an indication of a real difference between the methods that are being compared. Different statistical testing methods are available, depending on the characteristics of the experiments, e.g., sample size, theoretical distribution of the trait being measured, sampling method, etc. While normal distribution and random sampling are more common and tests are generally straightforward, this is not always the case.

A few studies related to data mining have supported their experiments with the use of SHT. [Halkidi et al., 2002 a] included HST as a basis for certain cluster validity approaches. The null hypothesis is that the data is randomly structured in the different clusters, and if this can be rejected, then that indicates good clustering. [Jahanian et al., 2004] proposed a statistical method to limit the error incurred in determining whether an object is a strong member of a cluster or not. It also uses cross-correlation analysis in order to limit the feature space so as to optimize the performance of Fuzzy C-means (FCM) in this domain. HST was also utilized in [Baron and Spiliopoulou, 2003] to detect strong evolution that may occur in web usage patterns over time.

2.5 Prediction-based Performance of Collaborative Filtering Systems

In the context of web recommendations, a Collaborative Filtering (CF) system, given an active session, searches in the dataset for the "best" *k* sessions matching it, and then provide suggestions from the most N popular pages in these matches, called Top-N pages. As discussed earlier in the introduction to collaborative filtering, there are two main categories of CFs: Memory-Based and Model-Based. Memory-Based CF techniques normally use the k-nearest neighbor (KNN) approach to make extensive scan over all elements in the dataset in order to find the best matches to the active session. [Pennock et al., 2000] reports the first versions of the memory-based CF were Tapestry [Goldberg et al., 1992], which performed CF manually, and for automatic CF systems, it cites GroupLens [Resnick et al., 1994] and Ringo [Shardanand and Maes, 1995]. However, these approaches are inefficient and do not scale up especially with the increase in number of products and services, as well as the web users.

Model-Based CF's, on the other hand, take advantage of the knowledge discovered through data mining techniques - in our case the fuzzy clustering algorithms – in order to reduce the search complexity and improve efficiency. From now and on, we use the term Model-Based CF to indicate the systems described in [Mobasher et al., 2002], and briefly discussed in the following section.

2.5.1 Prediction Systems

In this thesis, we consider and compare two prediction systems, namely Model-Based CF (MB) and Fuzzy-Hybrid CF (FH), which use clustering as a web usage mining model. Following are their descriptions and highlights.

2.5.1.1 Model Based Collaborative Filtering

Using well established clustering algorithms: [Mobasher *et* al., 2002], [Mobasher *et* al., 1999], [Azman and Ounis, 2004], Aggregate Usage Profiles (UPs, see subsection 2.6) are extracted offline from the clusters. Then in the online/test phase, the UPs are ranked based on their degree of matching with the active session. The match between a UP and a session is determined using equation (20). Finally, recommendation of pages in these UPs is positively related to both the popularity/weight of pages in the UPs and to the degree of matching between the UP and the active session, as described below in equation (19).

$$Rec(S,p) = (f_p^C \times match\ (S,C))^{1/2} \tag{19}$$

where

$$match(S,C) = \frac{\sum_k f_k^C \times S_k}{\left(\sum_k (S_k)^2 \times \sum_k (f_k^C)^2\right)^{1/2}} \tag{20}$$

where S is the active session, C is a UP, f_k^C is the weight of the k^{th} page in C, and S_k is its weight in S (binary in case of clickstream data).

2.5.1.2 Fuzzy-Hybrid Collaborative Filtering

[Suryavanshi et. al., 2005] proposed a technique Fuzzy Hybrid (FH) which uses the model and applies memory-based CFs. This provides prediction accuracy comparable to that of the memory based CF, while having the efficiency close to MB techniques. Clusters are built in the learning phase. Based on the K-nearest prototype approach [Keller et al., 1985], the online search proposed in FH is more efficient by limiting it to the K clusters that have cluster centers nearest to the active session. A memory-based-like search is then performed within these clusters, based on the assumption that sessions similar to the active session would have similar memberships to such clusters. At the end, the recommendation rate of a page is related to the similarity between the sessions accessing it with the active session, and to the number of occurrences of such page in these sessions.

2.5.2 Prediction Quality

The prediction quality of collaborative filtering has been defined in different ways. [Herlocker et al., 2004] investigated several aspects of the prediction quality of collaborative filtering recommender systems. These include some often neglected attributes such as the coverage of target data set, the novelty or serendipity of recommendations, the learning rate of the system and others, all of which may fall into one category addressing the suitability of the prediction system to its users. But the most important aspect of prediction quality perhaps is accuracy. This quality can be measured by a variety of methods depending on the type of data at hand. Here we distinguish two types of data: user rating or preference, and user choice, or what is generally referred to in the web domain as clickstream data. For the first type of data, user rating, a variety of prediction accuracy measures are proposed, including Correlation Coefficient [Herlocker et al., 2004], and Average Absolute Deviation (AAD) [Breeze et al., 1998] [Pennock et al., 2000] and MAE. For the second type, which is considered in this thesis, we propose and use three quality evaluation metrics: F1, a crispified version of MAE, and R, described as follows.

2.5.2.1 F1

This is a well-known measure which fairly combines 'recall' and 'precision', and widely used in the context of web usage mining, e.g., [Mobasher et al., 1999], [Azman and Ounis 1999], [Suryavanshi *et* al., 2005 b], [Sarwar et al., 2002]. It is defined as follows:

$$\text{F1} = 2 \times \text{Recall} \times \text{Precision}/(\text{Recall}+\text{Precision}) \qquad (21)$$

Recall is the proportion of correctly recommended items to the total number of items that should be recommended (all items that would likely be picked by the active user).

In the process of evaluating prediction systems, the whole data set is divided into a training set and a test set. Part of each of each object (session) making up the test set is hidden (the hidden URL set) while the remaining part is used as active user which the filtering system- devised based on the training set- uses to make predictions, of which the Top-N (recommended set) are compared with the hidden set. Under this scenario, recall is defined as following ratio:

$$\mathrm{Re}call = \frac{|\,recommended\_set \cap hidden\_set\,|}{|hidden\_set|} \tag{22}$$

The size of the recommendation set is often termed as TOP-N, so TOP-5 stands for the 5 pages most recommended by the system.

Precision is the proportion of correctly recommended items to the total number of recommended items, and is expressed as the following ratio:

$$\mathrm{Pr}ecision = \frac{|\,recommended\_set \cap hidden\_set\,|}{|recommended\_set|} \tag{23}$$

The ideal situation is having high values for both Recall and Precision, but because of their relation when Recall increases, Precision decreases, and vice-versa, as is evident from their relation to TOP-N, F1 is maximized for an optimum combination of Recall and Precision values.

2.5.2.2 MAE

Mean Absolute Error (MAE) is described by the following formula:

$$\mathrm{MAE} = (1/n)\sum_{i=1}^{n}|f_i - y_i| = (1/n)\sum_{i=1}^{n}|e_i| \tag{24}$$

where f_i is the i^{th} predicted value, and y_i is the i^{th} actual value, and n is the size of the union of predicted and actual sets. Thus MAE measures the error in prediction and therefore smaller values indicate better accuracy. As described in [Sarwar et al., 2002], [Herlocker et al., 1999], and [Herlocker et. al, 2004], actual values are range scores. But as our data is web usage data, we assigned the value 1 to the predicted value f_i if the visitor has accessed the page, and 0 otherwise. The same holds for the actual values y_i.

2.5.2.3 R measure

The R measure has been introduced by [Mobasher et al., 2002], defined as the ratio of Recall over the size of the recommendation set (*i.e. Recall / |recommended_set|*). They used it to compare three usage profile extraction techniques, and argued to be more stringent than the metric F1 for providing additional information. It was also used before in [Mobasher et al., 1999] to evaluate a transaction clustering technique through measuring its prediction quality.

2.5.3 Related Work in Evaluating Prediction Systems

Due to the importance of prediction and recommendation systems in web usage applications, numerous research studied ways to build be performing systems. [Sarwar et al., 2000] compared recommender systems based on three collaborative filtering (CF) algorithms: a memory based, an association rule model based, and a dimensionality-reduction algorithm. The model-based CF was used to study scalability issue and ways to improve it. The dimensionality-reduction CF was chosen for the same purpose, and also to deal with sparseness and synonymy nature of the web data. The authors utilized F1 as prediction accuracy metric. The comparison involved two different data sets, one on item-transactions, and the other on item-ratings. Two different similarity measures were used; the cosine function for the first data set and Pearson's correlation coefficient for the second. The results were in favor of the dimensionality reduction based CF, as it scaled well to large data sets while maintaining good recommendation quality.

[Breeze et al., 1998] compared several collaborative filtering algorithms using different methods and tools, including two types of evaluation metrics, one for prediction accuracy and the other for estimating the probability of a list of recommendations returned to the user.

[Pennock et al., 2000] proposed a recommendation technique, called Personality Diagnosis (PD), which is based on a probabilistic model. It studied scalability for newly-added data. The technique was compared to two memory-based and two model-based recommenders, using two different prediction accuracy metrics, namely F1 and AAD. The results, supported by significance hypothesis testing, showed superior performance of PD over other recommender systems.

We next discuss representation of the clustering results and quality, referred to as the Aggregate Usage Profiles.

2.6 Aggregate Usage Profiles

2.6.1 Definition

In web usage mining, Aggregate Usage Profiles (or UPs) are summarized forms of usage patterns, extracted through several mining techniques, including probabilistic latent semantic analysis (PLSA) [Jin et al., 2004], association rules, and page [Mobasher et al., 1999] and usage clustering [Mobasher et al., 2002].

In the clustering context, a UP is a representation of a cluster describing frequent 'items' and their corresponding frequencies in that cluster.

Our study of web usage focuses on clickstream data. The preprocessing step involved is similar to those described in Section 2.1.

A *frequency* f_j^a of a URL$_i$ in a cluster C_a is considered as the probability of accessing URL$_i$ in a session $Sess_k$ belonging to C_a, and is defined as follows [Nasraoui et al., 1999 a]:

$$f_j^a = \frac{|\{Sess_k : (Sess_k \in C_a) \wedge (URL_j \in Sess_k)\}|}{|C_a|} \qquad 1 \leq k \leq \mathrm{N} \qquad (25)$$

where N is the total number of sessions. A page item is considered frequent in a cluster C_a if its frequency is above a certain threshold (*MIN_F*), normally determined heuristically. Only frequent items of C_a belong to any UP. This distinguishes strong, representative pages in a cluster from noise pages [Nasraoui et al., 2002]. The UP is defined more formally as follows:

$$\mathrm{UP_a} := \{<\mathrm{URL}^a_j, f_j^a> \mid j{:}1..|\mathrm{UP_a}|,\ f_j^a > MIN\_F\} \qquad (26)$$

where URL^a_j , f_j^a , $|\mathrm{UP_a}|$ are the jth URL, its weight, and the number of pages present respectively in UP$_a$, with $0 \leq f_j^a \leq 1$. Another representation of a UP binarizes the frequencies into 1 and 0. This enables UPs to play the role of cluster centers in clustering applications, but detailed UP information is then lost. In the literature, the terms *popularity* [Suryavanshi et al., 2005 a] and *weight* are also often used synonymously with the term *frequency*.

2.6.2 Notion of UP Quality

UP quality in general can be perceived as the meaningfulness and soundness of the knowledge carried in a single UP or in a group of UPs. As UPs form a representation of the clustering results, different clustering quality attributes including cluster compactness or separation are expectedly inherited and reflected by the corresponding UPs. Therefore in general, good clustering should yield more meaningful UPs, so that careful observation or assessment of the inherent goodness of UPs derived from clustering offers a good and efficient way of assessing the quality of clustering itself.

In fact, the inheritance of UP quality from clustering quality has been an implicit assumption in evaluating several clustering techniques. For example, [Nasraoui et al., 1999 b] considered quality of usage profiles as the primary criterion for evaluation of clustering quality.

2.6.3 Use and Quality Evaluation of UPs

UPs were frequently used instead of cluster centers as input to model-based prediction systems [Mobasher et al., 2002], [Mobasher, 1999], [Mobasher et al., 1999], [Azman and Ounis, 2004]. They were also used to infer usage associations between web pages. They were preferred over association rules in deriving session-like associations [Nasraoui et al., 2002], and were validated based on the possibility of co-accessing different pages within the same UP initially without consideration of page frequencies [Perkowitz and Etzioni, 1998]. Subsequently, further evolution of this approach took into account page frequencies and a measure called 'weighted average visit per page' (WAVP) was used to evaluate UPs produced by different web mining techniques, including clustering [Mobasher et al., 2002] and PLSA systems [Jin et al., 2004]. Similar criterion was used in [Nasraoui and Saka, 2007], measuring the match between each session in the evaluated population and its most "similar" UP.

The traditional and most common approach for evaluating UPs is through human observations and judgment. Often, experts "manually" analyze the relevance of URLs within each UP based on their knowledge of the website, and interpret joint occurrences [Joshi et al., 1999] [Mobasher et al., 2002]. This often resulted in neglecting the so-called meaningless UPs from the evaluation, since the objective of such evaluation was proving the effectiveness of a clustering technique, and evaluating and reporting all UPs in the whole set was not feasible [Mobasher, 1999], [Nasraoui et al., 1999 a], [Joshi et al., 1999], [Mobasher et al., 2002]. Continuing on this same issue, different fuzzy clustering techniques have been compared based on the manual analysis of their respective UPs [Joshi et al., 1999] [Krishnapuram et al., 1999]. In other studies, observations made on UPs derived from the clustering technique RFC-MDE led the authors to infer good clustering quality for their technique. This was further confirmed by the high inter-cluster and low intra-cluster distances [Nasraoui et al., 1999 a], [Nasraoui et al., 1999 b], [Nasraoui et al., 2002]. In their evaluation of a competitive agglomeration-based clustering (CARD) technique, [Nasraoui et al., 1999 a] have combined manual assessment of UPs with cluster validation and found that meaningful UPs corresponded to highly valid clusters and vice versa.

As UP evaluation has frequently been based on human judgment, it may not be free of subjective bias. Furthermore, as the size of the data to be clustered increases, and the number and content of UPs also become larger, human evaluation becomes tedious or even infeasible. In general, human evaluation of clustering results (including UPs) suffers from high cost in terms of time and efforts, lack of reproducibility, and risk of subjectivity and bias [Crabtree et al., 2005].

3 PERFORMANCE ANALYSIS OF FUZZY CLUSTERING

FUZZY clustering is a popular technique recommended and used for modeling web usage data with applications such as usage prediction, recommender systems, web site restructuring, etc. Fuzzy clustering is a process which categorizes elements, typically usage clicks or usage sessions into groups, where each element can belong to several groups with different degrees of membership. A number of fuzzy clustering techniques have been proposed with their performance usually demonstrated through results obtained from implementation experiments using not so large of sample data sets [Oliveira and Pedrycz, 2007]. ARCA [Corsini et al., 2005], FCMdd [Krishnapuram et al., 1999], and RFSC [Suryavanshi et al., 2005 a] are representative techniques in this category. When introduced first, these techniques demonstrated performance using a small set of samples, justified by the fact that it is often difficult to acquire an adequately large number of representative web usage data sets. This became more difficult with requiring much effort and time to carry out experiments and gather results for many different usage data sets. However, the significance of results would be an assumption for a dependable comparison between these techniques. In this Chapter we present dependable results from a comprehensive set of experiments carried out to assess the performance of the above mentioned fuzzy clustering techniques. Our basis for dependability is founded on three aspects of our experiments. First, as in the default approach used for demonstrating the performance of such techniques, we compute cluster goodness or cluster validity indices (VIs). Clearly, the objectivity or unbiasedness of such VIs is crucial. Therefore, through appropriate experimentation, we ensure that the cluster validity index which is finally chosen is least biased with respect to parameters of the clustering algorithms analyzed. Second, we integrate each clustering technique into the same application, an instance of a model-based prediction system, and measure prediction quality and efficiency as the second criterion for our comparison. Making application quality a basis for comparison surely provides more ground truth to the performance of clustering. Prediction for web personalization is an important and frequently used application of clustering [Mobasher et al., 2002]. Thirdly, the above comparison results are subjected to statistical hypothesis testing (SHT) to increase confidence that the results obtained are not the outcome of mere chance. This is certainly beneficial, given the fact that large and varied, real world data are rather difficult to collect, and that the number of experiments that can be conducted is limited due to resource constraints. We consider the aforementioned three issues in our study as forming a basis for dependable performance analysis of fuzzy clustering techniques and illustrate this through numerous experiments.

The rest of this Chapter is organized as follows. In section 3.1, we describe our sampling procedure and parameter settings for the different clustering procedures being compared. In section 3.2, we assess the objectivity of some well-known VIs, and select the most appropriate one as a basis for the clustering comparison. In sections 3.3 and 3.4, we analyze the performance of the chosen fuzzy clustering techniques for their clustering goodness and prediction-based performance, respectively. In section 3.5, we discuss the results of our experiments and their statistical significance.

3.1 Pre-Evaluation Settings

To evaluate the fuzzy clustering techniques, we compose appropriate input data, set the clustering parameters for each technique, run the data on the individual implementations, and then compute the metrics upon which the comparisons are based.

On the input date, we carried out preprocessing and cleaning, computed pageview similarity, and performed sessionization previously described in 2.1. The sampling and parameter setting processes, however, are specific to this set of experiments.

3.1.1 Sampling

From a population of 65,000 sessions, we composed 9 training sets, well sampled using the following technique.

Since usage patterns are assumed to be time dependent, sessions were sorted by time of the last visited pageview in each session. The sessions in each training set were chosen to be continuous, but the beginning session point for each data set was chosen randomly. The sizes of the training sets chosen in terms of the sessions they contained were: 100, 160, 200, 240, 300, 360, 400, 440, 500 sessions. We were limited in the choice of the maximum size because of excessive computation time requirements of our ARCA implementation (see clustering time results for ARCA in Table II). We implemented the clustering techniques, carefully choosing their parameters for maximal performance as described in the following paragraphs.

3.1.2 Parameter Settings for the Evaluated Fuzzy Clustering Techniques

3.1.2.1 ARCA

For ARCA, the fuzziness coefficient *m* was set to 1.2, as it showed best results. ARCA also takes the number of clusters C as an input. Although in the original proposal of ARCA, *e* (see list of symbols) was 0.001, it took a long time for ARCA to terminate using this value. We therefore set *e* = 0.1. The best ARCA results were chosen, as suggested in [Corsini et al., 2005], by varying C from 2 to N/3, where N is the size of the sample to be clustered. We then run ARCA for each value of C, and choose the C value which the clustering metric XB index Section 2.3.1 is minimal.

3.1.2.2 FCMdd

We set the same values for the common parameters of both versions, LFMCdd and RFCMdd. For comparing these clustering algorithms, we found that values for some parameters are not specific. The initialization of medoids has several ways, and trying many different initializations for better results as recommended in [Joshi and Krishnapuram, 2000], [Nasraoui et al., 2002], is not practical for simple comparison experiments. Also we did not find any justification, such as overspecified C and minimal cardinality, in the literature for the choice of parameters in FCMdd for picking C. The value C certainly affects the comparison not only in terms of quality but also efficiency. To overcome these difficulties successfully, we have chosen the second type of medoids initialization as stated in [Nasraoui et al., 2002], which basically starts by choosing the most "popular" session as the first medoid, then keeps adding the session that is the farthest from existing medoids. Based on this interpretation of the FCMdd specifications of C, we overspecified C to 50, and set *MIN_Cardinality* to C/25. It is also proposed in [Nasraoui et al., 2002] that the best value of fuzzifier m to be between 1 and 1.5, so we considered $m = 1.2$. Also, maximum number of iterations, *MAX_ITER,* is set to 100, since a higher value turned out to be inefficient in our experiments. In [Nasraoui et al., 2002], it was proposed to set p to less than N/C and s to be N/2. In our experiments, we thus considered p= 0.75*N/C and s= N/2 (see list of symbols).

3.1.2.3 RFSC

For the Relational Fuzzy Subtractive Clustering (RFSC) algorithm, we proceeded similar to FCMdd, *i.e.*, overspecifying C to 50, and setting *MIN_Cardinality* to C/25.

3.2 Evaluating Validity Indices Based on Their Objectivity

While critics of cluster validity indices have pointed out their objectivity or bias, we do not always see this aspect given the importance it deserves when demonstrating the performance of various techniques. As mentioned earlier in section 2.3, most comparison studies of fuzzy clustering techniques use cluster validity indices and conclusions are drawn on this basis. Yet, several of these indices have been judged inappropriate because of their dependence on clustering parameters, including the number of clusters (section 2.3 and Appendix A).

Table I. Best C values for LFCMdd, ARCA, and RFSC

Sample Method\	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
LFCMdd	18	24	39	35	34	37	38	38	42
RFCMdd	25	34	41	42	44	41	46	50	47
ARCA	32	52	5	8	6	7	131	6	137
RFSC	20	29	40	42	48	50	46	50	48

We investigated the objectivity of some well-known VIs in order to choose the most suitable one(s) for our comparison. We realize that no validity index can be claimed to be best for all circumstances [Pal and Bezdek, 1995]. Table I shows the C values for best clustering as per their individual goodness criteria for RFCMdd, LFCMdd, ARCA, and RFSC. In this table, Train 100 stands for the training set of 100 sessions, Train 200 for set of size 200, and so on.

3.2.1 Xie-Beni Index (XB)

This VI is described in Section 2.3.1. Due to suspected relation with number of clusters, C, questions are often raised about the objectivity (absence of bias) of VIs as indicators of goodness. We therefore conducted experiments to shed light on this issue. We ran the 4 clustering algorithms on a training set of size 300, varying C from 2 to 40. Fig 3.a and 3.b demonstrate the relationship between C and XB values derived from clustering results of these techniques (we show ARCA results in a separate figure due to their much greater range).

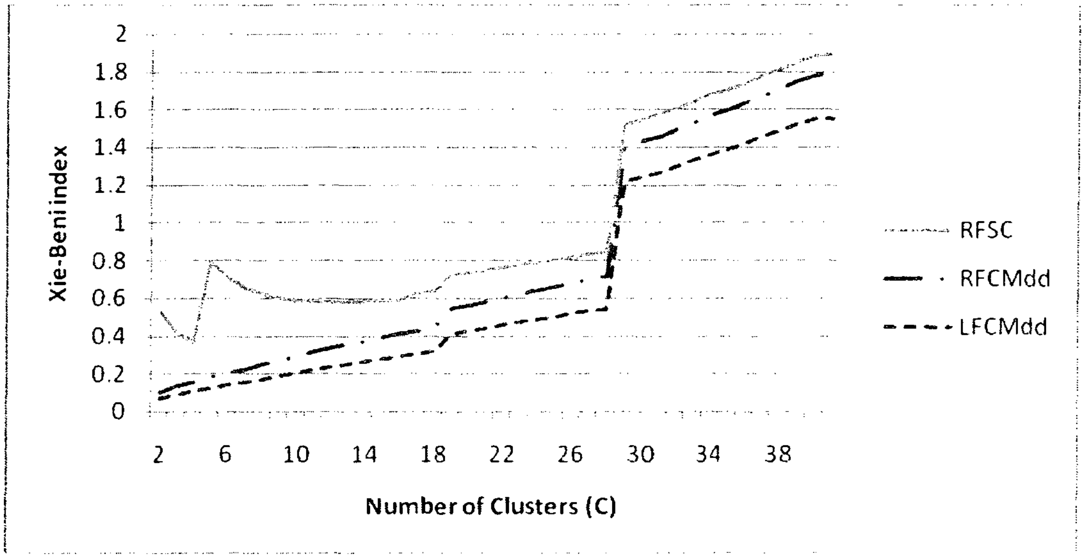


Fig. 3.a. Variation of Xie-Beni (XB) for increasing number of clusters (C) in 3 clustering techniques

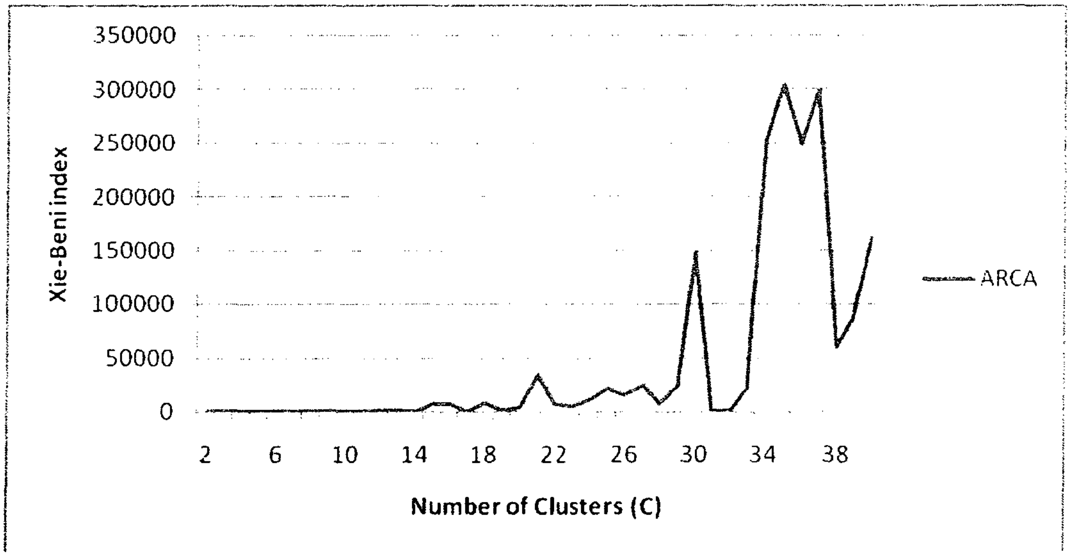


Fig. 3.b. Variation of XB for increasing C in ARCA

Figure 3.a shows a clear relationship of the index XB to cluster number C for three clustering techniques RFSC, RFCMdd and LFCMdd with an increasing trend as C increases. We find that the relation is even evident at low values of C, in contrast to earlier reports of a relation between XB index and C in the higher C range only

[Bouguessa et al., 2006], [Xie et al., 2002], as is the case of ARCA (Fig 3.b.). This result indicates that XB index cannot be a reliable metric for cluster validation of the 4 algorithms. We therefore do not use XB index for comparison of clustering techniques.

3.2.2 Partition Coefficient

This coefficient also has been introduced and defined in section 2.3.1. The smaller the coefficient, the better is the clustering. Figure 4 shows the variation of PC for increasing C. It can be observed that starting from C = 10, RFSC and both versions of FCMdd show a monotonic relation of PC with C. In fact, we are not interested in smaller values of C for these three techniques applied on train set 300 (see Table 1). Such dependency of PC on C was also confirmed in [Bouguessa et al., 2006]. Such behavior of PC makes this index an unreliable metric for assessing clustering validity. Therefore, PC is not used any further by us to compare the various clustering techniques.

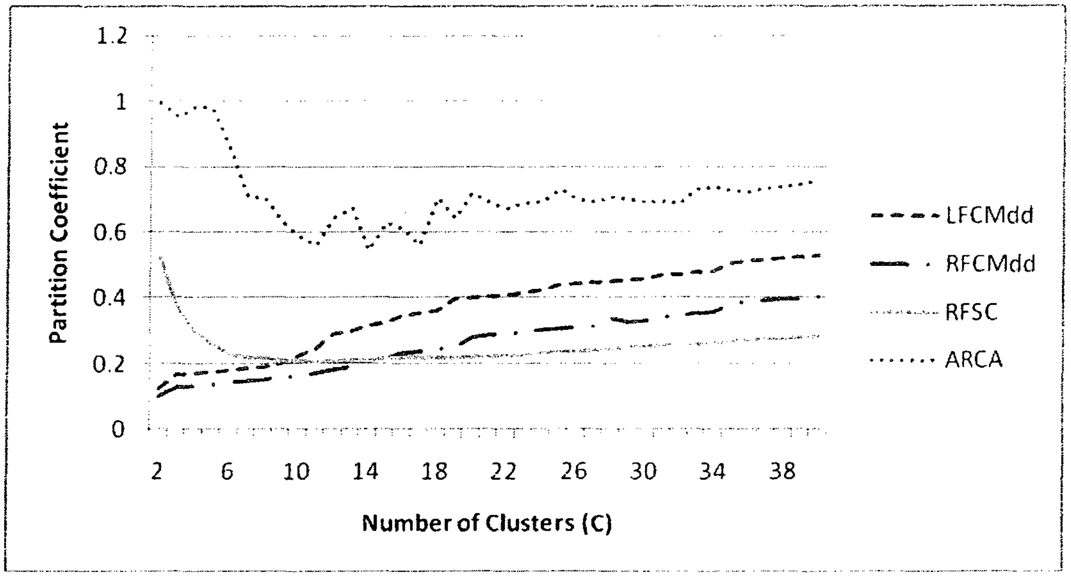


Fig. 4. Variation of Partition Coefficient (PC) for increasing number of clusters (C) in the 4 clustering techniques.

3.2.3 Xie-Y Index

This index too is defined in section 2.3.1. Here we study the objectivity of Xie-Y index and decide on its suitability as an acceptable criterion for cluster validation.

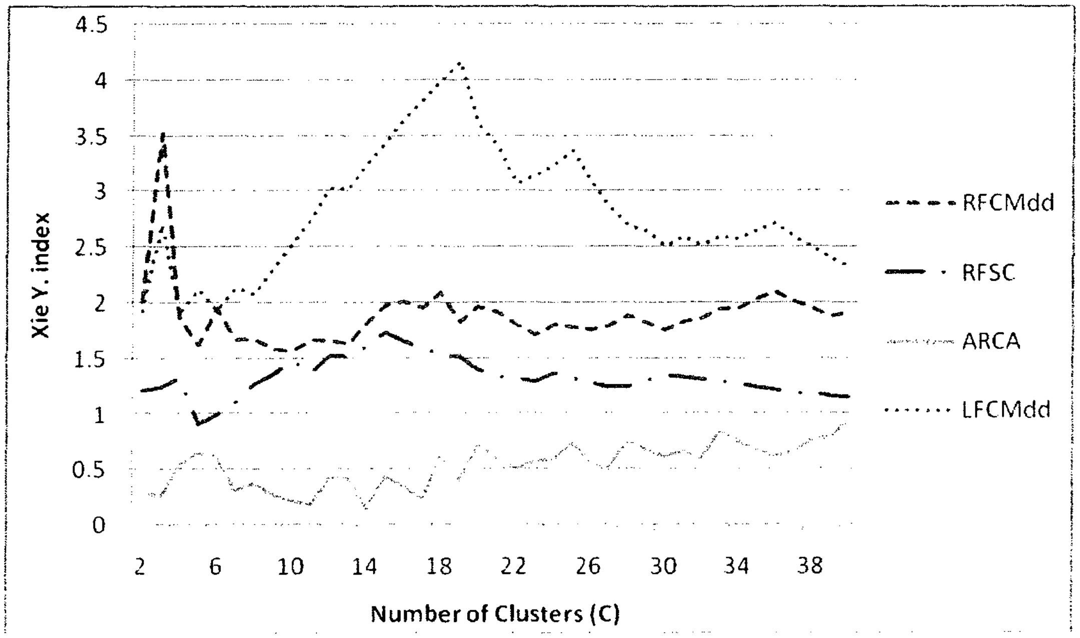


Fig. 5. Variation of Xie-Y for increasing number of clusters (C) in the 4 clustering techniques

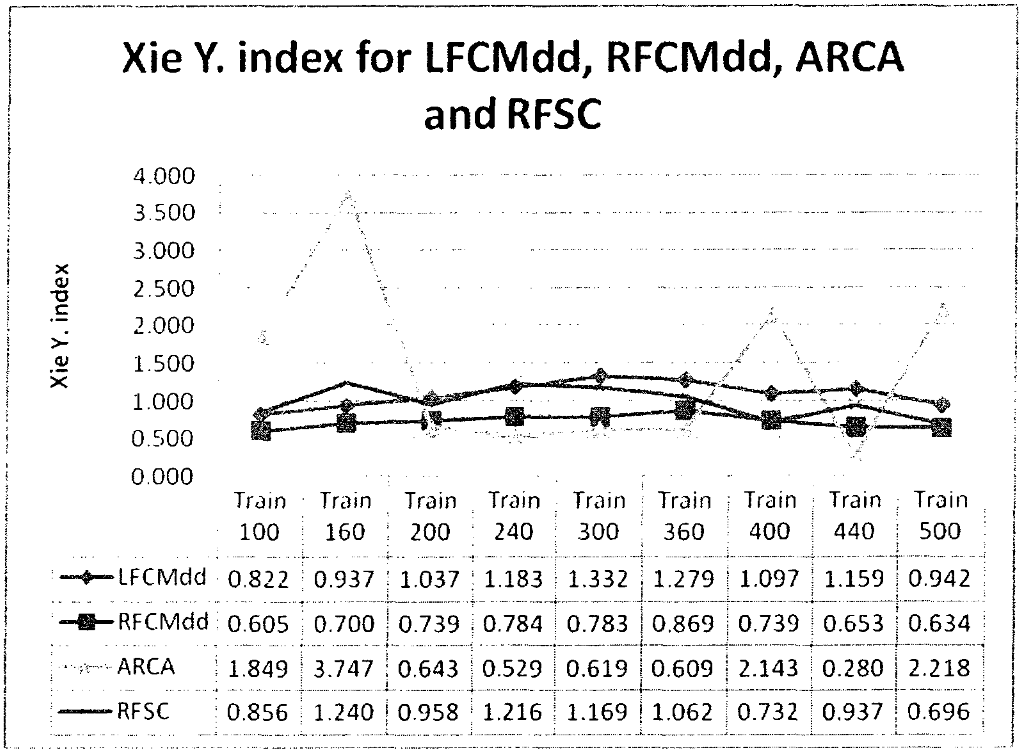
Results (Figure 5) show an up-and-down fluctuation of Xie-Y index across the range of C values. The ranking of the four compared clustering techniques has not changed over the whole range of C, with lowest Xie-Y index recorded for ARCA, intermediate for RFSC and highest for the 2 FCMdd versions. There is no monotonic increase of Xie-Y with C, and therefore Xie-Y is judged as the least biased among the 3 tested cluster validation criteria. We have therefore picked this as the most suitable among the 3 indices for use in comparative studies of clustering techniques.

3.3 Clustering-based Performance

This performance evaluation is based on two metrics, clustering validity and computation time.

3.3.1 Comparison of Clustering Validity

Figure 6 shows the results of our experiments for Xie-Y. ARCA indicates highest or lowest values for this index, in comparison to the other three techniques, and LFCMdd seems to have an overall better results, then follows RFSC and then RFCMdd.



	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
LFCMdd	0.822	0.937	1.037	1.183	1.332	1.279	1.097	1.159	0.942
RFCMdd	0.605	0.700	0.739	0.784	0.783	0.869	0.739	0.653	0.634
ARCA	1.849	3.747	0.643	0.529	0.619	0.609	2.143	0.280	2.218
RFSC	0.856	1.240	0.958	1.216	1.169	1.062	0.732	0.937	0.696

Fig. 6. Variation of Xie-Y for increasing number of clusters (C)

3.3.2 Comparison of Clustering Time

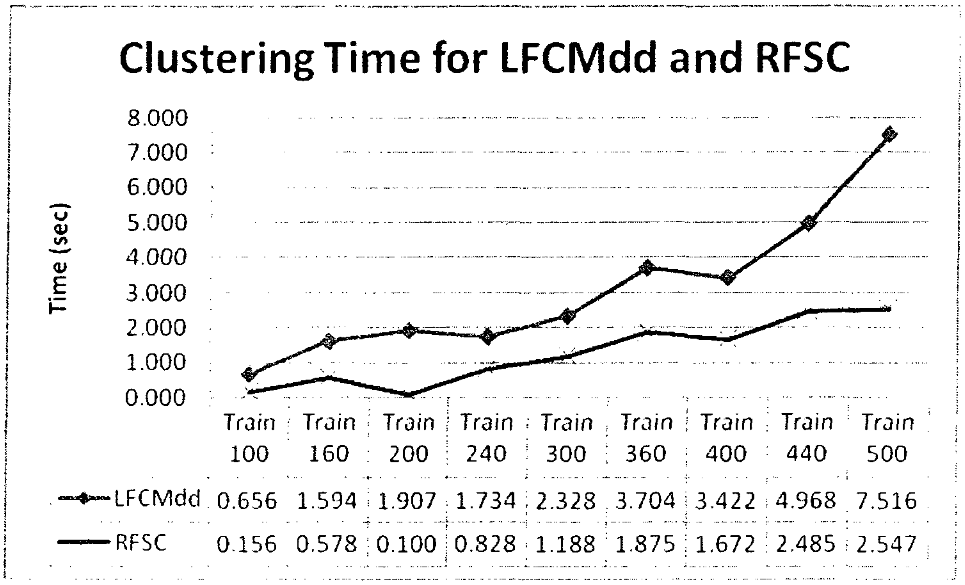
For our experiments, we used a typical desktop computer Pentium 4 CPU of 2.7GHz and

1GB RAM. Our results indicate that ARCA and RFCMdd are considerably slower in performance, so we show their related information separately in Table II.

Table II. Clustering time (in sec) for RFCMdd and ARCA

Sample Method\	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
RFCMdd	17	41	61	60	84	94	122	172	166
ARCA	305	2069	9879	10050	2.E+6	7.E+4	1.E+5	2.E+5	3.E+5

Figure 7 compares RFSC and LFCMdd for efficiency. As can be seen, RFSC revealed better performance across all samples. Here, it is worth noting that in iteratively searching for best C, the time required by ARCA grows rapidly as N increases. On the other hand, our experiments seem to reconfirm the scalability claims of RFSC [Suryavanshi et al., 2005]. For example, the time RFSC requires to cluster much larger sample sizes of 30,000 and 51,624 sessions were 2189 and 6176 seconds, respectively. Also, the claimed LFCMdd superior efficiency over RFCMdd was proven here.



	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
LFCMdd	0.656	1.594	1.907	1.734	2.328	3.704	3.422	4.968	7.516
RFSC	0.156	0.578	0.100	0.828	1.188	1.875	1.672	2.485	2.547

Fig. 7. Clustering time for LFCMdd and RFSC in the 9 experiments

3.4 Application Based Performance

The second aspect of our dependable performance analysis is comparison based on use of the cluster model in an application -- in our case an instance of a model based usage prediction system. We measure and compare prediction accuracy and efficiency. This was motivated by the fact that, as for many other data sets, fuzzy clustering of web log data has rarely any ground truth, *i.e.*, there is normally no oracle that can judge the semantic correctness of fuzzy clustering.

3.4.1 Input data

In order to ensure the correspondence of the patterns between the training and the test sets, each test set was composed of a sequence of sessions starting from the end of its training set in chronological order. The size of the test sequence was chosen as a quarter of the size of the corresponding training set, for each of the 9 samples used in our experiments.

3.4.2 Prediction System and Parameters

We used the results of each of the four clustering techniques in a prediction system [Suryavanshi et al., 2005 b]. We set the following parameters: TOP 5 predictions - as it seems a reasonable number for a real website. For each test session, we hid 20% of its total pageviews, ran the remaining part of such session on the prediction system, and obtained the 5 predicted pages for each test session as an output. We then measured metrics for prediction accuracy and efficiency.

3.4.3 Comparison criteria

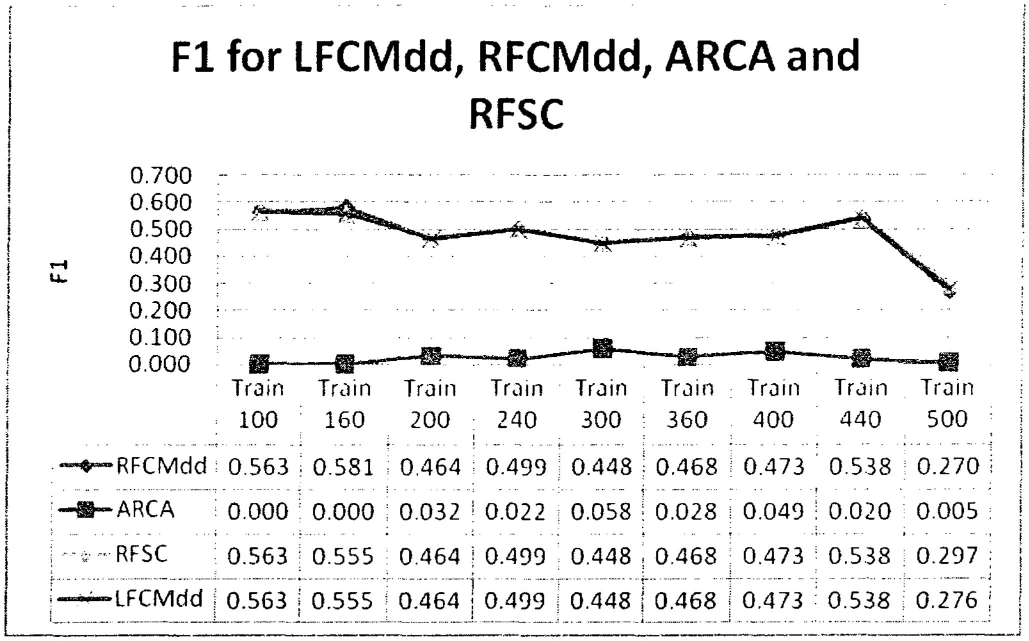
3.4.3.1 Prediction Accuracy:

We measured the quality of the predictions in terms of F1 (Eq. 21) and MAE (Eq. 24).

As can be seen from the experimental results in Figure 8, while RFCMdd, LFCMdd, and RFSC look identical in their F1 scores, the values are weaker for ARCA across all samples.

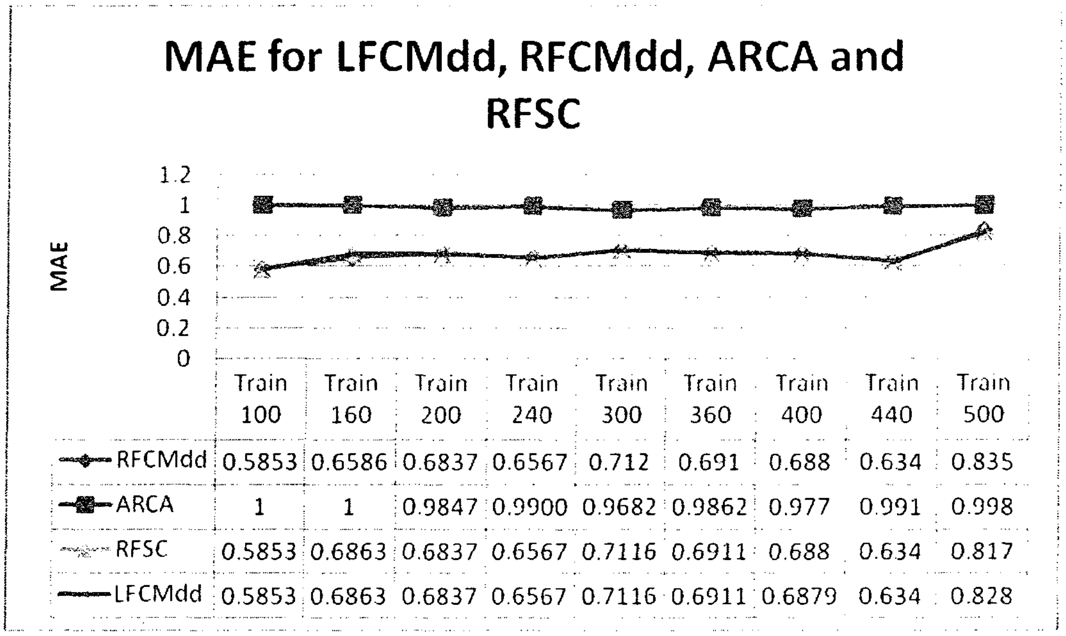
Figure 9 shows MAE scores for the four techniques across all the 9 samples. Once again, LFCMdd, RFCMdd, and RFSC showed practically identical scores across all samples.

Since the differences are too small, significance testing between both versions of FCMdd and RFSC is not needed for F1 and MAE results. In contrast, the difference between these three techniques and ARCA is evident.



	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
RFCMdd	0.563	0.581	0.464	0.499	0.448	0.468	0.473	0.538	0.270
ARCA	0.000	0.000	0.032	0.022	0.058	0.028	0.049	0.020	0.005
RFSC	0.563	0.555	0.464	0.499	0.448	0.468	0.473	0.538	0.297
LFCMdd	0.563	0.555	0.464	0.499	0.448	0.468	0.473	0.538	0.276

Fig. 8. F1 values for LFCMdd, RFCMdd, ARCA, RFSC across 9 samples



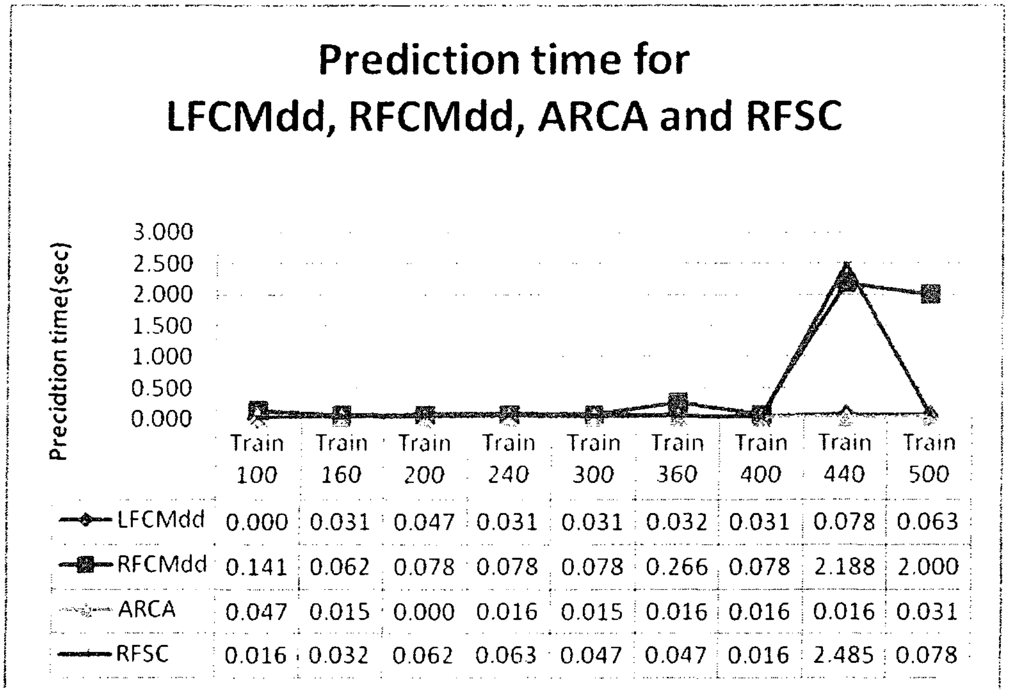
	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
RFCMdd	0.5853	0.6586	0.6837	0.6567	0.712	0.691	0.688	0.634	0.835
ARCA	1	1	0.9847	0.9900	0.9682	0.9862	0.977	0.991	0.998
RFSC	0.5853	0.6863	0.6837	0.6567	0.7116	0.6911	0.688	0.634	0.817
LFCMdd	0.5853	0.6863	0.6837	0.6567	0.7116	0.6911	0.6879	0.634	0.828

Fig. 9. MAE values for LFCMdd, RFCMdd, ARCA and RFSC across 9 samples

3.4.3.2 Prediction Time

This is the computation time required for running the versions of prediction algorithm.

Here, the four systems seem somewhat similar for small sample sizes but different for larger sizes (Figure 10). Overall, ARCA is the fastest in this respect, followed by LFCMdd, RFSC, and then RFCMdd. Numerically small differences need statistical testing for significance. We recall here that, because model-based prediction systems perform their search on the clusters level (first), the prediction time is expected to depend on C, which in turn explains the small prediction time of ARCA for small C values raising doubts in the prediction efficiency excellence of ARCA (Table I). Yet, what would remove such doubts is that ARCA sustain its high prediction efficiency even when its C values are the highest over all other techniques (specifically Training sets Train 400 and Train 500).



	Train 100	Train 160	Train 200	Train 240	Train 300	Train 360	Train 400	Train 440	Train 500
LFCMdd	0.000	0.031	0.047	0.031	0.031	0.032	0.031	0.078	0.063
RFCMdd	0.141	0.062	0.078	0.078	0.078	0.266	0.078	2.188	2.000
ARCA	0.047	0.015	0.000	0.016	0.015	0.016	0.016	0.016	0.031
RFSC	0.016	0.032	0.062	0.063	0.047	0.047	0.016	2.485	0.078

Fig. 10. Prediction time for LFCMdd, RFCMdd, ARCA and RFSC, for the 9 samples

3.5 Significance Testing

We used the guidelines in [Groebrner et al., 2006] and [Black, 2008] to develop the following statistical testing plan.

We apply statistical significance testing to the observed differences among the 4 techniques for both clustering (using Xie-Y and clustering time) and prediction (F1, MAE, and prediction time) processes. It is noted that higher values are desired for Xie-Y and F1, and lower values are desired for other metrics.

3.5.1 Methodology

The general statistical testing methodology involves the following steps:

(i) Design a null hypothesis (H_o) of no difference between two compared techniques versus an alternative hypothesis (H_a), where one technique has superior performance compared to the other.

(ii) Adopt an appropriate testing procedure, depending on the experimental situation, which involves the computation of a test statistic and finding the critical value.

(iii) Compare the computed value of the test statistic to its critical value, and reject H_o, at the chosen level of significance (denoted as α) depending on the magnitude of computed vs. critical values; do not reject H_o otherwise.

3.5.2 Application in our Context

We compare two clustering techniques at a time. The null hypothesis states "no difference" between the two clustering techniques, based on the metric while the alternative hypothesis states a superior performance of one over the other. The test is mainly unilateral and the significance level (α) is chosen as α= 0.05 (a confidence level of 95%), or α=0.01 (confidence level of 99%). An exception to this is the hypothesis no. 1

which is bilateral and the confidence level is 95%.

The statistical testing procedure is based on *Wilcoxon matched pairs, signed ranks* test. This is a non-parametric statistical test, which requires no assumption about the metric distribution, in contrast to the *student-t* test for matched pairs that requires a normal distribution, which is not necessarily met in our case. The implementation of the Wilcoxon rank test involves the computation of the absolute difference for the 9 pairs of the metric values, their ranking from lowest (rank 1) to highest (rank 9), restitution of the difference sign to the corresponding rank, and the summation of like-sign ranks. This yields two sums: the sum of positive ranks, and the sum of negative ranks. The smaller of these two sums, denoted *T*, is compared to a critical tabulated value for the chosen significance level and sample size. The null hypothesis is rejected if "T" is smaller or equal to the tabulated critical value. In our case, the critical value is equal to 8, for a total number of 9 pairs. When a difference between the pair members is null, that pair is dropped from the sample (this occurred once in our experiments).

3.5.3 Test results

Table III. Tested hypotheses and related metrics for comparison of four clustering algorithms

Hypothesis	Metric	Tested hypotheses (M_d: metric difference between the 2 techniques)	Calculated *T*-value
1	Xie Y. index	H_0: There is no difference between the qualities of LFCMdd and RFSC $\{M_d = 0\}$; H_a: There is a difference $\{M_d <> 0\}$	11^{ns}
2	Xie Y. index	H_0: RFSC has no better quality than RFCMdd $\{M_d = 0\}$; H_a: RFSC is better $\{M_d < 0\}$	0*
3	Xie Y. index	H_0: LFCMdd has no better quality than ARCA $\{M_d = 0\}$; H_a: LFCMdd is better $\{M_d < 0\}$	15^{ns}
4	Xie Y. index	H_0: ARCA has no better quality than RFCMdd $\{M_d = 0\}$; H_a: ARCA is better $\{M_d < 0\}$	15^{ns}
5	Cluster. Time	H_0: RFSC has no better performance than LFCMdd $\{M_d = 0\}$; H_a: RFSC is better $\{M_d > 0\}$	0**
6	Predict. Time	H_0:LFCMdd has no better performance than RFSC $\{M_d =0\}$; H_a: LFCMdd is better $\{M_d < 0\}$	2**
7	Predict. Time	H_0:RFSC has no better performance than RFCMdd $\{M_d =0\}$; H_a: RFSC is better $\{M_d > 0\}$	8*
8	Predict. time	H_0:ARCA has no better performance than RFSC $\{M_d =0\}$; H_a: ARCA is better $\{M_d < 0\}$	3*
Critical *T* value, for 2-sided and 95% confidence; n=9, α=0.05			6
Critical *T* value, for 1-sided and 95% confidence; n=9, α=0.05			8
		α=0.01	3
Critical *T* value, for 1-sided and 95% confidence; n=8, α=0.05			6
		α=0.01	2

* Significant at 0.05 | ** Significant at 0.01 | ns: Non significant at 0.05

Note: In hypothesis no. 8, there was a tie for the values of ARCA and RFSC (on Train400), and n is taken as 8, for which the critical *T* value is 6 for a significance level of 0.05.

Table III lists the hypotheses tested on various metrics for pairs of clustering techniques.

We have calculated the *T* values for the 8 hypotheses presented in Table III.

The test results displayed in Table III show statistical significance for testing hypotheses 2, 5, 6, 7 and 8. In these cases, the null hypothesis is rejected in favor of the alternative.

More explicitly the results can be summarized in the following lists, ordered by decreasing performance:

Xie-Y: LFCMdd, RFSC, ARCA, RFCMdd, with ARCA not significantly different from the other three techniques,

Clustering Time: RFSC, LFCMdd, RFCMdd, ARCA,

Prediction Time: ARCA, RFCMdd, LFCMdd, RFSC,

MAE: (RFSC, LFCMdd, RFCMdd), ARCA,

F1: (RFSC, LFCMdd, RFCMdd), ARCA.

Statistical tests were performed, and confirm the trend observed in Figures 8 and 9, with no difference within the group (RFSC, LFCMdd, RFCMdd) and highly significant difference between that group and ARCA for both MAE and F1.

We note that statistical testing has helped us in realizing that the apparent superiority of LCMdd over RFSC in terms of Xie-Y index is not significant and may be due to sampling or experimental error.

When we compared RFCMdd with RFSC or with LFCMdd, we realized that the lower performance of RFCMdd in clustering validity was not reflected in the prediction phase. This indicates that low clustering validity does not always lead to low prediction.

We also see that better clustering accuracy of ARCA did not help to achieve superior prediction accuracy. We thus discovered that low accuracy of prediction is not necessarily the result of low clustering compactness or separation. As such, we may conclude that considering prediction accuracy is essential in evaluating clustering techniques in the context of web personalization.

3.6 Conclusion

The following should be taken into consideration when reviewing and summarizing the results of the present study:

1. It is easily seen that there is no direct correlation between clustering validity index values and prediction quality. In other words, good prediction quality of model-based prediction systems does not necessarily imply good validity of the underlying clustering system; neither does good clustering validity necessarily lead to good prediction. Both aspects are important for dependable analysis, namely unbiased choice of cluster goodness index and performance of the usage model when used in an application.

2. Sample sizes are small compared to real web usage data. However, we were limited in our choice for sample sizes to accommodate ARCA's computational requirements (Table II). In fact, our training samples are already much larger in comparison to data set sizes used previously for reporting performance of ARCA [Corsini et al., 2005].

3. Statistical significance increases dependability of the analysis by ensuring that the experimental conclusions derived are not by mere chance.

4. No single method ranks best for all comparison criteria. For example, while ARCA ranked last for clustering time and prediction quality (F1 and MAE), it scored best for prediction time. In contrast, RFSC was best for clustering efficiency and prediction quality, but ranked last for prediction time. The remaining techniques scored in between, LFCMdd being closer in performance to RFSC while RFCMdd being closer to ARCA.

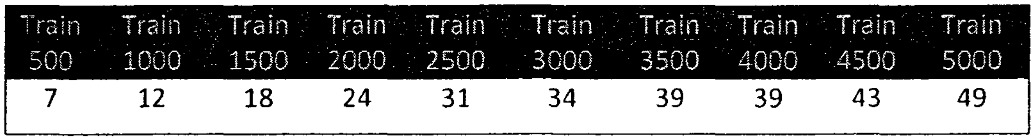
4 COMPARISON OF WEB PREDICTION SYSTEMS

In this chapter we compare two prediction systems based on Model-Based (MB) Collaborative Filtering and Fuzzy-Hybrid (FH) Collaborative Filtering (CF), previously described in Section 2.5. In [Suryavanshi et al., 2005 b], experiments were conducted to compare the accuracy of the three techniques: Fuzzy-Hybrid, Model-Based, and Memory Based Collaborative Filtering. Their results showed superior performance of FH, although no statistical significance was reported. No provision was made for varying sample size or number of clusters (C) of the underlying model. In this work, we conduct experiments comparing FH and MB across samples with different sizes and for different C values. We also study the performance consistency for these prediction systems. All results are subjected to statistical hypothesis testing.

4.1 The Prediction Model

The model on which these techniques are based is RFSC [Suryavanshi et al., 2005 a], as it has been shown in Chapter 3 to provide good accuracy both in terms of clustering validity and in prediction. The settings of this model are similar to those given in Chapter 3, except for *MIN_CARDINALITY* and overspecified_C, which we set to 25 and N/25, respectively. Table IV shows the number of clusters in the results, suggested by this model for different training samples.

Table IV. C values for RFSC across different sample sizes



Train 500	Train 1000	Train 1500	Train 2000	Train 2500	Train 3000	Train 3500	Train 4000	Train 4500	Train 5000
7	12	18	24	31	34	39	39	43	49

Section 4.2 reports a comparative analysis of the two prediction techniques using the conventional measures (F1, MAE, R). Section 4.3 investigates the performance consistency of the two techniques across different sample sizes. Statistical hypothesis testing is performed in Section 4.4 for the experiments conducted in Sections 4.2 and 4.3. Concluding remarks are provided in Section 4.5.

4.2 Conventional Analysis

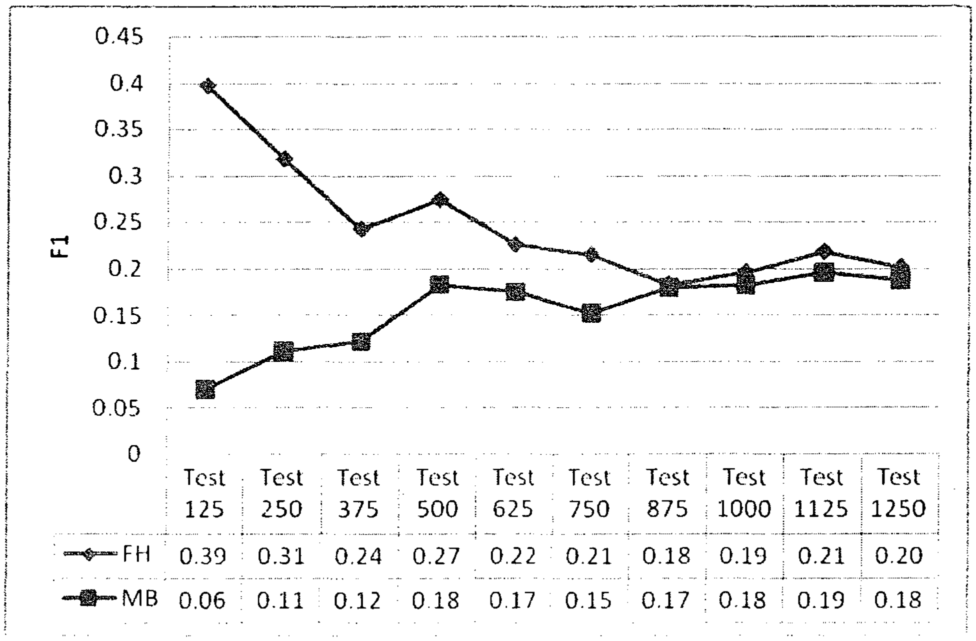
In this section, we compare the two prediction techniques in terms of their prediction effectiveness using 10 different sample sizez and the metrics F1, MAE, and R described in Section 2.5. The analysis is described as "conventional" in reference to the "random samples" used, in contrast to "designed samples" used below for consistency analysis.

4.2.1 Experimental Settings

In this set of experiments, we considered 10 training samples with varying sizes of 500, 1000, 1500, ..., 5000, with matching 10 test samples of sizes equal to 1/4 the training samples, *i.e.,* 125, 250, 375, ...,1250, respectively. This yields Test 125, Test 250, etc, similar to what we designated in Chapter 3. In contrast to samples in Section 4.3 later, the samples tested in this section are randomly selected in the same manner as in Chapter 3. We set TOPN to 5 and set K to 20.

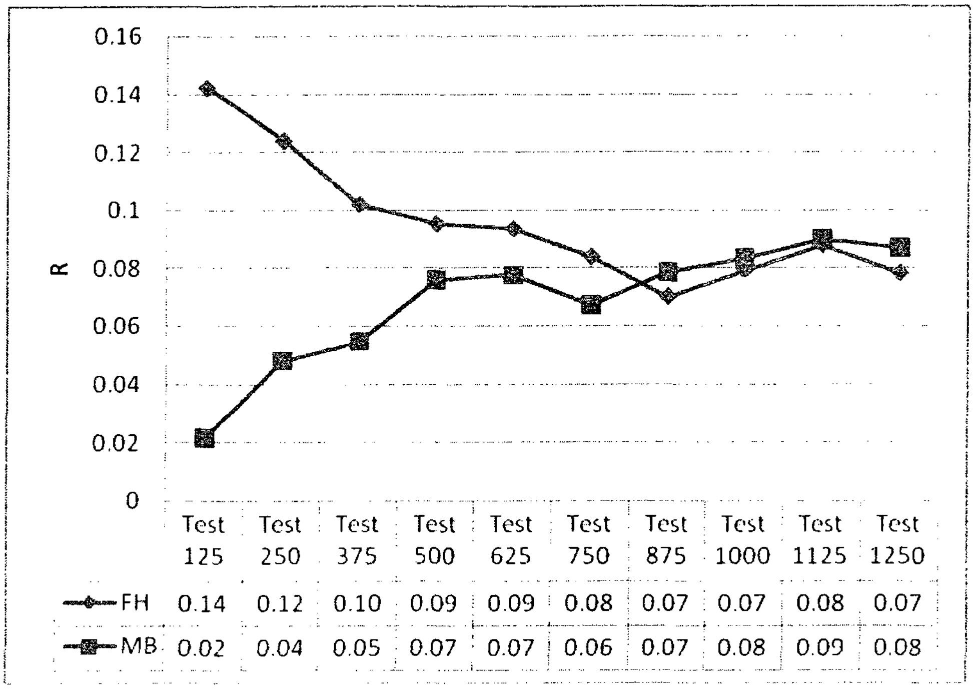
4.2.2 Results and Analyses

We applied the two prediction techniques on the ten samples and calculated the three prediction accuracy measures F1, MAE, and R. Results for these 3 measures are shown in Figures 11, 12 and 13 respectively.



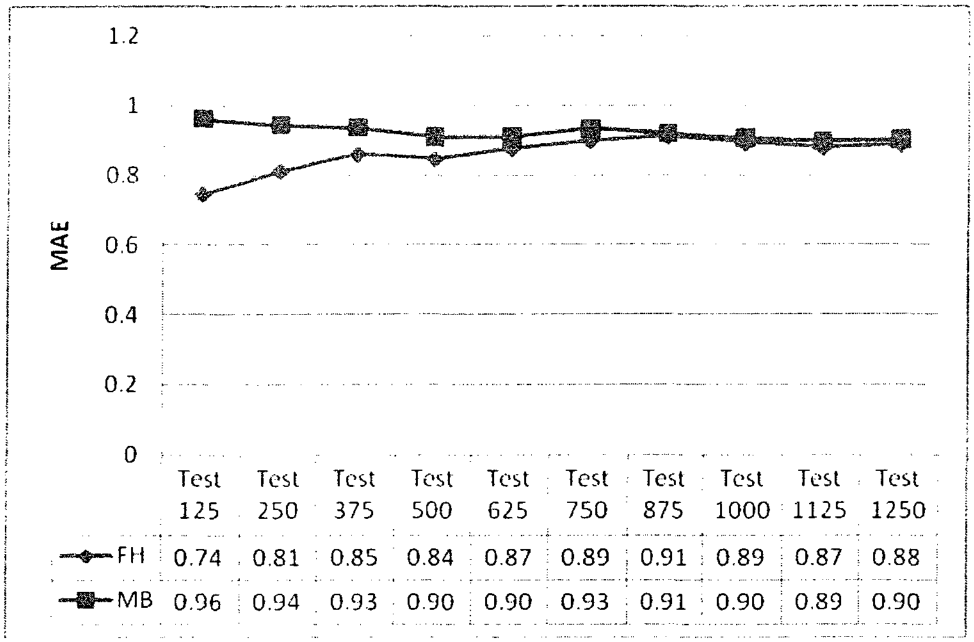
	Test 125	Test 250	Test 375	Test 500	Test 625	Test 750	Test 875	Test 1000	Test 1125	Test 1250
FH	0.39	0.31	0.24	0.27	0.22	0.21	0.18	0.19	0.21	0.20
MB	0.06	0.11	0.12	0.18	0.17	0.15	0.17	0.18	0.19	0.18

Fig. 11. F1 values for Fuzzy Hybrid (FH) and Model-Based (MB) CF



	Test 125	Test 250	Test 375	Test 500	Test 625	Test 750	Test 875	Test 1000	Test 1125	Test 1250
FH	0.14	0.12	0.10	0.09	0.09	0.08	0.07	0.07	0.08	0.07
MB	0.02	0.04	0.05	0.07	0.07	0.06	0.07	0.08	0.09	0.08

Fig. 12. R values for Fuzzy Hybrid (FH) and Model-Based (MB) CF



	Test 125	Test 250	Test 375	Test 500	Test 625	Test 750	Test 875	Test 1000	Test 1125	Test 1250
FH	0.74	0.81	0.85	0.84	0.87	0.89	0.91	0.89	0.87	0.88
MB	0.96	0.94	0.93	0.90	0.90	0.93	0.91	0.90	0.89	0.90

Fig. 13. MAE values for Fuzzy Hybrid (FH) and Model-Based (MB) CF

It appears that both prediction techniques are sensitive to variation in sample size, (and implicitly, cluster number C = N/25). We can see that performance of FH decreases when the sample size grows while MB shows increasing performance and hence better scalability of MB. Reasons beyond this for MB is that having more clusters and UPs helps it find better matches for the active session. It can also be noted that FH seems to outperform MB in small-to-medium size samples, while the difference between the two techniques decreases as the sample size increases, becoming trivial for sample sizes around 875. Statistical significance of these differences between the two prediction algorithms is tested in Section 4.4 below. However, the fact that MB shows slightly better performance as measured by R at larger sample sizes prompted us to investigate the consistency of relative performance of the two algorithms. We refer the readers to [Suryavanshi et al., 2005 b] for further conventional comparison between the two techniques.

4.3 Consistency Analysis

We emphasize that our comparisons are conducted to assess if one method predicts better than the other across different scenarios. The fact that MB performed slightly better (for R with large samples) and differences between the two methods are negligible for other effectiveness metrics at larger sample sizes (>875) may point to the possibility of inconsistent relative performance if sample composition changes drastically. For example, although the samples used were random, suppose they comprised a majority of sessions that favored FH. What if MB performs better for specific groups of sessions and such groups were missing in the samples used in the experiments? In such a situation, FH would have been unjustifiably favored.

4.3.1 Experiments and Results

To address the above concerns, we designed an experiment to test how each of the two techniques performs on the sessions for which the other technique yielded "good' predictions. By good prediction (or recommendation) we mean at least one page of the hidden part of the test session was returned in the predicted pages. In this section, we compare the performance consistency of the two prediction systems in terms of this goodness, measured as accuracy of recommendation, using the metric F1.

In a prediction experiment, two categories of test objects are distinguished: the successful category comprising the objects for which the prediction system has given good predictions, and the unsuccessful category which only has objects for which the system made no good predictions (i.e. sessions that satisfy: |hidden set ∩ recommended set| = 0). We have subjected a random sample of 12,905 sessions to each of the two systems, and classified the resulting sessions from each system within the successful and unsuccessful categories. Then we composed a new population of 1000 objects in which there are 100 (10% of the population) objects randomly coming from the successful category of the FH system, and the rest (90%) coming from the successful category of the MB system. Then we subjected the new population to the two recommender systems and calculated the accuracy metric F1 for TOPN = 5. We repeated the experiment for different population combinations, with 10% increments, leading to 9 populations in total with a composition of 10%FH-90%MB; 20%FH-80%MB; 30%FH-70%MB,....;90%FH-10%MB. The results are shown in Figure 14.

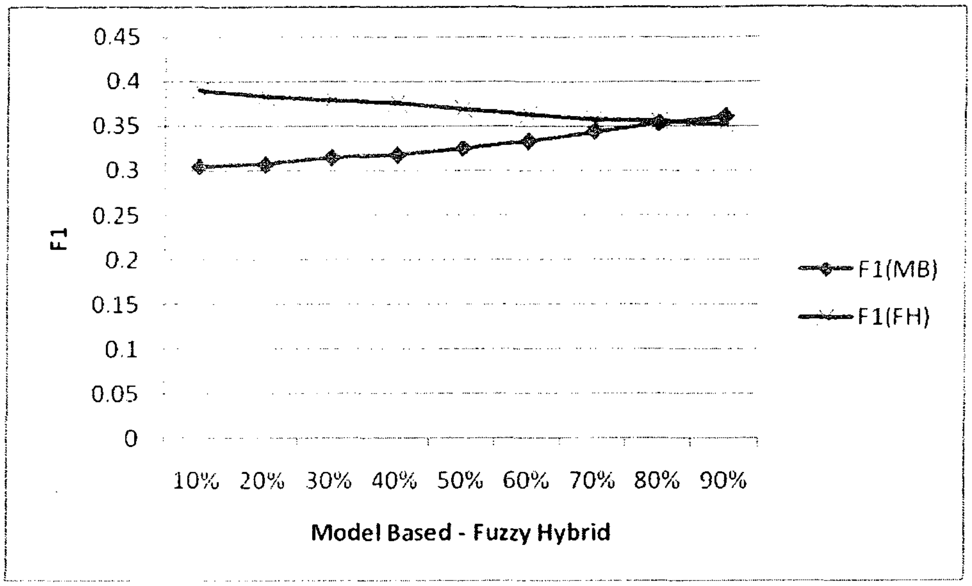


Fig. 14. Fuzzy Hybrid vs. Model Based Tolerant Combination

It is clear that FH is performing better than MB even on those samples where we know MB predicted accurately for most of the sessions. The MB method is successful only in samples in which more than 80% sessions come from MB. This confirms validity of the previous results (Section 4.2) showing a visibly superior performance of FH based recommender system. Statistical testing follows.

4.4 Significance Testing

In Section 4.2, two prediction systems, FH and MB, are applied to 10 sample sets of random sessions and compared for prediction effectiveness, using the metrics F1, R, and MAE. In Section 4.3, the two systems are applied to 9 designed 'random' samples and compared for prediction quality using the metric F1.

In both cases, the comparison involves paired observations, 10 in the first case, and 9 in the second. The statistical procedure is that of paired comparison, using the *Wilcoxon,* *signed ranks* test described previously in Chapter 3.

For the conventional test, we computed the values of each of the metrics F1, R, and MAE for each of FH and MB systems and compared the 10 pairs of values for each metric in the first experiment [(F1(FH) versus F1(MB); R(FH) versus R(MB); MAE(FH) versus MAE(MB)]. We formulated the relevant H_0 and H_a hypotheses as shown in Table V. As is clear from the Table, we have a situation of a unilateral test, as implied by the alternative hypothesis which states that FH is better than MB for each of the metrics. Recall that desired are higher values for F1 and R, and lower values for MAE.

The value "*T*" of the test statistic of the *Wilcoxon signed rank test* is obtained and reported in the last column of Table V and compared to the critical value $T_{0.05} = 11$, for 5% significance level, and $T_{0.01} = 5$ for significance level 1%.

The test is similarly applied to the performance consistency experiment, with 9 samples only, and a critical value of the test statistic, of $T_{0.05} = 8$, and $T_{0.01} = 3$ at the 5% and 1% significance levels, respectively.

4.4.1 Results

Results of the comparison of the two methods based on the criteria F1, R, and MAE showed a significantly superior performance of FH technique over MB. It can be concluded that FH system makes consistently better prediction than MB system, although the difference between the two prediction techniques becomes small for larger sample sizes.

Table V Tested hypotheses and metrics for comparison of the prediction algorithms

Hypothesis	Metric	Tested hypotheses (M_d: metric difference between the 2 techniques, *i.e.* $M_d = M_{FH} - M_{MB}$)	Calculated *T*-value
1	F1 Conventional	H_0: FH CF provides no better prediction than MB CF $\{M_d = 0\}$; H_a: FH CF is better than MB CF$\{M_d > 0\}$	0**
2	R Conventional	H_0: FH CF provides no better prediction than MB CF $\{M_d = 0\}$; H_a: FH CF is better than MB CF$\{M_d > 0\}$	10*
3	MAE Conventional	H_0: FH CF provides no better prediction than MB CF $\{M_d = 0\}$; H_a: FH CF is better than MB CF$\{M_d < 0\}$	0**
4	F1 Consistency	H_0: The two techniques perform equally when each is applied on a sample biased towards the other. $\{M_d = 0\}$; H_a: Fuzzy Hybrid CF performs better $\{M_d > 0\}$	1**
Critical *T* value, n=10; $\alpha = 0.05$			11
		$\alpha = 0.01$	5
Critical *T* value, n=9; $\alpha = 0.05$			8
		$\alpha = 0.01$	3

* Significant at 0.05 | ** Significant at 0.01 | ns: Non significant at 0.05

4.5 Summary and Conclusion

We conducted a dependable comparison between the two prediction systems FH and MB, covering different aspects. First, a conventional comparison was performed using well known metrics, namely F1, R, and MAE. We observed that as the sample size gets larger, FH's performance gets lower and MB's performance improves and gets close to FH.

However, the overall comparison of the two methods in all cases was in favor of FH. Second, we conducted a consistency experiment to test whether it is safer to always use one technique. To answer this, we tested each of the two techniques on datasets for which we obtained better accuracy using the other technique. We noticed that for most of these experiments, FH outperformed MB. Statistical testing confirmed the superiority of prediction by FH over MB. This result is consistent over a wide range of sample data sets of sessions. However, the difference between the two techniques becomes smaller as the sample size increases.

5 GOODNESS CRITERIA FOR WEB USAGE PROFILES

The exploitation of usage profiles (UPs) for applications such as clustering depends on the soundness and meaningfulness of the knowledge they carry. In Section 2.6 we provided a background on UPs and their quality. Early assessment of usage profile quality as mentioned relied on direct human judgment which suffers from being inefficient, not easily reproducible, and often subject to personal bias. To address this issue, we introduce here a new set of UP quality criteria coupled with corresponding computable measures to identify and quantitatively assess UP goodness. This should make UP evaluation more efficient and less susceptible to human subjective assessment. Our goodness criteria are defined so as to emulate viewpoints of human decision makers in evaluating meaningfulness and soundness of the UP set.

As an application and validation of these criteria, we deploy them in a post-classification procedure to find the best subset of clusters from the results of a clustering process. Using a carefully designed questionnaire, we then obtained expert opinions on the UPs generated from the subset of chosen clusters, which show a close match with what experts would expect to see as making up the user profiles. Details of our solution methodology and results are provided in this Chapter.

The rest of this chapter is organized as follows. In Section 5.2 we present our proposed UP quality criteria and their formulation. Section 5.3 studies and compares existing methods for cluster evaluation. Section 5.4 illustrates how UP quality can be used to identify best clusters in a post-clustering procedure. In Section 5.5, we describe the procedure used for expert evaluation of the results. We conclude in Section 5.6. For completeness, the UP evaluation questionnaire and the UPs analyzed by human experts are provided in Appendix B.

5.1 Goodness Criteria for Usage Profiles

5.1.1 Coherence

UP evaluation is mostly based on the existence of semantic justification for grouping together items (e.g. URLs, in our context) into UPs. We thus consider a UP to be coherent if it contains highly similar pages. For example, [Joshi et al., 1999] favored a clustering technique over another if it produced more "compact" clusters. Each UP derived from such compact clusters contained pages that dealt with one subject/topic. In contrast, pages within UPs derived from non-compact clusters dealt with varied subjects. This forms the first criterion and its associated measure is termed as *coherence*.

A suitable formula for coherence of a usage profile UP_a should consider similarities between all the pages it includes, for which we consider the following:

$$Coh_a = [\textstyle\sum_{pi \in \mathrm{UP}a} \sum_{pj \in \mathrm{UP}a} sim(p_i, p_j)]/(|\mathrm{UP}_a|\times(|\mathrm{UP}_a|-1)/2) \qquad (27)$$

where UP_a refers to pages present in UP_a, $1 \leq i \leq |\mathrm{UP}_a|-1$, $2 \leq j \leq |\mathrm{UP}_a|$, $j > i$, and $sim(p_i,p_j)$ is the similarity between those pages p_i and p_j in UP_a with frequencies greater than *min_F*. Our formulation of coherence reflects the overall similarity between all pairs of pages in UP_a. High Coh_a indicates coherent knowledge in the cluster. A UP that contains only one page has the highest coherence of 1.

5.1.2 Distinctness

The UPs should be fairly different from each other in terms of item (page) composition. UPs with much overlaps are not desirable [Mobasher et al., 2002] as it may indicate ineffective clustering. Also, [Joshi et al., 1999] disfavored a clustering technique because it produced different UPs having pages dealing with the same subject. This indicates that in a good clustering, different UPs should contain dissimilar pages dealing with different subjects. In the context of the clustering technique RFC-MDE by [Nasraoui et al., 2002], goodness of UPs is argued to reflect "distinct" user interests. Similar arguments were made in [Nasraoui et al., 1999 a] which proposed relational data tolerance in the clustering algorithm CARD and analyzed the UPs produced. Similarly, [Suryavanshi et al., 2006] compared ARCA and RFSC through qualities of the usage profiles yielded, and criticized UPs of ARCA since the number of similar profiles it generated was more than RFSC.

We propose two levels (or types) of distinctness, *crisp* and *fuzzy*. The former, defined below, refers to reduced overlap between two UPs, i.e., the more common items shared by UPs, the lower is their crisp distinctness.

$$CDist_{ab} = (|UP_a - UP_b| + |UP_b - UP_a|) \,/\, |UP_a \cup UP_b| \qquad (28)$$

where "-" and "∪" are set difference and union operations. Fuzzy distinctness on the other hand considers both the overlapping of page sets of different UPs as well as the similarity between pages belonging to different UPs, defined as follows:

$$FDist_{ab} = 1 - [\sum_{pi \in \mathrm{UP}a}\sum_{pj \in \mathrm{UP}b} sim(p_i, p_j)] / (|\mathrm{UP}_a|\times|\mathrm{UP}_b|) \quad (29)$$

where $FDist_{ab}$ is fuzzy distinctness between two usage profiles UP_a and UP_b, and $sim(p_i, p_j)$ is the similarity between pages p_i and p_j . Higher value of $FDist_{ab}$ indicates more distinct are UP_a and UP_b.

The two criteria of coherence and distinctness are also inherent in the definition of clustering, namely grouping similar or related objects and separating dissimilar or non-related ones.

5.1.3 Strength

The notion of UP strength can be perceived as the overall frequency of items within a UP. High overall frequency indicates many sessions within a cluster share same pages, which in turn indicates a high degree of similarity among the sessions that belong to the same cluster and hence reflects good clustering. This criterion was used by [Nasraoui et al., 2002] for favoring FCTMdd over FCMdd, because some FCTMdd clusters yielded "stronger" profiles than those by FCMdd. It has also been used by [Joshi and Krishnapuram, 2000] to discover invalid (low quality) UPs through their low overall weights. [Nasraoui et al., 2002] highlight the importance of distinguishing between strong and weak profiles, and emphasizes the role of "robust" (high frequency) pages in drawing such distinction. Similar remarks were made earlier by [Nasraoui et al., 1999 a] and [Nasraoui et al., 1999b] in the context of the fuzzy clustering techniques of CARD and RFC-MDE, respectively, arguing that high and low weights of a derived UP determined whether the UP was real or invalid. The strength of UP_a is directly derived from the weights of its pages, as follows:

$$Str_{\mathrm{a}} = [\textstyle\sum_{Pj \in UPi} f_j^a] \,/\, |\mathrm{UP_a}| \qquad (30)$$

Again, we emphasize that we only include frequent pages in representing the UP and calculating its strength.

5.1.4 Coverage

In our experiments in clustering, we have noticed that sometimes important items present in the web log were left out merely because their usage frequency was lower in comparison to others, or simply because the clustering failed to capture the correct gathering of data. The notion of coverage of a UP_a relates to the number of pages it contained which do not appear elsewhere in the UP set. We consider a set of UPs to be better than another if it covers more items (pages). Certainly, covering more items means revealing more relationships, which in turn indicates a potential to discover further knowledge about the data set. For a UP to have a high coverage, it needs to have a certain number of new items/pages, that is, pages present in this UP which are not already covered elsewhere in the UP set. After binarizing the weights, the item coverage of $\mathrm{UP_i}$ can be computed using the following formula:

$$Cov_{\mathrm{i}} = |\mathrm{P_i}|/|\mathrm{P_n}| \qquad (31)$$

where $\mathrm{P_n}$ is the set of all pages present in the sample sessions, and $|\mathrm{P_i}|$ is the number of new pages in $\mathrm{UP_i}$. Among the four UP quality criteria we proposed above, except for coverage which is mainly based on our own intuition, the other three are based on extensive study in the literature discussing manual methods of evaluating UPs.

WAVP [Mobasher et al., 2000] and F1 in [Nasraoui and Saka, 2006] are previously formulated measures for usage profile quality. Both of these are applicable in validating the process of summarizing a population into a few elements, whereas our criteria validate the goodness in partitioning this population into different groups. So for example, the previous criteria do not require that two groups should be different.

It is also important here to note that our criteria defined above are such that usage profile assessment is now internal rather than external, i.e., there is no requirement for external inputs or human effort.

5.2 Comparison with Cluster Evaluation Criteria

The concepts of recall and precision and the notion of matching between evaluated and gold clustering are virtually the basis for most ideal clustering criteria, including F measure, Purity, Entropy, and Rand Index [Manning et al., 2008]. An exception is in Mutual Information which has also information theoretical basis.

At the abstract level, there is certainly an overlap between these criteria and our definition of UP coherence. This overlap is evident when all pages in one cluster belong to one topic (or ideal cluster). In this case this will not only show high recall and precision, but also high coherence. Of course, this is assuming that the sense of belonging to one topic is reflected accurately in the page similarity matrix. The difference arises from two factors. Firstly, and very importantly, UP coherence does not require any external input, i.e., the ideal clustering to function, but rather uses only page-to-page similarity. Clearly such external input requires expertise and time. Moreover, "expertise" depends on the expert who creates the ideal clustering. Also different experts may give different inputs. Secondly, UP coherence is defined for significant (frequent) pages, and not for all pages in a cluster.

We believe a page-to-page similarity is more appropriate than page-to-topic relationships in reflecting goodness of UPs. If we were to evaluate clustering results directly and an ideal clustering was available, then ideal clustering criteria surely reflect the goodness of clustering better than UP coherence and distinctness criteria. But in evaluating UPs, we need to bear in mind that a UP represents a navigational pattern, and pages in a good pattern do not necessarily need to share one common topic. Users may start with a topic but end their navigation in a different topic. A good navigation is where pages accessed are similar. More specifically, if we represent pages as nodes and similarities between pages as lines connecting them, then a sound navigation is when we have a connected graph, and not necessarily a complete graph.

It is understandable that internal evaluation criteria reflect common clustering goodness features, usually compactness and separation. As mentioned earlier, there is a strong connection between validity indices and the first two UP criteria; coherence corresponds to compactness and distinctness to separation, however their formulation have been tailored to UP contents. As such, UP goodness criteria look at clustering from the viewpoint of features of the clustered objects rather than look at the objects directly. UP criteria also look at the clustering from an aggregate level rather than from the object level. Furthermore, in order to decrease the impact of noise in the evaluation process, UP criteria look only at significant (i.e., highly frequent) features of clustering, and not all features of each clustered object.

It must be pointed out that there is also a difference between UP coherence and distinctness, and the other two goodness criteria.

Strength is specific to UP quality, since the notion of frequency is not present in clustered objects. The term coverage as used in [Crabtree et al., 2005] is different from ours. Their use reflects the number of pages covered by clusters relative to a topic (or ideal cluster), and specifically with respect to the clustering precision. As a goodness criterion, our reference to coverage is relative to the whole page population. In a sense, our coverage reflects the main purpose of knowledge mining through web usage clustering rather than the validity. As mentioned in the coverage section, more new pages covered in UPs means more relationships and knowledge revealed. While coherence, distinctness, and strength reflect validity of such new pages in the UP, coverage in our sense indicates "how many items of the overall population which are not represented otherwise (in any cluster) are present in a UP." Perhaps the closest concept in page clustering to such coverage is cluster cardinality (number of pages). However, cardinality does not measure in any way the new knowledge in the cluster, and is usually limited to crisp clustering.

5.2.1 Applicability of Goodness Criteria

We designed these four quality criteria, realizing the importance of each one as well as their inter-relationships which may have a bearing on the overall UP quality. For example, if we decrease the threshold value for frequency and more pages are revealed in a UP, the coverage could increase. Such increase may reduce coherence, and reduce the chances to have a distinct UP. Also adding such "low" frequency pages would surely decrease UP strength.

However, it is clear that coverage is dependent on the number of existing clusters. The relationship is not necessarily monotonic. Initially, as the number of clusters and correspondingly UPs increases, more pages will be covered. But as this number increases, the chance that new UPs would cover new pages reduces. So when comparing two UPs, the number of clusters should also be taken into account. The frequency threshold should also be fixed in the comparison, since it has an impact on the measures. We do not claim completeness or perfection of our criteria in covering all aspects of UP quality. More work is required in improving the reflection of our measures to current and future quality aspects.

5.3 Using 'UP' Quality to Identify Best Clusters

This section discusses the use of the aforementioned goodness criteria for UPs in extracting best set of clusters from the results of a web usage clustering technique. We modify the RFSC clustering technique [Suryavanshi et al., 2005 a] to choose clusters on the basis of their UP quality. It is important to note here that most clustering techniques incorporate a process of choosing the best set of clusters, either by measuring the cluster validity index, or by thresholds for cluster cardinality, etc. Applicability of our goodness criteria is thus independent of the particular web usage clustering technique used.

5.3.1 Clustering Technique

The relational fuzzy subtractive clustering technique (RFSC) was chosen for its good accuracy and scalability to large web usage data [Suryavanshi et al., 2005 a, 2006], and also because a robust implementation of it was readily available.

Different approaches were adopted to determine C, the number of clusters. In the original

RFSC [Suryavanshi et al., 2005 a], objects and their potentials are required to satisfy certain criteria according to predetermined accept ratio ($\overline{\in}$) and reject ratio ($\underline{\in}$) to be chosen as cluster centers. Choosing the best set of clusters involved a search to find the values for $\overline{\in}$ and $\underline{\in}$ for which the goodness index of RFSC is minimum. Another approach is to run the clustering with a varied value for C, and then to choose the C for which evaluation criteria show best clustering. Another approach is to first run the algorithm with a fix value for C that is generally much larger than the expected number of clusters, and then "low" cardinality clusters are pruned. A crispification is required in order to calculate the cardinality of fuzzy clusters. Our proposed version of RFSC "RFSC with UP-based cluster filtering" is a modification of the last approach. It chooses a slightly larger value for C, and as the initial C, it gets the one produced by RFSC. Then it selects the best set of clusters based on the quality of UPs, as explained in detail in the next section.

5.3.2 Selection of Best Set of Clusters

Our procedure is based on the following steps:

1. Run the RFSC algorithm for a value of C that is slightly larger than expected range of C.
2. Generate the UPs for the C clusters obtained from RFSC.
3. Order UPs based on their combined strength and coherence (after calculating them).
4. Scan the UPs starting from the set that includes strongest and most coherent UP; keep adding the selected UP to the best set of UPs that have all the following conditions;

the thresholds mentioned below are set heuristically and may be changed as per the data set/domain:

Distinctness > 0.9, Coherence > 0.25, $|P_i| > 2$ (for coverage), Strength > 0.5.

The fuzzy version of distinctness is adopted in this procedure. Distinctness for a UP here stands for the minimal distinctness value between this UP and previously selected UPs. "Coverage" for a UP is again redefined in terms of previously selected UPs instead of the complete UP set. The clusters associated with UPs in the final selected set are the best set of clusters.

5.3.3 Advantages of Cluster Selection by UP Goodness Criteria

At the end of any fuzzy clustering, what concerns users the most is the semantic information contained in the usage profiles (rather than cluster centers and memberships). The C clusters chosen via the use of a validity index may or may not contain the right semantic indication. Thus, validity indices only cover the aspect of clustering soundness, reflected by compact and well separated clusters. Validity indices often are susceptible to bias because of their relationship to C [Ketata et al., 2009]. Choosing C on the basis of our goodness criteria surely improves clustering performance for web usage mining. More specifically, the procedure results in an added semantic dimension to cluster quality, *i.e.* the meaningfulness of clusters, in addition to the soundness of clustering.

Our procedure also saves computation time while focusing on cluster quality. Finding the best set of clusters by varying C, say from 2 to N/3 as suggested by ARCA or by varying accept and reject ratios as done in RFSC, requires computation of validity indices for each set of clusters. For large data sets, this can be computationally prohibitive. In contrast, in our solution approach, the clustering algorithm runs only once for a slightly larger value of C, and requires computation of goodness criteria only for significant items in a cluster, which corresponds to summarizing a group of sessions into one virtual session (the UP). This definitely improves computational time.

5.3.4 Experimental Settings

In these experiments, we performed the same preprocessing steps mentioned in Section 2.1, including URL similarity, sessionization, session similarity, etc.

We implemented the modified clustering system and run on data of 500 sessions, with an over-specified C value of 50. We set the following values for UP quality thresholds: Strength= 0.5, Coherence= 0.5, Distinctness= 0.9, and P_i for Coverage= 2. The algorithm provided five groups of UPs, one of which (named "A") comprised UPs that met all quality thresholds. Each of the 4 remaining UP groups failed to meet the threshold requirement of a distinct quality parameter, called as non-coherent UP set or category ("C"), non-distinct UP category ("B"), Weak UP category ("D"), and non-covering UP category ("E"). The first set A is called the acceptable UP set. We had to limit to 5 the number of displayed UPs in each set, simply for easy reference later when subjecting the cluster contents for manual analysis by experts. For the acceptable UP set, we picked the five UPs that were selected first by the procedure described earlier in Section 5.2. For identification of the UP, we used the ID of the corresponding cluster, which is a number between 0 and 49; for example, UP_4 stands for the UP of the 4th cluster, and so on (See Table VI and Appendix B).

5.3.5 Results and Observations

Appendix B shows 5 UPs from each of the 5 UP categories. Each URL is followed by its frequency (weight).

Within the acceptable-UP category, each UP seems to show coherence by handling one subject:

UP_0: COMP218.

UP_29: Professor #1 (P1) teaching COMP444.

UP_31: Professor #2 (P2) teaching COMP354.

UP_43: COMP352.

UP_2: Professor #3 (P3) teaching COMP321.

Also the UPs look very distinct from one another. In addition, the pages in each UP seem to cover well certain usages of the website. Finally, the UPs show relatively high overall weight of their pages.

In the non-distinct UP category B (See Appendix B), the field Closest_UP in Table VI states the ID of the closest UP to each non-distinct UP. It is clearly seen that the assigned closest UP is appropriate. UP_27 and UP_5 are not much different from UP_0, as all of them handle pages related to COMP218. Similarly, UP_6 and UP_49 cover pages of COPM354 already covered by UP_31. UP_43 covers principal pages in COMP352, and this is mainly what UP_7 covers also.

As regards to the non-coherent UP set C (See Appendix B), each of the following UPs seem to capture very different subjects:

UP_38:COMP353, Professor#4 (P4) main folder, and Help issues.

UP_1:COMP346 and COMP354 (we consider the page current_students relevant to the *courses).*

UP_24: COMP442, COMP444, COMP354, COMP238, and COMP346.

UP_25: SOEN337, SOEN341, SOEN347, and SOEN384.

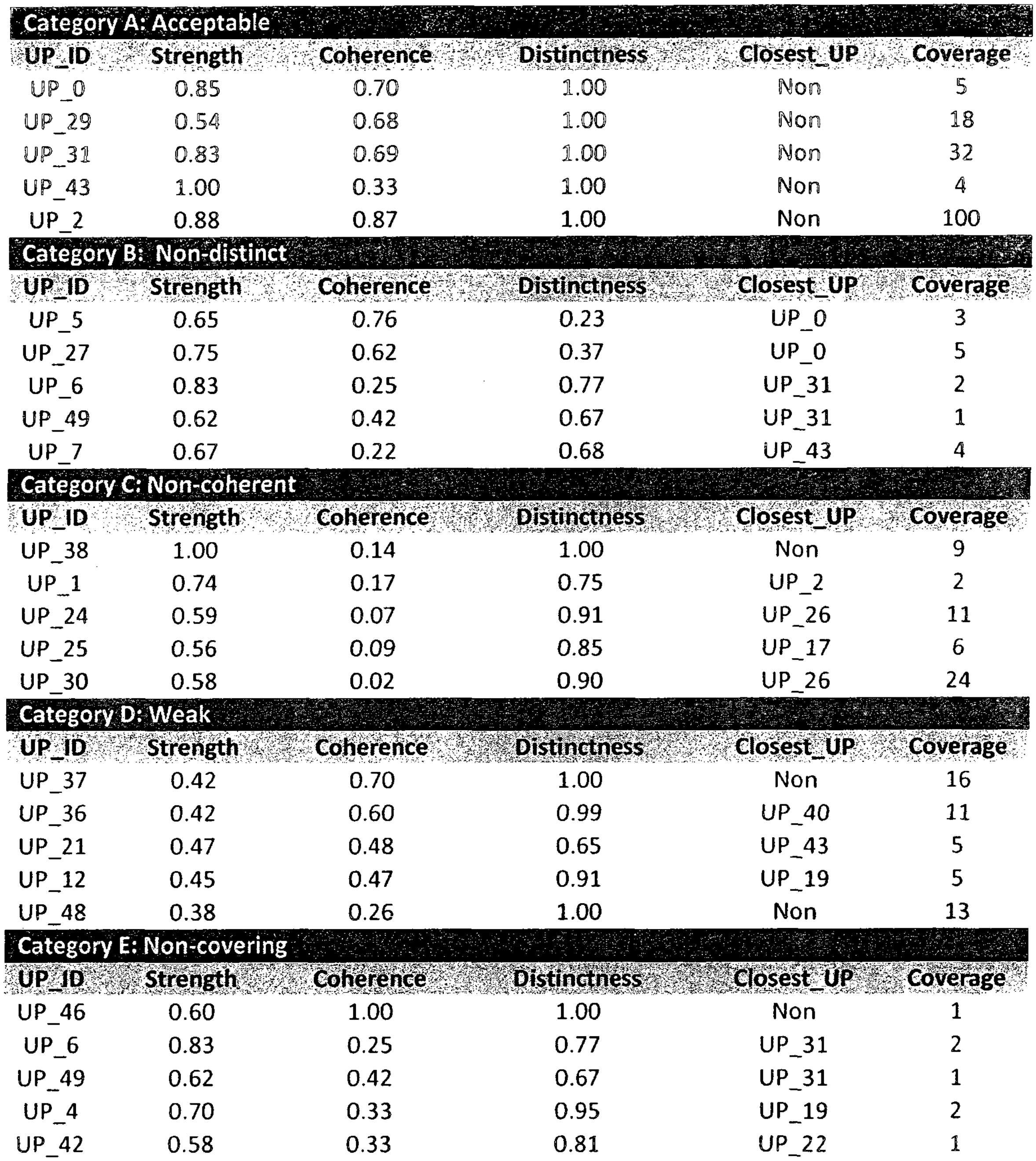
UP_30: About 25 different professors.

For the weak UP category D, it can be observed that, except for UP_21, these UPs have a relatively low average weight (< 0.5). It may be recalled that pages with weights less than 0.25 are not taken into account. For UP_21, there are two pages with relatively high weights 0.8 and 0.7, but the average weight of UP_21 pages is still small (See Table VI).

Finally, in the non-covering UP category E, each UP does not have adequate number of pages not already present in A.

From the above, it does seem that our technique does provide promising results by appropriate detection of desirable UPs. However, we have observed that further improvements are possible. For example, when analyzing the coherence of UPs, we discovered a limitation of URL similarity between pages in its inability to capture different aspects of the similarity that can be otherwise captured by human observation. Example is the UP_34, where a person can eventually conclude that the whole UP describes interests of prospective students. A semantic similarity, although possibly expensive, would certainly help in identifying more meaningful UPs. Usage-based similarity could also be added. Yet another possibility is to give more weight to higher folders within URLs, so that for example: "/~comp442/courseNotes" becomes more similar to "/~comp442/Assignments" than to "/~soen552/courseNotes."

Table VI. Different UP quality measures for 5 UPs of each of the five UP categories



Category A: Acceptable					
UP_ID	Strength	Coherence	Distinctness	Closest_UP	Coverage
UP_0	0.85	0.70	1.00	Non	5
UP_29	0.54	0.68	1.00	Non	18
UP_31	0.83	0.69	1.00	Non	32
UP_43	1.00	0.33	1.00	Non	4
UP_2	0.88	0.87	1.00	Non	100
Category B: Non-distinct					
UP_ID	Strength	Coherence	Distinctness	Closest_UP	Coverage
UP_5	0.65	0.76	0.23	UP_0	3
UP_27	0.75	0.62	0.37	UP_0	5
UP_6	0.83	0.25	0.77	UP_31	2
UP_49	0.62	0.42	0.67	UP_31	1
UP_7	0.67	0.22	0.68	UP_43	4
Category C: Non-coherent					
UP_ID	Strength	Coherence	Distinctness	Closest_UP	Coverage
UP_38	1.00	0.14	1.00	Non	9
UP_1	0.74	0.17	0.75	UP_2	2
UP_24	0.59	0.07	0.91	UP_26	11
UP_25	0.56	0.09	0.85	UP_17	6
UP_30	0.58	0.02	0.90	UP_26	24
Category D: Weak					
UP_ID	Strength	Coherence	Distinctness	Closest_UP	Coverage
UP_37	0.42	0.70	1.00	Non	16
UP_36	0.42	0.60	0.99	UP_40	11
UP_21	0.47	0.48	0.65	UP_43	5
UP_12	0.45	0.47	0.91	UP_19	5
UP_48	0.38	0.26	1.00	Non	13
Category E: Non-covering					
UP_ID	Strength	Coherence	Distinctness	Closest_UP	Coverage
UP_46	0.60	1.00	1.00	Non	1
UP_6	0.83	0.25	0.77	UP_31	2
UP_49	0.62	0.42	0.67	UP_31	1
UP_4	0.70	0.33	0.95	UP_19	2
UP_42	0.58	0.33	0.81	UP_22	1

Further evaluation of the results of our technique through an expert feedback study is presented in the next section.

5.4 Usability Evaluation Questionnaire

This questionnaire aims at using expert opinions to confirm that UP quality characteristics derived through our technique match users' observations regarding UP composition. The questionnaire is provided in Appendix B. It consists of three parts: an introductory text explaining the procedure for answering the questions, a table of statements (questions), and an Excel sheet containing different UPs. Responses were sought from 11 users including web master, system analysts, student advisors, teaching associates and faculty members, of which 8 responded.

Comments on the questionnaire:

Statements 1 and 2 reflect on coherence. As mentioned in Section 5.2.1, if pages are concerned with one common subject, that indicates they are conherent. Although the term "coherence" bears a relative connotation, the least requirement for coherent pages is that the overall similarity between pages is reasonable.

Statements 3, 4 and 5 reflect on distinctness. We recall from Section 5.2.2 that, UPs are considered non-distinct if they have pages relating to a common subject. Two pages are considered related to a common subject if they are similar. Statement 3 verifies that each UP_i of category B is not distinct from category A. On the other hand, distinctness among a set of UPs is indicated either: (a) through non-overlap (crisp distinctness), i.e. no sharing of common pages, which is verified in statement 4, or (b) through page dissimilarity (fuzzy distinctness), which is reflected in statement 5.

We admit that with a distinctness threshold of 0.9, the technique may not produce as distinct UPs as claimed by the statement. If one wants UPs that have 0 overlap or 0 overall similarity between their pages, higher threshold would be required.

Statements 6 and 7 reflect on coverage. Statement 6 investigates whether the evaluators agree on the low coverage of category E. By setting the coverage threshold to 2, we wanted to make sure decision makers will not miss knowledge about several pages by pruning out a low coverage UP. This is based on the assumption that the word "several" means at least 3. As such, statement 6 tests whether evaluators find several pages in a UP of category E missing in category A. If that was the case, then this indicates miss-configuration of our technique, mainly in the coverage threshold. We recall that the threshold is set heuristically, and is highly dependent on the dataset to be clustered. We recommend to set the thresholds which best suit the needs. For example, some may need all the pages in the sample to be present in UPs, in which case the coverage threshold should be set to 1, and so on. Similarly, statement 7 makes sure that each UP in category A (considered by our technique as highly covering) actually provides new knowledge about several pages. This is satisfied if such UP had new pages not present in other UPs in A.

Statements 8 and 9 reflect on strength. Statement 8 verifies whether category A provides "strong" UPs. A strong UP would be generally composed of "strong" pages, i.e., high weight pages. Similarly, statement 9 verifies whether UPs of category D are "weak" through questioning whether their pages have generally low weights.

While the criteria used in the method are purely quantitative, the questionnaire has some degree of fuzziness (using such words as appear, seem, generally) similar to the type of questions normally raised to a UP inspector evaluating a new set of UPs. It must be noted that the UP evaluator has normally no "standard" to compare against the set of UPs; he/she just uses his/her own judgment based on his/her knowledge and background and what he/she considers a good UP set.

It is also noted that the objective of the questionnaire is not to evaluate the goodness of the clustering, but rather the ability of our technique to discriminate between good UPs (selected) from low-quality UPs (rejected) through their relative comparison.

Evaluators' responses were as expected for statements 1, 2, 3, 5, 7, 8, and 9. Statement 5 on distinctness lacked accuracy on "number" of pairs of UPs in A, and led to ambiguity noted by picking “don’t know” by 2 of the 8 respondents.. For question 4 on distinctness, it seems the statement was too demanding, as one page may be common to two UPs that are still reasonably distinct. Again the statement does not specify the number of common pages and this may be the main source of discordance; also the heuristic threshold could be raised further to perhaps reduce the chance for ambiguity. Statement 6 on coverage is perhaps the least satisfactory among all for its ambiguity which stems from the word (several), a specific number, say “more than 2”, would have helped interpret the question better.

5.5 Conclusion

Usage profiles are used extensively ranging from their use for evaluation of clustering to their use in various applications as highly concise representations of clusters themselves. While the methodologies for evaluation of clustering are well established, the quality of a set of usage profiles has been judged manually so far by experts. The principal contribution of this work is the specification of a comprehensive set of four goodness criteria with computable measures for defining the quality of a set of usage profiles representing a given usage data set. These four criteria termed coherence, distinctness, strength and coverage can be computed internally, with no requirement of any external input, whatsoever. The criteria have been formulated by exhaustively studying the literature to accurately reflect the different judgments made manually by experts in deciding on the goodness of usage profiles. While coherence and distinctness are closely related to the well known cluster related measures of compactness and separation respectively, the other two measures, namely, strength and coverage are specific to usage profiles and are being defined for the first time in this work. An important point to note is that all these measures are needed to be specifically defined for usage profiles, because the very intent of usage profiles is to capture navigational patterns. In contrast, many cluster evaluation measures such as purity, F index, etc. are more concerned with semantic consistency. Thus, items belonging to different topics may be part of the same navigation pattern, but would tend to devalue cluster quality. We have implemented the use of these goodness criteria for selection of a good subset of clusters. Expert evaluation of the selected clusters does support the fact that these goodness criteria work well.

6 UP BASED COMPARISON OF CLUSTERING TECHNIQUES

In this chapter we compare the quality of usage profiles derived from two different clustering techniques: RFSC [Suryavanshi et al., 2005 a] and FCMdd [Krishnapuram et al., a 1999]. Again, our choice of RFSC was due to its claimed accuracy, scalability, and the availability of its implementation. We have chosen the version RFCMdd of FCMdd for its robustness and better performance [Krishnapuram et al., 1999]. While the experiments in Chapter 5 measure the quality of individual UPs, it is intended here to assess the overall quality of the whole UP set, on the basis of which, a judgment will be made on the relative value of the underlying clustering technique. The overall quality of the UP set from a given clustering technique will be computed as an aggregate value of the corresponding quality metric for each of the individual UPs for that technique. This is explained in more detail in section 6.3. Sections 6.4 and 6.5 address the experimental settings and results respectively. A conclusion is provided in section 6.6. In computing the aggregate overall value of the quality metric, due consideration will be given to frequency of involved items. In the next two sections, we show how to derive overall coherence and overall distinctness, considering the frequency of related pages.

6.1 Weight-Aware Coherence

To compute the coherence of a certain usage profile UP_i, we take into consideration page weights and build the following series of page pairs:

$$PP_i = \{(p_j, p_k) | \; j \neq k \; ; \; p_j, p_k \in UP_a \; (f_j^i > \alpha \; ; f_k^i > \alpha) \; \} \qquad (32)$$

where p_j and p_k are the j^{th} and k^{th} pages, f_j^i and f_k^i are their frequencies in UP_i, respectively.

We then calculate two quantities for each pair (p_j, p_k) in each PP_i:

1. sim_{jk}: The semantic/structural similarity between p_j and p_k

2. $sf_{jk}^i = 1 - |f_j^i - f_k^i|$: the similarity between the frequencies of p_j and p_k in UP_i, or frequency closeness.

From the above, two pages in a UP are similar not only if their structures are so, but also if they have close frequencies in that UP.

The coherence Coh_i of UP_i consists of two quantities shown below:

$$Coh_i = Coh_{i1} \times (Coh_{i2} + 1)/2 \qquad (33)$$

where:

$$Coh_{i1} = \sum_{p_j \in \mathrm{UP}_i} \sum_{p_k \in \mathrm{UP}_i, k>j} sim_{jk} \times sf_{jk}^i \Big/ \left(|UP_i| \times (|UP_i| - 1)/2\right) \ ; \ 0 \leq Coh_{i1} \leq 1 \qquad (34)$$

and

$$Coh_{i2} = \frac{\sum_{p_j \in \mathrm{UP}_i} \sum_{p_k \in \mathrm{UP}_i} \left(sim_{jk} - \overline{sim^i}\right) \times \left(sf_{jk}^i - \overline{sf^i}\right)}{S^2_{sim^i} + S^2_{sf^i} + \left(\overline{sim^i} - \overline{sf^i}\right)^2} ; \qquad 0 \leq Coh_{i2} \leq 1 \qquad (35)$$

Coh_{i2} represents the coefficient of concordance correlation [Kuei, 1989], [Nickerson, 1997], [Lin, 2000] between the two parameters sim_{jk} (page semantic similarity) and sf_{jk}^i (page frequency closeness).

The similarity sim_{jk} between the j^{th} and the k^{th} pages in UP_i can be binary if there is no other form of similarity between different pages/items.

For the two vectors of UP_i, similarity and frequency closeness, Coh_{i1} reflects how high the values of such vectors are. The higher the similarities between pages, and the frequency closeness for pages, the more coherent UP_i is. If there is no semantic similarity between pages, then Coh_{i1} becomes 0. On the other hand, Coh_{i2} reflects two aspects: (a) how correlated the two vectors are, i.e., whether frequency closeness follows the ups and downs of similarity, or it contradicts it. This is mostly reflected by the numerator in Coh_{i2}; and (b) how close such vectors are on average, and this is reflected by $\left(\overline{sim^i} - \overline{sf^i}\right)^2$.

6.2 Weight-Aware Distinctness

As done above for coherence, we also build the page pair series, but this time two pages in a pair belong to two different UPs for which we wish to calculate distinctness. We derive the weight-aware distinctness of a pair of UPs (say UP_a and UP_b), and then proceed to derive an overall distinctness metric *DIST* for the whole UP set.

$$PP_i = \{(p_j,p_k)]|\ p_j \in UP_a,\ p_k \in UP_b\ (f_j^a > \alpha\ ; f_k^b > \alpha)\ \} \qquad (36)$$

So i ranges from 1 to $|UP_a|\times|UP_b|$. We then calculate two quantities for each pair (p_j,p_k) in each PP_i similar to the ones done in weight-aware coherence:

1. sim_{jk}: The semantic/structural similarity between p_j and p_k

2. $sf_{jk}^{ab} = 1 - |f_j^a - f_k^b|$: the closeness between the frequencies of p_j and p_k in UP_a and UP_b

The distinctness $Dist_{ab}$ between UP_a and UP_b consists of two quantities as shown below:

$$Dist_{ab} = 1 - rel_{ab1} \times (rel_{ab2} + 1)/2 \qquad (37)$$

where:

$$rel_{ab1} = \sum_{p_j \in UP_a} \sum_{p_k \in UP_b} sim_{jk} \times sf_{jk}^{ab} \ /(|UP_a| \times |UP_b|);\ 0 \le rel_{ab1} \le 1 \qquad (38)$$

and $$rel_{ab2} = \frac{\sum_{p_j \in \mathrm{UP_a}} \sum_{p_k \in \mathrm{UP_b}} \left(sim_{jk} - \overline{sim^{ab}}\right) \times \left(sf^{ab}_{jk} - \overline{sf^{ab}}\right)}{S^2_{sim^{ab}} + S^2_{sf^{ab}} + \left(\overline{sim^{ab}} - \overline{sf^{ab}}\right)^2}; \quad 0 \leq rel_{ab2} \leq 1 \qquad (39)$$

where sim^{ab} is the overall similarity between pages in UP_a and ones in UP_b.

rel_{ab2} also represents the coefficient of concordance correlation [Kuei, 1989], [Nickerson, 1997], [Lin, 2000] between the page semantic similarity sim_{jk} and page frequency closeness sf_{jk}^{i}.

6.3 Overall UP Quality

From the above definitions and formulae, we derive 4 "overall" UP quality criteria to describe a whole set of UPs generated from a particular clustering algorithm which we use to characterize the whole set of UPs. We later deploy these overall quantities to compare different clustering algorithms or techniques. The 4 overall UP quality criteria are computed as follows:

- Overall Coherence: The overall coherence *COH* of the whole set of UPs is the average of Coh_i, the individual coherence of the $\mathrm{UP_i}$s in that set.

COH = average of all $\mathrm{Coh_i} = \sum_i \mathrm{Coh_i/C}$ (40)

where $\mathrm{Coh_i}$ is the coherence for an individual $\mathrm{UP_i}$ and C is the number of clusters;

- Overall Coverage: $COV = |\,\mathrm{P_{cov}}\,|/(|\mathrm{P_n}|\times \mathrm{C})$ (41)

where $\mathrm{P_{cov}}$ is the set of pages covered by all UPs derived from a certain clustering technique, and $\mathrm{P_n}$ is the set of all pages in the sample sessions. It is clear that such *cov* will have a very small range.

- Overall Distinctness: $DIST = \sum_a \sum_b \text{Dist}_{ab}/(C\times(C-1)/2)$ (42)

with summation over all pairs of UPs

- Oveall Strength: $STRN = \sum_a \text{Str}_a/C$ (43)

is the strength of an individual UP_a as given in formula (30).

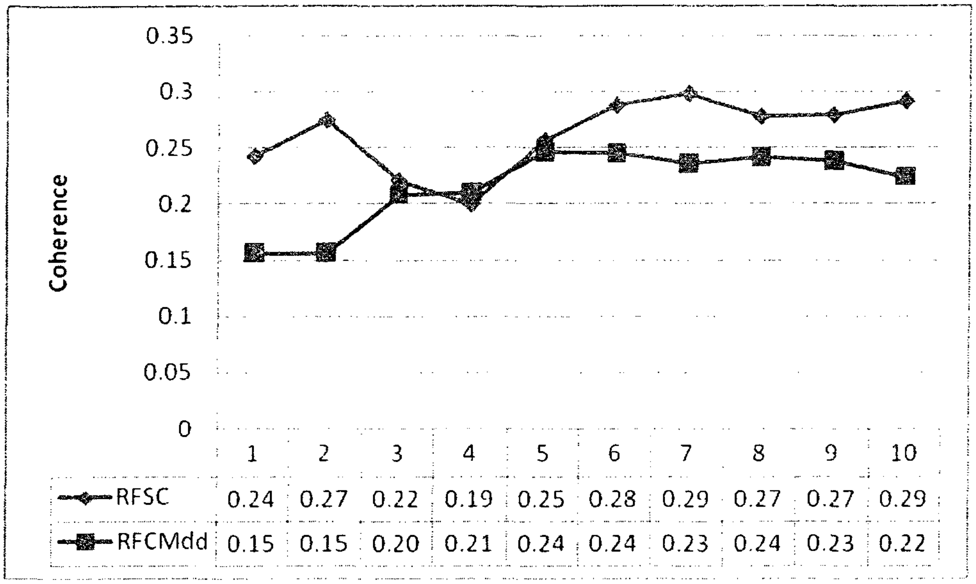
The values for all the 4 metrics lie within the range [0,1].

6.4 Experimental Settings

As done in Chapter 3, we used 10 samples but with sizes ranging from 500 to 5,000 sessions with a step size increment of 500. The settings are also similar to those in Chapter 3, including the values *overspecified_C* and *Min-Cardinality* for both techniques. These settings are chosen heuristically, with consideration to the large samples (large in comparison to the previous experiments). Choosing large samples is dictated by the need to better formulate the UPs and to simulate real web data set which are very large. ARCA was not included in the comparison because of the excessive time duration it would need (several days) if applied to such large size samples.

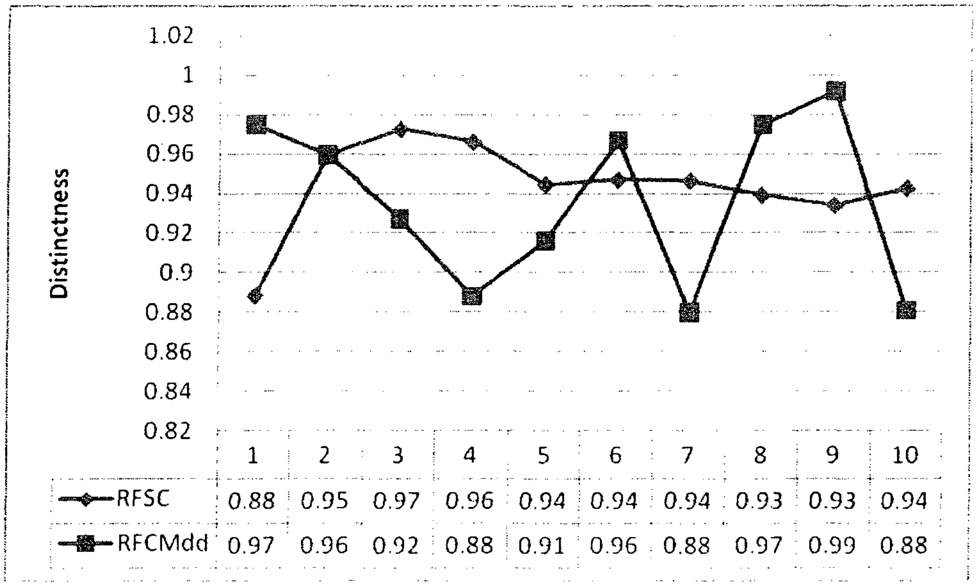
6.5 Experimental Results

We ran the two clustering techniques RFSC and RFCMdd on each of the 10 samples, and for each run we computed the four UP quality metrics for coherence, distinctness, strength and coverage. Figures 15, 16, 17, 18 illustrate the variation of the metric values for each of the clustering techniques across the different samples. Metric values for these clustering techniques across 10 sample sizes were compared and tested for statistical significance through the Wilcoxon signed rank test [Black, 2008].



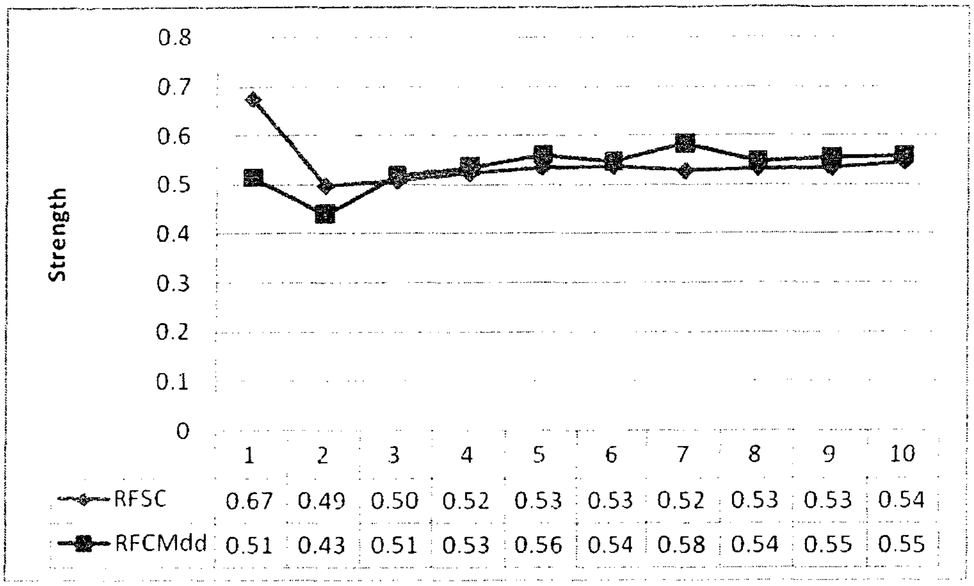
	1	2	3	4	5	6	7	8	9	10
—◆—RFSC	0.24	0.27	0.22	0.19	0.25	0.28	0.29	0.27	0.27	0.29
—■—RFCMdd	0.15	0.15	0.20	0.21	0.24	0.24	0.23	0.24	0.23	0.22

Fig. 15. UP Coherence values for RFSC and RFCMdd



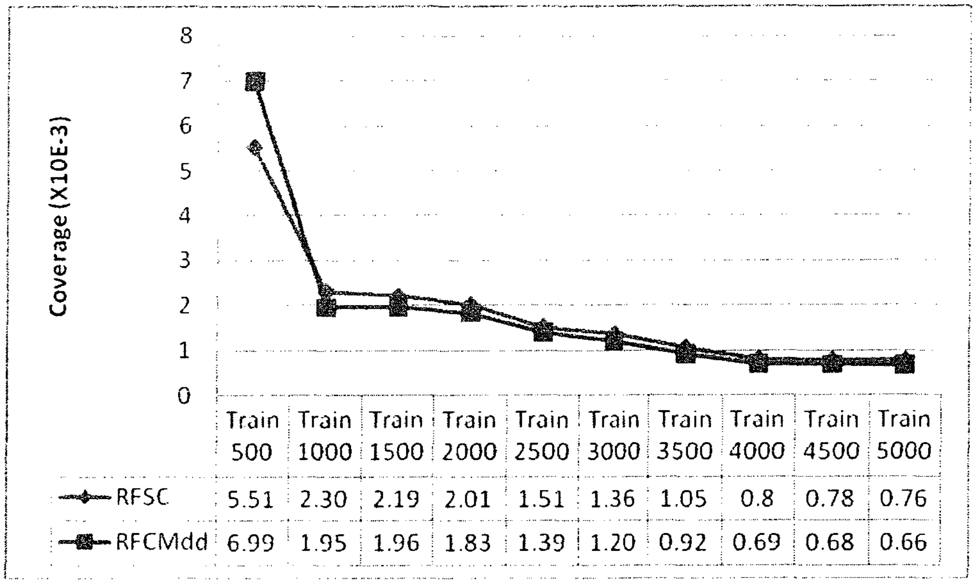
	1	2	3	4	5	6	7	8	9	10
—◆—RFSC	0.88	0.95	0.97	0.96	0.94	0.94	0.94	0.93	0.93	0.94
—■—RFCMdd	0.97	0.96	0.92	0.88	0.91	0.96	0.88	0.97	0.99	0.88

Fig. 16. UP Distinctness values for RFSC and RFCMdd



	1	2	3	4	5	6	7	8	9	10
RFSC	0.67	0.49	0.50	0.52	0.53	0.53	0.52	0.53	0.53	0.54
RFCMdd	0.51	0.43	0.51	0.53	0.56	0.54	0.58	0.54	0.55	0.55

Fig. 17. UP Strength values for RFSC and RFCMdd



	Train 500	Train 1000	Train 1500	Train 2000	Train 2500	Train 3000	Train 3500	Train 4000	Train 4500	Train 5000
RFSC	5.51	2.30	2.19	2.01	1.51	1.36	1.05	0.8	0.78	0.76
RFCMdd	6.99	1.95	1.96	1.83	1.39	1.20	0.92	0.69	0.68	0.66

Fig. 18. UP Coverage values for RFSC and RFCMdd

The results show consistently better (P<0.01) overall coherence for RFSC as compared to RFCMdd. In contrast, results for distinctness do not show consistent pattern of relative performance of the two clustering methods. In general, distinctness values are high for both methods, the lowest values being around 0.88. Despite the non-significance (P>0.05) of the differences between the two methods and their good performance in this respect, RFSC values were more stable in comparison to the fluctuating distinctness values for RFCMdd. Strength values noted were around 0.5 to 0.6, indicating no significant (P>0.05) differences between the two techniques. Coverage values were very small, perhaps because of the large number of pages involved in sample sessions. The differences in coverage between the two techniques were small, but consistently and significantly (P<0.05) in favor of RFSC. The overall better performance of RFSC compared to RFCMdd confirms previous results presented in Chapter 3 indicating superiority of RFSC based on clustering validity index and computation time.

It may also be noted that the values of all 4 UP quality metrics were different in the smallest sample size, but thereafter showed no apparent relationship with sample size or cluster number. That is with the exception of coverage that showed a steady decrease as sample size increased, because of the larger number of pages involved in large sample.

6.6 Summary and Conclusion

Using the UP goodness criteria introduced in Chapter 5, we derived a set of 4 UP goodness criteria to characterize a whole UP set generated from a particular clustering algorithm, taking into consideration page weights for coherence and distinctness as appropriate. We used the overall UP goodness criteria to assess the relative value of the underlying clustering technique. Based on such assessment, we compared two fuzzy clustering techniques: RFSC and RFCMdd. The results of our numerous experiments showed significantly better overall UP coherence for RFSC as compared to RFCMdd.

Distinctness values were generally high for both clustering techniques with no significant difference between the two, although the distinctness values were much less varying in RFSC than in RFCMdd. Strength values for the two methods were similar and in the range 0.5 to 0.6, while coverage was low in both, perhaps because of the large number of pages in the data set. There is no apparent relationship between the overall goodness criteria and sample size except for overall UP coverage that decreased for larger samples. Results from application of these 4 goodness criteria is in line with previous results, indicating the soundness of the methodology, which is simple and dependable and promises to be applicable in other profilable data domains, as well.

7 CONCLUSIONS AND FUTURE WORK

The ever-growing amount of information being offered and used through the Web has led to the development of various web usage mining technologies. The primary goal is better management of the gigantic information flow to enable the information providers reach their target clients and the users to get their needed information or service in the most efficient way. Among these applications are web systems for recommendation and personalization. These systems aim to assist the users in wasting no time browsing irrelevant pages. The systems first accumulate information on users' interests either explicitly from the users or implicitly through following their "footprints" as they traverse the Web. On the basis of the collected information, the systems offer a response or a recommendation on the basis of a prediction process. The better the prediction, the better the recommendation or personalization systems.

While several techniques are used to gather the needed information, we have focused in this work on collaborative filtering, an effective method of indirectly gathering information on likely user interests by matching the interests of like-minded group of past users of the web site. Although different web mining techniques are used to build collaborative filtering systems, we have chosen a model driven approach based on fuzzy clustering of web sessions for discovering the model in the form of usage patterns. After a review of fuzzy clustering, collaborative filtering, and usage profiling, we have focused on specific pertinent issues in these three areas. A special emphasis is given in our studies to a clustering algorithm (RFSC), and a related collaborative filtering technique (Fuzzy Hybrid), both of which have been recently developed at the Department of Computer

Science and Software Engineering of Concordia University. The web data used in our experimental studies are also obtained from the same Department.

In a comparative study of 4 fuzzy clustering techniques (RFSC, ARCA, RFCMdd, LFCMdd), in the context of web prediction, we reviewed several validity indices and tested 3 of them using the 4 clustering techniques with varying number (C) of clusters. We have shown the undesirable relationship of two of them - Partition Coefficient and XB – and show that XB correlated with C even at low C number. We found that the sensitivity of validity indices to increasing C depended on clustering technique. We have found that the Xie-Y index was least biased and adopted it therefore to assess the clustering effectiveness of the 4 clustering techniques. These were also compared for clustering efficiency, and for prediction effectiveness (accuracy) and efficiency, prediction being an end-use application of the usage model. ARCA scored poorest in clustering efficiency and prediction accuracy, but best in prediction efficiency. RFSC was fastest in clustering, confirming its scalability advantage. Other differences were non-significant or negligible. We found no tight relationship between clustering validity and prediction quality and concluded both attributes are essential in comprehensive studies of clustering quality in the context of prediction.

In another study, we compared the Fuzzy Hybrid Collaborative Filtering (FH) to a regular model-based collaborative filtering system (MB), both based on RFSC as clustering model. We applied the 2 systems on 10 random training samples for sizes from 500 to 5000, and 10 random test samples consisting of a fourth of those, and measured system prediction effectiveness using the metrics F1, MAE and R. We discovered a general significant trend favoring FH over MB, with differences being largest for smallest test samples and gradually decreasing with sample size to become negligible for samples of about 875 sessions. To further validate the trend, we applied the 2 systems on specially constructed samples, biased to different degrees towards one or the other system, and discovered that FH consistently maintained its superiority even in samples with a high bias favoring MB. We attributed such interesting trend to the exhaustive session scanning in selected clusters performed by FH, a feature inherited from the memory-based filtering. We noted that such superiority vanished in larger samples and C values and are interested to know the trend for even larger samples.

We then conducted a set of original studies on usage profiles. We first defined a new set of 4 criteria for evaluating UP quality and proposed a quantified measure for each criterion, while ensuring that the criteria semantics are reflected through their measures. We then applied these criteria in a post-clustering process to filter the best set of clusters. We have emphasized the importance of using such criteria to avoid reliance on potentially subjective human judgment on UP quality and enable the partial automation of best UP's identification, as derived from clustering. We then conducted a user survey to assess the practical relevance of these criteria and confirmed unambiguously the close match between evaluators' views and the semantics targeted by the criteria. We also extended the deployment of the criteria to cover whole sets of UPs derived from different clustering techniques, which enables the comparative evaluation of those techniques. This was applied in experiments with RFSC and RFCMdd that showed the better performance of RFSC, confirming previous results, emphasizing again the soundness of the procedure.

In all conducted experiments, we performed appropriate statistical testing on the comparative performance of tested procedures or algorithms. We believe that statistical significance must be systematically performed in all such studies to discriminate between real differences and those due to sampling or experimental errors, thus adding more confidence in the derived conclusions. This is especially crucial in web mining where data are inherently noisy and where large and varied real world data are difficult to collect and experiments are limited by resource constraints.

Future Work:

The following issues deserve further investigation:

1) Cluster validation:

-We believe that each validity index proposed in the literature has benefits over others in some conditions. It will be beneficial to exhaustively experiment and tabulate the objectivity of validity indices along with applicable dataset types and conditions.

-It will be useful to look into the possibility of devising web mining versions of various available validity indices that are not yet suitable for clickstream data.

2) Collaborative Filtering:

- We have seen that Fuzzy Hybrid performs better than Model-based procedure with small test samples, but the two procedures become similar at somewhat larger sample sizes. It will be interesting to pursue the research using much larger sample sizes.

3) Usage profiles:

- The computation of UP quality criteria (especially distinctness) is relatively time consuming. We plan to enhance the efficiency of calculating these qualities
- Investigate an innovative way of combining the criteria for UP quality evaluation into one aggregate criterion to allow one value comparison between clustering techniques.

- Although the syntactic page similarity seems to work fine, it doesn't seem to be sufficient. Devising a method capable of capturing the semantic similarity between pages would enhance the justification as to why certain pages are grouped into same UP's, and that would certainly help in assessing our UP approach.
- The more we understand how decision makers consider a set of UP's to be of high or poor quality, the better we can shape our measures to reflect such quality requirement. As such, we should aim at making decision makers become more engaged into improving the UP quality measures.

While many of the above questions remain to be answered, we believe that more will pop up as data volumes accumulate in web servers and across the Web, and the need becomes more urgent for technologies and algorithms that can meet the insatiable needs of a world-wide community of web users.

8 BIBLIOGRAPHY

[Abonyi and Feil, 2007] J. Abonyi and B. Feil: "Aggregation and Visualization of Fuzzy Clusters based on Fuzzy Similarity Measures," in: *Advances in Fuzzy Clustering and its Applications*, John Wiley & Sons, edited by J. V. de Oliveira and W. Pedrycz, accepted.

[Azman and Ounis, 2004] A. Azman and I. Ounis: "Discovery of aggregate usage profiles based on clustering information needs," Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, July 2004, Sheffield, United Kingdom, pp. 25-29.

[Bao et al., 2005] Y. Bao, H. Zou, and J. Zhang: "Using PACT in an e-commerce recommendation system", ICEC 2005, pp. 466-470.

[Baron and Spiliopoulou, 2003] S. Baron and M. Spiliopoulou: "Monitoring the Evolution of Web Usage Patterns," EWMF, 2003, pp. 181-200.

[Bensaid et al., 1996] A.M. Bensaid, L.O. Hall, J.C. Bezdek, L.P. Clarke, and M.L. Silbiger: "Validity-guided (re) clustering with applications to image segmentation". IEEE Trans. Fuzzy Systems. Vol. 4, 1996, pp. 112-123.

[Bezdek, 1974] J.C. Bezdek: *Cluster validity with fuzzy sets*, J. Cybernet. Vol. 3, 1974.

[Bezdek, 1981] C. J. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum Press, 1981.

[Black, 2008] K. Black: *Business Statistics: For Contemporary Decision Making.* John Wiley & Sons, 5th Edition, 2008. ISBN 978-0-471-78956-7.

[Bouguessa et al., 2006] M. Bouguessa, S. Wang, and H. Sun: "An objective approach to cluster validation", Pattern Recognition Letters Vol. 27(13), 2006, pp. 1419-1430.

[Breese et al., 1998] J. Breese, D. Heckerman, C. Kadie: Empirical analysis of predictive algorithms for collaborative filtering. In Proc. of UAI-98, 1998, pp. 43-52.

[Chen and Linkens, 2004] M. Y. Chen, D. A. Linkens: "Rule-base self-generation and simplification for data-driven fuzzy models," *Fuzzy Sets and Systems* Vol. 142 (2004) pp. 243–265.

[Connor and Herlocker, 1999] M. O'Connor, J. Herlocker: "Clustering items for collaborative filtering". In Proc. of ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, 1999.

[Cooley et al., 1999] R. Cooley, B. Mobasher, and J. Srivastava: "Data preparation for mining world wide web browsing patterns". *Knowledge and Information Systems* Vol. 1(1), 1999, pp. 5–32.

[Corsini et al., 2005] P. Corsini, B. Lazzerini, and F. Marcelloni: "A new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm," *Soft Computing*, Springer-Verlag, Vol. 9(6), 2005, pp. 439-447.

[Crabtree et al., 2005] D. Crabtree, X. Gao, and P. Andreae: "Standardized evaluation method for web clustering results," the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005, pp. 280-283.

[Dave, 1996] R.N. Dave: "Validating fuzzy partition obtained through c-shells clustering," Pattern Recognition Lett. Vol. 17, 1996, pp. 613–623.

[Dempster et al., 1977] A. Dempster, N. Laird, D. Rubin: "Maximum likelihood from incomplete data via the em algorithm". Journal of Royal Statistical Society **B**, Vol. 39, 1977, pp. 1–38.

[Dunn, 1974] J. C. Dunn: "Well Separated Clusters and Optimal Fuzzy Partitions," Journal of Cybernetics, Vol. 4, 1974, pp. 95-104.

[Davies and Bouldin, 1979] D. L. Davies and D. W. Bouldin: "Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1(2), 1979, pp. 95-104.

[Eirinaki and Vazirgiannis, 2003] M.Eirinaki and M.Vazirgiannis: "Web Mining for Web Personalization," ACM Transactions on Internet Technologies (ACM TOIT), Vol.3 (1), 2003.

[Fukuyama and Sugeno, 1989] Y. Fukuyama, M. Sugeno: "A new method of choosing the number of clusters for the fuzzy c-means method," Proc. of 5^{th} Fuzzy Systems Symposium, 1989, pp. 247–250.

[Gath and Geva, 1989] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," IEEE. Trans . Pattern Anal. Machine Intell., Vol. 7, 1989, pp. 773–781.

[Goldberg et al., 1992] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry: "Using collaborative filtering to weave an information tapestry". *Communications of the ACM*, Vol. 35(12), 1992, pp. 61–70.

[Gonzales,2005] M. Gonzales: "A Comparison In Cluster Validation Techniques," Master's Thesis, Mathematics Department, University of Puerto Rico at Mayaguez, 2005.

[Groebrner et al., 2006] D. F. Groebrner, P. W. Shannon, P. C. Fry, and K. D. Smith: *Business Statistics*, Pearson Prentice Hall, 2006. ISBN 0-13-153687-7.

[Jahanian et al., 2004] H. Jahanian, G. A. Hossein-Zadeh, and H. Soltanian-Zadeh: "Controlling the False Positive Rate in Fuzzy Clustering Using Randomization: Application to fMRI Activation Detection", *Magnetic Resonance Imaging*, Vol. 22, 2004, pp. 631–638.

[Jin et al., 2004] X. Jin, Y. Zhou, and B. Mobasher. "Web usage mining based on probabilistic latent semantic analysis". KDD '04: Proc. of ACM SIGKDD, 2004, pp. 197-205.

[Joshi et al., 1999] Joshi, A., Joshi, K., and Krishnapuram, R: "On Mining Web Access Logs," Technical Report, University of Maryland Baltimore County, 1999.

[Joshi and Krishnapuram, 2000] A. Joshi and R. Krishnapuram: "On mining web access logs," in Pro. Of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000), 2000.

[Halkidi et al., 2000] M. Halkidi, M. Vazirgiannis, and Y. Batistakis: "Quality Scheme Assessment in the Clustering Process", Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, 2000, pp. 265-276.

[Halkidi et al., 2001] M. Halkidi, Y. Batistakis, M. Vazirgiannis: "On Clustering Validation Techniques," Journal of Intelligent Information Systems, Vol. 17(2-3), 2001, pp. 107-145.

[Halkidi et al., 2002 a] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: part I," SIGMOD Record, Vol. 31(2), 2002, pp. 40-45.

[Halkidi et al., 2002 b] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Clustering Validity Checking Methods: Part II," SIGMOD Record, Vol. 31(3), 2002, pp.19-27.

[Halkidi and Vazirgiannis, 1996] M. Halkidi and M. Vazirgiannis: "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set," Proc. of ICDM 2001, 2001 pp. 187-194.

[Hathaway and Bezdek, 1994] R.J. Hathaway and J.C. Bezdek, "NERF c-means: Non-Euclidean relational fuzzy clustering," *Patt Recognit*, Vol. 27, 1994, pp. 429–437.

[Herlocker et al., 1999] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl: "An algorithmic Framework for Performing Collaborative Filtering," in Proc. of ACM SIGIR'99, ACM Press, 1999.

[Herlocker et al., 2004] Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; Riedl, J. T: "Evaluating collaborative filtering recommender systems", *ACM Trans. Inf. Syst.* Vol. **22** (1), pp. 5–53, doi:10.1145/963770.963772, ISSN 1046-8188.

[Hwang and Thill, 2007] S. Hwang and J. C. Thill: "Using Fuzzy Clustering Methods for Delineating Urban Housing Submarkets," in Proc. of 15th ACM Intl. Symp. on Advances in Geographic Information Systems, ACM GIS, 2007.

[Keller et al., 1985] J. Keller, M. Gray, and J. Givens, "A fuzzy k-nearest neighbor algorithm", IEEE Transaction on Systems, Man and Cybernetics, Vol. 15(4), 1985, pp. 580, 1985.

[Ketata et al., 2009] Ketata, A., Mudur, S., and Shiri. N. Dependable performance analysis for fuzzy clustering of Web usage data. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), (Nashville, TN, USA, March 30 - April 2, 2009). In press.

[Kim et al., 2004 a] Y. I. Kim, D. W. Kim, D. Lee, K. H. Lee, "A cluster validation index for GK cluster analysis based on relative degree of sharing," Inform. Sci. Vol. 168, 2004, pp. 225–242.

[Kim et al., 2004 b] D.W. Kim, K.H. Lee, D. Lee: "On cluster validity index for estimation of the optimal number of fuzzy clusters," Pattern Recognition, Vol. 37, 2004, pp. 2009–2025.

[Kohrs and M´erialdo, 1999] A. Kohrs, B. M´erialdo: "Clustering for collaborative filtering applications", in Proc. of the International Conference on Computational Intelligence for Modeling, Control & Automation (CIMCA'99), 1999.

[Krishnapuram et al., 1999] R. Krishnapuram, A. Joshi, and L. Yi: "A Fuzzy Relative of the k-Medoids Algorithm with Application to Web Document and Snippet Clustering," in Proc. IEEE FUZZIEEE99, 1999.

[Kwon, 1998] S.H. Kwon: "Cluster validity index for fuzzy clustering," Electronics Letters, Vol. 34(22), 1998, pp. 2176-2177.

[Legany et al., 2006] C. Legany, S. Juhasz, A. Babos: "Cluster validity measurement techniques" 2006. http://portal.acm.org/citation.cfm?id=1364328. pp. 388-393.

[Lin, 1989] L. I-K. Lin. "A concordance correlation coefficient to evaluate reproducibility". *Biometrics* Vol. **45** (1), 1989, pp. 255–268. doi:10.2307/2532051. PMID 2720055.

[Lin, 2000] L. I-K. Lin. "A Note on the Concordance Correlation Coefficient". *Biometrics* Vol. **56**, 1989, pp. 324–325.

[Lu et al., 2005] L. Lu, M. H. Dunham, Y. Meng, "Discovery of Significant Usage Patterns from Clusters of Clickstream Data," ACM SIGKDD Workshop on Knowledge Discovery in Web (WebKDD'05), 2005.

[Manning et al., 2008] C. Manning, P. Raghavan and H. Schütze: *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[Melegy et al., 2007] M. El-Melegy, E. Zanaty, W. Abd-Elhafiez and A. A. Farag: "On Cluster Validity Indexes in Fuzzy and Hard Clustering Algorithms for Image Segmentation," in Proc. of IEEE International Conference on Image Processing (ICIP'07), 2007, pp. VI-5-VI-8.

[Miller et al., 2004] B. N. Miller, J. A. Konstan, J. Riedl: "PocketLens: Toward a Personal Recommender System". ACM Transactions on Information Systems Vol. 22, 2004.

[Mobasher, 1999] B.Mobasher. "WebPersonalizer : A Server Side Recommender System Based on Web Usage Mining," in Proc. of the 9th workshop on Information Technologies and Systems (WITS'99), December,1999.

[Mobasher et al., 1999] B. Mobasher, R. Cooley, J. Srivastava: "Creating adaptive web sites through usage based clustering of URLs". In Proc. of KDEX'99, 1999.

[Mobasher et al., 2002] B. Mobasher, H. Dai, and M. N. T. Luo: "Discovery and evaluation of aggregate usage profiles for web personalization," Data Mining and Knowledge Discovery Vol. 6, 2002, pp. 61–82.

[Mobasher, 2007] B. Mobasher: "Data Mining for Personalization". In The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). Lecture Notes in Computer Science, Vol. 4321, Springer, Berlin-Heidelberg, 2007, pp. 90-135.

[Nasraoui et al., 1999 a] O. Nasraoui , H. Frigui, A. Joshi, and R. Krishnapuram: "Mining Web Access Logs Using Relational Competitive Fuzzy Clustering," in Proc. 8th International Fuzzy Systems Association Congress, 1999.

[Nasraoui et al., 1999 b] O. Nasraoui, R.Krishnapuram, and A. Joshi. "Relational clustering based on a new robust estimator with application to web mining". In Proc. of the North American Fuzzy Information Society, 1999, pp. 705-709.

[Nasraoui et al., 2000] O. Nasraoui, H. Frigui, R. Krishnapuram, and A. Joshi: "Extracting web user profiles using relational competitive fuzzy clustering". Intl. Journal on Artificial Intelligence Tools, Vol. 9(4), 2000, pp. 509–526.

[Nasraoui et al., 2002] O. Nasraoui, R. Krishnapuram, A. Joshi, and T. Kamdar: “Automatic web user profiling and personalization using robust fuzzy relational clustering”. In Segovia, J., Szczepaniak, P., Niedzwiedzinski, M., eds.: Studies in Fuzziness and Soft Computing, Vol. 105, Springer-Verlag, 2002, pp. 233–261.

[Nasraoui and Saka, 2006] O. Nasraoui, E. Saka: “Web Usage Mining in Noisy and Ambiguous Environments: Exploring the Role of Concept Hierarchies, Compression, and Robust User Profiles”. WebMine 2006, pp. 82-101.

[Nickerson, 1997] C. A. E. Nickerson: "A Note on "A Concordance Correlation Coefficient to Evaluate Reproducibility". *Biometrics* Vol. **53**(4) 1997 pp. 1503–1507. doi:10.2307/2533516.

[Oliveira and Pedrycz, 2007] J.V. de Oliveira and W. Pedrycz (Eds), *Advances in Fuzzy Clustering and its Applications*, John Wiley & Sons, 2007. ISBN 978-0-470-02760-8.

[Pakhira et al., 2004] M. K. Pakhira, S. Bandyopadhyay, U. Maulik: “Validity index for crisp and fuzzy clusters,” Pattern Recognition Vol. 37, 2004, pp. 487–501.

[Pakhira et al., 2005] M. K. Pakhira, S. Bandyopadhyay, U. Maulik: “A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification,” Fuzzy Sets Syst. Vol. 155, 2005, pp. 191–214.

[Pal and Bezdek, 1995] N. R. Pal and J. C. Bezdek: “On cluster validity for the fuzzy c-means model,” IEEE Trans. Fuzzy Syst., Vol. 3(3), 1995, pp. 370–379.

[Pal and Biswas, 1997] N. R. Pal and J. Biswas: “Cluster Validation using graph theoretic concepts”, Pattern Recognition, Vol. 30(4), 1997.

[Pennock et al., 2000] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles: "Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach," In *Proc.* of UAI-2000, 2000, pp. 473-480.

[Perkowitz and Etzioni 1998] M. Perkowitz and O. Etzioni: "Adaptive Web sites: Automatically synthesizing Web pages," in Proc. of AAAI-98, 1998.

[Resnick et al., 1994] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in Proceedings of the ACM Conference on Computer Supported Cooperative Work, 1994, pp. 175–186.

[Rhee and Oh, 1996] N. S. Rhee, K. W. Oh: "A validity measure for fuzzy clustering and its use in selecting optimal number of clusters," in IEEE International Conference on Fuzzy Systems, Vol. 2, 1996, pp. 1020-1025.

[Roubens, 1978] M. Roubens, *Pattern classification problems and fuzzy sets*, Fuzzy sets and Syst., Vol. 1, 1978, pp. 239-253.

[Sarwar et al., 2000] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl: "Analysis of recommender algorithms for e-commerce," In Proc. of the 2nd ACM E-commerce Conference, 2000.

[Sarwar et al., 2002] B.M. Sarwar, G Karypis, J. Konstan, and J. Riedl: "Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering," in Proc. 5th Intl. Conf. on Comp. and Inf, Tech. 2002.

[Schetcher et al., 1998] S. Schetcher, M. Krishnan, and M.D. Smith, "Using path profiles to predict HTTP requests," in Proc. of the 7th International World Wide Web Conference, Computer Networks, 30(1-7), 1998, pp. 457-467.

[Sharma, 1996] S. Sharma: *Applied multivariate techniques*, John Wiley & Sons, Inc., 1996.

[Shardanand and Maes, 1995] U. Shardanand and P. Maes: "Social information filtering: Algorithms for automating "word of mouth." in Proc. of Computer Human Interaction, 1995, pp. 210–217.

[Suryavanshi et al., 2005 a] B. S. Suryavanshi, N. Shiri, and S. P. Mudur: "An Efficient Technique for Mining Usage Profiles using Relational Fuzzy Subtractive Clustering," in Proc. of Int'l Workshop on Challenges in Web Information Retrieval and Integration (WIRI 05), ICDE, 2005.

[Suryavanshi et al., 2005 b] B. S. Suryavanshi, N. Shiri, and S.P. Mudur: "A Fuzzy Hybrid Collaborative Filtering Technique for Web Personalization," in Proc. of 3rd Workshop on Intelligent Techniques for Web Personalization (ITWP 05), Held at IJCAI 2005.

[Suryavanshi et al., 2006] B. Shankar Suryavanshi, N. Shiri, S. P. Mudur: "Analysis of Fuzzy Clustering Techniques Used for Web Personalization", In Proc. of NAFIPS 2006, IEEE press.

[Suryavanshi, 2006] B. S. Suryavanshi: "A new class of techniques for Web personalization" Master's Thesis, Department of Computer Science and Software Engineering, Concordia university, Montreal, 2006.

[Tang et al., 2005] Y. G. Tang, F. C. Sun, Z.Q. Sun: "Improved validation index for fuzzy clustering," in: American Control Conf., 2005.

[Theodoridis and Koutroubas, 1999] S. Theodoridis and K. Koutroubas: Pattern Recognition, Academic Press, 1999.

[Tonella et al., 2003] P. Tonella, F. Ricca, E. Pianta, C. Girardi, G. Di Lucca, A. R.

Fasolino, and P. Tramontana: "Evaluation Methods for Web Application Clustering," in Proc. of the 5th International Workshop on Web Site Evolution, 2003.

[Torra and Miyamoto, 2004] V. Torra and S. Miyamoto, "Evaluating Fuzzy Clustering Algorithms for Microdata Protection," Privacy in Statistical Databases, 2004, pp.175-186.

[Ungar and Foster, 1998] L. Ungar, D. P. Foster: "Clustering methods for collaborative filtering", in Proc. of the AAAI98 Workshop on Recommendation Systems, 1998.

[Windham, 1981] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," Fuzzy Sets Syst., 1981, pp. 177-185.

[Wang and Zhang, 2007] W. *Wang*, Y. *Zhang*: *"On fuzzy cluster validity indices," Fuzzy* Sets Syst., Vol. 158(19), 2007, pp. 2095-2117.

[Windham, 1985] M. P. Windham: "Numerical classification of proximity data with assignment measures," J Class, Vol. 2, 1985, pp. 157-172.

[Wu et al., 2005] K.L. Wu, M.S. Yang: "A cluster validity index for fuzzy clustering," Pattern Recognition Lett. Vol. 26, 2005, pp. 1275–1291.

[Yan et al., 1996] T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal: "From user access patterns to dynamic hypertext linking," in Proc. of the 5th International World Wide Web Conference, 1996.

[Xie and Beni, 199 X.L. Xie and G. Beni: "A validity measure for fuzzy clustering", IEEE Trans. on PAMI, Vol. 13(8), 1991, pp. 841-847.

[Xie et al., 2002] Y. Xie, V.V. Raghavan, and X. Zhao: "3M algorithm: finding an optimal fuzzy cluster scheme for proximity data," in Proc. of the FUZZ-IEEE Conf. IEEE World Congress on Computational Intelligence, 2002.

[Xu et al., 2006] G. Xu, Y. Zhang and X. Zhou: "Discovering Task-Oriented Usage Pattern for Web Recommendation," in Proc. of the 17th Australasian Database Conference (ADC'2006), 2006, pp. 167-174.

[Zamir, 1999] O. E. Zamir: "Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results," Doctoral thesis, University of Washington, 1999.

[Zhang et al., 2005] Y. Zhang, G. Xu, X. Zhou: "A Latent Usage Approach for Clustering Web Transaction and Building User Profile," in X. Li, S. wang, and Z.Y. Dong (Eds): ADMA 2005, LNAI 3584, 2005, pp. 31-42.

[Zhang et al., 2008] Y. Zhang, W. Wang, X. Zhang, Y. Li: "A cluster validity index for fuzzy clustering," Inf. Sci., Vol. 178(4), 2008, pp. 1205-1218.

9 APPENDIXES

9.1 Appendix A. List of Validity Indices, Formulas and References

In all the formulae given below, n is the sample size (number of sessions), c is the number of clusters, u_{ij} is the membership of the j^{th} session to the i^{th} cluster, c and cn both refer to the number of clusters. And d(x, y) = || x – y || is the dissimilarity (distance) between the two objects x and y. x_i refers to the i^{th} object to be clustered, and v_i is the i^{th} cluster center (prototype).

Index Name	Index Formula and references	Applicable to ClickStream data	Fuzzy(F) vs. Crisp(C)	Considers geometrical data features
PE	**Partition Entropy (PE):** proposed by [Bezdek, 1974]. [Zhang et al., 2008][Wang and Zhang, 2007] stated that PE has a monotonic negative correlation with C, and have experimentally shown that, for many experiments, PE failed to estimate C correctly. $V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log u_{ij}$	Y	F	N
WPE	**Windham Proportion Exponent (WPE):** Proposed by [Windham, 1981]. [Rhee and Oh, 1996] mentioned	Y	F	N

	WPE's lack of connection with the geometrical features of the dataset and its monotonic negative correlation with C. Such correlation was also confirmed by [Gath and Geva, 1989] and [Wang and Zhang, 2007]. $V_{\mathrm{WPE}} = -\log_e\left[\prod_{k=1}^{n}\left[\sum_{j=1}^{[u_k^{-1}]}(-1)^{j+1}\binom{c}{j}(1-ju_k)^{c-1}\right]\right]$ where $u_k = \max_i\{u_{ik}\}$.			
MPC	**Modified Partition Coefficient (MPC)**: Proposed by [Dave, 1996], as an adjustment to PC index bias, and reviewed by [Wang and Zhang, 2007] and [Zhang et al., 2008], as both have shown MPC successfully estimated C values for most experiments. $V_{\mathrm{MPC}} = 1 - \frac{c}{c-1}(1 - V_{\mathrm{PC}})$	Y	F	N
CE	**Classification entropy (CE):** Proposed by [Bezdek, 1981]. Also Reviewed by [Abonyi and Feil, 2007], [Rhee and Oh, 1996], both stated CE's monotonic negative correlation with C, and its lack of connection with the geometrical features of the dataset. $CE(U;c) = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\, log_a u_{ij}}{n}$	Y	F	N
KYI	**KYI index:** Proposed by [Kim et al., 2004 a], and reviewed by [Wang and Zhang, 2007] in which it has	N	F	N

	proven to correctly detect the number of clusters for several datasets. $V_{\mathrm{KYI}}(U, V ; X) = \frac{2}{c(c-1)} \sum_{p \neq q}^{c} S_{\mathrm{rel}}(F_p, F_q) = \frac{2}{c(c-1)} \sum_{p \neq q}^{c} \sum_{j=1}^{n} [c \cdot [u_{F_p}(x_j) \wedge u_{F_q}(x_j)] \cdot h(x_j)].$ where $h(x_j) = -\sum_{i=1}^{c} u_{F_i}(x_j) \log_a u_{F_i}(x_j)$ and u_{Fi} (x_j) is the membership of the jth object to the ith cluster.			
SC	**Partition index (SC)** proposed by [Bensaid et al., 1996], and reviewed by [Abonyi and Feil, 2007]. $SC = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} u_{ik}^{m} \lVert \boldsymbol{x}_k - \boldsymbol{v}_i \rVert_A^2}{n_i \sum_{t=1}^{c} \lVert \boldsymbol{v}_i - \boldsymbol{v}_t \rVert_A^2}$	Y	F	N
S	**Separation index (S)** Also proposed by [Bensaid et al., 1996] and reviewed by [Abonyi and Feil, 2007]. $S = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \lVert \boldsymbol{x}_k - \boldsymbol{v}_i \rVert^2}{n * \min_{i,t} \lVert \boldsymbol{v}_i - \boldsymbol{v}_t \rVert^2}$	Y	F	N
T	**T index:** Proposed by [Tang et al., 2005]. It was reviewed by [Wang and Zhang, 2007] and they have shown that T index failed to identify C for most experiments conducted by them. $V_T(U, V ; X) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{2} \lVert x_j - v_i \rVert^2 + \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{k=1, k \neq i}^{c} \lVert v_i - v_k \rVert^2}{\min_{i \neq k} \lVert v_i - v_k \rVert^2 + 1/c}$	Y	F	Y
I_RFSC	Proposed by [Suryavanshi et al., 2005 a] as an adaptation	Y	F	Y

	of XB to RFSC clustering technique. Our experiments show this index to be highly dependent on C. $\text{Compactness} = \frac{1}{C}\sum_{i=1}^{C}\left[\sum_{j=1}^{N} u_{ij}^2 * R_{c_i j}^2 \Big/ \sum_{j=1}^{N} u_{ij}\right]$ and Separation = $\min_{i \neq k} R_{c_i c_k}^2$. for i and k in $[1..C]$. where x_{c_i} and x_{c_k} are the i^{th} and k^{th} cluster centers. and $R_{c_i c_k}$ is the dissimilarity between these centers. $\text{Index of goodness} = \frac{Compactness}{Separation}$			
FD	Referenced by [Nasraoui et al., 2000]. Distance within clusters: $\overline{D}_{W_i} = \sum_{x^{(k)} \in X_i} \sum_{x^{(l)} \in X_i, l \neq k} d_{kl}^2 / \lvert X_i \rvert (\lvert X_i \rvert - 1)$ Distance between clusters $\overline{D}_{B_{ij}} = \sum_{x^{(k)} \in X_i} \sum_{x^{(l)} \in X_j, l \neq k} d_{kl}^2 / \lvert X_i \rvert \lvert X_j \rvert$ A combination of these two quantities would form a complete validity index. However, no such combination was found in the literature.	Y	F	Y
P	Proposed by [Chen and Linkens, 2004]. Also reviewed by [Wang and Zhang, 2007] where they tested it on several samples on which shown it failed to identify the correct C value. $V_P = \frac{1}{n}\sum_{k=1}^{n}\max_i(u_{ik}) - \frac{1}{K}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\left[\frac{1}{n}\sum_{k=1}^{n}\min(u_{ik}, u_{jk})\right]$ where $K = \sum_{i=1}^{c-1} i$	Y	F	N

DI	**Dunn Index (DI):** Reviewed by [Abonyi and Feil, 2007][Legany et al., 2006]. Its disadvantages, according to [Abonyi and Feil, 2007], are its computationally high costs, and its high sensitivity to noise. Different versions of this index have also been proposed [Theodoridis and Koutroubas, 1999][Pal and Biswas, 1997]. $D = \min_{i=1 \ldots n_c} \left\{ \min_{j=i+1, \ldots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1 \ldots n_c} (diam(c_k))} \right) \right\}$, where $d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\}$ and $diam(c_i) = \max_{x, y \in c_i} \{d(x, y)\}$	Y	C	Y
ADI	Alternative Dunn Index (ADI): reviewed by [Abonyi and Feil, 2007]. Designed to overcome the computational complexity of DI. $ADI(c) = min_{i \in c}\{min_{j \in c, i \neq j}\{\frac{min_{x_i \in C_i, x_j \in C_j}\|d(y, v_j) - d(x_i, v_j)\|}{max_{k \in c}\{max_{x,y \in C} d(x, y)\}}$	Y	C	Y
DB	Referenced by [Legany et al., 2006] and [Gath and Geva, 1989]. Described by [Gath and Geva, 1989] as providing botanically incorrect and other meaningless cluster numbers, on Iris dataset. $DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$, where $R_i = \max_{j=1, \ldots n_c, i \neq j} (R_{ij}),\ i = 1 \ldots n_c$ $R_{ij} = \frac{s_i + s_j}{d_{ij}}$ $d_{ij} = d(v_i, v_j),\ s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i)$	Y	C	Y

RMSSTD and RS	Referenced by [Legany et al., 2006]. This index cannot be applied directly to clickstream data, because it requires calculation of the dataset mean and variance. $RMSSTD = \sqrt{\frac{\sum_{i=1\ldots n_c,\ j=1\ldots d} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j}\right)^2}{\sum_{i=1\ldots n_c,\ j=1\ldots d} \left(n_{ij} - 1\right)}}$ $RS = \frac{SS_t - SS_w}{SS_t}$, where $SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} \left(x_k - \overline{x_j}\right)^2$, $SS_w = \sum_{i=1\ldots n_c,\ j=1\ldots d} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j}\right)^2$	N	C	Y
SD	Referenced by [Legany et al., 2006]. Because it requires calculation of the dataset mean and variance, SD index is not applicable for clickstream data. $SD = \alpha \cdot Scatt + Dis$ $Dis = \frac{\max_{i,j=1\ldots n_c} \left(\lVert v_i - v_j \rVert\right)}{\min_{i,j=1\ldots n_c} \left(\lVert v_i - v_j \rVert\right)} \sum_{k=1}^{n_c} \left(\sum_{z=1}^{n_c} \lVert v_k - v_z \rVert \right)^{-1}$ $Scatt = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\lVert \sigma(v_i) \rVert}{\lVert \sigma(x) \rVert}$ Variance of the dataset: Variance of a cluster: $\sigma_x^p = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^p - \overline{x^p}\right)^2$ $\sigma_{v_i}^p = \frac{1}{\lVert c_i \rVert} \sum_{k=1}^{n} \left(x_k^p - \overline{v_i^p}\right)$ $\sigma(x) = \begin{bmatrix} \sigma_x^1 \\ \vdots \\ \sigma_x^d \end{bmatrix}$ $\sigma(v_i) = \begin{bmatrix} \sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^d \end{bmatrix}$	N	C	Y
I_G	Defined by [Rhee and Oh, 1996]. $I_G = D/C$	N	F	Y

	$C = \frac{2}{n(n-1)} \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} \sum_{i=1}^{c} d^2(X_{j1}, X_{j2})\omega_1$ where w_l is defined as = min{u_{ij1}, u_{ij2}} which is the membership value of data points X_{j1} and X_{j2} belonging to C_i by definition of fuzzy theory. $D = \frac{1}{n^2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} d^2(X_{j_1}, X_{j_2})\omega_2$ $\omega_2 = \min\{\max_{i_1} u_{i_1 j_1}, \max_{i_2 \neq i_1} u_{i_2 j_2}\}$			
F_{HV}	Proposed by [Gath and Geva, 1989]. Reviewed by [Abonyi and Feil, 2007], [Wang and Zhang, 2007], and [Zhang et al., 2008]. Its hypervolume formula was described to have objective performance, as regards to the clusters number, and that it shows a clear extremum. Again it is shown by [Zhang et al., 2008] that for several experiments, F_{HV} failed to estimate C correctly. $V_{FHV} = \sum_{i=1}^{c} [\det(F_i)]^{1/2}$. where $F_i = \frac{\sum_{j=1}^{n} (u_{ij})^m (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{n} (u_{ij})^m}$	N	F	Y
K	Proposed by [Kwon, 1998]. Reviewed by [Zhang et al., 2008](uses data mean, so doesn't fit for clickstream data) is an extension to XB to solve its dependency on C, using a penalty function. The index was compared to XB	N	F	Y

	and proven to estimate the correct C, but also to be unbiased to *m*. Again it is shown by [Zhang et al., 2008] that for several experiments, K failed to estimate C correctly. $V_{\mathrm{K}} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{2}\|x_j - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{\min_{i \neq k}\|v_i - v_k\|^2}$ where $\bar{v} = \sum_{j=1}^{n} x_j / n$.			
PBM(F)	Proposed by [Pakhira, 2004], and reviewed by [Pakhira, 2005] and [Zhang et al., 2008]. The fuzzy version is called PBMF. In [Zhang et al., 2008] it was shown to successfully estimate C values for all experiments. m = 1.5 is the suggested value by [Pakhira, 2004] and [Pakhira, 2005]. $V_{\mathrm{PBMF}} = \left(\frac{1}{c} \times \frac{E_1}{J_m} \times D_c\right)^2 \quad E_1 = \sum_{j=1}^{n}\|x_j - v\|,$ $D_c = \max_{i,j=1}^{c}\|v_i - v_j\|$ $J_m(U,Z) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m\|x_j - v_i\|$	N	F	Y
W	[Zhang et al., 2008] proposed it and have shown it to successfully estimate C values for all experiments. $V_{\mathrm{W}}(V,U) = \frac{\mathrm{Var}^N(V,U)}{\mathrm{Sep}^N(c,U)}$ $\mathrm{Var}^N(V,U) = \frac{\mathrm{Var}(V,U)}{\mathrm{Var}_{\max}}$	N	F	Y

	$\mathrm{Sep}^N(c,U)=\dfrac{\mathrm{Sep}(c,U)}{\mathrm{Sep}_{\max}},\quad c=2,3\ldots,c_{\max}$ $\mathrm{Var}_{\max}=\underbrace{\max}_{c}\mathrm{Var}(V,U)\quad \text{and}\quad \mathrm{Sep}_{\max}=\underbrace{\max}_{c}\mathrm{Sep}(c,U)$ $\mathrm{Sep}(c,U)=1-\max_{i\neq j}S(F_i,F_j)=1-\max_{i\neq j}\max_{x_k\in X}\min(u_{ik},u_{jk})$ $S(F_p,F_q)=\underbrace{\max}_{x_k\in X}\min(u_{pk},u_{qk})$ $\mathrm{Var}(U,V)=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}d^2(x_j,v_i)/n(i)\right]*\left(\dfrac{c+1}{c-1}\right)^{1/2}$ $d(x,y)=\{1-\exp(-\beta\|x-y\|^2)\}^{1/2}$ $\beta=\left(\dfrac{\sum_{j=1}^{n}\|x_j-\bar{x}\|^2}{n}\right)^{-1}\quad \text{with } \bar{x}=\dfrac{\sum_{j=1}^{n}x_j}{n}$			
PCAES	Partition Coefficient and Exponential Separation (PCAES): Proposed by [Wu et al., 2005]. It was shown by [Zhang et al., 2008] that for several experiments, PCAES failed to estimate C correctly. $V_{\mathrm{PCAES}}=\sum_{i=1}^{c}\mathrm{PCAES}_i=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2/u_M-\sum_{i=1}^{c}\exp(-\min_{k\neq i}\{\|v_i-v_k\|^2/\beta_T\})$ where $u_M=\min_{1\leq i\leq c}\left\{\sum_{j=1}^{n}u_{ij}^2\right\},\quad \beta_T=\dfrac{\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{c}\quad \text{and}\quad \bar{v}=\sum_{j=1}^{n}x_j/n.$	N	F	Y
FS	Proposed by [Fukuyama and Sugeno, 1989]. It was also shown by [Zhang et al., 2008] and [Wang and Zhang, 2007] that for several experiments, FS failed to estimate C correctly. $V_{\mathrm{FS}}=J_m(U,V)-K_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|x_j-v_i\|^2-\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|v_i-\bar{v}\|^2$	N	F	Y

9.2 Appendix B. Questionnaire Explanatory Message, the Questionnaire, and List of Acceptable and Unacceptable Aggregate Usage Profiles

“It is assumed that you, the person answering this questionnaire, have a good idea of the website of our CSE department at Concordia and its users. The following are groupings of web pages in the CSE website. Each group represents a usage profile (UP), *i.e.*, a group of pages accessed by a single fictitious visitor of the website. A web page is represented by its unique URL.

The UP’s are classified into five categories, called: A, B, C, D, and E, each including 5 UP’s. These categories are provided in the attached Excel sheet, each on a single column (the sheet thus has 5 columns, labeled A to E). Each UP is identified and labeled as UP_i, where i is an integer.

Please look at the URLs (pages) composition/content in each UP within each category and fill out the following form for the 9 statements by choosing and marking one of the 3 cells: "Agree", "Disagree", or "Don't know."

UP Evaluation Questionnaire and Responses.

Quality	Agree	Disagree	Don't know	Statements
Coherence	6	1	1	1. The pages in each UP in C do not appear to be concerned with some common subject.
	8	-	-	2. Most pages in each UP in A appear to be concerned with some common subject.
Distinctness	8	-	-	3. For each UP in B, there is a UP in A that has similar pages.
	5	3	-	4. Each pair of UP's in A do not seem to share common pages.
	6	-	2	5. Pairs of UP's in A are distinct in their pages.
Coverage	3	4	1	6. No UP in E has (several) pages not found in A.
	8	-	-	7. Each UP in A has several pages that don't appear elsewhere in A.
Each page within a UP is given a "weight" describing its degree of membership to that UP (*i.e.* its frequency of occurrence in that UP). Please respond to the following statements:				
Strength	7	-	1	8. Pages in each UP in A generally have "high" weight values.
	7	-	1	9. Pages in each UP in D generally have "low" weight values.

Table VII. 5 Usage Profiles from the acceptable category (category A)

Category A	Weight
UP_0	
/~comp218/	1.00
/~comp218/Comp218/Comp218WebPage/Html_Files/Main.html	1.00
/~comp218/Comp218/Comp218WebPage/Html_Files/menu.html	1.00
/~comp218/Comp218/Comp218WebPage/Html_Files/ConU_Logo.htm	0.97
/~comp218/Comp218/Comp218WebPage/Html_Files/assignment.html	0.29
UP_29	
/~P1/comp444/	0.88
/~P1/comp444/border.html	0.75
/~P1/comp444/course_logo.html	0.75
/~P1/comp444/display.html	0.75
/~P1/comp444/main.html	0.75
/~P1/comp444/nav_bar.html	0.75
/~P1/comp444/assignments.html	0.63
/~P1/hobbit/	0.63
/cgi-bin/cgiwrap/~P1/	0.50
/~P1/comp444/assignments/	0.38
/~P1/bottom.html	0.38
/~P1/comp444/news.html	0.38
/~P1/comp444/syllabus.html	0.38
/~P1/	0.38
/~P1/low.html	0.38
/~P1/main.html	0.38
/~P1/top.html	0.38
/~P1/border.html	0.38
UP_31	
/~comp354/	1.00
/~P2/	1.00
/~P2/about_me.html	1.00
/~P2/comp354/	1.00
/~P2/comp354/project_deliv_1.html	1.00
/~P2/comp354/project_deliv_1_files/	1.00
/~P2/comp354/project_deliv_1_files/frame.html	1.00

/~P2/comp354/project_deliv_1_files/outline.html	1.00
/~P2/comp354/project_deliv_1_files/slide0001.html	1.00
/~P2/comp354/project_deliv_1_files/slide0009.html	1.00
/~P2/comp354/project_deliv_1_files/slide0048.html	1.00
/~P2/comp354/project_deliv_1_files/slide0053.html	1.00
/~P2/comp354/project_deliv_1_files/slide0059.html	1.00
/~P2/comp354/project_deliv_1_files/slide0060.html	1.00
/~P2/comp354/project_deliv_1_files/slide0061.html	1.00
/~P2/comp354/project_deliv_1_files/slide0062.html	1.00
/~P2/comp354/project_deliv_1_files/slide0063.html	1.00
/~P2/comp354/project_deliv_1_files/slide0064.html	1.00
/~P2/comp354/project_deliv_1_files/slide0065.html	1.00
/~P2/comp354_w2005r.html	1.00
/~P2/site_tree.html	1.00
/~P2/comp354/project_deliv_1_files/slide0050.html	0.50
/~P2/comp354/project_deliv_1_files/slide0051.html	0.50
/~P2/comp354/project_deliv_1_files/slide0052.html	0.50
/~P2/comp354/project_deliv_1_files/slide0049.html	0.50
/~P2/comp354/project_deliv_1_files/slide0054.html	0.50
/~P2/comp354/project_deliv_1_files/slide0055.html	0.50
/~P2/comp354/project_deliv_1_files/slide0056.html	0.50
/~P2/comp354_w2005pp.html	0.50
/~P2/comp354/project_deliv_1_files/slide0057.html	0.50
/~P2/comp354/project_deliv_1_files/slide0058.html	0.50
/current_students.shtml	0.50
UP_43	
/~cc/COMP352/announcements.html	1.00
/~cc/COMP352/index.html	1.00
/~comp352/2005w/index.shtml	1.00
/~comp352/2005w/Info/	1.00
UP_2	
/~P3/soen321/lhs-menu.shtml	0.91
/~P3/soen321/top.shtml	0.91
/~P3/soen321/	0.88
/~P3/soen321/main.shtml	0.88
/~P3/soen321/grades.shtml	0.88
/~P3/	0.84

Table VIII. 5 Usage Profiles from the non-distinct category (category B)

Category B	Weight
UP_5	
/~comp218/	0.91
/~comp218/Comp218/Comp218WebPage/Html_Files/ConU_Logo.htm	0.84
/~comp218/Comp218/Comp218WebPage/Html_Files/menu.html	0.84
/~comp218/Comp218/Comp218WebPage/Html_Files/Main.html	0.80
/~comp218/Comp218/Comp218WebPage/Html_Files/assignment.html	0.64
/~comp218/Comp218/Comp218WebPage/Html_Files/SectionS.htm	0.50
/~comp218/Comp218/Comp218WebPage/PDF_Files/	0.41
/~comp218/Comp218/Comp218WebPage/Html_Files/Slides.htm	0.27
UP_27	
/~comp218/Comp218/Tut01/step1.html	1.00
/~comp218/Comp218/Tut01/Tutorial01.html	1.00
/~comp218/Comp218/Comp218WebPage/Html_Files/Main.html	0.75
/~comp218/Comp218/Comp218WebPage/Html_Files/menu.html	0.75
/~comp218/	0.75
/~comp218/Comp218/Tut01/step2.html	0.75
/~comp218/Comp218/Comp218WebPage/Html_Files/ConU_Logo.htm	0.75
/~comp218/Comp218/Tut01/step3.html	0.50
/~comp218/Comp218/Comp218WebPage/Html_Files/tutorial.html	0.50
UP_6	
/~P3/comp346/main.shtml	1.00
/~P2/comp354_w2005r.html	0.94
/~P3/comp346/grades.shtml	0.89
/~P2/site_tree.html	0.83
/~P2/about_me.html	0.83
/~P2/	0.83
/current_students.shtml	0.72
/~comp354/	0.61
UP_49	
/~P2/	1.00
/~P2/about_me.html	1.00
/~P2/site_tree.html	1.00
/current_students.shtml	0.86

/~comp354/	0.43
/~P2/comp354_w2005r.html	0.43
/~P2/comp354_w2005pp.html	0.29
/~P3/comp346/main.shtml	0.29
/~P2/comp354/	0.29
UP_7	
/~cc/COMP352/index.html	0.92
/~cc/COMP352/announcements.html	0.83
/~comp352/2005w/	0.83
/~cc/COMP352/material/	0.75
/~comp352/	0.50
/programs/ugrad/courses.html	0.42
/programs/ugrad/cs/comp352.shtml	0.42

Table IX. 5 Usage Profiles from the non-coherent category (category C)

Category C	Weight
UP_38	
/~comp353/winter2005/	1.00
/~comp353/winter2005/SectionX/	1.00
/~P10/comp353/week3/	1.00
/~P10/comp353/week4/	1.00
/~P10/contact.html	1.00
/~P4/	1.00
/help/help.html	1.00
/help/homepage.html	1.00
/help/tutorials/homepage.html	1.00
UP_1	
/~P3/comp346/main.shtml	0.85
/current_students.shtml	0.84
/~P3/comp346/grades.shtml	0.76
/~comp354/	0.51
UP_24	
/~comp442/	1.00
/~comp442/2005W/	1.00
/current_students.shtml	1.00

/~P11/comp354.shtml	0.80
/~comp444/	0.80
/~comp354/	0.60
/~comp354/html/	0.40
/~comp354/html/project.shtml	0.40
/~comp238/2004F/	0.40
/~comp238/2005W/	0.40
/~comp239/	0.40
/~P3/comp346/main.shtml	0.40
/~comp346/	0.40
/~ormandj/comp354/2005/	0.40
/~comp238/	0.40

UP_25

/~soen337/	1.00
/~soen337/WINTER2005/	1.00
/~soen341/	0.67
/current_students.shtml	0.67
/~P12/soen384/common/	0.33
/~soen337/WINTER2005/Tutorials/	0.33
/~P12/soen384/W04/SOEN384_W04_FinalExamResults.htm	0.33
/~soen384/	0.33
/~comp346/	0.33

UP_30

/~P13/	0.75
/~P9/	0.75
/~P14/	0.75
/~P15/	0.75
/~P16/	0.75
/~P17/	0.75
/~P18/	0.75
/~P5/	0.75
/~P8/	0.75
/~P19/	0.75
/~P20/	0.50
/~P21/	0.50
/~P22/bcd-ideas.html	0.50
/~P23/research.html	0.50
/~P24/	0.50
/~P25/	0.50

/~P1/	0.50
/~P26/	0.50
/~P27/	0.50
/~P28/	0.50
/~P29/	0.50
/~P30/	0.50
/~P31/	0.50
/~P32/	0.50
/~P33/Banner.html	0.50
/~P34/contents.html	0.50
/~P11/main.html	0.50
/~P35/	0.50
/~P36/	0.50
/department/hiring.html	0.50

Table X. 5 Usage Profiles from the weak category (category D)

Category D	Weight
UP_37	
/~P7/JainSip/docs/	0.67
/~P7/JainSip/docs/allclasses-frame.html	0.67
/~P7/JainSip/docs/overview-frame.html	0.67
/~P7/JainSip/docs/overview-summary.html	0.67
/~P7/JainSip/docs/javax/sip/message/Request.html	0.33
/~P7/JainSip/docs/javax/sip/package-frame.html	0.33
/~P7/JainSip/docs/javax/sip/package-summary.html	0.33
/~P7/JainSip/docs/javax/sip/ServerTransaction.html	0.33
/~P7/JainSip/docs/javax/sip/SipException.html	0.33
/~P7/JainSip/docs/javax/sip/SipProvider.html	0.33
/~P7/JainSip/docs/javax/sip/SipStack.html	0.33
/~P7/JainSip/docs/javax/sip/Transaction.html	0.33
/~P7/JainSip/docs/javax/sip/TransactionState.html	0.33
/~P7/JainSip/docs/javax/sip/ClientTransaction.html	0.33
/~P7/JainSip/docs/javax/sip/message/package-frame.html	0.33
/~zanibbi/	0.33
UP_36	

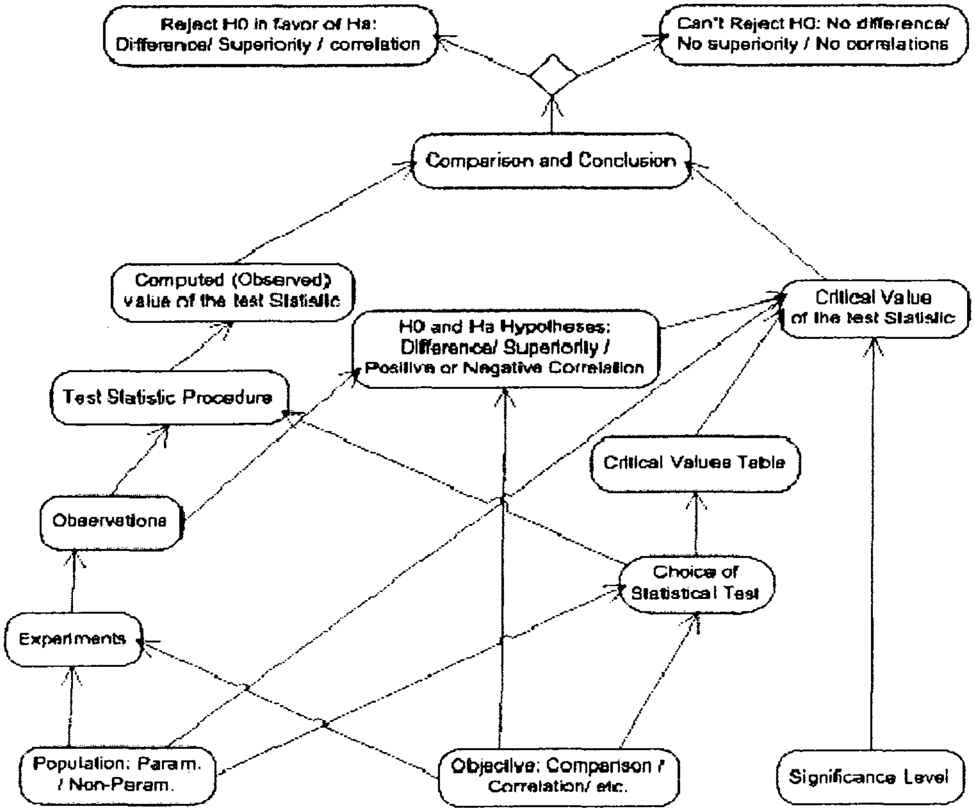
/~P8/	0.67
/people/faculty.html	0.67
/people/people.html	0.67
/~P8/Graphics/graphex.html	0.33
/~P8/people.html	0.33
/~P8/photo.html	0.33
/~P8/techpubs.html	0.33
/~P8/tunick.html	0.33
/~P8/tunick-pictures.html	0.33
/~P8/documentation.html	0.33
/~P8/family.html	0.33
UP_21	
/~comp352/2005w/	0.86
/~comp352/	0.71
/~comp352/2005w/Info/	0.57
/~comp352/2004f/Overheads/	0.29
/~comp352/2005w/Info/out.html	0.29
/~comp352/2005w/X.html	0.29
/current_students.shtml	0.29
UP_12	
/programs/ugrad/courses.html	0.69
/programs/ugrad/cs/cs.html	0.46
/programs/ugrad/soen/soen.html	0.46
/programs/ugrad/honours/honours.html	0.38
/prospective_students.html	0.38
/programs/ugrad/soen/curriculum.html	0.31
UP_48	
/~P9/	0.67
/~P9/hspl/	0.67
/~P9/hspl/honourroll.htm	0.33
/~P9/hspl/personnel.htm	0.33
/~P9/hspl/projects.htm	0.33
/~P9/hspl/publications.htm	0.33
/db/	0.33
/db/db/db_group.html	0.33
/db/db/index.html	0.33
/db/db/main.html	0.33
/db/dbdm/dm.html	0.33

/research/rescenters.html	0.33
/research/research.html	0.33

Table XI. 5 Usage Profiles from the non-covering category (category E)

Category E	Weight
UP_46	
/~P5/	0.60
UP_6	
/~P3/comp346/main.shtml	1.00
/~P2/comp354_w2005r.html	0.94
/~P3/comp346/grades.shtml	0.89
/~P2/site_tree.html	0.83
/~P2/about_me.html	0.83
/~P2/	0.83
/current_students.shtml	0.72
/~comp354/	0.61
UP_49	
/~P2/	1.00
/~P2/about_me.html	1.00
/~P2/site_tree.html	1.00
/current_students.shtml	0.86
/~comp354/	0.43
/~P2/comp354_w2005r.html	0.43
/~P2/comp354_w2005pp.html	0.29
/~P3/comp346/main.shtml	0.29
/~P2/comp354/	0.29
UP_4	
/current_students.shtml	0.98
/~comp335/2004F/	0.57
/~comp335/	0.55
UP_42	
/people/graduates.html	0.75
/~P6/	0.50
/people/people.html	0.50

9.3 Appendix C. Statistical Hypothesis Testing Components and Dependencies


Reject H0 in favor of Ha:
Difference/ Superiority / correlation
Can't Reject H0: No difference/
No superiority / No correlations
Comparison and Conclusion
Computed (Observed)
value of the test Statistic
H0 and Ha Hypotheses:
Difference/ Superiority /
Positive or Negative Correlation
Critical Value
of the test Statistic
Test Statistic Procedure
Critical Values Table
Observations
Choice of
Statistical Test
Experiments
Population: Param.
/ Non-Param.
Objective: Comparison /
Correlation/ etc.
Significance Level