Interval Set Clustering of Web Users with Rough K-means

Pawan Lingras Chad West

Abstract

Data collection and analysis in web mining faces certain unique challenges. Due to a variety of reasons inherent in web browsing and web logging, the likelihood of bad or incomplete data is higher than conventional applications. The analytical techniques in web mining need to accommodate such data. Fuzzy and rough sets provide the ability to deal with incomplete and approximate information. Fuzzy set theory has been shown to be useful in three important aspects of web and data mining, namely clustering, association, and sequential analysis. However, there is limited research on clustering based on rough set theory. Clustering is an important part of web mining that involves finding natural groupings of web resources or web users. Researchers have pointed out some important differences between clustering in conventional applications and clustering in web mining. For example, the clusters and associations in web mining do not necessarily have crisp boundaries. As a result, researchers have studied the possibility of using fuzzy sets in web mining clustering applications. Recent attempts have used genetic algorithms based on rough set theory for clustering. However, the genetic algorithms based clustering may not be able to handle large amount of data typical in a web mining application. This paper proposes a variation of the K-means clustering algorithm based on properties of rough sets. The proposed algorithm represents clusters as interval or rough sets. The paper also describes the design of an experiment including data collection and the clustering process. The experiment is used to create interval set representations of clusters of web visitors.

Index Terms

Clustering, Interval sets, K-means algorithm, Rough sets, Unsupervised learning, Web mining

Authors are with the Department of Mathematics and Computing Science, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3, Canada, Phone: (902) 420-5798, email: Pawan.Lingras@StMarys.ca

I. Introduction

WEB mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of users and information [1]. Logs of web access available on most servers are good examples of the data set used in web mining. Three important operations in web mining are clustering, association, and sequential analysis. This paper focuses on clustering, which is a process of identifying natural groupings of objects.

The clustering process is an important step in establishing user profiles. User profiling on the web consists of studying important charactersitcs of the web visitors. Due to the ease of movement from one portal to another, web users can be very mobile. If a particular web site doesn't satisfy the needs of the user in a relatively short period of time, the user will quickly move on to another web site. Therefore, it is very important to understand the needs and characteristics of web users.

Clustering faces several additional challenges in web mining, compared to traditional applications [1]. The clusters tend to have fuzzy or rough boundaries. The membership of an object in a cluster may not be precisely defined. There is a likelihood that an object may be a candidate for more than one cluster. In addition, due to noise in the recording of data and incomplete logs, the possibility of the presence of outliers in the data set is quite high. Joshi and Krishnapuram [1] argued that the clustering operation in web mining involves modeling an unknown number of overlapping sets. They proposed the use of fuzzy clustering [2], [3], [4] for grouping the web users. This paper proposes unsupervised rough set clustering using a modified K-means algorithm.

Any classification scheme can be represented as a partition of a given set of objects. Objects in each equivalence class of the partition are assumed to be identical or similar. In

web mining, it is not possible to provide an exact representation of each class in the partition [1]. Rough sets [5], [6] enable us to represent such classes using upper and lower bounds. Lingras [7] described how a rough set theoretic classification scheme can be represented using a rough set genome. The resulting genetic algorithms (GAs) were used to evolve groupings of highway sections represented as interval or rough sets. Lingras [8] applied the unsupervised rough set clustering based on GAs for grouping web users. The preliminary experimentation by Lingras [8] illustrated the feasibility of rough set clustering for developing user profiles on the web. However, the clustering process based on GAs seemed computationally expensive for scaling to a larger data set.

One of the most popular and efficient clustering algorithms in conventional applications is K-means clustering. In the K-means approach, randomly selected objects are used as the centroids of clusters. The objects are then assigned to different clusters based on their distance from the centroid. The newly formed clusters are then used to determine new centroids. The process continues until the clusters stabilize. Lingras and Huang [9] provided a theoretical and experimental analysis of various clustering techniques for two datasets of different sizes. They clearly illustrated the computational advantages of the K-means approach for large datasets. However, it is necessary to adapt the K-means algorithm for creating intervals of clusters based on rough set theory.

This paper proposes the adaptation of the K-means algorithm to create interval sets based on rough set theory. The proposed rough K-means algorithm was used to find cluster intervals of web users. The web site that was used for the experimentation catered to first year computing science students. The students used the web site for downloading classnotes and lab assignments; downloading, submitting and viewing class assignments; checking their current marks; as well as for accessing a discussion board. The web site logged as many as

30,000 entries during a busy week. Over the sixteen week period under study, the number of entries in the web log was more than 360,000. The web site was accessed from a variety of locations. Only some of the web accesses were identifiable by student ID. Therefore, instead of analyzing individual students, it was decided to analyze each visit. This also made it possible to guarantee the required protection of privacy. Each visit was determined based on continuous access from a given IP number.

A first year course consists of a wide variety of student behaviour. It will be interesting to study the behaviour pattern of the students over several weeks. Lingras [8] hypothesized that there are three types of visitors: studious, crammers, and workers. Studious visitors download notes from the site regularly. Crammers download all the notes before an exam. Workers come to the site to finish assigned work such as lab and class assignments. Generally, the boundaries of these classes will not be precise. The preliminary experiments based on GAs for two weeks worth of web logs showed the feasibility of using rough sets to represent the three classes. However, GAs were computationally intensive and hence couldn't be easily used for the larger sixteen week dataset. In this paper, the proposed adaptation of K-means based on rough set theory is used to classify the visitors from the entire sixteen week period into upper and/or lower bounds of the three classes mentioned above.

II. Review of Web Personalization and Clustering

The interest in web personalization can be traced back to the Firefly system (www.firefly.com), which was used to suggest music CDs that match users' interests. Similar attempts can also be seen on Amazon.com. When a user requests information about a book, the system provides a list of additional books. The list consists of books purchased by people who bought the book the user is interested in. Attempts at web personalization are increasing at a rapid rate. The increased interest is also leading to more formal frameworks for web personaliza-

tion [10]. Armstrong et al. [11] proposed the use of a tour guide approach. Perkowitz and Etzioni [10] attempted to formalize the concept of adaptive web sites. The adaptive web sites were defined as those that automatically improve their organization and presentation by learning from visitor access patterns. They suggested that much of the earlier work has been focused on fairly simple adaptations such as automatically creating shortcuts in the site, and customization of a web site to suit the needs of each individual user. Perkowitz and Etzioni proposed the use of sophisticated adaptations of web sites to users' needs, and aggregation of information gleaned from the user population to improve the navigation for a large number of users. Perkowitz and Etzioni's further extended their research [12] by focusing on clustering of web documents in anticipation of user needs. Joshi and Krishnapuram [1] discussed the issue of clustering for web mining in great detail and suggested that a fuzzy clustering approach may be more suitable than the traditional statistical approaches used by Perkowitz and Etzioni [12]. There is a significant amount of research related to fuzzy clustering [2], [3], [4] in literature. Rough set theory [5] has often been considered to be an alternative to the fuzzy set theory. In many cases, rough set theory has also been used to complement fuzzy set theory. However, there is limited unsupervisory clustering methodology based on rough set theory. This paper proposes an efficient K-means clustering approach based on rough set theory.

III. Review of Rough Sets

The notion of rough set was proposed by Pawlak [5]. This section provides a brief summary of the concepts from rough set theory essential for introducing the rough set theoretic K-means algorithm.

Let U denote the universe (a finite ordinary set), and let $R \subseteq U \times U$ be an equivalence (indiscernibility) relation on U. The pair $A = (U, R)$ is called an approximation space.

The equivalence relation R partitions the set U into disjoint subsets. Such a partition of the universe is denoted by $U/R = E_1, E_2, ..., E_n$, where E_i is an equivalence class of R. If two elements $u, v \in U$ belong to the same equivalence class $E \subseteq U/R$, we say that u and v are indistinguishable. The equivalence classes of R are called the elementary or atomic sets in the approximation space $A = (U, R)$. The union of one or more elementary sets is called a composed set in A. The empty set $\emptyset$ is also considered a special composed set. $Com(A)$ denotes the family of all composed sets.

Since it is not possible to differentiate the elements within the same equivalence class, one may not be able to obtain a precise representation for an arbitrary set $X \subseteq U$ in terms of elementary sets in A. Instead, any X may be represented by its lower and upper bounds. The lower bound $\underline{A}(X)$ is the union of all the elementary sets which are subsets of X, and the upper bound $\overline{A}(X)$ is the union of all the elementary sets which have a non-empty intersection with X. The pair $\left(\underline{A}(X), \overline{A}(X)\right)$ is the representation of an ordinary set X in the approximation space $A = (U, R)$, or simply the rough set of X. The elements in the lower bound of X definitely belong to X, while elements in the upper bound of X may or may not belong to X.

IV. Review of K-Means Approach

K-means clustering is one of the most popular statistical clustering techniques. The name K-means originates from the means of the k clusters that are created from n objects. Let us assume that the objects are represented by m-dimensional vectors. The objective is to assign these n objects to k clusters. Each of the clusters is also represented by an m-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing k objects as the centroids of the k clusters. The objects are assigned to one of the k clusters based on the minimum value of the distance $d(\mathbf{v}, \mathbf{x})$ between

the object vector $\mathbf{v}$ and the cluster vector $\mathbf{x}$. The distance $d(\mathbf{v},\mathbf{x})$ is given by:

$$d(\mathbf{v},\mathbf{x}) = \frac{\sum_{j=1}^{m}(v_j - x_j)^2}{m} \quad (1)$$

After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as:

$$x_j = \frac{\sum_{\mathbf{v}\in\mathbf{x}} v_j}{\text{Size of cluster } \mathbf{x}}, \text{where } 1 \leq j \leq m. \quad (2)$$

The process stops when the centroids of clusters stabilize, i.e. the centroid vectors from the previous iteration are identical to those generated in the current iteration.

V. Adaptation of K-means to Rough Set Theory

Rough sets were proposed using equivalence relations. However, it is possible to define a pair of upper and lower bounds $\left(\underline{A}(X),\overline{A}(X)\right)$ or a rough set for every set $X \subseteq U$ as long as the properties specified by Pawlak [5] are satisfied. Yao et al. [13] described various generalizations of rough sets by relaxing the assumptions of an underlying equivalence relation. Skowron and Stepaniuk [14] discuss a similar generalization of rough set theory. The present study uses such a generalized view of rough sets. If one adopts a more restrictive view of rough set theory, the rough sets developed in this paper may have to be looked upon as interval sets.

Let us consider a hypothetical classification scheme

$$U/P = \{X_1, X_2, \ldots, X_k\} \quad (3)$$

that partitions the set U based on certain criteria. The actual values of X_i are not known. The classification of web users is an example of such a hypothetical classification scheme. A set of visitors can be classified into three classes depending on the predominant usage:

studious, crammers, and workers. However, the actual sets corresponding to each one of these classes are not known. Let us assume that due to insufficient knowledge it is not possible to precisely describe the sets $X_i, 1 \leq i \leq k$, in the partition. However, it is possible to define each set $X_i \in U/P$ using its lower $\underline{A}(X_i)$ and upper $\overline{A}(X_i)$ bounds based on the available information. In this study, the available information consists of web access logs. Since the objects and clusters in the K-means algorithm are represented by vectors, we will use vector representations, $\mathbf{v}$ for an object and $\mathbf{x_i}$ for cluster X_i.

We are considering the upper and lower bounds of only a few subsets of U. Therefore, it is not possible to verify all the properties of the rough sets [5]. However, the family of upper and lower bounds of $\mathbf{x_i} \in \mathbf{U/P}$ are required to follow some of the basic rough set properties such as:

(C1) An object $\mathbf{v}$ can be part of at most one lower bound

(C2) $\mathbf{v} \in \underline{A}(\mathbf{x}_i) \Longrightarrow \mathbf{v} \in \overline{A}(\mathbf{x}_i)$

(C3) An object $\mathbf{v}$ is not part of any lower bound

$$\Updownarrow$$

$\mathbf{v}$ belongs to two or more upper bounds.

Note that (C1)-(C3) are not necessarily independent or complete. However, enumerating them will be helpful in understanding the rough set adaptation of the K-means algorithm.

K-means clustering is a process of finding centroids for all clusters, and assigning objects to each cluster based on their distance from the centroids. This process is done iteratively until stable centroid values are found. Incorporating rough sets into K-means clustering requires the addition of the concept of lower and upper bounds. Eq. 2 that is used to calculate the centroids of clusters needs to be modified to include the effects of lower as well as upper bounds. The modified centroid calculations for rough sets is then given by:

$$x_j = \begin{cases} w_{lower} \times \frac{\sum_{\mathbf{v} \in \underline{A}(\mathbf{x})} v_j}{|\underline{A}(\mathbf{x})|} + w_{upper} \times \frac{\sum_{\mathbf{v} \in (\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}))} v_j}{|\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})|} & \text{if } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \neq \emptyset \\ w_{lower} \times \frac{\sum_{\mathbf{v} \in \underline{A}(\mathbf{x})} v_j}{|\underline{A}(\mathbf{x})|} & \text{otherwise} \end{cases}, \quad (4)$$

where $1 \leq j \leq m$. The parameters w_{lower} and w_{upper} correspond to the relative importance of lower and upper bounds. It can be easily seen that eq. 4 is a generalization of eq. 2. If the upper bound of each cluster were equal to its lower bound, the clusters will be conventional clusters. Therefore, the boundary region $\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})$ will be empty, and the second term in the equation will be ignored. Thus, eq. 4 will reduce to eq. 2.

The next step in the modification of the K-means algorithms for rough sets is to design criteria to determine whether an object belongs to the upper or lower bound of a cluster. For each object vector, $\mathbf{v}$, let $d(\mathbf{v}, \mathbf{x}_i)$ be the distance between itself and the centroid of cluster X_i. The differences $d(\mathbf{v}, \mathbf{x}_i) - d(\mathbf{v}, \mathbf{x}_j)$, $1 \leq i, j \leq k$, were used to determine the membership of $\mathbf{v}$ as follows.

1. If $d(\mathbf{v}, \mathbf{x}_i) - d(\mathbf{v}, \mathbf{x}_j) \leq threshold$, for any pair (i, j), then $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$ and $\mathbf{v} \in \overline{A}(\mathbf{x}_j)$. Furthermore, $\mathbf{v}$ was not part of any lower bound. The above criterion gurantees that property (C3) is satisfied.
2. Otherwise, $\mathbf{v} \in \underline{A}(\mathbf{x}_i)$ such that $d(\mathbf{v}, \mathbf{x}_i)$ is the minimum for $1 \leq i \leq k$. In addition, by property (C2), $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$.

The rough K-means algorithm, described above, depends on three parameters w_{lower}, w_{upper}, and $threshold$. Experimentation with various values of the parameters is necessary to develop a reasonable rough set clustering. The following sections describe the design and results of such an experiment.

VI. Study data and Design of the Experiment

The study data was obtained from the web access logs of the introductory first year course in computing science at Saint Mary's University. The initial number of students in the course was 180. The number reduced over the course of the semester to 130 to 140 students. The students in the course come from a wide variety of backgrounds, such as computing science major hopefuls, students taking the course as a required science course, and students taking the course as a science or general elective. As is common in a first year course, students' attitudes towards the course also vary a great deal. It was hoped that the profile of visits will reflect some of the distinctions between the students. For the initial analysis, it was assumed that the visitors could fall into one of the following three categories:

1. Studious: These visitors download the current set of notes. Since they download a limited/current set of notes, they probably study classnotes on a regular basis.
2. Crammers: These visitors download a large set of notes. This indicates that they have stayed away from the classnotes for a long period of time. They are planning for pre-test cramming.
3. Workers: These visitors are mostly working on class or lab assignments or accessing the discussion board.

The rough set classification scheme is expected to specify lower and upper bounds for these classes.

It was hoped that the above mentioned variety of user behaviours would be identifiable based on the number of web accesses, types of documents downloaded, and time of day. Certain areas of the web site were protected and the users could only access them using their IDs and passwords. The activities in the restricted parts of the web site consisted of submitting a user profile, changing a password, submission of assignments, viewing the

submissions, accessing the discussion board, and viewing current class marks. The rest of the web site was public. The public portion consisted of viewing course information, a lab manual, classnotes, class assignments, and lab assignments.

If the users only accessed the public web site, their IDs would be unknown. Therefore, the web users were identified based on their IP address. This also made sure that user privacy was protected. A visit from an IP address started when the first request was made from the IP address. The visit continued as long as the consecutive requests from the IP address had sufficiently small delay.

The web logs were preprocessed to create an appropriate representation of each user corresponding to a visit. The abstract representation of a web user is a critical step that requires a good knowledge of the application domain. Previous personal experience with the students in the course suggested that some of the students print preliminary notes before a class and an updated copy after the class. Some students view the notes on-line on a regular basis. Some students print all the notes around important days such as midterm and final examinations. In addition, there are many accesses on Tuesdays and Thursdays, when the in-laboratory assignments are due. On and off campus points of access can also provide some indication of a user's objectives for the visit. Based on some of these observations, it was decided to use the following attributes for representing each visitor:

1. On campus/Off campus access.
2. Day time/Night time access: 8 a.m. to 8 p.m. was considered to be the day time.
3. Access during lab/class days or non-lab/class days: All the labs and classes were held on Tuesday and Thursday. The visitors on these days are more likely to be workers.
4. Number of hits.
5. Number of classnotes downloads.

The first three attributes had binary values of 0 or 1. The last two values were normalized. Since the classnotes were the focus of the clustering, the last variable was assigned higher importance. Previously, Lingras [8] used rough set theoretic genetic algorithms for developing intervals of clusters. The previous analysis was carried out for a short period of two weeks around the midterm examination. These two weeks logged 54,528 entries. The data preparation identified a total of 3,243 visits. The visitors that did not download any notes clearly fall in the worker category. Therefore, the clustering was restricted to those 1,264 visits, which downloaded at least one classnotes file. The initial analysis demonstrated the feasibility of developing rough set representation of clusters of the visitors to the web site. However, the approach could not be easily extended to the longer period of sixteen weeks. The present study used all 361,609 entries from the web log over the sixteen week period. There were a total of 22,996 visits. The visits that didn't download any classnotes were treated as workers. The clustering was applied to the remaining 8,442 visits.

The modified K-means algorithm was run for various values of $threshold$ and initial centroid vectors. The value of w_{lower} was set at 0.75 and w_{upper} was equal to 0.25. The resulting rough set classification schemes were subjectively analyzed.

VII. Results and Discussion

Table. I shows the average values of the five variables used in the clustering. It was possible to classify the three clusters as studious, crammer, and worker. The workers had the lowest number of hits per visit. The average hits for users in the lower bound of the workers class was 16. The users that may be from the worker class (upper bound) averaged 23 hits per visit. The users that were definitely workers downloaded only 1.3 documents on average, which probably corresponded to the sample programs from the lab manuals. Another interesting fact was that the users from the lower bound of the worker cluster

Cluster	On Campus	Day time	Lab/class day	Hits	Classnotes
$\underline{A}(Studious)$	0.07	0.5	0.02	61.1	21.6
$\overline{A}(Studious)$	0.55	0.67	0.32	25.3	4.69
$\underline{A}(Crammer)$	0.48	0.70	0.33	145.4	70.48
$\overline{A}(Crammer)$	0.44	0.68	0.27	96.9	46.10
$\underline{A}(Worker)$	1	1	1	16.2	1.31
$\overline{A}(Worker)$	0.65	0.74	0.47	23.2	3.89

TABLE I
VECTOR REPRESENTATION OF CLUSTERS

always came from on-campus locations, on lab/class days, and during the day time.

There was a significant overlap between the upper bounds of the worker and studious classes. This fact is also evident in the centroid vectors for the two upper bounds. The studious visitors had wider variations. Their average hits were 61.1 for the lower bound. Average notes downloads for users that were definitely studious was 21.6. The hits and notes for the lower bound of the studious cluster are rather high. This fact can be explained by the relatively small size of the lower bound with only 44 visitors. The upper bound was significantly larger with 6596 visitors. The large size of the upper bound provided more moderate values for hits and notes. The users that may be studious (upper bound) had a smaller number of hits (25.3) and downloaded fewer notes (4.69). The visitors for the lower bound of studious cluster contrasted sharply from the workers in the values of the first three variables. Studious visitors almost always came from off-campus locations. Half of them visited during the night time. Almost all of them visited during non-lab/class days.

The crammers, on the other hand, had the highest number of hits per visit. The lower bound of the crammer class averaged 145 visits and 70 classnotes. The upper bound had a slightly smaller number of hits and classnotes. However, these values were still significantly larger than those of the lower/upper bounds of any other class. It is possible that some of these visitors may in fact be various search engines. More investigations are necessary to

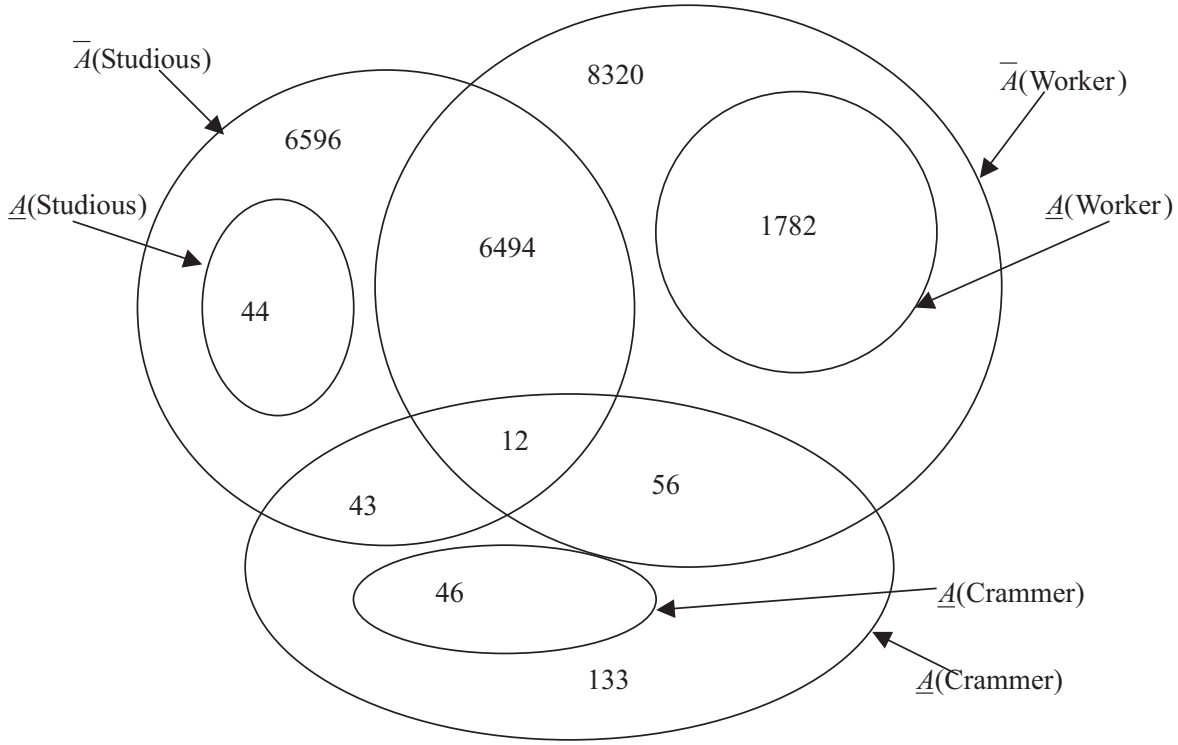

Fig. 1. Upper and lower bound cardinalities of the clusters

verify this hypothesis.

Due to the higher weights associated with the hits and classnotes, the rest of the attributes played a less significant role in distinguishing visitors. However, as mentioned before, workers always came from on-campus locations, on lab/class days, and during the day time. The studious visitors were significantly different. They almost always came from off-campus and during non-lab/class days. The rest of the upper and lower bounds are somewhat difficult to analyze for the first three variables, based on Table I. However, it is easy to see that non-workers tended to come more during the non-lab/class days. A more detailed study of these associations is necessary for more reliable conclusions.

There was an appreciable amount of overlap between various upper bounds. This seems to suggest that the proposed clustering satisfies one of the important criteria of overlapping cluster boundaries [1]. In order to verify the fact that some of the visitors belonged to multiple categories, a more detailed study of the cluster memberships was carried out. The cardinalities of the upper and lower bounds of the clusters are shown in Fig. 1.

Fig. 1 seems to indicate that there were a higher number of workers. The lower bound of workers consisted of 1782 visitors, while the upper bound contained 8320 visitors. Even though the lower bound of the studious cluster consisted of only 44 visitors, the upper bound had 6596 visitors. The crammer cluster was the smallest with 46 visitors in the lower bound and 133 in the upper bound. An additional analysis was also carried out by taking the intersections of all of the upper bounds. The intersections provide some indication of the visitors who couldn't be precisely classified. A total of 43 visitors were identified as either studious or crammers. There were 6494 visitors who could be classified as either workers or studious. This large overlap seems reasonable because the workers may also download some of the notes for study after taking care of their other work on the web site. It is quite likely that the distinction between studious and workers may be fuzzy, a possibility indicated by the large overlap between their upper bounds. Furthermore, 56 visitors may have been either crammers or workers. Finally, 12 visitors were present in all of the three upper bounds.

VIII. Summary and Conclusions

This paper proposed an adaptation of the K-means algorithm to develop interval clusters of web visitors using rough set theory. Web visitors for an introductory computing science course were used in the experiments. It was assumed that the visitors would be classified as studious, crammers, or workers. Since some of the visitors may not precisely belong to one of the classes, the clusters were represented using rough sets.

In order to develop interval clusters the K-means algorithm was modified based on the concept of lower and upper bounds. The experiment resulted in a meaningful clustering of web visitors. The study of variables used for clustering made it possible to clearly identify the three clusters as studious, workers, and crammers. It was interesting to note that the visitors from the lower bound of workers visited from on-site locations, on lab/class days, and during the day-time. Visitors from the lower bound of the studious cluster, on the other hand, came from off-campus locations, during the night. Even though the lower bounds of studious and worker clusters were significantly different from each other, there was a large overlap between their upper bounds. This seemed reasonable because many of the visiting patterns for regular study may be similar to those of workers. As expected, crammers had the highest number of hits and downloads per visit. A more detailed association analysis will be necessary to understand the implications of rough set clustering described in this paper. Results of such an analysis will be provided in subsequent publications.

Acknowledgment

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for their financial support. The help of Stephanie MacDonald during the programming is also appreciated.

References

[1] A. Joshi and R. Krishnapuram, "Robust fuzzy clustering methods to support web mining," in *Proceedings of the workshop on Data Mining and Knowledge Discovery, SIGMOD '98*, 1998, pp. 15/1–15/8.

[2] R. J. Hathaway and J.C. Bezdek, "Switching regression models and fuzzy clustering," *IEEE Transactions of Fuzzy Systems*, vol. 1, no. 3, pp. 195–204, 1993.

[3] R. Krishnapuram H. Frigui and O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation: Parts I and II," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 1, pp. 29–60, 1995.

[4] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[5] Z. Pawlak, "Rough sets," *International Journal of Information and Computer Sciences*, vol. 11, pp. 145–172, 1982.

[6] Z. Pawlak, "Rough classification," *International Journal of Man-Machine Studies*, vol. 20, pp. 469–483, 1984.

[7] P. Lingras, "Unsupervised rough set classification using GAs," *Journal Of Intelligent Information Systems*, vol. 16, no. 3, pp. 215–228, 2001.

[8] P. Lingras, "Rough set clustering for web mining," in *Proceedings of 2002 IEEE International Conference on Fuzzy Systems*, 2002.

[9] P. Lingras and X. Huang, "Statistical, evolutionary, and neurocomputing clustering techniques: cluster-based versus object-based approaches," *submitted to European Journal of Operational Research*, 2002.

[10] M. Perkowitz and O. Etzioni, "Adaptive web sites: an AI challenge," in *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

[11] T. Joachims R. Armstrong, D. Freitag and T. Mitchell, "Webwatcher: A learning apprentice for the world wide web," in *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[12] M. Perkowitz and O. Etzioni, "Adaptive web sites: Conceptual cluster mining," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

[13] Y.Y. Yao X. Li T.Y. Lin and Q. Liu, "Representation and classification of rough set models," in *Proceeding of Third International Workshop on Rough Sets and Soft Computing*, 1994, pp. 630–637.

[14] A. Skowron and J. Stepaniuk, "Information granules in distributed environment," in *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, Setsuo Ohsuga Ning Zhong, Andrzej Skowron, Ed., pp. 357 – 365. Springer-Verlag, Lecture notes in Artificial Intelligence 1711, Tokyo, 1999.