Effective web personalisation based on rough biclustering

H. Hannah Inbarani* and K. Thangavel

Department of Computer Science,
Periyar University,
Salem-636 011, Tamil nadu, India
Fax: 0427-2345124
E-mail: hhinba@gmail.com
E-mail: ktvelu@gmail.com
*Corresponding author

Abstract: Web personalisation systems include a new generation of recommender systems that integrate multiple online channels, are more scalable, are more adaptive and can better handle user interactivity. Efficient and intelligent techniques from artificial intelligence, machine learning, web mining and statistics are needed to mine this data for actionable knowledge, and to effectively use the discovered knowledge to enhance the users' experience. In this paper, we propose a rough biclustering approach for creating user model, based on which user profiles are constructed and the user profiles are matched with the active user session for web page recommendation. To determine the effectiveness of the proposed approach, it is compared with conventional biclustering, spectral co-clustering and CDK-means biclustering using the evaluation metric used for page recommendation. The experimental results show that the proposed rough biclustering algorithm outperforms the other approaches for web page recommender systems.

Keywords: co-clustering; biclustering; web personalisation; page recommendation; user modelling.

Reference to this paper should be made as follows: Hannah Inbarani, H. and Thangavel, K. (2013) ‘Effective web personalisation based on rough biclustering’, *Int. J. Granular Computing, Rough Sets and Intelligent Systems*, Vol. 3, No. 1, pp.59–84.

Biographical notes: H. Hannah Inbarani is currently working as an Assistant Professor, Department of Computer Science, Periyar University, Salem, India. She received her PhD in Computer Science from Periyar University in 2012. Her research interest includes rough sets, fuzzy systems, internet technologies, swarm intelligence and web mining. She has published several research papers in the areas like web usage mining, bioinformatics, rough sets, image processing and swarm intelligence techniques in reputed international journals.

K. Thangavel received his PhD degree in the area of Optimisation Algorithms from Gandhigram Rural Institute-Deemed University, Gandhigram, India in 1999. Currently he is working as a Professor and the Head in the Department of Computer Science, Periyar University, Salem, India. He has published more than 125 research publications in various international journals. His research interests include image processing, soft computing and bio-informatics. He is an awardee of Tamilnadu State award for the year 2009 and he received the Sir C V Raman award from His Excellency Dr. A.P.J. Abdul Kalam, former President of India, instituted by Periyar University, Salem for the year 2012.

1 Introduction

With today's increasing information problem overload, the area of recommender systems research remains to be more challenging than ever before. The continued explosion in the amount of content and the number for information sources available online is making the need for effective personalised content delivery more acute. This has resulted in a renewed interest in web personalisation as an indispensable tool for web-based organisations. Numerous approaches are introduced for personalisation system which can be categorised into two major groups, which are content-based filtering agents and collaborative filtering systems (Mobasher et al., 2000). The main limitation of content-based filtering is the lack of diversity in the recommendations. One research area that has recently contributed greatly for this problem is web mining.

The personalisation systems using web usage mining incorporate data mining techniques to discover interesting patterns in web usage data. The data source for web usage mining is generally the server access log, but sometimes a client-side agent collects data. The purpose of web usage mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns. More advanced data mining methods and algorithms adapted appropriately for use in the web domain include association rules, sequential pattern discovery, clustering (Hui et al., 2006), biclustering (Koutsonikola and Vakali, 2009) and classification. Clustering is used to group together items that have similar characteristics (Petridou et al., 2008).

1.1 Motivation

The primary motivation behind the use of clustering in collaborative filtering (Xu, 2008) and web usage mining is to improve the efficiency and scalability of the real-time personalisation tasks. In the context of web personalisation, this task involves clustering user sessions identified in the preprocessing stage. A variety of clustering techniques can be used for clustering similar users' sessions based on occurrence patterns of uniform resource locator (URL) references. User sessions can be mapped into a multidimensional space as vectors of URL references. Another approach for obtaining aggregate usage profiles is to directly compute (overlapping) clusters of page view references based on how often they occur together across user sessions (rather than clustering sessions, themselves). The usage profiles obtained in this way is called page cluster (Mobasher et al., 2002).

However, both user clustering (UC) and page view clustering (PC) are one-sided approaches, in the sense that they examine similarities either only between users or only between pages, respectively. This way, they ignore the clear duality that exists between users and items. Furthermore, UC and PC algorithms cannot detect partial matching of preferences, because their similarity measures consider the entire set of pages or users, respectively. Another limitation of user or page clustering algorithms is that number of clusters must be given as input based on the structure of input patterns. Hence, the first goal this work is to simultaneously cluster users and pages based on their URL references.

In order for a categorisation technique to be useful in an application such as web mining, it needs to satisfy five basic requirements (Joshi and Krishnapuram, 1998):

- The technique should be able to determine the appropriate number of components automatically, since a priori knowledge about the number of components is rarely available.
- It needs to be robust. By robustness, it is meant that the categorisation process (and hence the performance of a system) should not be affected drastically due to outliers (bad observations), provided there is enough good data to support the assumed model.
- The technique should be able to handle overlapping components.
- It needs to be scalable to extremely large high-dimensional datasets.
- Another key feature to be addressed in the developing web usage mining techniques is vagueness and imprecision inherent web usage data.

Hence, the second goal of this work is to propose a robust technique for handling overlapping objects which is scalable to large high dimensional datasets.

1.2 Contribution

The simultaneous clustering of users and pages discovers biclusters, which correspond to groups of users which exhibit high correlation on groups of pages. For page recommendation, biclusters allow the computation of similarity between a test user and a bicluster only on the pages that are included in the bicluster. Thus, partial matching of preferences is taken into account too. Moreover, a user can be matched with several nearest biclusters, thus to receive recommendations that cover the range of his various preferences. A simple and robust biclustering approach was already proposed (Hannah Inbarani and Thangavel, 2011) in our previous work for web page recommendation.

The biclusters may *overlap*, which means that several users or pages of the session matrix may participate in multiple biclusters. The degree of overlapping is not taken into account for each user and for each page in the biclustering techniques.

When there is insufficient knowledge to precisely define clusters as sets, rough sets are used. Web users visit pages without sufficient information about the websites. They wander between web pages without certainty. The rough set is a tool for handling vagueness and uncertainty inherent to decision situations. Hence, in this paper, we propose rough set-based biclustering technique is proposed for simultaneous clustering of users and pages.

The contributions of this paper are summarised as follows:

- To capture the range of the user's preferences and to handle uncertainty which prevails in the web navigation patterns we introduce for the first time, to our knowledge, the application of rough biclustering (RB) algorithm for web personalisation. The effectiveness of this approach is compared with spectral co-clustering approach proposed by Dhillon (2001) for co-clustering of words and documents, CDK-means approach proposed by Pensal et al. (2005) which is a K-means like approach for biclustering of categorical data and conventional biclustering (CB) approach proposed in our previous work (Hannah Inbarani and Thangavel, 2011) using recommendation evaluation metrics and the results are discussed in Section 7.

- A web user profiling approach and recommendation approach based on RB is also proposed for web page recommendation.

The rest of this paper is organised as follows. Section 2 summarises the related work, whereas Section 3 lists the research issues addressed in this paper, Section 4 describes the methodology for web page recommendation process and the proposed RB approach is explained in Section 5 and Section 6 discusses the comparative analysis of RB with CB, CDK-means and spectral co-clustering approaches.

2 Related work

Web clustering can involve either grouping of users who present similar browsing patterns or grouping of pages having related content based on information derived from different sources.

Specifically, UC approaches can be based on usage data recorded in web server log files and create web communities, i.e., groups of users with similar browsing behaviour (Pallis and Koutsonikola, 2006). On the other hand, in web page clustering approaches, information can be extracted from pages' content (Hammouda and Kamel, 2004), structure, i.e., links between web pages or pages' structure as described by the involved tags (Tanasa and Trousse, 2004), and usage data, i.e., which pages tend to be accessed by users with similar interests (Nakagawa and Mobasher, 2003). Moreover, the clustering results may be beneficial for a wide range of applications such as websites' personalisation (Nasraoui et al., 2008), web caching and prefetching (Li et al., 2007), search engines (Liu et al., 2005) and content delivery networks (Pallis and Koutsonikola, 2006). In addition, the clustering results can contribute to the enhancement of recommendation engines (Chi et al., 2008) and to the design of collaborative filtering systems (Srinivasa and Medasani, 2004).

In user/page clustering approaches, the exact user access patterns are not taken into account. Hence, recent studies have used biclustering approaches to disclose this duality between users and pages, by grouping them in both dimensions simultaneously (Liu et al., 2005; Koutsonikola and Vakali, 2009). The goal of these approaches is to identify groups of related web users and pages, which results from the tendency of some users to visit the same set of pages. This behaviour characterises users' interests as similar and highly related to the topic that the specific set of pages involves. The obtained results are particularly useful for applications such as e-commerce and recommendation engines, since relations between clients and products may be revealed. These relations are more meaningful than the one-way clustering of users or pages (Koutsonikola and Vakali, 2009).

Usually, the clusters (or biclusters) resulting from the web usage mining algorithms may not necessarily have crisp boundaries, rather they have fuzzy or rough boundaries (Hannah Inbarani and Thangavel, 2009). Koutsonikola and Vakali (2009) have proposed fuzzy biclustering approach to cluster users and pages simultaneously. The limitation of this two way clustering approach is that it is based on clustering and so the exact user access patterns cannot be obtained. Hence, it is not suitable for page recommendation as correlation between pages disappear as the user access patterns are merged in user and page clustering techniques.

The concept of biclustering has been used in Mirkin (1996) to perform grouping in a matrix by using both rows and columns. However, biclustering has been used previously in Hartigan (1972) under the name direct clustering. Recently, biclustering (also known as co-clustering, two-sided clustering, two-way clustering) has been exploited by many researchers in diverse scientific fields, towards the discovery of useful knowledge (Cheng and Church, 2000; Dhillon, 2001; Long et al., 2005). One of these fields is bioinformatics (Tang et al., 2001), and more specifically, microarray data analysis. The results of each microarray experiment are represented as a data matrix, with different samples as rows and different genes as columns. Other fields are text mining (Dhillon, 2001) and web mining (Koutsonikola and Vakali, 2009).

There are several approaches to deal with the biclustering problem. Many different algorithms for biclustering have already been proposed in the literature (Cheng and Church, 2000; Tang et al., 2001). In short, these methods can be classified by:

1 the type of biclusters they find

2 the structure of these biclusters

3 the way the biclusters are discovered.

The type of the biclusters is related to the concept of similarity between the elements of the matrix. For instance, some algorithms search for constant value biclusters, while others search for coherent values of the elements or even for coherent evolution biclusters (de Castro et al., 2007). The structure of the biclusters can be of many types. There are single bicluster algorithms, which find only one bicluster in the centre of the matrix; the exclusive columns and/or rows, in which the biclusters cannot overlap in either columns or rows of the matrix; arbitrary positioned, overlapping biclusters and overlapping biclusters with hierarchical structure. The way the biclusters are discovered refers to the number of biclusters discovered per run. Some algorithms find only one bicluster, others simultaneously find several biclusters and some of them find small sets of biclusters at each run. Rough and fuzzy biclustering approaches were applied in Nizar Banu and Hannah Inbarani (2011) but these approaches are based on K-means and fuzzy C-means and so the number of clusters must be given as input and it is a time consuming process.

In this paper, we propose a RB technique for simultaneous clustering of users and pages and the proposed approach is compared with co-clustering approach proposed by (Dhillon, 2001) for co-clustering word and documents and CDK-means proposed by (Pensal et al., 2005) which is a K-means like approach for biclustering of categorical data and the results are discussed in Section 5.

3 Research issues

In this section, we examine the issues of web page recommendation systems. Table 1 summarises the symbols that are used in the sequel.

Table 1 Symbols

Symbol	*Definition*
nb	Number of biclusters
K	Number of recommended biclusters
m	Number of users
n	Number of pages
UP	User profile matrix of size ($nb \times n$)
P	Pattern matrix of size ($nb \times n$)
BU	Users in bicluster ($nb \times m$)
BP	Users in bicluster ($nb \times n$)
μ_{ui}	User membership values
μ_{pij}	Page membership values
Nn	Active sub session size
N	Number of pages recommended
S	Session matrix/user access matrix
$p_1, p_2, \ldots, p_n$	Pages/URLs
$u_1, u_2, \ldots, u_m$	Users

3.1 Accuracy

Web personalisation is viewed as a data mining task. Hence, the accuracy of models learned for this purpose can be evaluated using a number of metrics that have been used in machine learning and data mining literature such as mean absolute error (MAE) and area under the ROC curve, depending on the formulation of the learning task.

In this work, MAE is used to measure the accuracy of web page recommendation results.

3.2 Scalability

The performance and scalability dimension aims to measure the response time of a given recommendation algorithm and how easily it can scale to handle a large number of concurrent requests for recommendations. Typically, these systems need to be able to handle large volumes of recommendation requests without significantly adding to the response time of the website that they have been deployed on. The proposed approach is scalable because the recommendation is performed online and user profile discovery is an offline process. The online parts concern the time it takes to create a recommendation list, based on the pages visited by the active user session. As proved in Symeonidis et al. (2008), online part of biclustering approaches take less execution time than user/page clustering approaches.

3.3 Sparsity

Sparsity refers to the fact that as the number of pages in a website increases, even the most prolific users of the system will only visit a very small percentage of all pages. As a result, there will be many pairs of customers that have no pages in common and even those that do will not have a large number of common pages. Sparsity can be handled well by selecting appropriate value for *K*.

3.4 Precision

Precision and recall are standard metrics used in information retrieval. While precision measures the probability that a selected item is relevant, recall measures the probability that a relevant item is selected. Precision and recall are commonly used in evaluating the selection task. Coverage measures the percentage of the universe of items that the recommendation system is capable of recommending. The F1-measure that combines precision and coverage has also been used for this purpose task. In this work, precision, coverage and F1-measure are the metrics used for measuring the prediction process.

3.5 Similarity measure

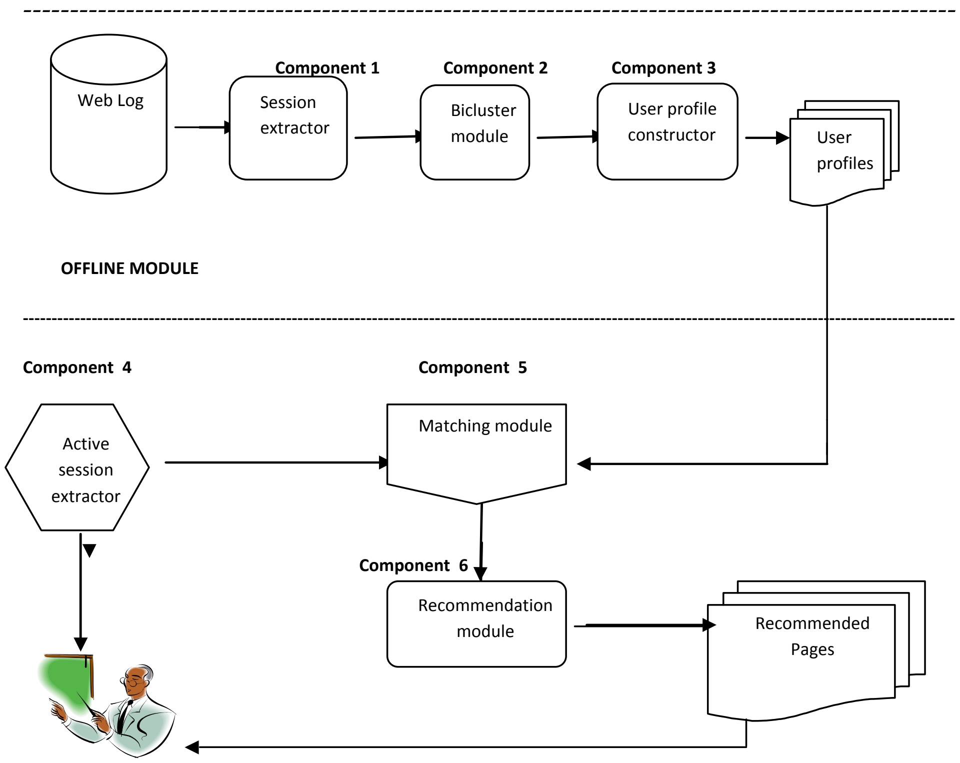
Similarity measure: The most extensively used similarity measures are based on correlation and cosine-similarity (Symeonidis et al., 2008). Specifically, user-based clustering algorithms mainly use Pearson’s correlation, whereas for PC algorithms, adjusted cosine measure is preferred. The adjusted cosine measure is a variation of the simple cosine formula that normalises bias from subjective ratings of different users. In this work, cosine similarity measure is to find the similarity of users with patterns.

4 Methodology

Web personalisation system based on web usage mining discovers web usage profiles, followed by a recommendation system that can respond to the users’ individual interests.

The architecture of the proposed system is shown in Figure 1.

Figure 1 System architecture (see online version for colours)



The recommendation process consists of the following four major steps.

1 preprocessing

2 biclustering of users and pages

3 user profiling

4 page recommendation.

4.1 Preprocessing

Data cleaning operation is performed as defined in Tanasa and Trousse (2004), which removes image files and style sheet files. The access log of a web server is a record of all files (URLs) accessed by users on a website. Each log entry consists of the information components such as remotehost, Rfc931, Authuser, date, request, status and bytes.

The sample entries in the web log file are listed in Figure 2.

Figure 2 Sample web log file

```
218.248.30.146 [21/Nov/2009:03:10:51 + 0530]  "POST/make_slides.php   HTTP/1.1" 200 740
216.104.15.130 [21/Nov/2009:03:20:37 + 0530]  "GET/messengerplus.php  HTTP/1.0" 200  15202
```

In the next step, using user session identification process, user sessions are identified and session matrix is created. User access matrix $S = \{S_{ij}\}$ where $S_{ij} = 1$ if page j has been visited by user i otherwise it is set to zero. The weight associated with each visited page is represented by $W = \{W_{ij}\}$ where each entry in the weight matrix specifies the number of hits on a specific page as defined in Claypool et al. (2001). For each user, the weight vector of each navigational session is represented as a sequence of visited pages with corresponding weights $\{w_{11}, w_{12}, w_{13}, \ldots, w_{1n}\}$ where w_{ij} denotes the weight for a page j visited in i^{th} user session. Each row of user access matrix is called a session vector/user access vector/transaction.

4.2 The biclustering process

The biclustering process on a user access matrix involves the determination of a set of clusters taking into account both users and pages. Each bicluster is defined on a subset of users and a subset of pages. Moreover, two biclusters may overlap, which means that several users or pages of the session matrix may participate in multiple biclusters. Another important characteristic of biclusters is that each bicluster should not be fully contained in another determined bicluster. Overlapping is allowed in order not to miss important biclusters.

4.3 User profiling

The first step in intelligent web personalisation is the automatic identification of user profiles. This constitutes the knowledge discovery engine. The discovered user profiles are used to recommend relevant URLs to old and new anonymous users of a website (Nasraoui and Petenes, 2003).

User profiling is the process of collecting information about the characteristics, preferences, and activities of web communities. An efficient and effective algorithm for web recommendations is the user profiling approach, which is on a basis of collaborative filtering techniques, a kind of commonly used algorithms in recommender systems.

This can be accomplished either explicitly or implicitly. Explicit collection of user profile data is performed through the use of online registration forms, questionnaires, and the like. The methods that are applied for implicit collection of user profile data vary from the use of cookies or similar technologies to the analysis of the users' navigational behaviour that can be performed using web log mining techniques (Liu and Wu, 2004). Mobasher et al. (2002) have proposed a potentially effective method profile aggregations based on clustering transactions (PACT) to generate aggregate profiles based on the centroids of each transaction cluster. However, the centroid of each cluster may represent the different groups of pages without much correlation. Hence, in this paper, a robust fuzzy biclustering approach is proposed to generate profiles which reveal the implicit relationship that exists between the pages and users. Discovery of aggregate profiles based on biclustering had been already proposed in our previous work (Hannah Inbarani and Thangavel, 2011). The procedure for obtaining user profiles is described in Figure 3.

Figure 3 Procedure for obtaining user profiles

```
Algorithm : Building user profile based on rough biclustering
[Input]:   A set and number of users and pages in lower and upper
           approximation biclusters, rough page membership values $\mu_{pij}$
           and Threshold $\alpha$, nb – number of biclusters
[Output]:  Set of user profiles UP = {$UP_j$}   j = 1, 2, …, nb
Step 1: Apply Rough biclustering algorithm for clustering users and pages
        simultaneously
Step 2: For j = 1 to nb
          Compute weight of each page in the bicluster using
              $|B_j| = |W_l * \underline{BU_i}| + W_u * |\underline{BU_i} - \overline{BU_i}|$
              WT ($P_i$, $B_j$) = |$B_j$| * $\mu_{pij}$
              where |$B_j$| is the number of users in bicluster j and $\mu_{pij}$
              is the membership of page $P_i$ in bicluster j
          End for
Step 3 : For j = 1 to nb
           $UP_j$ = { }
           Normalize page weights
           If the page weight value WT ($P_i$, $B_j$) is greater than threshold $\alpha$
             $UP_j$ = {$UP_j$} U {$P_i$}
         End for
Step 4: For j = 1 to nb
              Sort the pageviews of bicluster $UP_j$ with their weights in descending
              order and output the user profile
        End for
```

4.4 Recommendation process

Traditional recommendation process requires explicit user participation for providing his/her interest to the pages. However, despite their success, the explicit ratings may suffer from some limitations, such as additional user effort, user behaviour alteration and data sparsity. To overcome such problems, several researches (Claypool et al., 2001) have investigated the use of implicit interest indicators.

An important implicit indicator of the user's navigation path is the time spent or number of hits on different pages. In this paper, the time spent on different pages is used as an implicit indicator of page rating since implicit rating can be used for the analysis of any website for CB approaches. The complete recommendation process is described in Figure 4.

Figure 4 Recommendation engine for RB

```
Algorithm: Rough biclustering based recommendation engine:
Input: A set of NB profiles and a current session S
Output: Recommendation vector R.
Step 1: Apply rough biclustering approach to determine pages in
        Lower approximation and boundary region.
Step 2: Compute the similarity between user's sub session vector and
        the user profiles generated using rough biclustering using
        Similarity measure.
Step 3: Sort each row of similarity matrix in descending order based
        on weights.
Step 4: Recommend top K biclusters with high similarity with
        current set of visited pages and N pages sorted in
        descending order of weights not including URLs in the
        current session.
```

In order to provide recommendations, the biclusters containing users with preferences that have strong partial similarity with the test user have to be generated. This stage is executed online and consists of two basic operations:

- The formation of test users' neighbourhood, i.e., to find the *K* nearest biclusters.
- The generation of the top-N recommendation list of pages.

5 RB – proposed approach

5.1 Rough set preliminaries

Rough sets theory is a new mathematical tool to handle uncertainty and incomplete information. Polish Mathematician Pawlak.Z initially proposed rough sets in 1982. It obtained broad attentions after its successful use in knowledge discovery. Now, Rough sets have been applied to broad domains such as artificial intelligence, knowledge discovery in databases, pattern recognition, and failure detection. Rough sets can be approached by two accurate sets, the lower and upper approximation (Pawlak, 1995).

Lingras (Peters, 2006) defines the following properties for *rough clustering*.

Property 1 A data object can be a member of one lower approximation at most.

Property 2 A data object that is a member of the lower approximation of a cluster is also a member of the upper approximation of the same cluster.

Property 3 A data object that does not belong to any lower approximation is a member of at least two upper approximations belonging to the same.

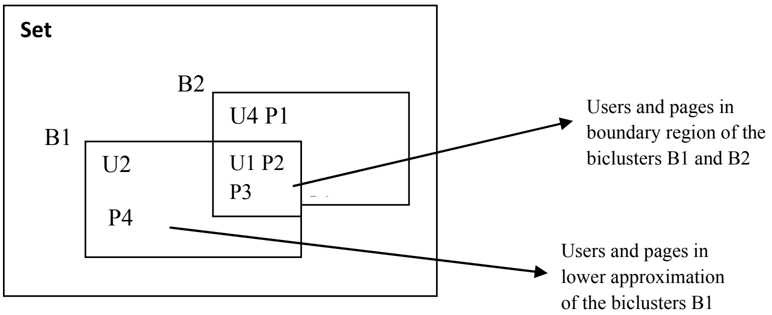
5.2 RB – proposed approach

In the framework of generalised rough sets, each bicluster is viewed as a generalised rough set which has two approximations, a lower and an upper approximation. The lower approximation is a subset of the upper approximation. The members (rows or columns) of the lower approximation belong to only one bicluster. However, the members of the upper approximation may belong to the bicluster as well as other biclusters. Therefore, the boundary region between the lower and upper approximation forms an overlapping part among corresponding biclusters. Given a data matrix *D*, for each object, there are possible three kinds of bicluster membership. They are as follows:

- It may not belong to any biclusters in *D*.
- Otherwise, if the object belongs to a bicluster, it is the key step which determines whether an object belongs to the lower approximation of the bicluster or upper approximation of the bicluster in *D*.

A two way RB method based on rough K-means is discussed in Wang et al. (2007). This algorithm generates overlapping biclusters and the degree of overlapping is determined using overlapping factor. But the two-way biclustering generates overlapping clusters of users and pages rather than overlapping biclusters. The two-way clustering methods apply standard clustering methods on the row and column dimension of data matrix separately and combine the results to obtain biclusters. The combined results exhibit similarity on either row dimension or column dimension; however, they may not well capture the overall coherence of both a subset of rows and a subset of columns.

Figure 5 Users and pages in lower and boundary region of biclusters


Set
B2
U4 P1
B1
U2
U1 P2
P3
P4
Users and pages in boundary region of the biclusters B1 and B2
Users and pages in lower approximation of the biclusters B1

The set of properties defined for rough clustering can also be extended for RB also. The biclustering, in data mining, is referred to the process of simultaneously finding clusters on the rows and columns of a matrix.

Property 1 A row data object or a column data object can be a member of only one lower approximation of the bicluster at most.

Property 2 A row data object or a column data object that is a member of the lower approximation of a bicluster is also member of the upper approximation of the same bicluster.

Property 3 A row data object or a column data object that does not belong to any lower approximation bicluster is member of at least two upper approximation biclusters.

Figure 6 RB approach

```
Algorithm: Rough biclustering algorithm
Input:   User access matrix (m, n)
             m – Number of users
             n – Number of pages
             minp – Minimum number of pages allowed in a bicluster
            Wu  – Weight for upper approximation
            $W_l$ – Weight for lower approximation

Output: NB biclusters
             Users and pages in lower approximation
             Users and pages in boundary region
             Identify distinct patterns of Ses and store it in matrix P
             /* P – set of distinct patterns
            /* nb is the number of distinct patterns/biclusters in Ses

Step 1: Extract all the nb distinct patterns using Hadamard product.
Step 2: Insert all the pages in the extracted Pattern  $l$ in Bicluster
          $BPi, \quad l = 1, 2, ..., nb$
Step 3: Compute Rough page membership $\mu_{ij}$ of each page in the bicluster as defined in (1.1)
Step 4: Insert pages in Lower or upper approximation of the bicluster as defined in (1.1)
        according to the conditions specified in (2)-(4)
Step 5:   for  I  =   1 to m
          for  j =    1 to nb
            Compute the Rough user similarity between user access vector i
            and pattern j using definition (1.3)
          end
          end
Step 6: Compute rough user membership as defined in (1.3)
Step 7: Insert users in Lower or upper approximation of the bicluster based on rough
        membership
         value as defined in (1.2)
Step 8: For each bicluster
            Output users in lower approximation  $\underline{BU_i}$
            Output pages in lower approximation  $\underline{BP_i}$
            Output users in boundary region if $\left(\underline{BU_i} - \overline{BU_i}\right) \neq Ǿ$
            Output pages in boundary region if $\left(\underline{BP_i} - \overline{BP_i}\right) \neq Ǿ$
```

A vague concept description can contain boundary-line objects from a universe, which cannot be with absolute certainty classified as satisfying the description of a concept. Uncertainty is related to the idea of membership of an element to a set. From rough sets perspective, a set membership function can be defined, which is related to the rough sets concept. This can be considered as another numerical measure of imprecision (uncertainty). The rough membership function of an object *x* to a set *X* is defined by (1).

Definition 1: Rough membership

The measure characterising a degree of uncertainty of membership of an element x in universe to the set *X* with respect to the possessed knowledge (in an information system) is defined by (1).

$$\mu_x(x) = \left(Cardinality\left([x]_B \cap X\right)\right) \ \left(Cardinality\left([X]_B\right)\right. \tag{1}$$

It is possible in rough sets to find a strict connection between vagueness and uncertainty. Vagueness is related to sets of objects (concepts), whereas uncertainty is related to elements of sets. Rough sets show that vagueness is defined in terms of uncertainty.

The rough set membership function can be used to define the lower and upper approximation of a set and the boundary region as in (3.19)

$$\underline{BU}_i = \{X \in U : \mu_x B(x) = 1\} \tag{2}$$

$$\overline{BU}_i = \{X \in U : \mu_x B(x) > 0\} \tag{3}$$

$$\underline{BU}i - \overline{BU}_i = \{X \in U : 0 < \mu_x B(x) < 1\} \tag{4}$$

Definition 2: Hadamard product:

Hadamard product (named after French mathematician Jacques Hadamard, also known as the entry wise product. Note also that both *A* and *B* need to be the same size, but not necessarily square.

Formally, for two matrices of the same dimensions:

$$A, B \in \mathbb{R}^{m\times n}$$

The Hadamard product $A \cdot B$ is a matrix of the same dimensions

$$A \circ B \in \mathbb{R}^{m\times n}$$

with elements given by

$$(A \circ B)_{i,j} = (A)_{i,j} \cdot (B)_{i,j} \tag{5}$$

Definition 3:

Given two biclusters *A* and *B*, the Rough similarity or overlapping degree *R* of the two biclusters is defined as $R = | A \cap B | / | A \cup B|$, where $| A \cap B |$ is the volume of $A \cap B$ and $| A \cap B |$ is the volume of $A \cup B$.

5.2.1 Pattern extraction

A pattern v can be extracted by the Hadamard product (as in Definition 2) of each row [considered as a user access vector (1xp)] with other rows of user access matrix denoted by $S_i \cdot S_j$ where $S_i = \{S_{i1}, S_{i2}, \ldots, S_{in}\}$and $S_j = \{S_{j1}, S_{j2}, \ldots, S_{jn}\}$.

Example 2.2: Consider the user access matrix S where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ and $P = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$. Suppose that the matrix that represents S is.

Four biclusters extracted from the user access matrix defined in Table 2 are shown in Table 3.

Table 2 User access matrix

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	0	0	0	0	0	1	1
u_2	0	0	0	0	0	1	1
u_3	0	0	0	1	1	0	0
u_4	0	0	0	1	1	0	0
u_5	0	0	1	1	1	0	0
u_6	1	1	1	1	0	0	0
U_7	1	1	1	0	0	0	0

Table 3 Biclusters obtained from user access matrix

Bicluster	*Users*	*Pages*
B1	BU1 = $\{u_1, u_2\}$	BP1 = $\{p_6, p_7\}$
B2	BU2 = $\{u_3, u_4, u_5\}$	BP2 = $\{p_4, p_5\}$
B3	BU3 = $\{u_5, u_6\}$	BP3 = $\{p_3, p_4\}$
B4	BU4 = $\{u_6, u_7\}$	BP4 = $\{p_1, p_2, p_3\}$

Table 4 Rough biclusters

Bicluster	*Users in lower approximation*	*Users in upper approximation*	*Pages in lower approximation*
B1	$\underline{BU}1 = \{u_1, u_2\}$	$\overline{BU}1 = \{u_1, u_2\}$	$\underline{BP}1 = \{p_6, p_7\}$
B3	$\underline{BU}2 = \{u_3, u_4\}$	$\overline{BU}2 = \{u_3, u_4, u_5\}$	$\underline{BP}2 = \{p_5\}$
B3	$\underline{BU3} = \{u_5, u_6\}$	$\overline{BU3} = \{u_5, u_6\}$	$\underline{BP3} = \{\}$
B4	$\underline{BU}4 = \{u_7\}$	$\overline{BU4} = \{u_6, u_7\}$	$\underline{BP}4 = \{p_1, p_2\}$

	Pages in upper approximation	*Pages in the boundary region*
B1	$\overline{BP}1 = \{\}$	$\overline{BP}1 - \underline{BP}1 = \{\}$
B3	$\overline{BP}2 = \{p_4, p_5\}$	$\overline{BP}2 = \underline{BP}2 = \{p_4\}$
B3	$\overline{BP3} = \{p_3, p_4\}$	$\overline{BP3} - \underline{BP3} = \{p_3, p_4\}$
B4	$\overline{BP3} = \{p_1, p_2, p_3\}$	$\overline{BP}4 - \underline{BP}4 = \{p_3\}$

Four rough biclusters formed from the user access matrix in Table 3 are shown in Table 5.

5.2.2 Patterns

The various patterns extracted by Hadamard product are:

Table 5 Patterns

Patterns	*Pages*
P1	$\{p_6, p_7\}$
P2	$\{p_4, p_5\}$
P3	$\{p_3, p_4\}$
P4	$\{p_1, p_2, p_3\}$

5.2.3 Rough page membership

Rough page membership of each page is computed as defined in equation (1).

For calculating the rough membership of p_3 in BP3, for $B = \{p_3\}$, i.e., for those patterns in which p_6 is visited,

$$[x]_B = \{\{p_3, p_4\}\{p_1, p_2, p_3\}\}$$
$$X = \{p_3, p_4\}$$
$$\mu_x(x) = 1/2 = 0.5$$

For calculating the rough membership of p_3 in BP4, for $B = \{p_3\}$ i.e., for those patterns in which p_6 is visited,

$$[x]_B = \{\{p_1, p_2, p_3\}\{p_3, p_4\}\}$$
$$X = \{p_1, p_2, p_3\}$$
$$\mu_x(x) = 1/3 = 0.33$$

For calculating the rough membership of p_6 in BP1, for $B = \{p_6\}$ i.e., for those patterns in which p_6 is visited,

$$[x]_B = \{p_6, p_7\}$$
$$X = \{p_6, p_7\}$$
$$\mu_x(x) = 2/2 = 1$$

5.2.4 Rough user membership

User membership is computed using rough similarity (as defined in 1.3) with the extracted patterns.

Membership of u_1 in $B1$ is $= \left|\{p_6, p_7\}\right| / \left|\{p_6, p_7\}\right| = 2/2 = 1$

6 Experimental evaluation

6.1 Experimental setup

6.1.1 Data source

Dataset 1

The web access log from http://www.techcmantix.com is used for the experiments. After preprocessing and removing references by image files and style sheet files, a total of 2,599 (MAXSIZE) transactions were produced using the transaction identification process. The total number of URLs representing page views was 362 and after eliminating the image files, style sheet files, the total number of remaining page view URLs in the training and the evaluation sets is 113. Approximately 25% of these transactions were randomly selected as the testing set, and the remaining portion was used as the training set for page recommendation.

6.1.2 Evaluation metrics

The performance of each method is measured using three different standard measures, namely, precision, coverage, and the F1-measure as defined in Mobasher et al. (2002). These measures are adaptations of the standard measures, precision and recall, often used in information retrieval. In this context, precision measures the degree to which the recommendation engine produces accurate recommendations. On the other hand, coverage measures the ability of the recommendation engine to produce all the page views that are likely to be visited by the user. The precision measure represents the ratio of matches between the recommendation set and the target set to the size of recommendation set. The coverage measure represents the ratio of matches to the size of the target set.

6.2 Recommendation results of RB

The recommendation engine takes a collection of user profiles as input and generates a recommendation set by matching the current user's activity against the discovered patterns. A fixed-size sliding window over the current active session is used to capture the current user's history depth. Thus, the sliding window of size *nn* over the active session allows only the last *mm* visited pages to influence the recommendation value of items in the recommendation set. This sliding window is called as active session window.

In each iteration, each transaction *t* in the evaluation set was divided into two parts. The first *n* page views were used for generating recommendations, whereas, the remaining portion of *t* (target set) was used to evaluate the generated recommendations. For the recommendation process, a session window size of 2 was chosen. The recommendation results are given in Table 6 for the sample path.

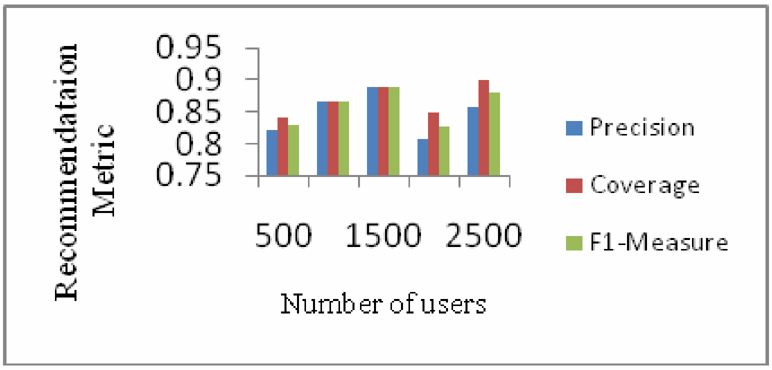
Table 6 Recommended pages

Input pages	*Recommended pages in the lower approximation*	*Recommended pages in the boundary region*
/iphonesimulator.php	/sms-software-unlimited-sms-from-pc-any-mobile.php	/infra.html
/downloads.php	/software-development-company.php	
	/contact.php	
/messengerplus.php	/downloads.php	/sms-software-unlimited-sms-from-pc-any-mobile.php
/downloadmsg.php	/portfolio.php	
	/contact.php	
	/website-design-services.php	/software-development-company.php
	/billing-automation-with-accounts.php	
	/school-management-system.php	

6.3 Performance analysis for RB

The required input of the algorithm is minimum number of pages to be included in the bicluster. In order to discover the best biclusters, it is important to fine-tune this input variable. Figure 7 illustrates the values of F1 measure, precision and coverage for varying sizes of datasets.

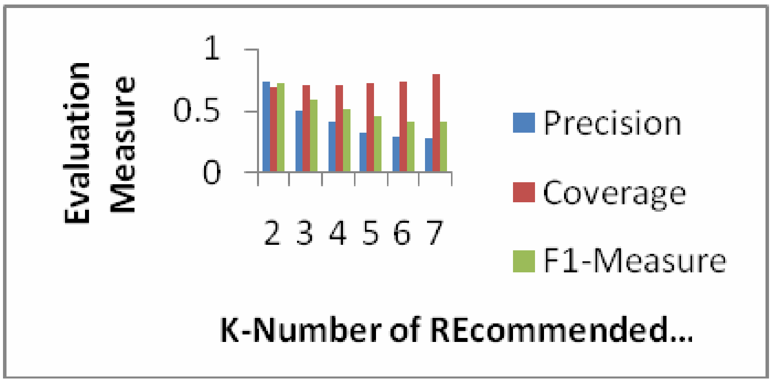
Figure 7 Precision, coverage and F1-measure for various sizes of datasets (see online version for colours)



6.3.1 Impact of Number of biclusters recommended

The impact of number of biclusters recommended for MAXSIZE data is illustrated in Figure 8. Precision becomes low when the number of biclusters K recommended is increased. But coverage becomes high when the number of biclusters recommended is decreased. F1-measure attains its maximum value at $K = 2$. Hence, optimum performance is obtained at $K = 2$.

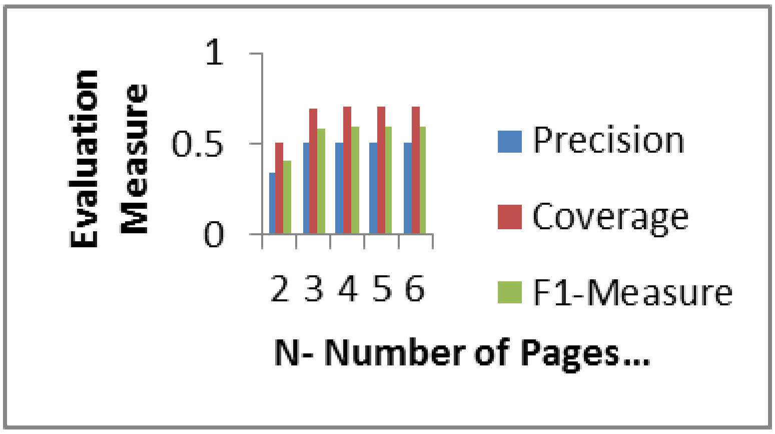
Figure 8 Precision, coverage and F1-measure for various values of *K* (see online version for colours)



6.3.2 *Impact of recommendation list's size*

The impact of recommendation list size *N* for MAXSIZE data is shown in Figure 9. As expected, with increasing *N*, coverage, precision and F1-measure increases and stable value is reached for these measures at *N* = 4. In real applications, *N* should be kept low, because it is impractical for a user to see all recommendations when their number is large. Hence, the optimum value for *N* is taken as 4.

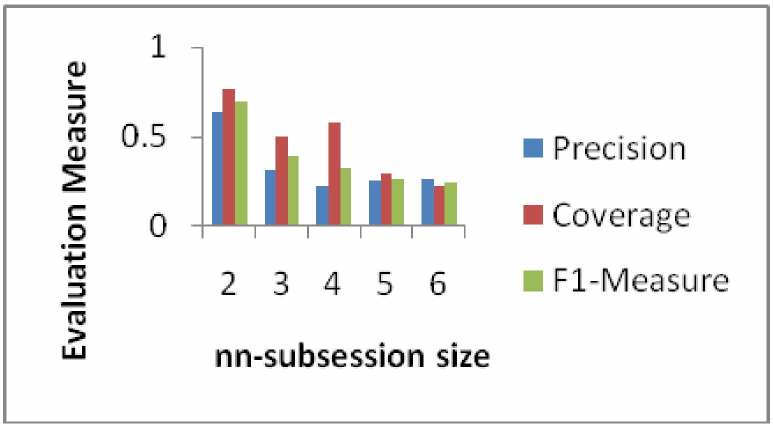
Figure 9 Precision, coverage and F1-measure for various values of *N* (see online version for colours)



6.3.3 *Impact of subsession size nn*

The impact of active session size for MAXSIZE data is illustrated in Figure 10. The evaluation measures show optimum performance at *nn* = 2. When the value of active session size is increased, the evaluation measures exhibit poor performance. Hence, the optimum value for *nn* = 2.

Figure 10 Precision, coverage and F1-measure for various values of subsession size *nn* (see online version for colours)



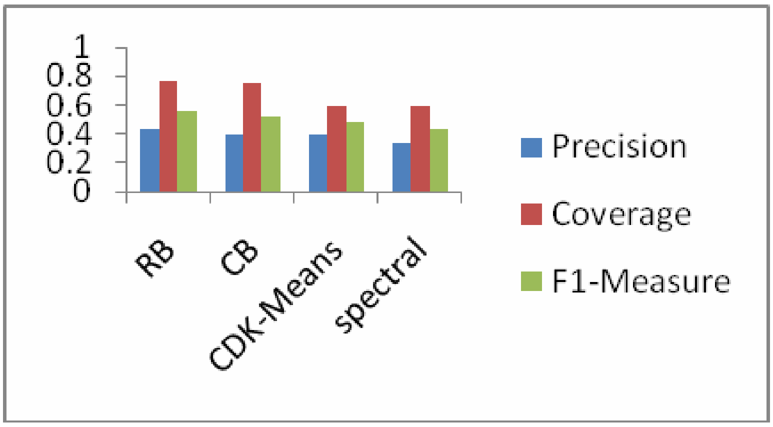
6.3.4 Parameter setting

The minimum number of pages and users in a bicluster is set to 2. For CB, the implicit rating obtained from the hits of the users in different pages are used as weights and the weight of each page in the bicluster is determined as per the user profiling algorithm discussed in Section 2. Unless otherwise specified, the default values for the parameters are $K = 4$, $N = 4$, $nn = 2$. These optimum values are selected so that the value of F1-measure is maximised as depicted in Figures 8, 9 and 10.

6.4 Comparative results for effectiveness

In this section, the performance of RB, CB, CDK-means and spectral co-clustering are compared for page recommendation. The performance comparison of RB, CB, CDK-means and spectral co-clustering using F1-measure, precision and coverage for the maximum dataset size is shown in the Figure 11.

Figure 11 Precision, coverage and F1-measure for various approaches (see online version for colours)



Spectral co-clustering exhibits poor performance in terms of precision but shows improved performance than CDK-means in terms of coverage. RB outperforms all the other approaches both in terms of precision and coverage since RB captures overlapping between user access patterns and only the optimised weights are used for recommendation of web pages.

Dataset 2

The second dataset is the web access log of www.bbminfo.com. This dataset is referred to as *dataset 2*. After preprocessing and removing references by web spiders, image files and style sheet files, a total of 13,612 transactions were produced using the transaction identification process. The total number of URLs representing page views was 145 and the image files and style sheet files were eliminated and the total number of remaining page view URLs in the training and the evaluation sets is 83. Approximately 25% of these transactions were randomly selected as the testing set, and the remaining portion was used as the training set for page recommendation. The total number of remaining page view URLs in the training and the evaluation sets is 83.

6.5 Recommendation results of RB

The goal of the recommendation process is to generate dynamic recommendations given the user's access pattern so far. The recommendations are in the form of a set of URLs and their corresponding recommendation scores are given by weights. Each actual completed session, *S*, is treated as ground-truth; a subset of this session is treated as incomplete current sub session, *nn*, and the Biclustering based recommendations are treated as predicted complete session Si. The recommendation result of RB is shown in Table 7.

Table 7 Recommended pages

Input pages	*Recommended pages in the lower approximation*	*Recommended pages in the boundary region*
/services/services.html	organisation/membersreport.html	/travels/index.html
Organisation/convener.htm	/hr/index.html	/checkbrowser.html
/organisation.htm	/bbm/aboutbbm.html	/hr/index.html
	/hr/employment.html	/newsletter.php
	/bbm/contact.html	/support.html
	/bbm/index.html	
	/bbmsoftsolutions/services.html	
/bbm/aboutbbm.html	/hr/candireview.htm	/support.html
/company.html	/hr/makingCV.html	/contact.php
/hr/employment.html	/checkbrowser.html	/bbmsoftsolutions/ bussiness.htm
	/hr/index.html	
	/services/services.html	
	/newsletter.php	

6.6 *Performance analysis of RB based recommendation*

Figure 12 illustrates the values of F1 measure, precision and coverage for varying sizes of dataset 2.

6.6.1 *Impact of number of biclusters recommended*

The impact of number of biclusters recommended for MAXSIZE data is illustrated in Figure 13. Precision becomes low when the number of biclusters K recommended is increased. But coverage becomes high when the number of biclusters recommended is decreased. F1-measure attains its maximum value at $K = 6$. Hence, best performance is obtained at $K = 2$.

Figure 12 Precision, coverage and F1-measure for various sizes of datasets (see online version for colours)

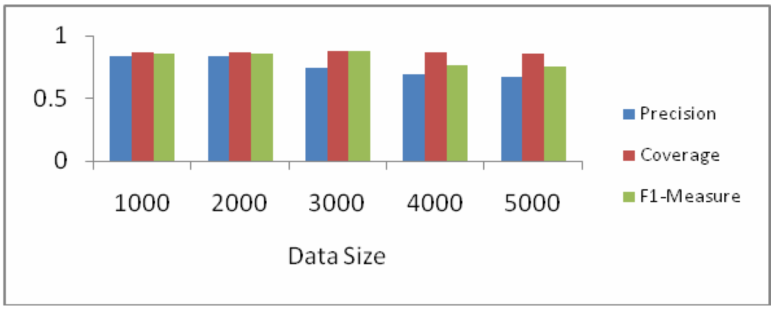
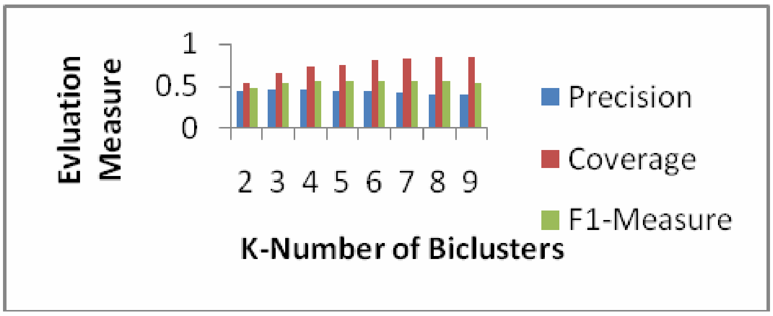


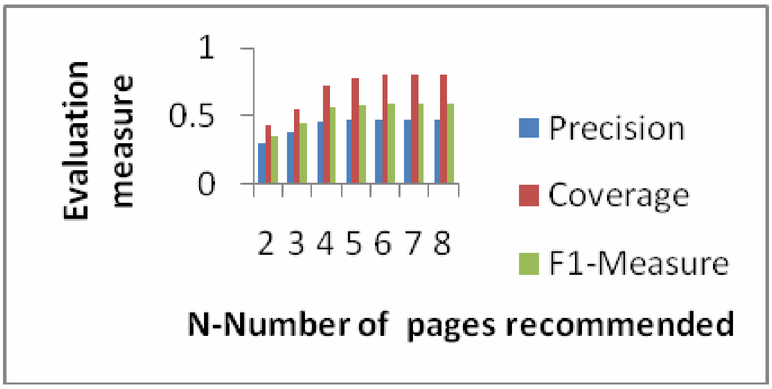
Figure 13 Evaluation measures for various values of K (see online version for colours)



6.6.2 *Impact of recommendation list's size*

The impact of recommendation list size N for MAXSIZE data is shown in Figure 14. As expected, with increasing N, coverage, precision and F1-measure increases and stable value is reached for these measures at $N = 6$. In real applications, N should be kept low, because it is impractical for a user to see all recommendations when their number is large. Hence, the optimum value for N is taken as 6.

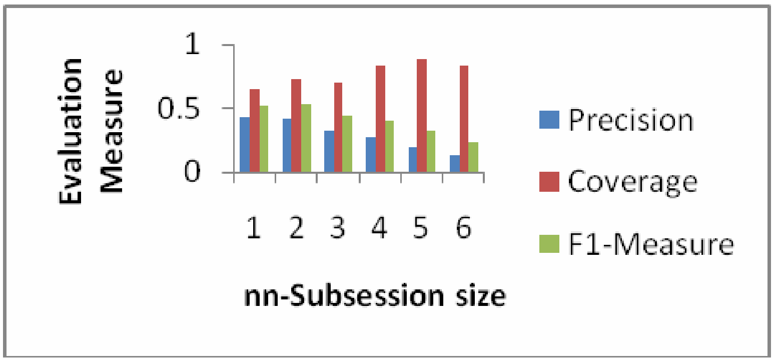
Figure 14 Evaluation measures for various values of *N* (see online version for colours)



6.6.3 Impact of subsession size nn

The impact of active session size for MAXSIZE data is illustrated in Figure 15. The evaluation measures show optimum performance at *nn* = 2. When the value of active session size is increased, F1-measure exhibits poor performance. Hence, the optimum value for *nn* = 2.

Figure 15 Evaluation measures for various values of *nn* (see online version for colours)



6.7 Parameter setting

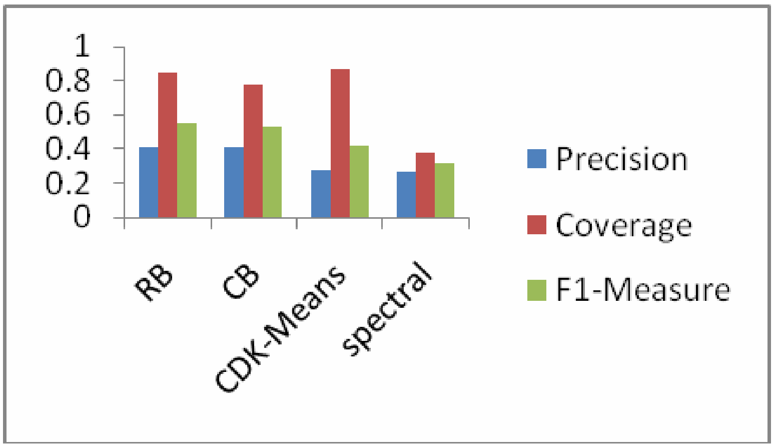
The minimum number of pages and users in a bicluster is set to 2. Unless otherwise specified, the default values for the parameters are *K* = 6, *N* = 6, *nn* = 2 for dataset 2. These optimum values are selected after several runs based on sensitivity analysis for the best performance in terms of F1-measure.

6.8 Comparative analysis

In this section, the performance of RB is compared with CB, CDK-means and spectral co-clustering for page recommendation.

Figure 16 illustrates the values of F1-measure, precision and coverage for varying sizes of dataset 2. The RB algorithm results in high precision, coverage and F1-measure as overlapping between biclusters is captured by lower approximation and upper approximation biclusters. The performance comparison of CB, CDK-means and spectral co clustering using F1-measure, precision and coverage for the maximum dataset size is shown in the Figure 16.

Figure 16 Precision, coverage and F1-measure for various approaches (see online version for colours)



It can be observed from the figure that the performance of RB algorithms is better than the performance of CB, CDK-means and spectral co-clustering. It is obvious that biclustering algorithms outperform user and page clustering algorithms.

7 Conclusions

The target of personalisation based on web usage data is to compute a recommendation set for the current user session based on user's past navigation patterns. In this paper, a new personalised recommendation method based on rough sets and biclustering is proposed to improve the web-personalised recommendation. It captures overlapping biclusters and the users and pages are placed in lower and upper approximation. We performed experimental comparison of the RB algorithm with biclustering algorithms CB, CDK-means and spectral co-clustering. Our extensive experimental results illustrate the effectiveness of the proposed RB algorithm.

Finally, it is concluded that RB based recommendations are very intuitive, deal with natural overlap in user interests. This makes rough recommendations suitable for real-time recommendations of huge websites.

The following observations are highlighted from the experiments carried out for this study.

- Traditional clustering approaches cannot take different preferences of users as the user is assigned to exactly one cluster.

- Biclustering approach identifies the co-occurrence patterns of pages within usage data for the Site.
- The biclusters may *overlap*, which means that several users or pages of the session matrix may participate in multiple biclusters. The degree of overlapping is not taken into account for each user and for each page in CB algorithms. Hence, the RB approach finds the overlapping degree and places users and pages lower and upper approximation accordingly.

References

Cheng, Y. and Church, G.M. (2000) 'Biclustering of expression data', in *Proceedings. of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp.93–103.

Chi, C-C., Kuo, C-H., Lu, M-Y. and Tsao, N-L. (2008) 'Concept-based pages recommendation by using cluster algorithm', in *Proceedings of Eighth IEEE International Conference on Advanced Learning Technologies*, Santander, Spain, 1–5 July, pp.298–300.

Claypool, M., Le, P., Wased, M. and Brown, D. (2001) 'Implicit interest indicators', in *Proceedings of Sixth International Conference on Intelligent User Interfaces*, ACM Press, pp.33–40.

de Castro, P.A.D., de França, F.O., Ferreira, H.M. and Von Zuben, F.J. (2007) 'Applying biclustering to text mining: an immune-inspired approach', in *ICARIS*, Springer, Vol. 4628, pp.83–94.

Dhillon, I.S. (2001) 'Co-clustering documents and words using bipartite spectral graph partitioning', in *Proceedings of the ACM SIGKDD Conference*.

Hammouda, K.M and Kamel, M.S. (2004) 'Efficient phrase-based document indexing for web document clustering', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 10, pp.1279–1296.

Hannah Inbarani, H. and Thangavel, K. (2009) 'Rough set based user profiling for web personalization', *International Journal of Computer Science*, Academy Publishers, Vol. 2, No. 1, pp.103–107.

Hannah Inbarani, H. and Thangavel, K. (2011) 'A robust biclustering approach for effective web personalization', *Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications*.

Hartigan, J.A. (1972) 'Direct clustering of a data matrix', *Journal of the American Statistical Association*, Vol. 67, No. 337, pp.123–129.

Hui, Z., Bin, P., Ke, X. and Hui, W. (2006) 'An efficient algorithm for clustering search engine results', *International Conference on Computational Intelligence and Security*, November, Guangzhou, China, pp.1429–1434.

Joshi, A. and Krishnapuram, R. (1998) 'Robust fuzzy clustering methods to support web mining', *Proc. Workshop in Data Mining and knowledge Discovery, SIGMOD*, pp.1–15.

Koutsonikola, V.A. and Vakali, A.I. (2009) 'A fuzzy bi-clustering approach to correlate web users and pages', *Int. J. Knowledge and Web Intelligence*, Vol. 1, p.2.

Li, H.Y., Xie, C.S. and Liu, Y. (2007) 'A new method of prefetching I/O requests', in *Proceedings of International Conference on Networking, Architecture and Storage*, Guilin, China, July, pp.217–224.

Liu, J-G. and Wu, W-P. (2004) 'Web usage mining for electronic business applications', in *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, August, pp.57–63.

Liu, X., He, P. and Yang, Q. (2005) 'Mining user access patterns based on web logs', in *Proceedings of Canadian Conference on Electrical and Computer Engineering*, May, Saskatoon Inn Saskatoon, Saskatchewan, Canada, pp.2280–2283.

Long, B. and Zhang, Z. and Yu, P.S. (2005) 'Co-clustering by block value decomposition', in *Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM Press, New York, pp.635–640.

Mirkin, B. (1996) *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Dordrecht.

Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2002) 'Discovery and evaluation of aggregate usage profiles for web personalization', *Data Mining and Knowledge Discovery*, Vol. 6, pp.61–82.

Mobasher, B., Dai, H., Luo, T., Sun, Y. and Zhu, J. (2000) 'Integrating web usage and content mining for more effective personalization', in *Proceedings of the First International Conference on Electronic Commerce and Web Technologies*, pp.165–176.

Nakagawa, M. and Mobasher, B. (2003) 'A hybrid web personalization model based on site connectivity', in the *Fifth International WEBKDD Workshop: Web Mining as a Premise to Effective and Intelligent Web Applications*, pp.59–70.

Nasraoui, O. and Petenes, C. (2003) 'An intelligent web recommendation engine based on fuzzy approximate reasoning', in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp.1116–1121.

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R. (2008) 'A web usage mining framework for mining evolving user profiles in dynamic websites', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 2, pp.202–215.

Nizar Banu, P.K. and Hannah Inbarani, H. (2011) 'Analysis of click stream patterns using soft biclustering approaches', *IJITSA*, Vol. 4, No. 1, pp.53–66.

Pallis, G. and Koutsonikola, V. (2006) 'A.: insight and perspectives for content delivery networks', *Communications of the ACM*, Vol. 49, No. 1, pp.101–106.

Pawlak, Z. (1995) 'Vagueness and uncertainty: a rough set perspective', *Computational Intelligence*, Vol. 11, No. 2, pp.277–232.

Pensal, R.G., Robardet, C. and Boulicaut, J-F.C. (2005) 'A bi-clustering framework for categorical data', in Jorge, A. et al. (Eds.): *PKDD 2005, LNAI*, Springer Verlag, Berlin, Heidelberg, Vol. 3721, pp.643–650.

Peters, G. (2006) 'Some refinements of rough K-means clustering', *Pattern Recognition*, Vol. 39, pp.1481–1491.

Petridou, S., Koutsonikola, V., Vakali, A. and Papadimitriou, G. (2008) 'Time-aware web users clustering', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 5, No. 20, pp.653–667.

Srinivasa, N. and Medasani, S. (2004) 'Active fuzzy clustering for collaborative filtering', in *Proceedings of IEEE International Conference on Fuzzy Systems*, Budapest, Hungary, July, pp.1607–1702.

Symeonidis, P., Nanopoulos, A., Papadopoulos, A.N. and Manolopoulos, Y. (2008) 'Nearest-biclusters collaborative filtering based on constant and coherent values', *Information Retrieval*, Vol. 11, No. 1, pp.51–75.

Tanasa, D. and Trousse, B. (2004) 'Advanced data preprocessing for intersites web usage mining', *IEEE Intelligent Systems*, pp.59–65.

Tang, C., Zhang, L., Zhang, I. and Ramanathan, M. (2001) 'Interrelated two-way clustering: an unsupervised approach for gene expression data analysis', in *Proceedings of the 2nd IEEE Int. Symposium on Bioinformatics and Bioengineering*, pp.41–48.

Wang, R., Miao, D., Li, G. and Zhang, H. (2007) 'Rough overlapping biclustering of gene expression data', *Proceedings of IEEE Transactions on Bioinformatics and BioEngineering*, pp.828–834.

Xu, G. (2008) *Web Mining Techniques for Recommendation and Personalization*, PhD thesis.