Circuits and Systems, 2016, 7, 2663-2675
Published Online July 2016 in SciRes. http://www.scirp.org/journal/cs
http://dx.doi.org/10.4236/cs.2016.79230

Scientific Research Publishing

An Approach for Content Retrieval from Web Pages Using Clustering Techniques

R. Manjula, A. Chilambuchelvan

Department of CSE, R.M.K Engineering College, Chennai, India
Email: manjularesearch@gmail.com, chill97@gmail.com

Received 22 April 2016; accepted 15 May 2016; published 28 July 2016

Abstract

Mining the content from an information database provides challenging solutions to the industry experts and researchers, due to the overcrowded information in huge data. In web searching, the information retrieved is not an appropriate, because it gives ambiguous information for the user query, and the user cannot get relevant information within the stipulated time. To overcome these issues, we propose a new methodology for information retrieval EPCRR by providing the top most exact information to the user, by using the collaborative clustered automated filter which makes use of the collaborative data set and filter works on the prediction by providing the highest ranking for the exact data retrieved. The retrieval works on the basis of recommendation of data which consists of relevant data set with highest priority from the cluster of data which is on high usage. In this work, we make use of the automated wrapper which works similar to the meta crawler functionality and it obtains the content in the semantic usage data format. Obtained information from the user to the agent will be ranked based on the Enabled Pile clustered data with respect to the metadata information from the agent and end-user. The information is given to the end-user with the top most ranking data within the stipulated time and the remaining top information will be moved to the data repository for future use. The data collected will remain stable based on the user preference and works on the intelligence system approach in which the user can choose any information under any instances and can be provided with suitable high range of exact content. In this approach, we find that the proposed algorithm has produced better results than existing work and it costs less online computation time.

Keywords

Collaborative Filter, Automated Wrapper, Clustering, Information Retrieval, Data Repository

1. Introduction

Content Mining plays a vital role in the information retrieval to the user accordingly to the given query or re-

How to cite this paper: Manjula, R. and Chilambuchelvan, A. (2016) An Approach for Content Retrieval from Web Pages Using Clustering Techniques. *Circuits and Systems*, **7**, 2663-2675. http://dx.doi.org/10.4236/cs.2016.79230

quest. For the past decades, we have noticed a vast growth in the data access through the web [1]. Even though through the web we have got information that is not a relevant data to the user, since the information obtained may be overcrowded. That is information with lot of contents or the irrelevant information to the given query is mismatched to the query, so the user cannot get the relevant data for the request [2]. The information obtained may be overlooked or overload. A traditional information retrieval (IR) technique has provided solutions to the fundamental issues. IR-based systems are not describing explicitly about how the systems can act like users and it is not supporting to obtain knowledge from large data sets to answer what users really want [3]. In data mining, it is challenging task to know what is to be performed to get the relevant information with content mining [4].

For a short while, many mining methodologies have been proposed to give the solution on approximation with this challenge. Unfortunately, the user based systems and agent based systems can only show the architectural proposal for the information gathering and management [5]. They cannot provide the novel approaches to these challenging issues. Some of the mining methodology provides the exact content to the maximum but not within the stipulated time and it has been accumulated by the most adherent relevant information [6]. To overcome these issues exact content mining techniques has been proposed and named as EPCRR (Enabled Pile Clustered exact content retrieval and repository). This work is similar but advanced web content mining which can be viewed as the use of data mining techniques with the advancement towards the automatic data retrieval [7]. It facilitates the web mining procedure namely usage, structure, content and user profiles. In this facilitation, we add the content mining with the exact data information to the end-user by framing the appropriate data cluster set with the pile approach method and ranking the data with the hierarchy and maintain the time factor for to give the information with exact content and the top most overlooked obtained information will be stored in the data repository. Meta crawlers use meta data stored in the data repository to deliver mined content to the user [5].

Mining the data is a tedious process based on the user request since each user will look for different information. There are different kinds of user will expect the exact information from the mining process, for the given input the database may contain ocean of information and retrieval process will be tedious, since it will be in a confusion to give out the output irrespective of this issue it has to give exact information to the given input. So it ranks the top most information with overcrowded and overlooked data as an output. The traditional informational retrieval techniques will deliver the content without the verification and overwhelmed data [7] [8]. The content is not devised using the characteristics and usage. Overwhelmed information is given and timing factor is not considered. To address these issues, the proposed system has given the relevant process for the problem.

A native solution has been provided for the issues faced. In the proposed work we have used a Pile filter process and data set is set formed by using the meta data information, which is maintained by the agent. An automated wrapper will do the desired functionality for framing the data format for the given request and the advanced meta crawler will facilitate to provide the information desirably with the web data format within the time stipulation. The exact content and remaining hierarchy of data will be moved to the repository so that any change in the information can be updated and used for the future purpose [9] [10].

The rest of this paper is organized as follows. In Section 2, related work is discussed and in Section 3 briefly overview of the proposed work explained and proposed Architecture design was discussed. In Section 4, the data set formation and wrapper agent is explained. In Section 5, EPCRR (Enabled Pile Clustered exact content retrieval and repository) is explained. In Section 6, several theoretical and experimental research results are discussed. In Section 7, efficiency of the proposed system with Justification has been given. At last, some conclusions and future work discussed.

2. Related Work

Collaborative Filtering with clustering technique have been extensively studied by some researchers. Rong Hu *et al*. [11] proposed a concept which uses the description and functionality information as metadata to measure the characteristic similarities between services. It used AHC Algorithm in which the results depend strongly on the choice of number of clusters K, and initial value of K is not known. This system also suffers from cold start and data sparsity problem.

Mai *et al*. [12] designed a clustering collaborative filtering algorithm based on neural network in e-commerce recommendation system. With the data from web visiting message, the cluster analysis gathers users with similar characteristics. However, it was difficult to find the user's preference on web visiting is relevant to prefer-

ence on purchasing.

Mittal *et al*. [13] proposed a work to calculate the user predictions by first minimizing the size of the item set and it was explored by the user. K-means clustering algorithm was applied to partition movies based on the query requested by the user. However, it has a drawback that each object must belong to exactly one group which leads to the limitation that all group must have at least one member.

Li *et al*. [14] proposed a concept to incorporate multidimensional clustering into a collaborative filtering recommendation model. In first stage, the user and item profiles were collected and clustered using the proposed algorithm. Then the clusters with poor similarity features were removed and the appropriate clusters were selected based on cluster pruning. In third stage, an item was predicted by performing a weighted average of deviations from the neighbour's mean. This approach was increasing the diversity of recommendations while maintaining the accuracy of recommendations.

Zhou *et al*. [15] represented a data-providing service in terms of vectors and it considered the relation between given input, expected output, and semantic relations among them. Refined fuzzy C-means algorithm was used to cluster the vectors .Through merging similar services into a same cluster, the capability of search engine service was significantly improved, especially in large Internet-based service repositories. However, in this approach, it assumed that domain ontology exists for facilitating semantic interoperability. Besides, this approach is not suitable for some services which lack parameters.

Pham *et al*. [16] proposed a concept to use network clustering technique on social network of users to identify their neighbourhood, and then use the traditional CF algorithms to generate the recommendations.

Simon *et al*. [11] used a high-dimensional divisive hierarchical clustering algorithm and it requires feedback on past user history implicitly and to discover the relationships within the users. Products of high interest were recommended to the users based on the clustering results. However, the implicit feedback was not providing sure information about the user's preference.

3. Overview of the Proposed Work

We proposed a new semantic content mining process by forming the user query as a new data for the mining execution. The mining execution begins with the data set formation from the user information, the data set is formed by looking in to the identification of the data after the data is identified the refining process begins by forming the clusters of data for the given information to avoid the repetition of the content and also to extract the exact information content for the given query. once the clusters has been formed ,we are going to check for the similarity between each and every words and content the cluster can also emulates the similarities between two words and four words till to the end count of the given input. Then the clustered data will move on to the collaborative automated filtering process to obtain the hierarchy for the obtained data [17]. The clustering process begins by checking the data with the function similarity, characteristics similarity and description similarity and then the advanced Agglomerative Clustering algorithm is used for finding the similarities between words with respect to the measured factor, once the cluster of data is formed this will be act as wrappers to give out the output for the user query. It is going to give exact information data for the obtained process with advanced mining process with the automated collaborative filter which in turn will be acting wrapper to give out the specified output data which is a mined content information [10]. Then the obtained information will pass on to the web with the snippets then before giving out the information the agent will supervise the information and form the list for the data given along with the ranking the listed data will be moved to the data repository and from the repository the information will be given to the user based on their priority. If the user the user got the information from the mining process to their exact that information will act as a Meta data for the future retrieval of other data in the data repository, this repository will be maintained by the agent and not by the user.

Architecture Diagram proposed in **Figure 1** will give a layout for the exact mining process used to get the information. Each block used in the diagram shows the systematic relationship between the processes and shows the schematic representation of flow process between them. It is characterized by the working procedure of the system by showing how the schedule is carried from the tenure process. In the proposed process to cross check the hierarchy *i.e*. to provide the ranking after the filtering and refinement process we induce an agent to check the refined content is aligned according to the hierarchy. This is agent will do the process of third party expert system *i.e.*, it uses the rating techniques for to get the content from the collaborative filter used, while providing the ranking for the data it also consider the user preferences for the period of time to get delved [17]. To confirm

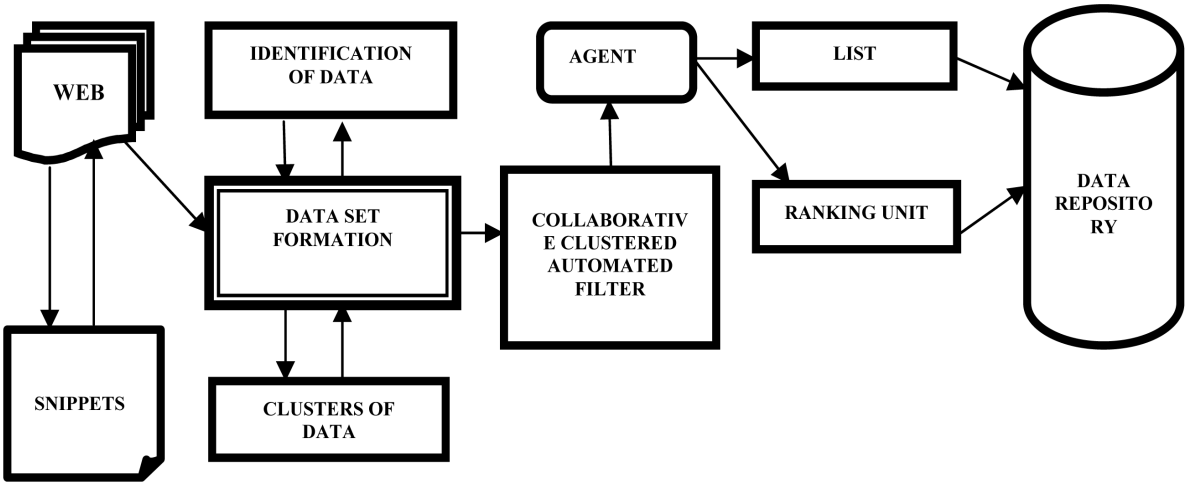

Figure 1. Architecture diagram of proposed system.

the process we can make use genetic algorithm to check match between the filtered contained from the filter and also check the search with the given user query.

In the proposed architecture design we have four blocks, first block will discuss about the data set formation block and second block will illustrate the filtering process and the third block will discuss about the monitoring and storage of data set and the final block will enumerate the details about the content retrieval. The proposed design will facilitate the content mining process in an efficient manner and also it deals about the how fast they can able to get the data within the stipulated period of time. This system also avoids the data overlapping and data mismatching of the information to the end user. The clustering algorithm used will give out the exact content retrieval. The comparative has been made with the popular search engines and the effective comparative study has been made and analysed. The following diagram **Figure 2** depicts the steps involved in the proposed work.

4. Data Set Formations

Many systems uses the recommender systems to form the dataset, the recommender system also uses the direct recommender algorithm to give out the data in the form of services [7]. Mostly the data set formed will not look in to the description value of the data instant it checks only for the migrant value of the data based on the usage of the data information. The data set recommended also will be maintained by user and not by the external affairs so that the data set recommended will give an enormous value accordingly to the different user for the same input of data the variation in the data and mismatch of the data content was formulated in the data set recommender system and also it will be avoided by the measures induced in the filtering phase used by the different collaborators [12] [14].

In our proposed work the data set is formed and it has been send to the wrapper and that wrapper will be maintained by the external agents so that the information lost can be avoided. The data set is formed by collecting all the information about the given data and the data is identified by the characteristics, similarity based on functionality and description, then after the comparison process the clustered data set is formed once the cluster of data is emulated the advanced collaborative clustered automated filter is used to filter out the unwanted content information [18] [19]. The data set is formed by giving out the unique id and value for each data unit, based on the characteristics and similarities and also it look for the user preferences in framing the clustered data unit which is a pile in our proposed system. when the uniqueness is found in the data set using the advanced collaborative filter the data set will be formulated the advanced collaborative filter works based on the aagglomerative Clustering Algorithm includes selecting the rating similarity computation and predicted rating computation, once the data set framed is given to the web the snippets along with the web will formulate the dataset accordingly to the desired format [8]. In this paper we make use of real data set based on recovery services in which each service has its own service description, functionality and user privileges in the form of ratings.

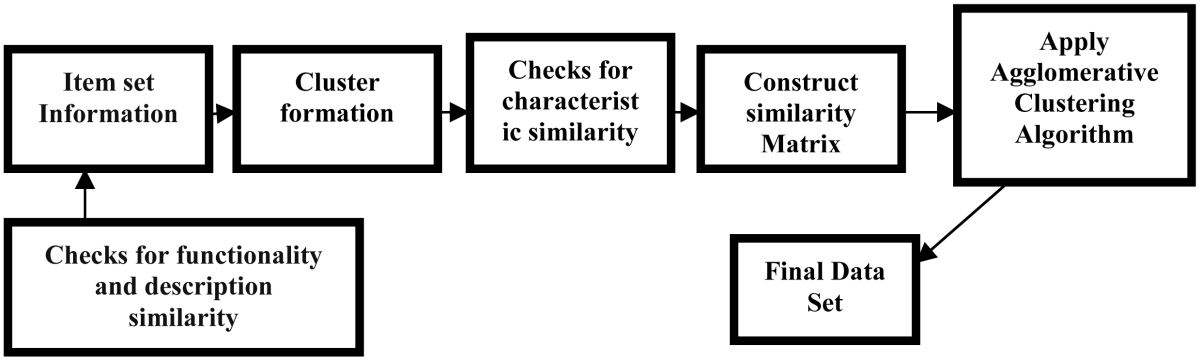

Figure 2. Diagram for similarity and functionality calculation.

Characteristic Similarity and Functionality Calculation

Description and functionality similarity are computed using Jaccard similarity coefficient (JSC) is the statistical measure for calculating similarity between samples sets [9]. For both sets, the JSC is defined as a cardinality of intersection is divided by the cardinality of their union. Concretely the formula for computing similarity between and b is [19],

$$D_{sim}(a,b) = \frac{|D_a \cap D_b|}{|D_a \cup D_b|} \tag{1}$$

This can be inferred from this formula that the larger $D_a \cap D_b$ is, the more similar the two services are. From the above Division $D_a \cup D_b$ is the scaling factor that ensures that description similarity is between 0 and 1. Similarly the Functionality similarity is calculated as given below.

$$F_{sim}(a,b) = \frac{|F_a \cap F_b|}{|F_a \cup F_b|} \tag{2}$$

The weighted sum of description similarity and functionality similarity is used to compute characteristic similarity between *a* and *b*.

$$C_{sim}(a,b) = \alpha \times D_{sim}(a,b) + \beta \times F_{sim}(a,b) \tag{3}$$

In this formula, $\alpha \in 0, 1$ is the description similarity weight and $\beta \in 0, 1$ is the weight of functionality similarity. The relative importance between these two expressed using weight. In the recommender system, for the total *n* services provided, calculate the characteristic similarities of every pair of services and $n \times n$ characteristic similarity matrix M is formed. An entry $m_{a,b}$ in M represents the characteristic similarity between *a* and *b*.

5. Enabled Pile Clustered Exact Content Retrieval and Repository

Clustering methods are partitioned the set of objects into clusters and a cluster contains more similar objects and dissimilar objects are in different clusters according to some defined criteria. In huge data store cluster analysis algorithms have been utilized [20].

Clustering algorithms is divided into either hierarchical or partition based. Some standard partition based approaches like K-means suffer from several limitations: 1) results are dependent on the cluster value since initially they don't know the value of K; 2) cluster size is not subjected to monitoring process; 3) algorithms converge to a local minimum [21]. Hierarchical clustering methods are classified in to two types based on bottom-up or top-down approach namely agglomerative or divisive clustering.

Pile Clustering Process

In this paper, we present a pile Clustering based Collaborative Filtering approach for big data applications and it is relevant to recommendation. Services are merged into some clusters via an Agglomerative Clustering algorithm before Collaborative Filtering technique is applied and, the rating similarities between services are com-

puted for single cluster. There is less number of services in a cluster than the whole system, this approach costs less online computation time. Moreover, as the relevant ratings of services are grouped in the same cluster and dissimilar are in other clusters. Predictions of the ratings services in the same cluster are more accurate than the dissimilar services in other clusters. This approach provides a better solution for data sparsity and cold start problem. The clustering of services are explained in **Figure 3** and the below algorithm.

Input: *1.An array of services s[n]*
2.A characteristic similarity matrix m[i][j]
Output: *A cluster_Set array with K clusters*

Algorithm:

```
for each i, Cluster[i]=s[i];
for each i,j , dc[i][j]=m[i][j];
for each i from 1 to k ,Cluster_Set[i]= Cluster [i];
max_matrix=maxVal(dc[i][j]);
while(max_matrix>=0.1)
do
        service_1=argmaxi(dc[i][j]);
        service_2=argmaxj(dc[i][j]);
        service_1=joinServiceCluster(Cluster [service_1], Cluster [service_2]);
        for each Cluster[temp] belong to S
                if ((Cluster[temp]!= Cluster [service_1] ) and (Cluster[temp]!= Cluster [service_2] ))
                        temp=Cluster[temp];
                        serv_1=Cluster [service_1];
                        serv_2=Cluster [service_2];
                        dc[serv_1][ temp]=avg(dc[serv_1][ temp], dc[serv_2][ temp]);
                end if
        end for
        S=S - {Cluster [service_2]};
        max_matrix=maxVal(dc[i][j]);
end while
```

Many current clustering systems use the agglomerative hierarchical clustering because of their clustering strategy, best performance. Furthermore, it does not require the number of clusters as input [22]. Therefore, we

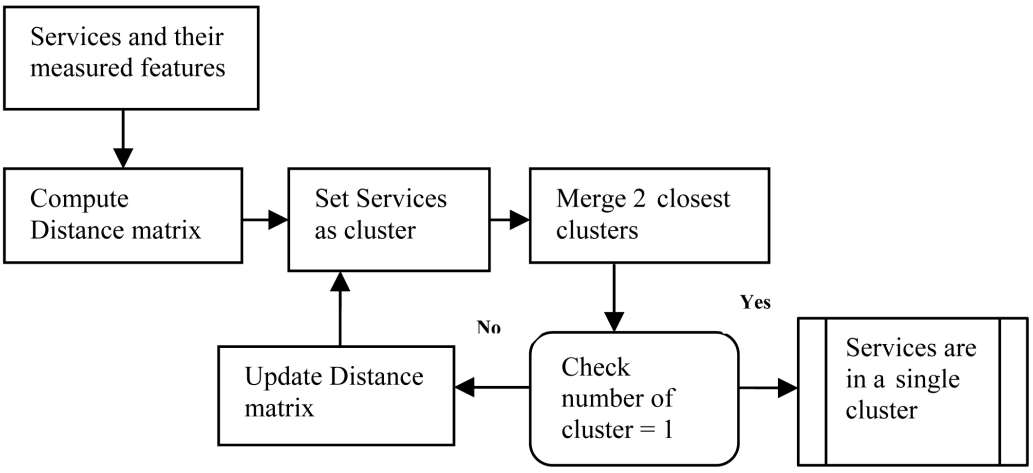

Figure 3. Diagram for clustering the services.

use an agglomerative clustering algorithm for service clustering as follows.

Once the pile clustered is formed it has been given to the repository for the content retrieval by the filtering process under the supervision of the wrapper agent. The agent will authenticate and retrieve the information on demand, once the content is given to the end user, and query information output values of the data set are stored in the repository, so that the user can able to get the filtered mined content with the stipulated period of time.

Future research can be done with respect to service similarity, such that semantic analysis may be performed on the description information of service. By this way, more semantically similar services may be clustered together, which will increase the overall coverage of recommendations. Research can also be done to mine the implicit interests of the user based on usage records and reviews. The semantic measure of obtaining the relevant query is to search through different search. The overall steps of the semantic meta search engine includes process like: 1) using semantic similarity measures relevant query are formed; 2) based on the relevant query web documents are extracted; 3) ranking of web documents. Here, input query and neighbours extracted from ontology is used to select the most suitable query and then, ranking of web pages obtained from the different search engine was done using QSPR measure. With different set of queries the experimentation was done and the results performance was analyzed with the help of precision, F-measure. From the experimental results, we found that the proposed Meta search engine has performed better than existing work by achieving the precision of 0.8. Finally an expert system is introduced to rank the documents that are retrieved. Experts are to be authenticate to rate the page that are displayed that, when the same query is given again, the ranking is based on the experts preference.

It also suffer from a major disadvantage that each object must belong to exactly one group which leads to the limitation that all group must have at least one member. Fuzzy clustering produces results which include too much noise which affects the accuracy.

6. Experimental Results

For experimental verification a comparison is done with other search engines for example say, google, yahoo and altavista. The comparison is performed within these search engines and the proposed system. The proposed system uses the normal search engines as API and uses the proposed system as enhanced filters. The experimental results are calculated based on the results from the distribution hypothesis for the clustering of the documents and the genetic algorithm for the identification of the similarity between the contents. Later a sample experts' preference is given and consolidated with the ranking system where the search results has enhanced to a certain level. The initial data of precision and recall is collected from mined data records in different webpages and the comparison is given in **Table 1**.

For the purpose of experimentally verifying the recommendation process, a real data set is processed using EPCRR algorithm. Recommender systems based on Agglomerative clustering and collaborative filtering involves two stages. In the first stage, characteristic similarities between various services are first computed. Then, all services are merged into clusters using Agglomerative clustering Algorithm. In the second stage, rating similarities between services that belong to the same cluster are computed. Then some services whose rating similarities with the target service exceed a threshold are selected as neighbours of the target service. At last, the predicted rating of the target service is computed. Generally, a recovery service is described with some tags and contains certain functionality. As an experimental case, ten recovery services are considered and the corresponding description tags and functionality are listed in **Table 2**.

First description and functionality similarities between recovery services are computed. For instance, there are four same stemmed tags among the six different stemmed tags in s_2 and s_3 and the functionality of the two services are similar. Therefore, $D_{sim}(s_2,s_3) = 4/6$ and $F_{sim}(s_2,s_3) = 1$. Characteristic similarity is calculated using the weighted sum of the description similarity and functionality similarity. The description similarity weight α is set to 0.5. Then the characteristic similarity between s_2 and s_3 is computed as $C_{sim} = (0.5 \times 4/6) + (0.5 \times 1) = 0.833$.

Table 1. Comparison result analysis.

	Google	Yahoo	Bing	system
Precision	1.593	1.545	1.490	1.612
recall	0.882	0.058	0.059	0.132

Three digits after the decimal point are retained for the computation results. Characteristic similarities between all the recovery services are all computed by the same way, and the results are shown in **Table 3**.

Now the agglomerative clustering algorithm is applied. Initially individual services are considered as clusters and based on characteristic similarity clusters are combined.

The reduction step of the Algorithm is described as follows:

Step 1: More similar pair is searched in the similarity matrix and merged to form a cluster.

Step 2: New similarity matrix is created and using the average values the similarities between clusters are calculated.

Step 3: The similarities are stored.

Step 4: Proceed with step1 until the similarity is negligible.

The reduction steps are illustrated in **Table 4**.

After some reduction process now there are only 4 clusters remaining and the algorithm is terminated. By using this algorithm, the ten recovery services are merged into four clusters, where s_1 and s_6 are merged into a cluster named C_1, services s_2, s_3, s_5, s_7 and s_9 are merged into a cluster named C_2, service s_4 is separately merged into a cluster named C_3 and services s_8 , s_{10} are merged into a cluster named C_4.

Table 2. Example of different recovery services in computer system.

No	Name of Services	Functionality	Description Information	Stemmed Tags
s_1	Disk Utility	Consistency Checker	Consistency, Checker, DOS, Windows	Consistent, Check, DOS, Window
s_2	CD Roller	File Recovery	Recovery, Data, Optical, Disks	Recover, Data, Optic, Disk
s_3	Iso Buster	File Recovery	Recovery, Data, Hard, Drives, Optical, Disks	Recover, Data, Hard, Drive, Optic, Disk
s_4	Copy CatX	Imaging Tools	Backup, Images, Damaging, Data	Backup, Image, Damage, Data
s_5	Recuva	File Recovery	Recovers, Data, Windows, Hard, Drives	Recover, Data, Window, Hard, Drive
s_6	CHKDSK	Consistency Checker	Consistency, Checker, MAC	Consistent, Check, MAC
s_7	Data Life Saver	File Recovery	Booting, Data, Recovery, Files	Boot, Data, Recover, File
s_8	Data Rescue 3	Bootable	Recovers, DVD, HFS, Systems	Recover, DVD, HFS, System
s_9	CDRecover	File Recovery	Recovery, Data, Optical, Disks	Recover ,Data, Optic, Disk,
s_{10}	Knoppix	Bootable	Linux, DVD, Recovery, Systems, Utility	Linux, DVD, Recover, System, Utility

Table 3. Similarity matrix (initial stage).

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	*	0	0	0	0.063	0.7	0	0	0	0
C_2	0	*	0.833	0.071	0.643	0	0.667	0.071	**1**	0.063
C_3	0	0.833	*	0.056	0.786	0	0.625	0.056	0.833	0.05
C_4	0	0.071	0.056	*	0.063	0	0.071	0	0.071	0
C_5	0.063	0.643	0.786	0.063	*	0	0.643	0.063	0.643	0.056
C_6	0.7	0	0	0	0	*	0	0	0	0
C_7	0	0.667	0.625	0.071	0.643	0	*	0.071	0.667	0.063
C_8	0	0.071	0.056	0	0.063	0	0.071	*	0.071	0.75
C_9	0	**1**	0.833	0.071	0.643	0	0.667	0.071	*	0.063
C_{10}	0	0.063	0.05	0	0.056	0	0.063	0.75	0.063	*

Table 4. Algorithm reduction (clusters = 4).

	(C_1,C_6)	(C_2,C_3,C_5,C_7,C_9)	C_4	(C_8,C_{10})
(C_1,C_6)	*	0.008	0	0
(C_2,C_3,C_5,C_7,C_9)	0.008	*	0.068	0.064
C_4	0	0.068	*	0
(C_8,C_{10})	0	0.064	0	*

Suppose there are four users (u_1, u_2, u_3, u_4) who rated the ten recovery services. A rating matrix is shown in **Table 4**. The ratings are on 5-point scales and 0 means the user did not rate the recovery service. As u_3 does not rate s_5 (a not-yet-experienced item), u_3 is regarded as an active user and s_5 is looked as a target recovery. By computing the predicted rating of s_5, it can be determined whether s_5 is a recommendable service for u_3. Furthermore, s_2 is also chosen as another target recovery. Through comparing the predicted rating and real rating of s_2, the accuracy of proposed system will be verified in such case. Since s_5 and s_2 are both belong to the cluster C_2, rating similarity is computed between recovery services within C_2 by using formula (4). The rating similarities between s_5, s_2 and every other recovery service in C_2 are listed in **Table 5**.

Rating similarity is computed using Pearson correlation coefficient and it ranges in value from −1 to +1. The value of −1 indicates perfect negative correlation and the value of "+1" indicates positive correlation. Without loss of generality, the rating similarity threshold γ in formula (5) is set to 0.5. Since the rating similarity between s_5 and s_2 is 0.544 and the rating similarity between s_5 and s_3 is 0.736 which are both greater than γ, s_2 and s_3 are chosen as the neighbours of s_5, *i.e.*, $N(s_5) = s_2, s_3$.

Since the rating similarity between s_2 and s_3 is 0.839 and the rating similarity between s_2 and s_5 is 0.544 which are both greater than γ, s_3 and s_5 are chosen as the neighbours of s_2, *i.e.*, $N(s_2) = s_3, s_5$. According to formula (6), the predicted rating of s_5 for u_3 is 1.97 and the predicted rating of s_2 for u_3 is 1.06. Thus, s_5 is not a good recovery service for u_3 and will not be recommended to u_3. In addition, as the real rating of s_2 given by user u_3 is 1 (**Table 6**) while its predicted rating is 1.06, it can be inferred that proposed system may gain an accurate prediction.

7. Efficiency of the Proposed System

A hybrid system will make use of the combination of collaborative content based filter and it restricts itself with the collaborative filtering strategies. Before the content based filtering begins it accepts dataset formed in the process and with the hierarchy of the user preference the recommended data set are formulated and the ranking will be given based on the user preferences [18]. In a content-based system, keywords are used to describe the items, besides the user profile is built to indicate the keyword item to which the user specified their desires. In other words, the algorithm proposed try to give the most relevant data in the hierarchy for the recommended user and also it make sure that there is no similarity in the user retrieved information and further unused and used information will be stored in the repository for future use [22]. Generally the cold start problem, data sparsity may affect the system performance and here we discussed how the proposed system overcomes these problems to improve the performance of the system.

Accuracy of the Proposed Recommendation: To evaluate the accuracy of this algorithm, Mean Absolute Error (MAE), which is a measure of the deviation of recommendations from their true user-specified ratings, is used in this paper. The recommendation quality is measured using the mean absolute error (MAE) and sometimes it is also called absolute deviation. This method takes the mean of the absolute difference between each prediction and all ratings of users in the test set. MAE is computed as follow:

$$\text{MAE}=\frac{1}{\text{n}}\sum_{u,i}\left|p_{u,i}-r_{u,i}\right| \tag{4}$$

In this formula, n represents the number of rating-prediction pairs, $r_{u,i}$ is the rating that an active user u gives to a recovery service i, $P_{u,i}$ denotes the predicted rating of i for u.

For each test recovery service in each fold, its predicted rating is calculated based on traditional system and proposed system approach separately. The recovery services considered as the real data set is experimented with

Table 5. Rating matrix of cluster of services.

	C_1		C_2				C_3	C_4	
	s_1	s_6	s_2	s_3	s_5	s_7	s_4	s_8	s_{10}
u_1	4	5	4	3	3	1	2	5	4
u_2	2	3	4	5	4	4	3	1	0
u_3	5	0	1	2	0	2	1	4	4
u_4	2	1	5	5	5	1	5	5	4

Table 6. Rating similarity between selected services.

	Recovery services pair	**Rating similarity**
Rating similarity with respect to s_5	(s_5,s_2)	0.544
	(s_5,s_3)	0.736
	(s_5,s_7)	0
	Recovery services pair	Rating similarity
Rating similarity with respect to s_5	(s_2,s_3)	0.839
	(s_2,s_5)	0..544
	(s_2,s_7)	−0.187

both the concepts and the MAE is calculated. Therefore, without loss of generality in our experiment, the value of K is set to 4, 5, 6, 7, 8 respectively. Furthermore, rating similarity threshold γ is set for two cases. Under these parameter conditions, the predicted ratings of test services are calculated by proposed system and Traditional system. Then the average MAEs of Proposed system and Traditional system can be computed using formula (6). The comparison results are shown in **Figure 4**.

While the rating similarity threshold $\gamma < 0.5$, MAE values of proposed recommendation decrease as the value of K increases. The services are divided into clusters, and the services in a cluster will be more similar with each other. Furthermore, target service neighbours are chosen from the cluster of that the target service belongs to. Therefore, these neighbours might be more close to the target service and it results more accurate prediction.

While $\gamma = 0.5$, MAE values of Proposed system and Traditional system both increase. The intermediate results of these two approaches were checked and if the rating similarity threshold is set to 0.5 then the test services have only few or no neighbours, when neighbours have to be selected from a smaller cluster. It results large deviations between the predicted ratings and the real ratings.

Computation Time for the Proposed Recommendation

The time complexity of this approach involve two parts namely the offline cluster formation with agglomerative clustering algorithm and the online collaborative filtering. There are two main computationally expensive steps in this algorithm. The first step is the computation of the pair wise similarity between all the services. The number of services in the recommender system is n, and the complexity of this step is generally O (n^2). The second step is to repeat the selection of the pair of most similar clusters or the pair of clusters that optimizes the criterion functionality. A naive way of performing this step is to merge each pair of clusters after each level of the agglomeration then re-compute the gains achieved and select the most promising pair. If the number of the target service's neighbours reaches to the maximum value, then its worst case time complexity of item-based prediction is O (n_k). Since $n_k \ll n$ and $m_k \ll m$, the cost of computation will decrease significantly [18]. In order to evaluate the efficiency of proposed recommender system, the online computation time of proposed recommender system is compared with that of traditional recommender system, as shown in **Figure 5**.

In all, proposed recommender system spends less computation time than traditional Item-based Collaborative Filtering. Since the number of services in a cluster is less than the available services, and the time of rating similarity computation between every pair of services will be reduced. The rating similarity threshold γ increase,

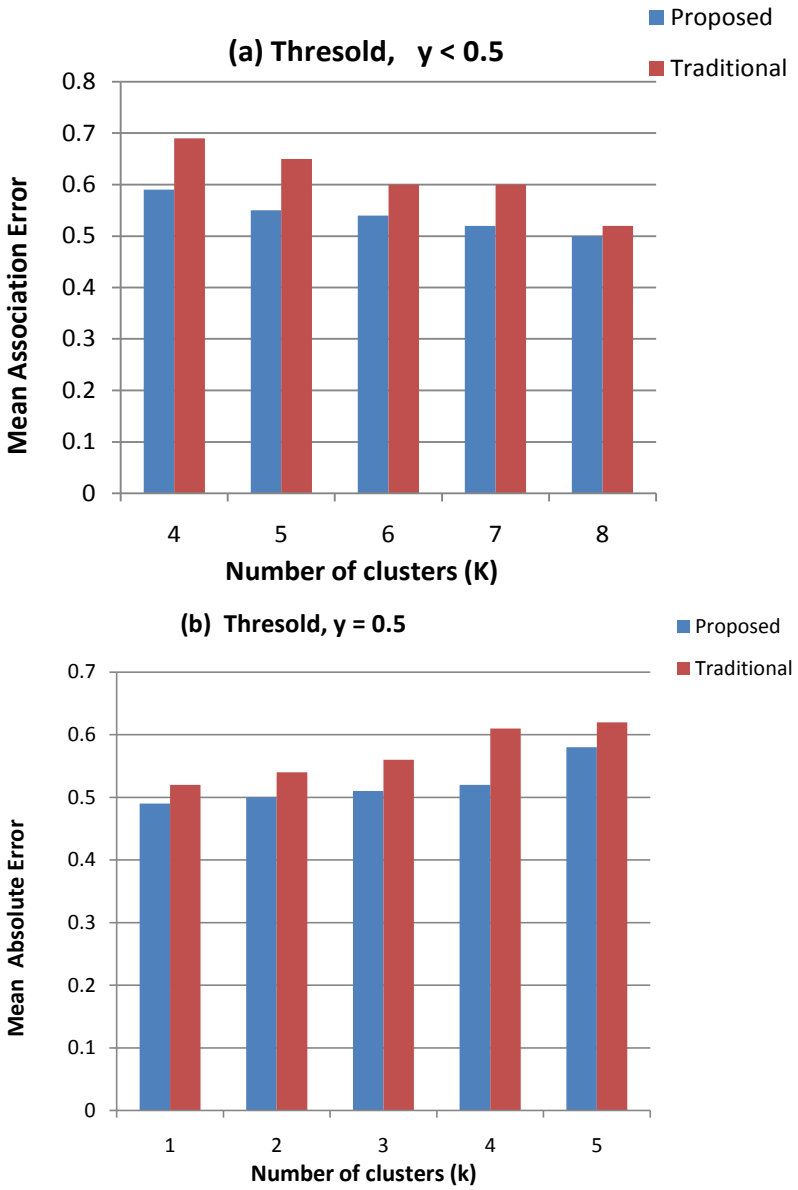

Figure 4. Comparison of MAE with proposed and traditional recommendation systems.

then the computation time of proposed recommender system decrease. It is due to the number of neighbours of the target service decreases when γ increase.

8. Conclusion and Future Work

We conclude our work by proposing an exact semantic search engine which gives preference to the user with highest priority of data content retrieval and it works on the data using agglomerative clustering. This work extends with the filter which works under the user agent without any supervision. In pile clustering, the ranking hierarchy is provided to the relevancy of the data, and the user will get the content in the highest order with efficient mining process. The mining process constitutes refining the information content process which has been clustered in the data set which has highest affinity towards relevance of the information. So the user in any environment can able to get exact information accordingly to their desires. Next acquired information which is not used will be moved to repository for future use. This procedure simulates that information retrieval works on

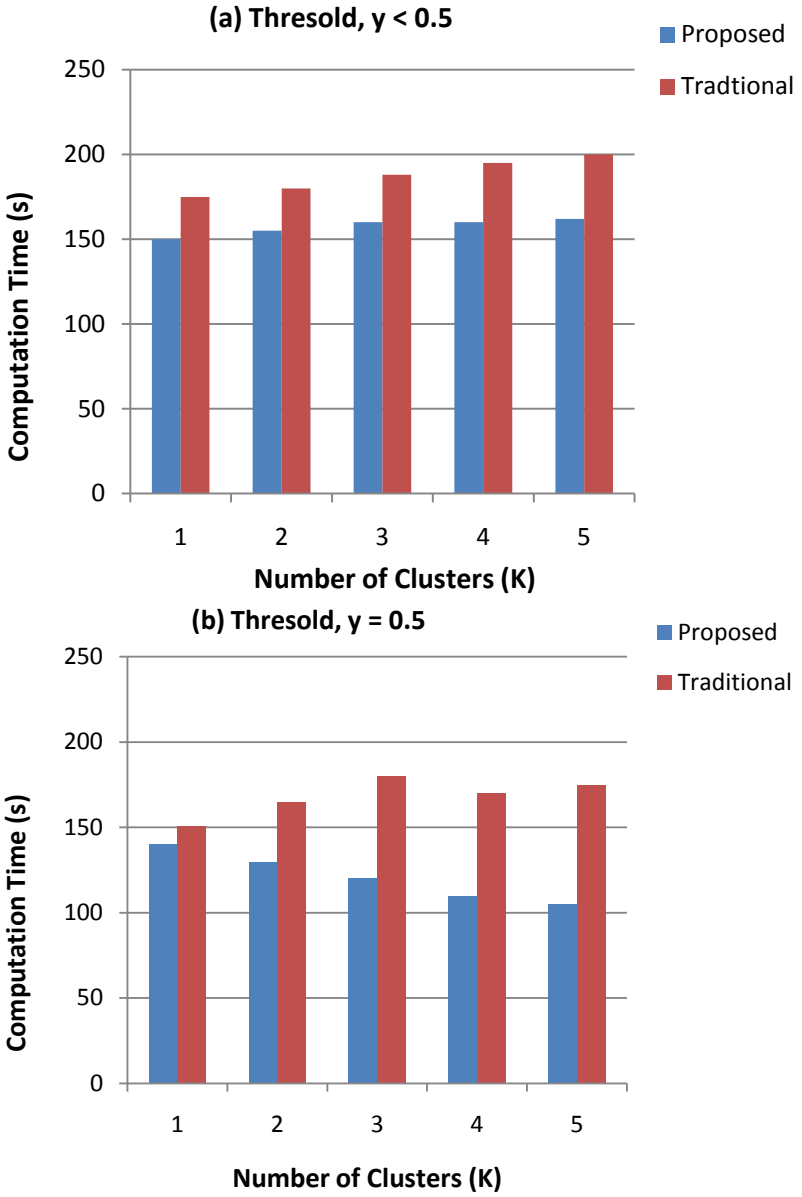

Figure 5. Comparison of computation time with proposed and traditional recommendation system.

intelligent system, but actually the data recommendation is used to give justification for the intelligence. In future, the same work can be extended purely on expert system without any intervene from the external user to obtain the content in the absence of mining.

References

[1] Gupta, V. and Lehal, G.S. (2013) A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages. *Journal of Emerging Technologies in Web Intelligence*, **5**, 157-161. http://dx.doi.org/10.4304/jetwi.5.2.157-161

[2] Bellogí, A., Cantador, I., Dí, F., *et al*. (2013) An Empirical Comparison of Social, Collaborative Filtering, and Hybrid Recommenders. *ACM Transactions on Intelligent Systems and Technology*, **4**, 1-37. http://dx.doi.org/10.1145/2414425.2414439

[3] Cafarella, M.J., Wu, E., Halevy, A., Zhang, Y. and Wang, D.Z. (2008) Web Tables: Exploring the Power of Tables on

the Web. VLDB.

[4] Sleiman, H.A. and Corchuelo, R. (2013) A Survey on Region Extractors from Web Documents. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 9.

[5] Bisht, S.S. and Bansal, S. (2013) Optimization of Web Content Mining with an Improved Clustering Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, **3**, 479-483.

[6] Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[7] Li, M.J., Ng, M.K., Cheung, Y.M., *et al*. (2008) Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters. *IEEE Transactions on Knowledge and Data Engineering*, **20**, 1519-1534. http://dx.doi.org/10.1109/TKDE.2008.88

[8] Michael, R., Lyu, F. and Wang, J.M. (2013) QoS Ranking Prediction for Cloud Services Zibin Zheng, Member, IEEE, Xinmiao Wu, Yilei Zhang, Student Member. *IEEE Transactions on Parallel and Distributed Systems*, **24**, 1213-1222.

[9] Zheng, Z., Ma, H., Lyu, M.R., *et al*. (2011) QoS-Aware Web Service Recommendation by Collaborative Filtering. *IEEE Transactions on Services Computing*, **4**, 140-152. http://dx.doi.org/10.1109/TSC.2010.52

[10] Mai, J., Fan, Y. and Shen, Y. (2009) A Neural Networks-Based Clustering Collaborative Filtering Algorithm in E-Commerce Recommendation System. *International Conference on Web Information Systems and Mining*, Shanghai, June 2009, 616-619. http://dx.doi.org/10.1109/WISM.2009.129

[11] Rajaraman, A. and Ullman, J.D. (2012) Mining of Massive Datasets. Cambridge University Press, Cambridge.

[12] Mittal, N., Nayak, R., Govil, M.C., *et al*. (2010) Recommender System Framework Using Clustering and Collaborative Filtering. 2010 3*rd International Conference on Emerging Trends in Engineering and Technology*, Goa, November 2012, 555-558. http://dx.doi.org/10.1109/icetet.2010.121

[13] Li, X. and Murata, T. (2012) Using Multidimensional Clustering Based Collaborative Filtering Approach Improving Recommendation Diversity. 2012 *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, December 2012, 169-174. http://dx.doi.org/10.1109/WI-IAT.2012.229

[14] Zhou, Z., Sellami, M., Gaaloul, W., *et al*. (2013) Data Providing Services Clustering and Management for Facilitating Service Discovery and Replacement. *IEEE Transactions on Automation Science and Engineering*, **10**, 1-16. http://dx.doi.org/10.1109/TASE.2012.2237551

[15] Pham, M.C., Cao, Y., Klamma, R., *et al*. (2011) A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. *Journal of Universal Computer Science*, **17**, 583-604.

[16] Thilagavathi, G., Srivaishnavi, D., Aparna, N., *et al*. (2013) A Survey on Efficient Hierarchical Algorithm Used in Clustering. *International Journal of Engineering*, **2**, 2553-2556.

[17] Yamashita, A., Kawamura, H. and Suzuki, K. (2011) Adaptive Fusion Method for User-Based and Item-Based Collaborative Filtering. *Advances in Complex Systems*, **14**, 133-149. http://dx.doi.org/10.1142/S0219525911003001

[18] Platzer, C., Rosenberg, F. and Dustdar, S. (2009) Web Service Clustering Using Multidimensional Angles as Proximity Measures. *ACM Transactions on Internet Technology*, **9**, Article No. 11.

[19] Zhao, Y., Karypis, G. and Fayyad, U. (2005) Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**, 141-168. http://dx.doi.org/10.1007/s10618-005-0361-3

[20] Ortega-Mendoza, R.M., Villaseñor-Pineda, L. and Montes-y-Gómez, M. (2014) Using Lexical Patterns for Extracting Hyponyms from the Web. Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, México.

[21] Ferraraa, E., De Meob, P., Fiumarac, G. and Baumgartnerd, R. (2014) Web Data Extraction, Applications and Techniques: A Survey. *Knowledge-Based Systems*, 301-323.

[22] Ye, H.W. (2011) A Personalized Collaborative Filtering Recommendation Using Association Rules Mining and Self-Organizing Map. *Journal of Software*, **6**, 732 -739.