Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# COMPARATIVE APPROACH FOR RANKING BY SCRUTINIZING THE WEB OPINION USING SDCA

**[1]N.DEEPTHI, [2]R.SEETHALAKSHMI**

[1]M.Tech-Advanced Computing, School of Computing, SASTRA University, Tamil Nadu, India-613402
[2]Professor, School of Computing, SASTRA University, Tamil Nadu, India-613402

E-mail:[1]deepthi.nats@gmail.com, [2]rseetha123@cse.sastra.edu

**ABSTRACT**

Web opinions are available for the social sites due to the advancement in the technology. The web opinion acts as an interface between the internet users and web. It allows internet users to communicate and express their opinions. But analyzing and clustering of the web opinion is a challenging task. The clustering of the typical document differs from clustering of the web opinion. The web opinions are taken from the social networks. Analyzing the social network helps in preventing the crimes, terrorists activities etc. The scalable density based algorithm enables the identification of themes within discussions in web social networks and their development. The predefined set for clustering is useful in web opinion clustering. The ranking methods that are being used for clustering are the XMLTF*IDF and Bin Ranking.

**KEYWORDS:** *Clustering, Scalable Density Based Clustering, Web opinion, Ranking, DBSCAN.*

## 1. INTRODUCTION

The web is providing an opportunity to the users to express their views which normally contains thousands of comments especially on the agitative topics. When the number of opinions increases for a particular topic they indirectly indicate the importance of the particular topic.
Clustering these comments is helpful in preventing the terrorists' attacks and some unnecessary social activities. But clustering these comments is not an easy task. Figure 1 and Figure 2 represents the situations before and after clustering.

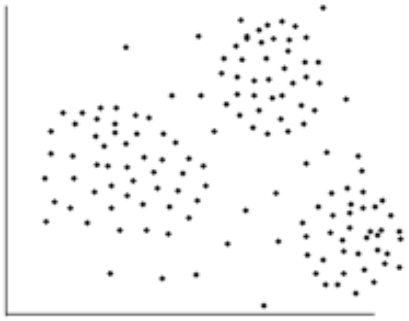

*Figure1. Before Clustering – with noise*

The comments are usually referred to as web opinions. The clustering process is done to group the similar opinions. This clustering is done for the easy retrieval of the information ie the opinion about a particular topic in order to prevent the unnecessary activities like crimes. There are two challenges in clustering: 1.There is no link between the web opinion as in the web pages or blogs. 2. More over the opinions are short texts. [3]

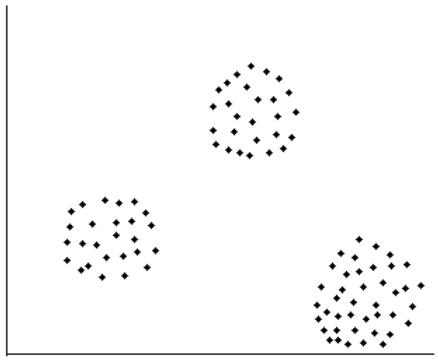

*Figure2. After Clustering – noises are removed*

The traditional algorithm fails in clustering the web opinion because they virtually vote their opinion on the particular topic.[5] The link for the person who hosts the information is not present here and their terms in the opinion are meagre and they contain some new-fangled words which are not in the ontology or typical dictionary. These are the special properties of the web opinion. To overcome these

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

drawbacks one uses unsupervised learning approach where no predefining of the sets is not necessary for clustering. [1]

The ranking is the core concept for clustering. There are several ranking schemes. One such method is XMLTF*IDF and Bin Ranking. [4] [9] .The XMLTF*IDF prevents the occurrence of unrelated terms in the searches. When compared to the bin ranking methodologies this method is more effective in ranking and is more accurate. Figure 3 shows the block diagram of the web opinion based clustering. These things are done to improve the accuracy of the clustering.

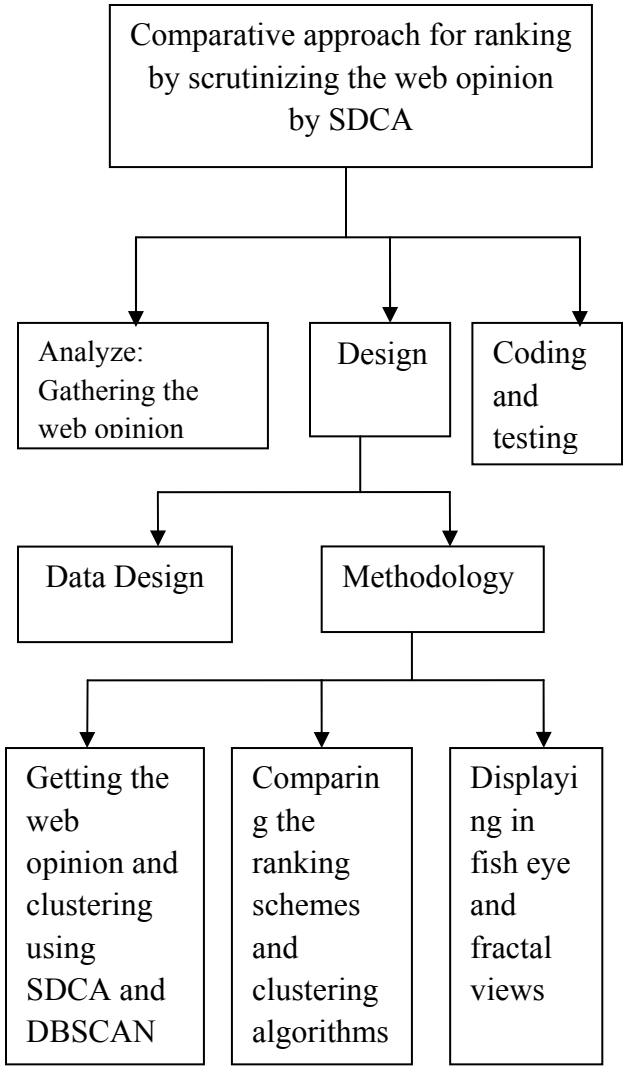


*Figure3.Block Diagram of the web opinion.*

## 2. RELATED WORK

### 2.1. K-Means Algorithm

In 1967 James Mac Queen proposed the K-Means algorithm. The K-Means is a prototype based clustering. They select K initial centroids. For each point it finds closest centroids and assigns that point to the centroid. This forms a K cluster. The K- means is a partitional clustering algorithm. This is fast for low dimensional data and also it finds the exact sub clusters if large numbers of clusters are specified. We cannot use this in clustering web opinion because this cannot handle data of varying size and densities.

More over the outliers are not identified and this is more cramp to the data which has a centroid. The web opinions keep on increasing and more over they don't rely on the centroid.

### 2.2. EM Clustering Algorithm

In 1977 Arthur Dempster, Nan Laird, and Donald Rubin proposed the expectation maximum algorithm. An expectation-maximization (EM) algorithm is for discovering the maximum likelihood or maximum a posterior (MAP) which estimates the parameters in statistical models. This model is for unobserved latent variables. EM is an iterative method. The EM algorithm finds the expectation of the log likelihood which is evaluated using the current estimate of parameters and a maximum step for finding the parameters maximized during the expectation of the log likelihood. This method requires the predefined set for clustering. But in web opinion this not quite possible since one cannot predict the number of clusters as they keep on rowing.

### 2.3. Clustering Short Texts

The Banerjee et al. described clustering of short texts in Wikipedia. One of the problems that occur is the information overload. In this the sources usually deliver large number of items periodically. The solution for this is clustering similar items in the feed reader for utilizing the information to its complete extent and also it must be more manageable for the user. This is a provoke task since a part of the original data is only received. Clustering akin items provide an interface to the user. They remove the repeated and similar items. [2] The table1 gives the comparative details about other methods and specifies that the graph method is more efficient. This method is also not successful in clustering the web

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

opinion because their size keeps on increasing and it contains noisy data and the accuracy is not meet.

| | Direct | Aggloremative | graph |
|---|---|---|---|
| Baseline | 67.05 | 22.03 | 81.62 |
| Wiki method | 82.66 | 83.88 | 89.56 |

*Table1.Different Clustering Accuracy in percentage*

### 2.4. Comments on the Clustering

The YeXin Wang, Li Zhao, YanZhang performed the clustering analysis for the web comments. They state that the traditional algorithms fail to work on the web opinions. They don't have any relationship between the hosted information and they are short messages. They proposed two phase method for clustering the short texts. In this method they convert the short messages into longer texts and then they cluster using the graph. But they fail to dynamically add the original comment and adjust their threshold. This method cannot be used for the clustering web opinion. These problems can be rectified by scalable distance based algorithm.[6]

### 2.5. Tree Data Model

In 2007, C. Sun, C.Y. Chan, and A.K. Goenka**,** lowest common ancestor is used to find the XML nodes which contains the query terms within the sub trees. Then they proposed the smallest lowest common ancestor which combines the operations of AND and OR Boolean operators. XSeek is used for finding the return query but it doesn't deal with the key word ambiguity problem.[4] And the results that are retrieved for the given query is with the irrelevant data terms. The XML used the functional shipping concepts. There are many ranking schemes for this. But none of the algorithm is effective in retrieving the relevant terms. [14]

## 3. CONTENT CLUSTERING

The DBSCAN and Scalable distance based clustering algorithm are known as the content clustering algorithm. The content clustering is a two phase approach. In the first phase it identifies the parallel threads to obtain the synopsis from the clusters. In the second phase it reveals the topic similarity. Their objective is to cluster similar terms without any predefined set and not to cluster any noisy terms.

### 3.1. DBSCAN

In 1996 Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu proposed the DBSCAN algorithm. [3] The DBSCAN finds the number of clusters which starts from the predictable density distribution of parallel nodes. This consists of two parameters eps and minimum points required to form a cluster. This method starts with an arbitrary point that has never been visited. Once the neighbourhood is retrieved and if it contains sufficient points, the clustering process is started. The density in the noisy region is lower than the density in the normal regions. [11] The boundary is known as the eps and the minimum eps is the neighbourhood radius. This requires a predefined set of clusters which is not possible in web opinion. So the DBSCAN is not applicable for web opinion.

**DBSCAN Algorithm**

```
1.  DBSCAN(D, eps, minPts)
2.  C=0
3.  For all unvisited point in dataset
4.  mark point as visited
5.  M = getNeighbors (R, eps)
6.  if sizeof(M) < minPts
7.  mark point as NOISE
8.  else
9.  C = next cluster
10. expandCluster(point, M, C, eps, minPts)
11. add R to C
12. for each R'  in N
13.  if R' is not visited
14.  mark R' as visited
15.  N' = regionQuery(R', eps)
16.  if sizeof(n') >=minPts
17.  N =N joined with N'
18.  if R' is not yet a member of any cluster
19.  add R' to C
20.  add point to C
```

### 3.2. Shared Nearest Neighbor

The shared nearest neighbor (SNN) is the enhanced method for density based clustering. The SSN and DBSCAN differ in the definition of the similarity between points in pairs. [5]SNN defines the similarity of the points in the pairs as the number of nearest neighbors the two points share. The density

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is measured by the sum of the similarities of the nearest neighbors of the point. Points that have high density are selected as the core points, and that with the low density are identified and removed as the noise. All the points similar to the core point are cluster together. They form cluster in elongated shapes so that they can overcome the different problems that occur in DBSCAN algorithm. It is reported that SNN performs well when compared to the DBSCAN. But SNN cannot be used in the web opinion since it is fast growing. [10]

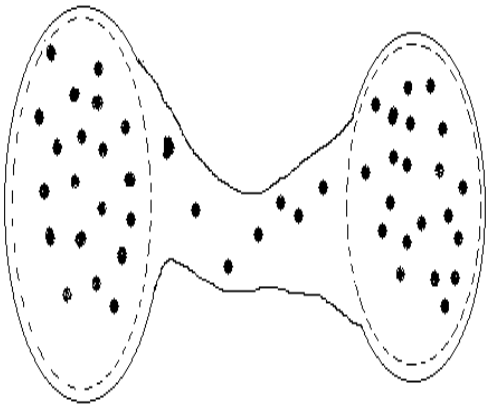

*Figure4.Clustering by SNN*

### 3.3. Scalable Density Based Clustering

The scalable density based clustering doesn't require any predefined set for clustering and this removes noise in a better way. This is illustrated using solid clusters in the initial step. When the size of the cluster keeps on increasing they are represented using the dotted lines. For every iteration a new circle in the dotted format grows. The points that are density reachable directly are not included since their size keeps on increasing. The points in the cluster are close to one another with a reasonable distance and this is not valid to the direct density reachable.

**SDC Algorithm**

1. Create instances of all points as unclassified points S={s1,s2……sn}
2. Repeat
3. Randomly select a point Pi in S
4. The number of points in eps-neighbourhood of Pi ≥ minpts
5. Create the initial cluster Cj by including the eps-neighbourhood points
6. S = S - Cj
7. Else Pi is classified as X
8. S = S – Cj
9. Until S = ϕ
10. For each initial cluster Cj
11. Repeat
12. Find the centroid
13. eps = eps - ▲eps
14. insert points from X in which the distance from the centroid of the cluster is larger than eps
15. until no other points are found
16. the points that are found in X are considered as noise

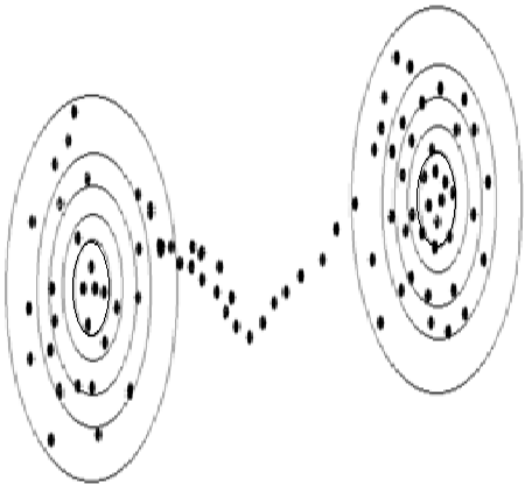

*Figure5.Clustering by SDC*

## 4. CONCEPT FOR SELECTING A CLUSTER

The web opinions found in the web forum are usually noisy. For selecting the noisy data the three clustering concepts have been defined.

### 4.1. Ranking

There are several ranking methods for ranking an XML page. The most popular of these are the XML TF*IDF and Bin Ranking.

#### 4.1.1. XML Term and Inverse Document Frequency

The XMLTF*IDF is one of the best suited method for ranking the noisy data. This takes into consideration of three issues .The first issue is about the effective identification of the target nodes. The second issue is about the identification of the node type and the third issue is to rank the query results. These issues are the problems in ranking a noisy content. [7] [8] The usage of this method prevents the presence of irrelevant content in the ranked content.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To check this method we compare this method with the Bin ranking scheme. The noisy content is present in the bin ranking whereas it has been omitted in the XMLTF*IDF. The keyword ambiguity problems are also resolved here.[4] [12][13]

**TF*IDF Algorithm:**

Input: Query

Output: Ranked list

1. Let max = 0; $T_{for}$ = null
2. List $L_{for}$ = getallnodetypes()
3. for each tn ϵLfor do
4. $C_{for}(t_n,Q)$ = getsearchforconfidence(tn,Q)
5. If($C_{for}(t_n)$ > max) then
6. max = Cfor($t_n$); tfor = tn
7. rank link list
8. $N_{for}$=getnext($T_{for}$)
9. While(!end(IL[1])||….||(!end(IL[m])))) do
10. Node a= getMin(IL[1],IL[2]….,IL{m})
11. If(!is ancestor($N_{for}$,a)) then
12. $P_s(Q,aN_{for})$=getsimilarity($N_{for}$,Q)
13. rankedList.insert($N_{for}$,$P_s(Q,N_{for})$)
14. $N_{for}$=getNext($t_{for}$)
15. If(is ancestor($N_{for}$,a)) then
16. Else
17. $P_s(Q,a)$=0
18. For each two neighboring ordered results r1 and r2in ranked list do
19. If(($P_s(r_1,Q)$-$P_s(r_2,Q/P_s(r_2,Q)$ < σ) then
20. For each such $r_i$ do
21. $P(Q,r_i)$=get popurarity($r_i$,Q,CT,L)
22. Re-rank those $r_i$ in ranked list according to their $P(Q,r_i)$
23. Return ranked list;

**4.1.2. Page Ranking**

Page rank algorithm use web page link structure to assign global importance to web pages. This uses the random web surfer method which refers to the process of starting at a random page and goes to through the link with uniform probability. The page rank has dynamic versions. [19] The versions are personalized web page rank and object rank methods.

**4.1.2.1. Personalized Web Page Ranking**

The personalized web page ranking has a preference set that the user likes. When a preference set is given it performs the fixed point iterative computation over the given set and generates the personalized search results. The scalability is more attractive. In this method the time taken for the computation of the given query is high.[15]

**4.1.2.2 Object Ranking**

Perform keyword search in data base.[17] [18] This method uses the query term posting list as a set of random walk starting point and this continues the walk on the graph. The high recall search method is used. This requires the multiple iterations over all the nodes and this links the entire data base. The object rank has two operating modes such as the online mode and off line mode. [16]

The online ranking mode performs the process once the query is given. This increases the retrieving process time for the given query. The offline ranking mode performs the pre computation for the top k results in the advance. This method is expensive and requires lots of storage space. This is not feasible for all the terms in the data dictionary.

**4.1.3 Bin Ranking**

The bin ranking system employs a hybrid approach. The hybrid approach refers to the combination of both personalized page rank and object rank mechanisms. This reduces the time and the storage space. The same object rank process is used but it is used in the small graphs instead of the full data graph. The sub graphs are pre computed in offline. This pre computation can be parallelized with linear scalability. This method scales to large clusters by distributing the sub graphs between the nodes of the clusters. This can be used to store more number of sub graphs in RAM. This reduces the average query execution time. But when this ranking scheme is compared with the XMLTF*IDF the problem arises in the case of the accuracy and the amount of time taken for the retrieval of the query.

**4.2 Exclusion of terms**

The second most important criterion is to exclude the terms that occur repeatedly. For example same person may vote for three or four times then it refers to the process of redundancy creation. To overcome this we use our algorithm. This refers to the process of excluding the similar terms and this may also omit the perfect terms that occur in the top N search but those will be included in the future clusters. The non comparable terms doesn't have any effect in the comparison process.[1]

**4.3 Bigraphs**

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The two word terms must also be a part of the document vectors. [1] The natural language processing becomes a specific tool for the identifications of the nouns and verbal phrasal's which is done in a more specific way than the single word terms. But the natural processing fail because of the non edited nature and conversional style of the forum messages. To overcome this we form bigrams by joining two adjacent words without any punctuation or any stopping term between them. The TFIDF formula is used to score the extracted bigrams or monograms.

## 5. EXPERIMENTAL ANALYSIS

The SDC and DBSCAN algorithms are implemented in .NET. The ranking schemes used are the XML TF*IDF and Bin Ranking. The figure 6 represents the selection of the system. The figure 7 represents the process of creating the database. The figure 8 is for searching the database. The figure 9 is for ranking scheme. The figure 10 is the comparative analysis for the DBSCAN and SDCA. This represents SDCA is more effective than DBSCAN.


*Figure7. Creating the data base*


*Figure 6. Choosing the system*

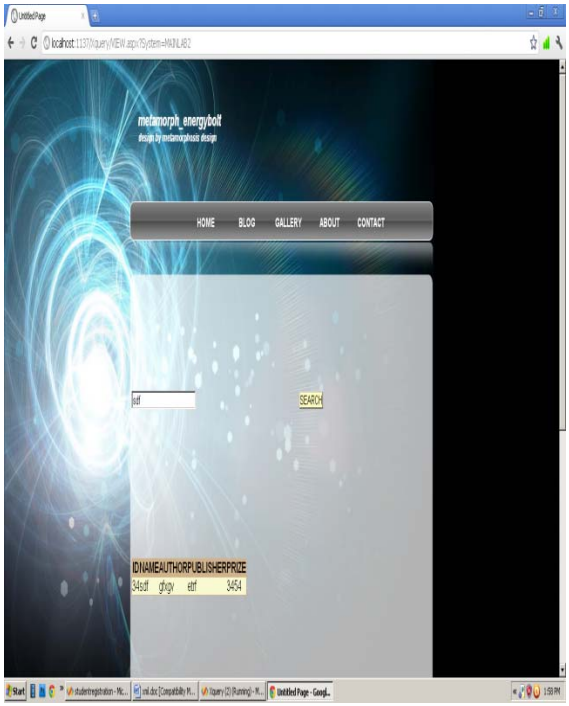


*Figure8. Searching in the database*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Figure9. Clustering by using ranking*

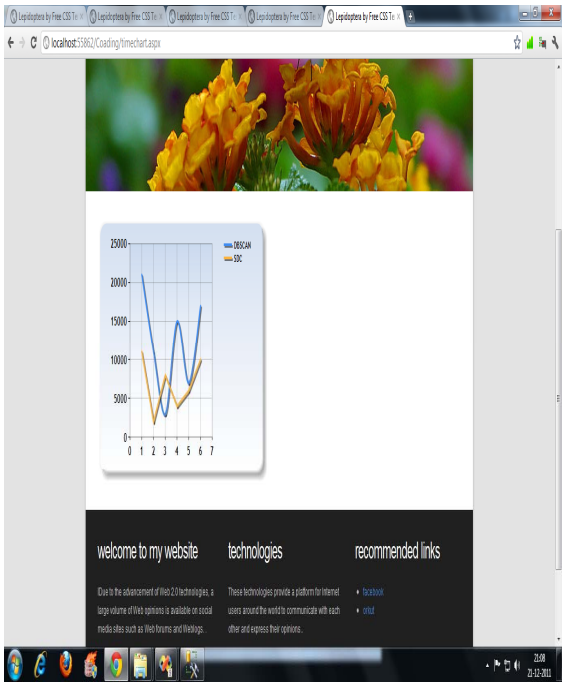


*Figure10.Comparing the SDC and DBSCAN algorithm*

## 6. CONCLUSION

The web opinions are available for the social sites. These make the users in the site to communicate with each other. The scalable distance method is used for clustering the web opinion and the density based clustering is used only for typical documents. When there is no effective clustering algorithm the web opinion becomes isolated messages. The advantage of SDC is it groups the less relevant clustering into small groups when they are density- reachable. For the ranking method used in clustering one use XMLTF*IDF and bin ranking. But bin ranking fails to satisfy the condition. The XMLTF*IDF ranks only the relevant terms and omits the unnecessary terms. Due to this the accuracy is increased and is more effective. Finally a time chart is given for SDC and DBSCAN where SDC takes only less time span for clustering the web opinion and there by increases the accuracy. The limitations of this is only limited macro and micro accuracy is achieved. To improve the accuracy more one can perform the association along with the clustering.

## 7. REFERENCES

[1] Christopher C. Yang and Tobun Dorbin Ng, "Analyzing and visualizing the web opinion development and social interactions with density based clustering", *IEEE, 2011*.

[2]Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using Wikipedia," Proc. ACM SIGIR, Amsterdam, The Netherlands, 2007, pp. 787–788.

[3] B. Bicici and D. Yuret, "Locally scaled density based clustering," Proc.ICANNGA, 2007, pp. 739–748.

[4] Zhifeng Bao, Jiaheng Lu, Tok Wang Ling, Senior Member, IEEE, and Bo Chen,"Towards an effective XML keyword search", IEEE transactions on knowledge and data engineering, vol. 22, no. 8, august 2010.

[5] D. Bollegala, Y. Matsuo, and M. Ishizjka, "Measuring semantic similarity between words using Web search engines," Proc. Int. WWW Conf., 2007, pp. 757–766.

[6] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," IEEE Trans. Syst., Man, Cybern.A, Syst., Humans, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[7] Z. Bao, B. Chen, T.W. Ling, and J. Lu, "Effective XML Keyword Search with Relevance Oriented

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Ranking," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 517-528, 2009.
[8] D. Carmel, Y.S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer, "Search XML Documents via XML Fragments," Proc. ACM SIGIR, pp. 151-158, 2003.
[9] Heasoo Hwang, Andrey Balmin, Berthold Reinwald, and Erik Nijkamp," BinRank: Scaling Dynamic Authority-Based Search Using Materialized Subgraphs" IEEE transactions on knowledge and data engineering, vol. 22, no. 8, august 2010.
[10] M. Ester, H. Kregel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proc. Int.Conf. Knowl. Discov. Data Mining (KDD), 1996, pp. 226–231.
[11] L. Ertöz,M. Steinbach, and V. Kumar, "Finding clusters of different sizes,shapes, and densities in noisy, high dimensional data," Proc. 2nd SIAM Int. Conf. Data Mining, San Francisco, CA, 2003, pp. 47–58.
[12] L. Guo, F. Shao, C. Botev, and J. Shanmuga sundaram, "XRANK : Ranked Keyword Search over XML Documents," Proc. ACM SIGMOD Conf., 2003.
[13] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv, "XSEarch: A Semantic Search Engine for XML," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 45-56, 2003.
[14] V.Hristidis, N. Koudas, Y. Papakonstantinou, and D. Srivastava, "Keyword Proximity Search in XML Trees," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 525-539, Apr. 2006.
[15] T.H. Haveliwala, "Topic-Sensitive PageRank," Proc. Int'l World Wide Web Conf. (WWW), 2002.
[16] G. Jeh and J. Widom, "Scaling Personalized Web Search," Proc. Int'l World Wide Web Conf. (WWW), 2003.
[17] A. Balmin, V. Hristidis, and Y. Papakonstantinou, "ObjectRank:Authority-Based Keyword Search in Databases," Proc. Int'l Conf.Very Large Data Bases (VLDB), 2004.
[18] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma, "Object-Level Ranking: Bringing Order to Web Objects," Proc. Int'l World Wide Web Conf.(WWW), pp. 567-574, 2005.
[19] S. Chakrabarti, "Dynamic Personalized PageRank in Entity- Relation Graphs," Proc. Int'l World Wide Web Conf. (WWW), 2007.