ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Social Network Aggregator using Crawler and Document Clustering

Neetu, Deepti Sharma

M. Tech(pursuing), Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

Head, Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

ABSTRACT: Online social networking data include rich collections of objects and vast community networks. Databases are essential to store the dramatically growing amount of such interconnected data. In these circumstances, the database management systems have to provide a natural way of managing, processing, and analyzing these complex, heterogeneous, temporal and voluminous un-structured data. Database modeling has been one of the major themes of database research over the past decades. However, a comprehensive review of recent years activity in database and data mining conferences, shows that database support for online social networks, based on a complete, efficient and scalable data model, which can facilitate inter-operability of social networking, remains an open issue. Hardly any progress has been made in order to design a database model for storing and retrieving online social-network-related data. Standard data models, query languages and access methods, such as the relational model and SQL, are often inefficient as they do not accurately capture the inherent structure of data and lack native support for large corpus. This less-than-ideal situation calls for a new scheme to store and manage huge amounts of heterogeneous data produced in the social networking sites. We believe that such a system would greatly ease the development and management of advanced online social networking applications, as well as facilitate efficient retrieval of rich information from the huge amounts of data. In short, the way we represent, store and query the online social network should allow a more semantic view of the whole structure and content using crawling and document clustering.

KEYWORDS: Social Web, Web Aggregators, Document Clustering, Focused Crawlers.

I. INTRODUCTION

The main objective is to develop an approach that helps the user in finding the needed information easily and quickly. The first approach is to use the focused crawler to retrieve the respective documents and form the cluster with the query links snippets. Search engine retrieve hundred of web pages as a result of search query to find his needed information. The second approach is to remove all the irrelevant snippet data .the third approach is to create a text summarization by using Document Clustering. The main goal is to find the accuracy of summarization of the content retrieved.

Versatile centered focused crawlers are entering components in customizing the human-PC communication. Conventional non-versatile centered crawlers are reasonable for groups of clients with shared interests and objectives that don't change with time. For this situation, it is anything but difficult to perceive the asked for points and begin an engaged creep to recover assets from the Web. The versatile crawler's preference is the capacity to learn and be receptive to potential modifications of the portrayals of client needs. This could happen when clients don't know precisely what they are searching for, or on the off chance that they choose to refine the inquiry during the execution if the outcomes are not regarded intriguing. In this manner, versatile centered crawlers are more appropriate for customized seek frameworks that incorporate a superior model of the data needs, which monitors client's interests, objectives, inclinations, and so forth. As a result, versatile crawlers are typically prepared for single clients and not for groups of individuals. A further favorable position of versatile crawlers is the affectability to potential modifications in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

the search engine. Website pages are continually being refreshed, and also the related hyperlink structure. In this way, an engaged crawler ought to have the capacity to screen any adjustment keeping in mind the end goal to search for new and intriguing data. The related areas that can truly profit by the engaged slithering are the purported vertical gateways and studies on Web development. The previous space is identified with Web locales that give data on a moderately limit scope of merchandise and enterprises for groups of clients. Recovering significant, dependable and breakthrough data re-hotspots for the client is a regular reason for these entrances, and offering customized electronic daily papers, individual shopping specialists or meeting checking administrations. Inquire about exercises on the investigation of the advancement of Web hyper textual environment more often than not requires a consistent checking of particular bits of Web and of related changes amid a given time interim, e.g., busty groups of online journals . Both spaces could truly profit by utilizing centered creeping frameworks to recover data that meets the given information requirements, finding properties that consolidate the topical substance of pages and the linkage connection between them. The rest of the part is organized as takes after: research on the WWW and issues identified with crawlers' improvement are talked about incorporates references to centered crawlers where the adaptively highlight is not generally unequivocally included.

The idea of clustering is to group similar objects into their classes. As far as multi documents are concerned, these objects refer to sentences and the classes represent the cluster that a sentence belongs to. By looking at the nature of documents that address different subjects or topics in the documents, some researchers try to incorporate the idea of clustering into their study. Using the concept of similarity, sentences which are highly similar to each other are grouped into one cluster, thus generating a number of clusters. The most common technique to measure similarity between a pair of sentences is the cosine similarity measure where sentences are represented as a weighted vector of invert document frequency. Once sentences are clustered, sentence selection is performed by selecting sentence from each cluster. Sentence selection is then based on the closeness of the sentences to the top ranking in that cluster. Those selected sentences are then put together to form the final summary

II. LITERATURE SURVEY

Cliff Lampe, Nicole Ellison, Charles Steinfield [1] depicts that, large numbers of college students have become avid Facebook users in a short period of time. In this paper, we explore whether these students are using Facebook to find new people in their offline communities or to learn more about people they initially meet offline. Our data suggest that users are largely employing Facebook to learn more about people they meet offline, and are less likely to use the site to initiate new connections.

Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi [2] depecits that, more and more web users keep up with newest information through information streams such as the popular micro-blogging website Twitter. In this paper we studied content recommendation on Twitter to better direct user attention. In a modular approach, we explored three separate dimensions in designing such a recommender: content sources, topic interest models for users, and social voting. We implemented 12 recommendation engines in the design space we formulated, and deployed them to a recommender service on the web to gather feedback from real Twitter users. The best performing algorithm improved the percentage of interesting content to 72% from a baseline of 33%. We conclude this work by discussing the implications of our recommender design and how our design can generalize to other information streams.

Uldis Bojars, Alexandre Passant, John Breslin, and Stefan Decker [3] depicts that, social network and data portability has recently gained a lot of interest as one of the issues for social media sites on the Web. In this paper, we will show how Semantic Web technologies and especially the FOAF and SIOC vocabularies can be used to model user information and user-generated content in a machine-readable way. Thus, we will see how data and network information can be reused among various services and applications, at almost zero-cost for developers of such tools.

Francesca Carmagnola, Fabiana Vernero, and Pierluigi Grillo Sonars [5] depicts that, social networks are playing an important role in personal as well as corporate environments. However, perceived issues and evolving challenges may hinder further expansion of social networks to meet new opportunities. In this paper, we review inherent concepts and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

properties of social networks and highlight major analytical evaluation criteria, which are used to identify key findings that reveal degrees of benefits and shortcomings of social networks. We also discuss some proposed solutions related to decentralized social networks in the context of business implications as well as their effects on privacy, identity and trust issues.

III. PROPOSED ALGORITHM

A. *Focused Crawler Algorithm:*

1. Get the URL.
2. Dump the URL into a data structure called queue.
3. Go to that URL scan the entire page find out if any links are present, if any URL's are present dump them into that linked list..
4. All the URL's present in the linked list are called as child URL's and the one present in queue are called as parent.
5. Now pick first child URL from linked list dump it into the queue this URL then becomes the parent repeat step 1.
6. Repeat the process for each and every child URL present in the linked list.
7. Keep on doing so till the depth mentioned at the start of the code is reached.

B. *Document Clustering Proposed Algorithm:*

Document clustering is the part of Partitioning Clustering analysis which aims to form k groups from the n data points taken as an input. This partitioning happens due to the data point associating itself with the nearest mean.

Classical Approach - The main steps of k-means algorithm are as follows:

1. Randomly select k data points to represent the seed centroids.
2. Repeat steps 3 and 4 until cluster membership stabilizes- either number of iterations specified by the user, or the dimensions of centroid does not change.
3. Generate a new partition by assigning each data point to its closest cluster center - assigning happens based on the closest mean.
4. Compute new cluster centers - calculating new centroids using the mean for multidimensional data-points

IV. PSEUDO CODE

Here's a pseudo code summary of the algorithm that can be used to implement a focused web crawler:
Ask user to specify the starting URL on web and file type that crawler should crawl.
Add the URL to the empty list of URLs to search.

1. While not empty (the list of URLs to search)
{

Take the first URL in from the list of URLs
Mark this URL as already searched URL.

2. If the URL protocol is not HTTP then
break;
go back to while

3. If robots.txt file exist on site then
If file includes .Disallow. statement then
break;

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

go back to while

4. Open the URL

5. If the opened URL is not HTML file then
   Break;
   Go back to while

   Iterate the HTML file

6. While the html text contains another link {

   If robots.txt file exist on URL/site then
   If file includes .Disallow. statement then
   break;
   go back to while

   If the opened URL is HTML file then
   If the URL isn't marked as searched then
   Mark this URL as already searched URL.

   Else if type of file is user requested
   Add to list of files found.

   }
}

Here's a pseudo code summary of the algorithm that can be used to implement a document cluster:

1. Begin with n clusters, each containing one object and we will number the clusters 1 through n.
2. Compute the between-cluster distance D(r, s) as the between-object distance of the two objects in r and s respectively, r, s =1, 2, ..., n. Let the square matrix D = (D(r, s)). If the objects are represented by quantitative vectors we can use Euclidean distance.
3. Next, find the most similar pair of clusters r and s, such that the distance, D(r, s), is minimum among all the pairwise distances.
4. Merge r and s to a new cluster t and compute the between-cluster distance D(t, k) for any existing cluster $k \neq r, s$. Once the distances are obtained, delete the rows and columns corresponding to the old cluster r and s in the D matrix, because r and s do not exist anymore. Then add a new row and column in D corresponding to cluster t.
5. Repeat Step 3 a total of $n - 1$ times until there is only one cluster left.

Inputs : Data: X := (x1, x2, . . . , x`) ⊂ IRn , Number of classes: k
Initialization: Choose random centers c1 . . . , ck
Solution: for i = 1, . . . , k do
$Ci = \{x \in X | i = \arg\min 1 \leqslant j \leqslant k \| cj - xk\ 2\}$
for i = 1, . . . , k do
$ci = \arg\min z \in Rn\ P\ x \in Ci\ k\ \|z - x\|\ k\ 2$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

V. SIMULATION RESULTS

For experiment in the scheme 200 documents and 10 queries are raised. The experiment was done using c# programming language and NUnit standard package.. To store intermediate and final result Microsoft sql-server is used. Performance Measurement Parameter: In this scheme, two performance parameters are defined to evaluate the proposed approach. These two performance parameters are listed as follows. (Precision): It is the fraction of retrieved documents that are relevant.

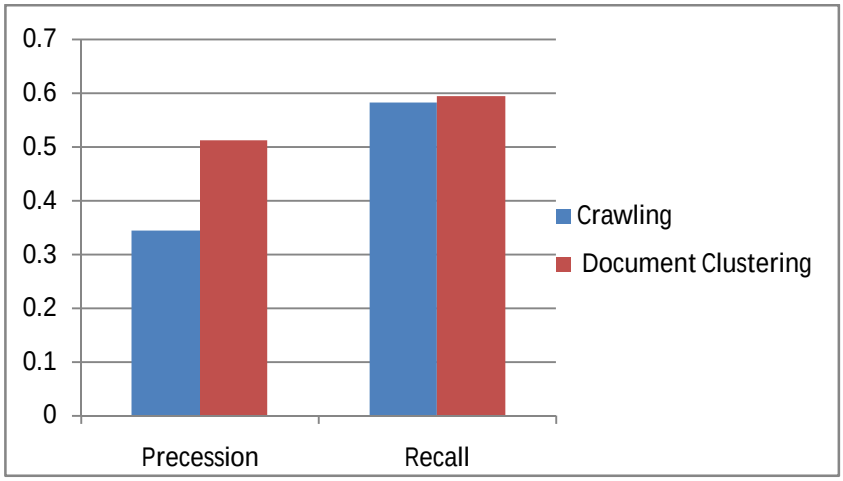
Precession=#relevant_items_retrieved /#retrieved_items
Recall(R) is the fraction of relevant documents that are retrieved
Recall=relevant_items_retrieved /relevant_items

All the stages of the proposed model has been implemented and executed by taking 10 queries and on 200 no of documents. TABLE I contains comparison result. The result shows that precision and recall value of the proposed work is increased. This means proposed model is more efficient to retrieve the relevant documents:

Model	Precession	Recall
Crawling	0.34435	0.583442
Document Clustering	0.51222	0.594534

Table 1 : Comparison results based on precision and recall depicting time (in milliseconds) took for precession recall over the regression and classification time over crawling and document clustering

Graphical Comparison of Result : The graph 1 shows scheme proposed model works in as under with respective results gained using precession recall



Graph 1: Bar Chart depicting Comparison results based on precision and recall vide time (in milliseconds) took for precession recall over the regression and classification time over crawling and document clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

VI. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better with the total precision recall mechanism. The proposed algorithm provides effective and productive scenarios for focus crawling and enhances the appropriate retrieval of information using respective query vide document clustering. In design considerations the performance of the proposed algorithm can be compared with other generic algorithms and the proposed scheme depicts the effective and appropriate results and information. However, for the future work the same can be implemented using hadoop or parallel computing with Map and Reduce along-with fully automated model and even using certain ontology parameters the respective query can correlate the document relation or features patters in specialized manner and can defined automation for the same for prompt document reference and retrieval of information from social media world.

REFERENCES

1. Cliff Lampe, Nicole Ellison, and Charles Steineld. A face (book) in the crowd: Social searching vs. social browsing. In Proceedings of the ACM Special Interest Group on Computer-Supported Cooperative Work, 2006.
2. Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In CHI '10: Proceedings of the 28th international conference on Human factors in computing systems, pages 1185{1194, New York, NY, USA, 2010. ACM
3. Uldis Bojars, Alexandre Passant, John Breslin, and Stefan Decker. Social network and data portability using semantic web technologies. In Proceedings of the Workshop on Social Aspects of the Web, 2008.
4. Petter Brandtzaeg and Jan Heim. Why People Use Social Networking Sites. In Ant A. Ozok and Panayiotis Zaphiris, editors, Online Communities and Social Computing, volume 5621, chapter 16, pages 143-152. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
5. Francesca Carmagnola, Fabiana Vernero, and Pierluigi Grillo. Sonars: A social networks-based algorithm for social recommender systems. In Proceddings of the 17th International Conference on User Modeling, Adaptation, and Personalization, 2009
6. Michael Chisari, The future of social networking. In Proceedings of the W3C Workshop on the Future of Social Networking, 2009
7 . Rohani, Vala Ali; Ow Siew Hock (2010). "On Social Network Web Sites: Definition, Features, Architectures and Analysis Tools".Journal of Advancesin Computer Research (2): 41-53. Retrieved 2011-06-11.
8. Heckmann, D., Schwarzkopf, E., Mori, J., Dengler, D., & Kroner, A. (2007). The user model and context ontology gumo revisited for future web 2.0 extensions. In Proceedings of the international workshop on contexts and ontologies: Representation and reasoning
9. Erétéo, G., Buffa, M., Gandon, F., Leitzelman, M., & Limpens, F. (2009). Leveraging social data with semantics. In Proceedings of the W3C workshop on the future of social networking.
10 Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending content from information streams. In CHI '10: Proceedings of the 28th international conference on human factors in computing systems (pp. 1185–1194).
11. S. Catanese, P. De Meo, E. Ferrara, and G. Fiumara. "Analyzing the Facebook Friendship Graph." In Proceedings of the 1st Workshop on Mining the Future Internet, pages 14 - 19, 2010.
12 Resnick, P., Lacovou, N., Suchak, M., Bergstrom, P., Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the ACM conference on computer supported cooperative work