Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/4589
DOI URL: http://dx.doi.org/10.21474/IJAR01/4589
ISSN NO. 2320-5407

RESEARCH ARTICLE

QUERY INFORMATION RETRIEVAL PROCESS IN SOCIAL MEDIA.

Simran Goyal and *Charu Pujara.
Department of Computer Science and Engineering Manav Rachna International University, Faridabad-121004, India.

Manuscript Info

Manuscript History

Received: 24 April 2017
Final Accepted: 26 May 2017
Published: June 2017

Abstract

Now days, it is difficult for the users of social network to keep a record of the social friendships and different activities amongst multiple networks as the huge amount of social data is available. A user-centric system is desired that has the ability to aggregate all the social data from different SNSs and allowing users and the system to resolve the user queries with high precision. This paper suggests an open vision of challenges faced to design an intelligent social network database system. The system suggests the set query language related to the social networks as naive building blocks, and the importance of an intelligent machine learning process as the retrieval processor to supervise and improvise upon the user's implementations.

Introduction:-

Social Networks like Facebook, Bebo, Delicious etc. are popular networks that provide web services to the users. Users create profiles, publish data by Sharing, scrapping, tweeting etc. interact among others and generate online content on the network [1]. Now days, users communicate, connect and collaborate using services of social network. Numerous researches have been done to study the structure of social networks and its properties. User can represent his query in a formal language and knows the output of the query but will fail to represent the best solution and rank the users. It is beyond the capability of a user or an expert of the domain to syntax the machine learning process and set keywords for the most accurate and precise output [2].

Social networks include a large amount of dataset with the exponential rise in the number of users. Every Social network has its own protocols and query language to provide the expected results to the user resourcefully and precisely. Facebook has developed FQL (Facebook Query Language) [5] which is a similar language to SQL. Twitter provides user search based on words, particular people, places, and adding further properties related to tweet. LinkedIn provides job and functionalities to search based on people. Each of the specified social networks is restricted to their own network and provides specialized solution to the social network. To process a query on the social network is a major challenge in the industry as well as in academia. There are tools available to analyze the user's network, closeness, clique and other properties of the network. Social Network can be seen as graphs and specialized databases are designed for querying graph based structure [3]. It is evident from the research that there is a need to provide an intelligent Social network database which has the ability to query the graph based database and analyze the properties of nodes.

The graph model needs a language to process the query efficiently. A new language for the queries, which can be understood by them better has been introduced. One such language that has been discussed in this paper is SociQL,

Corresponding Author:- Charu Pujara.
Address:- Department of Computer Science and Engineering Manav Rachna International University, Faridabad-121004, India.

a query language designed to answer relevant queries in social-network analysis and also to collaborate data from different networks to a common network, that is, the role of a social network aggregator. Alex Patriquin of Compete.com reported on the overlapping of users having accounts at various online social network services

A 2009 study of 11,000 users reported that the majority of MySpace, LinkedIn, and Twitter users also have Facebook accounts [13].

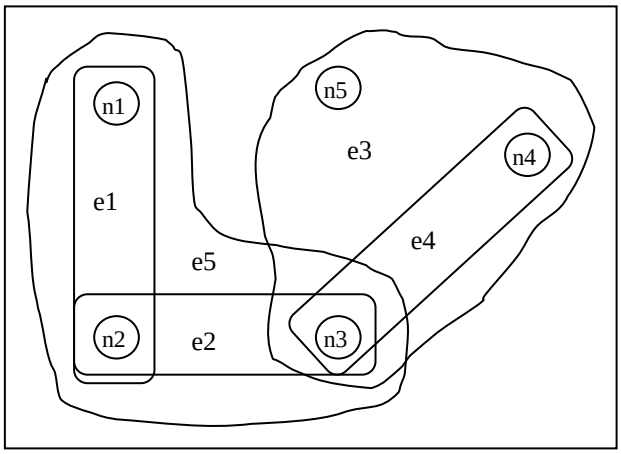
Table 1:- Survey of user having account at multiple sites

Site	LinkedIn	Facebook	Friendster	Bebo	Orkut	Plaxo	Ning	MySpace	Hi5
LinkedIn	100	42	8	4	3	3	8	32	2
Facebook	2	100	2	4	1	9	1	64	2
Friendster	6	23	100	5	1	0	2	49	4
Bebo	1	25	2	100	0	0	1	65	3
Orkut	8	26	4	3	100	1	2	29	7
Plaxo	54	48	8	5	4	100	14	34	2
Ning	19	35	6	6	2	2	100	44	1
MySpace	0	20	1	3	0	0	0	100	1
Hi5	1	24	4	7	2	0	0	69	100

This paper is divided into four sections: Section 1 gives an introduction; section 2 discusses the Related Work. Section 3 throws light on major challenges in designing an intelligent social network database. Section 4 concludes the paper.

Related work:-

A social network can be represented as a collection of nodes and edges where each user represent a node and an edge between two nodes represents a relation like friends or followers depending upon the protocol of social network or group of people. The nodes can have various attributes like user name, display name, location etc. Edges are hyper-edges to easily represent groups of people and the interactions between them [4]. In a weighted graph, the system may generate the weight as per the interaction of the user to the group or another user. The edges can also be linked with the public and private attributes of the user and its corresponding value [5]. Queries refer to the structure of graph and query processor refers to the engine which will extract the result for the database as per the desired interest of the user with the defined keywords of the network hiding the information of the underlying data storage [6]. The notion of social network and the methods of social networking analysis has been a huge challenge. In the current scenario, the autonomous social units and the relational ties between these units for transfer or flow of resources, are the two major elements that have been identified for such issue.



Nodes:-

n1	Name = Henry Age = 45
n2	Name = Ben High-School = BAHigh
n3	Name = Candy
n4	Name = Bob Age = 32
n5	Name = Ella Hobby = Painting

Edges:-

e1	({n1},{n2})	Label = follows
e2	{n2,n3}	Label = friends
e3	{n3,n4,n5}	Label = co-authors Title = Fun Maze ISBN = 23415
e4	{n3,n4}	Label = friends
e5	({n3},{n1,n2)}	Label = email Content = Check out my resume…

Suppose two nodes u and v that are adjacent in graph G, the events in which u becomes active are correlated with the events of v being active.

There are two explanations that tell about this social correlation- one is influence and the other is homophily.
Influence refers to the change done by the social context to a person's behavior. In a social network, it can be recognized as the tendency of a group of people to exhibit similar behavior to the source of influence. Like, when a poet uses a keyword because one of his/her friends have recently adopted it. Homophily refers to the tendency of individuals to associate and bond with similar others. Individuals in hemophilic relationships share common characteristics (values, behaviors, etc.) that make communication and relationship formation easier. It also takes into account the external influence of elements in the environment, such as family ties and geography [5].

For an example, there are four poets with their publications (poems) and their affiliation to different institutions. At the same time, the poems are linked to the convocation where they were presented and also the keywords contained in the publication.
Now in SQL, the name of the poets can be retrieved who have collaborated their poems with a given poet and the corresponding poems.

Q1: SELECT writes (m1, n1)
FROM poem n1, poet m1, poet m2
writes (m1, n1), writes (m2, n1)
WHERE m2.name = 'Sam' AND
n1.title>< 'DREAM'

The main construct provided but this query is 'select-from-where'. In Q1, we use m2 to denote Sam as our focal actor, and then we form a network in the FROM with two poets (m1 and m2) and a poem in common (n1). The result of the query is composed by the network formed by poems (n1) and poets (m1), linked by the 'writes' relationship, after applying all the predicates in the WHERE.

SociQL [3] is a simple model that has been developed for capturing the semantics of the information contained in, and extracted from, social networks. A social network can be defined as a 4-tuple (O, R, PO, PR).
O is an abject set, which is a set of social objects.
R is a relation set, which is a set of links between the objects in O that represents the flow of information or materials, like the relation 'writes' or 'affiliates'.

PO and PR are the properties that define every object and relation respectively, such as the title and the year of publication for a paper.

Once a query is received from an application (such as a query editor), the analyzer converts the query into a sequence of objects, that then is analyzed by the parser (a program to analyze these objects) to determine the validity of the grammatical structure of the query and creates an internal representation of it. Then, an execution plan is found for the query, specifying the order in which the statements are executed, and, specially, the order in which data from external sources will be requested. At the query processing step, all external data (if any) used by the query is fetched and stored locally first, then the actual SociQL query is translated into an SQL expression that is executed on the local relational database containing all data needed by the query. Finally, an additional step may be required, when the ORDER BY clause is present, the objects in the query are arranged according to their importance. The query planning step is concerned with identifying the ordering in which data is requested from external social networking sites (which are used to interlink relations). The goal is to find a strategy to answer queries that minimizes the amount of external data that is retrieved.

The first step attempt to define the query at the lowest level, in other words, special constructs like semantic network elements or path relationships will be translated to simple elements. Once we have all the elements at the lowest level, it is time to plan for the path relationships. For queries over path relationships, the planner performs a breadth first search on every path relation, up to a maximum number of levels defined in the query. Then, every path will create an independent query. Basically, the planner attempts to find the most cost effective plan, by examining the costs of the various sub queries. Local queries, for example, queries directly translatable to SQL and issued to a database on the same intranet are assumed to be the least expensive. Since local queries are most preferable, the planner makes a sub query with only the local actors, in order to get an initial subset of possible results and, based on it, to narrow the queries to the external resources.

Numerous attempts have been made to model the database, each of which has its own underlying theory related rules, principles, terms and level of development. However, according to study, only few have a potential to stand by the requirements of online social networks database [3]. A relational data model is not suitable for representing a relation that exists between users of the online social network. A graph is a noble approach to represent the interrelationships over a set of data entities and can be taken as a conceptual tool to represent network related data. Graph theory plays an important role in many areas like databases, computer science and other disciplines. Generally, graph models are motivated by real-life applications where information about data inter-connectivity is more important as the data. Therefore, despite the wealth of social network structures and analysis, there is still a need for new designs, methods and specifically data management systems. Graph data models and related query technologies [1] offer notable advantages to discover relationships in huge data sets and can be the base for many new functions which can be used for the future online social network efficiently.

Dries *et. al.[11]* associates the analyzing , clustering and querying proficiency of social network. SoQL [3], SociQL[4] is grounded on relational tables majorly focusing on centrality measure of social network. SNQL [5] is founded on GraphLog considering querying and manipulation of data. SocialScope[11] ranks the output of the query on the communal associations.

Major Challenges:-

The high level of personal information of a user and the inter-communication between the user can be extracted from the social network. Structuring graphs as a theoretical graph data model defines a useful database. Graphs are the foundation in mathematics and computer science. Graph theory is supported by a huge amount of formal study and analysis. The system can use effective graph related algorithms which are specially designed to utilize the discussed graph data structures. Graphs can represent online social network services easily and the graph databases are sufficient to meet the present and the future needs for social web [2]. A database system that can answer the query written in a natural language for a social network is extremely challenging. The database systems should be capable of managing, processing, and analyzing complex, heterogeneous, temporal and voluminous graph-structured data. A more flexible and dynamic system is required which provides the applications and user to augment novel facts and associations as they want. Designing of a new system that can handle the voluminous data and manage the heterogeneous resources effectively in online social network requires reconsidering about all aspects of a DBMS, including data modeling, storage management, indexing, and query processing and optimization [2].

Some of the main issues that can serve the answer of the user's query specified in natural language are listed below:
1. Semantic Query Mapping: Mapping the semantic meaning of the user's query to the attributes of the social network is a major task. Researchers have done study to map the semantic meaning to the query but to map the meaning of an informal language and slangs that are used in the social network is a challenging task. User on the network post about their birthdays, events, daily happenings etc.. Moreover, on the social network user uses short context to convey their message or tweets. There is a need of a system that can modularize the intention of user mentioned in the short context. The system that process the query needs to understand the intention of the user and also able to identify the false information provided by the user. The system also needs some machine learning mechanism as well as intelligence to compute the query efficiently. Several work have already been done which studied the systematic implementation within a social network [7][8][9][10] but none of the solution is viable which can process a query in natural language.

2. Machine Learning Process: It has been evident that there is a need of an optimal operational leaning process and accomplishing it effectively in a query processor.

Table 2:- Comparison of different social networks

Networks→ Features↓	SocConnect	Flock	Xeeme	HootSuite	People Aggregator
Analytics	No	No	No	Yes	No
Scheduling	Yes	No	yes	Yes	Yes
Team Collaboration	Yes	Yes	yes	Yes	No
Group friends	Yes	No	No	No	No
Rating of activities & Friends	Yes	No	No	No	No

Providing an extra bit of information to the user other than the expected output is another major challenge. User's previous results feedback for a particular problem and improving the results for the current problem is another important concern. It is important to cache the results of the query for other users to save time, memory and machine learning. Graph Query processing, pattern matching and graph indexing are to design in order to achieve more novel approach to map the problem of processing pattern of social network.

3. Privacy: With the privacy concern of users, online social networks have given rights to the user to restrict the visibility of the information to friends, friends of friends, group, all or none. Attackers can relate the information available on the social network and can extract some meaningful information about the user. There is a need to develop a better privacy mechanism that can restrict the usage of valuable information about the user by others. Developing a formal solution to this privacy problem is a bit difficult with considering all the built-in functions in use without exposing private information [5] of the user.

4. Query Evaluation: Due to the enormous size of social network and the perpetual growth of the embryonic network, there is a need of an operative mechanism to demonstrate the closeness of two nodes, measuring centrality etc.

5. Query optimization: To provide an optimal, flexible dynamic and minimal time for executing a query is another important task.

6. Advanced features: The hypothetical questions, management of time and sources within a network, considering collaboration into the query language and updating the coordination of the networks is another hurdle to cross over. Displaying effective results of different statistics over the network is another challenge to face.

There are social network aggregators available that integrates the services provided by the multiple online social networks like Hoot suite, flock etc. There are various solutions available to integrate the social network but no one has tried to integrate the information available within multiple social networks. None of the aggregator has mined the multiple social networks and extracted some useful information after collecting data from different Social Networks in a natural language. A Comparative study of various services of social network aggregators is explored and presented in table 2.

Flock can get updates from friends, status updates and photos submitted at Multiple Social networks whereas SocConnect users can creates a personalized social and semantic contexts for their social data. Users can combine and cluster friends and rate Network. HootSuite aggregates organizations and businesses to collaboratively execute promotions across multiple social networks and XeeMe Organizes Social presence, discovers new network and people. It organizes the entire social presence of the user, determine new networks and people and develop their presence and influence. But none of the aggregator has mined the multiple social networks and extracted some useful information after collecting data from different Social Networks. All the social network aggregators discussed can integrate Facebook, LinkedIn and Twitter accounts but none of the network can handle the query written in a natural language.

Conclusion:-

User Personalization and user interaction is a high level of information that can easily be extracted from the social network. No single model of data exists which is appropriate for all the users and problem domain. Currently, SociQL allows expressing many useful queries, but there are several enhancements that could improve the applicability of the language. Path expressions used in SociQL queries are useful for general problems, but it is difficult to define complex regular expressions over the structure of the network. Thus, considering the modern nature of communication of users via social network, there is a need to build a more realistic query processor that can answer the user's query in a natural language and overcomes the traditional strategies to adopt the high level of user's communication. Graphs are the best solution to represent the complicated data structure of the online social network. In this paper, we have discussed the major issues to model the intelligent database to provide a novel solution to the user's query.

Acknowledgement:-

Authors would like to express the gratitude to the Research Mentors of Accendere Knowledge Management Services Pvt. Ltd..

References:-

1. Cohen, S., Ebel, L., & Kimelfeld, B. (2013). A Social Network Database that Learns How to Answer Queries. In *CIDR*.
2. YANG, S. (2011). *Data Modeling and Query Processing for Online Social Networking Services* (Doctoral dissertation).
3. Ronen, R., & Shmueli, O. (2009, March). SoQL: A language for querying and creating data in social networks. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on* (pp. 1595-1602). IEEE.
4. San Martın, M., Gutierrez, C., & Wood, P. T. (2011). SNQL: A social networks query and transformation language. *cities*, *5*, r5.
5. Serrano Suarez, D. F. (2011). SociQL: a query language for the social web.
6. Holland, P. W., & Leinhardt, S. (1977). Social Networks: A Developing Paradigm.
7. Zou, L., Chen, L., & Özsu, M. T. (2009). Distance-join: Pattern match query in a large graph database. *Proceedings of the VLDB Endowment*, *2*(1), 886-897.
8. Williams, D. W., Huan, J., & Wang, W. (2007, April). Graph database indexing using structured graph decomposition. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 976-985). IEEE.
9. Tong, H., Faloutsos, C., Gallagher, B., & Eliassi-Rad, T. (2007, August). Fast best-effort pattern matching in large attributed graphs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 737-746). ACM.
10. Mitra, S., Bagchi, A., & Bandyopadhyay, A. K. (2009). Design of a data model for social network applications. *Database Technologies: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, *1*.
11. A. Dries, S. Nijssen, and L. De Raedt. A query language for analyzing networks. In CIKM, 2009.
12. Staworko, S., & Wieczorek, P. (2012, March). Learning twig and path queries. In Proceedings of the 15th International Conference on Database Theory (pp. 140-154). ACM.
13. Cohen, S., & Cohen-Tzemach, N. (2013, June). Implementing link-prediction for social networks in a database system. In Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks (pp. 37-42). ACM.
14. Amer-Yahia, L. V. S. Lakshmanan, and C. Yu.Socialscope: Enabling information discovery on social content sites. In CIDR, 2009.
15. Cli 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.