International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.2, Issue.4, July-Aug. 2012 pp-2839-2842 ISSN: 2249-6645

Detecting Identification Anomalies in Social Networking with Cluster based re-ranking and Slink Algorithms

Salina Adinaryana[1], Dr. G. Samuel Vara Prasada Raju[2] , Allam Mohan[3]
*(Assoc Prof, Dept of IT, Sri Vishnu Engineering College for Women, Bhimavaram, A.P, India)
**(Professor, Dept of CSE, SDE, Andhra University, Visakhapatnam, A.P, India)
***(Asst. Prof, Dept of IT, Sri Vishnu Engineering College for Women, Bhimavaram, A.P, India)

ABSTRACT: *In today's Fast growing commercial word Social network Websites (SNW) like FaceBook, Twitter etc, are the major source for maintain social communication, doing e-business. Now a day's dynamic data discovery is a part of it's activity for many organizations in banking sector, content providing sites and ecommerce websites. Interestingly on the other side we can see that most of the users are having more than one registration on social networks with two fold effect both positive and negative.*
Many students and young people are using these social networking websites for adolescents or to download some study materials, music etc[10,11]., from different blogs. This leads to lot of fake user registrations[7]. Just receiving invitation to their emails, creates some of the user profiles and it pops the user data of the mail account. The profile created may not be correct and this was copied may lead to false or duplicate information on the network domain. This became a bitter problem to find many facts like user identity, number of users registered and visitors tracking etc. there by it leads to ambiguity.
As one person may have more than one profile created by him intentionally or unknowingly generated by anyone way specified above. This became a common identification problem to every organization and most researchers are mainly concentrating on providing efficient techniques for linking of records. There are many solutions of this kind, to propose the best we are attempting to analysis and identify the optimization techniques with clustering and algorithms for fake user data discovery in SNW.

Keywords: *Social network Web sites, social communication, PAN, query data base, cluster based analysis, SLINK, MST*

I. INTRODUCTION

Record linkage for remote database is a commonly identified problem from many years, which is used to analyze remote data supporting a variety of decisions in different organizations. How ever, since heterogeneous databases are usually designed and managed-where all the records are available either locally or remote by the organizations. They can be accessed by using some common key like email address or PAN number etc.. Although it may be possible to use they would involve transferring the entire remote relation. As a matching data from various sources in a batch. The key question here is one of record. The databases exhibiting entity heterogeneity are distributed, and it is not possible to create and maintain a linkage: given a record in a local database (often called the central data repository or warehouse where pre computed enquiry record), how do we find records from a remote linkage results can be stored. A centralized solution database that may match the enquiry record? Traditional maybe impractical for several reasons. First, if the record linkage techniques, however, are designed to link databases span several organizations, the ownership and cost allocation issues associated with the ware-house. Even if the warehouse could be developed, it would be difficult to keep it up-to-date. As updates occur at the operational databases, the linkage results would become stale, if they are not updated immediately. This staleness maybe unacceptable in many situations. The systems may agree to transmit incremental changes to the data warehouse on a real-time basis. Even if such an agreement is reached, it would be difficult to monitor and enforce it. The staleness of the linkage tables and limiting their usefulness due to certain overhead delays and time consuming due to many databases, each with many records, undergoing real-time changes. This is because the warehouse must maintain a linkage table for each pair of sites, and must update them every time for the associated databases changes.
. The participating system allows controlled sharing of portions of their databases using standard database queries, but they do not allow the processing of scripts, stored procedures, or other application programs from another organization. Here our intention is not to discuss on ability of existing systems but to suggest a system that is more efficient for record linkage with clustering and MST.

1.1 INITIATION

When I am searching for one popular personality on twitter, linked-in and FaceBook[8]. This situation Initiated me to concentrate on this work. In this work I identified problems like: (1) Identifying a person having more than one account (2) real active identity. These things encouraged me to identify a system that can determine user genuinely. And also I understood that most of the users are not in a position to readily available to project their correct information onto SNW and many youngsters may have more than one account for different purposes. So how do we get the exact user data? This gave a way to propose a solution of that kind where both the end users will be satisfied. As a part of this process I had come across many similar problems like e-banking, downloading e-books, ecommerce with SNW login etc[11,12]., and understood that there is a need for better system to handle user data.

In the Data discovery process, cluster analysis has been used to create groups of documents with the goal of improving the efficiency and effectiveness of data retrieval. The terms in a document collection can also be clustered to show their relationships.

A recent review (Willett 1988) provides a comprehensive summary of research on term-based document clustering [2]. Terms may be clustered on the basis of the documents in which they co-occur, in order to aid in the construction of a thesaurus or in the enhancement of queries (e.g., Crouch [1988]). If the collection to be clustered is a dynamic one, the requirements for update must be considered.

II. RELATED WORK

This paper consists of two parts where first part is addressing the efficiency and effectiveness of data retrieval using database re ranking based on cluster analysis, and then algorithm best fit to implement the system and finally results base on the system implemented to check the feasibility.

2.1 MATHEMATICAL APPROACHES

In this part, my approach is to identify and avoid fake users on social network by 1.) Database re-ranking based on cluster analysis[1] and 2. Record linkage with Slink.

As the degree of matching of evidences in data bases is higher, the two database entries are more similar. In database clustering, databases with similar entries or more likely are classified as one cluster[4]. Therefore, relevant db table entries are in the same cluster according to the cluster hypothesis (van Rijsbergen) which states that relevant documents tend to be more similar to each other than to non-relevant documents.

The documents in a cluster have effect on cluster centroid. The cluster centroid for a pair of clusters Ci and Cj is given by:

$$\frac{miCi + mjCj}{mi + mj}$$

where m is the size of a cluster.

The same query-cluster similarity value is applied to all the database tables in the cluster at the re-ranking stage. In this way, the databases in a cluster can affect one another through calculation of cluster centroids so that context retrieval is possible, due to the interaction of evidences contained in databases.

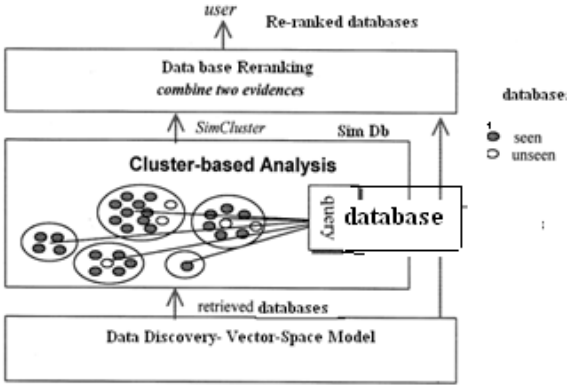

Fig 1.Clustering based data base re-ranking

In record linkage process the single link method merges at each stage the closest previously unlinked pair of points(here records) in the data set. Since the distance between two clusters is defined as the distance between the closest pair of records each of which is in one of the two clusters, no cluster centroid or representative is required, and there is no need to recalculate the similarity matrix during processing. This shows how clustering efficiency. So that this makes the method more attractive for non redundant computation[4] and a storage perspective wise. Database re ranking avoid similar entries in the clusters and Slink will improves record linkage of data discovery process in Vector Space model.

A number of algorithms for the single link method have been reviewed by Rohlf (1982), including related minimal spanning tree algorithms. The computational requirements range from $O(NlogN)$ to $O(N^5)$. Many of these algorithms are not suitable for information retrieval applications where the data sets have large N and high dimensionality.

2.2 ALGORITHMIC APPROACH

2.2.1. VAN RIJSBERGEN ALGORITHM

Van Rijsbergen (1971) developed an algorithm[5] to generate the single link hierarchy that allowed the similarity values to be presented in any order and therefore did not require the storage of the similarity matrix. It is $O(N^2)$ in time and $O(N)$ in storage requirements. It generates the hierarchy in the form of a data structure that both facilitates searching and is easily updated, and was the first to be applied to a relatively large collection of 11,613 documents (Croft 1977).

2.2.2 SLINK ALGORITHM[5]

The SLINK algorithm (Sibson 1973) is optimally efficient, $O(N^2)$ for computation and $O(N)$ for time, and therefore suitable for large data sets. It is simply a sequence of operations by which a representation of the single link hierarchy can be recursively updated; the dendrogram is built by inserting one point at a time into the representation.

The hierarchy is generated in a form known as the pointer representation, which consists of two functions Π and Λ for a data set numbered 1..N, with the following conditions:

- $\Pi(N) = N$
- $\Pi(i) > i$
- $\Lambda(N) = \infty$
- $\Lambda(\Pi(i)) > \Lambda(i)$ for $i < N$

In simple terms, $\Lambda(i)$ is the lowest level (distance or dissimilarity) at which i is no longer the last (i.e., the highest numbered) object in its cluster, and $\Pi(i)$ is the last object in the cluster it joins at this level; a mathematical definition for these parameters is provided by Sibson (1973).

In the pseudo code for SLINK below, three arrays of dimension N are used: *ptr* (to hold the pointer representation), *dp* (to hold the distance value associated with each pointer), and *distance* d(to process the current row of the distance matrix); *next* indicates the current pointer for a point being examined.

```
ptr [0] = 0;
dp[0] = MAXINT;
/*iteratively add the remaining N-1 points to the hierarchy
*/
for (i = 1; i < N; i++)
{ ptr [i] = i;dp[i] = MAXINT;
/* calculate and store a row of the dist_matrix for i */
for (j = 0; j < i-1; j++) d[j] =calc_distance(i,j);
for (j = 0; j < i-1; j++)
{ next = ptr[j];
```

```
if (dp[j] < d[j])
d[next] = min(d[next],d[j]);
else
{ d[next] = min(dp[j],d[next]);
ptr[j] = i;dp[j] = d[j];}}
/* relabel clusters if necessary */
for (j = 0; j <i-1; j++)
{ next = ptr [j];
if (dp[next] < dp [j])
ptr[j] = i;
}}
```

For output in the form of a dendrogram, the pointer representation can be converted into the *packed representation*. This can be accomplished in $O(N^2)$ time (with a small coefficient for N^2) and $O(N)$ space.

2.2.3 MINIMAL SPANNING TREE ALGORITHM

Another attempt is with an MST. A minimal spanning tree (MST) which is a sub graph(known as tree here) generated from the given graph linking N objects with N - 1 connections without any cycles and the sum of the N - 1 dissimilarities is minimized. It can be shown that all the information required to generate a single link hierarchy for a set of points is contained in their MST (Gower and Ross 1969). Once an MST has been constructed, the corresponding single link hierarchy can be generated in $O(N^2)$ operations; or the data structures for the MST can be modified so that the hierarchy can be built simultaneously (Rohlf 1982).

The Prim-Dijkstra algorithm[5] (Dijkstra 1976) consists of a single application of principle 1, followed by N - 1 iterations of principle 2, so that the MST is grown by enlarging a single Cluster: Let records unique values(record key) are taken as points for MST.

1. Place an arbitrary record in the MST and connect its nearest neighbor to it.
2. Cluster will grow in size by finding the record not in the MST, closest to any record in the MST and add it to the cluster.
3. If a record remains that is not in the cluster, return to step 2.

Prim-Dijkstra algorithm is provided by Whitney (1972). The algorithm here uses arrays *npoint* and *ndistance* to hold information on the nearest in-tree neighbor for each record, and *notintree* is a list of the *nt* unconnected records. *Lastestpt* is the latest record added to the tree.

```
/*  initialize lists */
/*define  infinity =9999 */
for (i = 0; i < n; i++)
{ ndistance[i] = infinity; notintree[i] = i;}
/* arbitrarily place the Nth point in the MST */
latestpt = n; nt = n-1;
/*  grow the tree an object at a time */
for ( i = 0; i < n-1;i++)
{
/*consider the latestpt in the tree for the NN list */
for (j = 0; j < nt; j++)
{ D = calculate_distance(lastpoint, notintree[j]);
if  (D < ndistance[j]
{  npoint[j] =lastestpt;
ndistance[j] = D;}}
/* find the unconnected point closest to a point in the  tree
*/
nj = index_of_min(ndistance);
/* add this point to the MST; store this point and their
clustering level      */
lastpoint = notintree[nj];
store_in_MST ( lastpoint, npoint[nj], ndistance[nj]);
/* remove lastpoint from notintree list;  */
/* close up npoint and ndistance lists    */
notintree[nj] = nt;
npoint[nj] = npoint[nt];
ndistance[nj] =ndistance[nt];
nt = nt - 1;}}
```

2.3 PROPOSED SYSTEM

The clustering architecture was designed and implemented with an assumption that it will be invoked, when a user creates a registration on the SNW it automatically starts monitoring user inputs by giving initial ranking and maintains a dataset at a **query database** as a pivotal elements like PAN, Adhar card and Driving license which are issued by government departments and had a unique id number., once the data is inserted it considers the PAN number or email-ID and mobile number as set of keys to identify the user and data is mined from different databases available on internet and they are maintained as different clusters. Finally the query database is re ranked done using cluster analysis which is show in fig 1.

III. RESULTS

After the user gets registered in SNW

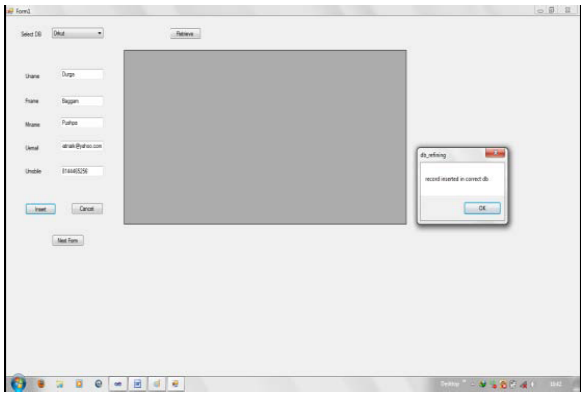

Fig 2.user details inserted in Query DB

The details gets inserted in the query database and initial ranking is given for the records which is show in Fig.2. while querying user data from **query db** for user records similarity checking to form clusters the data queried is shown in fig.3.

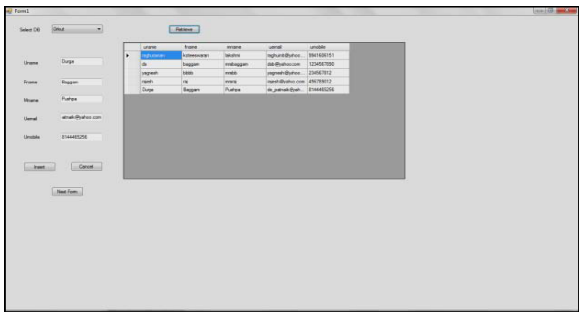

Fig.3 queried user data for clustering

IV. CONCLUSION

Research on social networking and their implications are attracting more researcher's because it is a social area depends on behavior of large set of people in different age groups of different demographic sizes and their activities result in different areas like dynamic change in Database size, complexity in analysis for OLAP, redundancy and many more DB issues. Some of the most interesting factor that motivated me is to identify a persons by avoiding unauthorized entry for adolescent action and to provide security for e-transaction associated with banking[8] and ebay, snap-deal and flipkart[9] etc., like shopping websites.

This is our preliminary work to have a better understanding of the system. As a result we had studied many papers related to it and accessed to many SNW's by registering on entering fake user data interestingly found that most of the sites are not processing any verification other than email-id which can be created by anyone as they like by giving any details(invalid or fake). We proposed a clustering architecture which acts as a preliminary gateway for any user registration by giving initial ranking. The complexity here is all registration pages should undergo this process. Future enhancement for this technique is to implement a internal indexing with R-trees for query database so that retrieval from query database becomes faster so that it may improves the overall performance.

REFERENCES

[1] Kang, H. K. (1997). Two-level document ranking methods using mutual information in natural language information retrieval. Ph.D. thesis, Department of Computer Science, Korea Advanced Institute of Science and Technology. Kemong (1992). The Kemong Company new encyclopedia. Seoul: Kemongsa Publishing Co.

[2] Cluster analysis- http://en.wikipedia.org/wiki/Cluster_analysis

[3] Park, Y. C. (1997). Building word knowledge for information retrieval using statistical information. Ph.D. thesis,Department of Computer Science, Korea Advanced Institute of Science and Technology.

[4] Using Interdocument Similarity Information in Document Retrieval Systems Alan Griffiths, H. Claire Luckhurst, and Peter Willett Department of Information Studies, University of Sheffield, Western Bank, Sheffield SIO 2TN, United Kingdom

[5] Slink Algorithm http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap16.htm [6]. twitter user registration - http://twitter.com/#!/

[7] Face book login- https://www.facebook.com/login.php

[8] Face book login for database transactions http://www.facebook.com/pages/State-Bank-of-India/104565649587176

[9] Face book shopping http://www.facebook.com/eGardenOnlineShopping

[10] mp3skull.com/mp3/4shared_com.html

[11] Free ebooks download www.**facebook**.com/pages/Free-**download-books**/128576447214493