CHAPTER 5

SOCIAL MEDIA ANALYSIS GROUPING ALGORITHM

The Social Media is a readymade data bank for companies looking to learn more about customers. It is a media to spread messages or feedbacks. User interactions also help discover communities, where their links are based on the user’s contacts. This chapter presents a novel data mining technique for grouping similarity in people’s thoughts called the Social Media Analysis Grouping Algorithm (SMAGA) using social databases.

5.1 INTRODUCTION

Social media mining can provide data to analyze. Studies have also conducted based on user likes or dislikes. Social media analysis is being taken seriously by companies to understand their customer preferences or detect potential problems in business. The tools needed to recognize requires keywords in posts or comments on social networking sites and their analysis remains an inexact science. The results can be distorted due to spelling errors, thus making interpretation difficult. References to allied dictionaries are needed to overcome this complexity. A social network community is a set of users with similar criteria. A community is also defined by a group with similar dislikes or likes. The highest level of data abstraction is the view level. A view is a way of presenting data to particular group of users. Social network community study is on the rise. Community detection approaches based on like Node, Group, Network and Hierarchy have been used [55]. Analyzing similar users or grouping them based on likes and dislikes gives a better vision, thus leads to many applications. Users recommend or share their thoughts in their links. The links are also used to answers questions posed by users for solving problems. Tweet texts can be used to filter the required information based on any criteria.
5.2 GROUPING DATA

Data can be grouped by selecting a group category at the top of the report. For example, Group results by query strings that users searched for on Google. Only searches that returned the specified site will be included. The Query list can be reviewed for expected keywords. When expected keywords don't appear, it implies a site might not have enough useful content relevant to those keywords. Group results by individual page on a property returned by search results can change the metric calculation. Results of a web page can also be grouped by Country or device. The data between two exact values in any grouping category can be compared. Some useful comparisons are sorting by difference to see queries with significant new activity and comparing total searches on the site to mobile searches. For example search results for "pets" returns only the following three results, all from the same site, and that users can click each of them with equal frequency:

<table>
<thead>
<tr>
<th>Google Search Results</th>
<th>Metrics Aggregated by Site</th>
<th>Metrics Aggregated by Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.pets.com/Dogs">www.pets.com/Dogs</a></td>
<td>Click-through rate: 95%</td>
<td>Click-through rate: 38%</td>
</tr>
<tr>
<td><a href="http://www.pets.com/Cats">www.pets.com/Cats</a></td>
<td>All clicks for a site are combined</td>
<td>5 pages shown, 1/5 of clicks to each page</td>
</tr>
<tr>
<td><a href="http://www.pets.com/Parrots">www.pets.com/Parrots</a></td>
<td>Average position: 2</td>
<td>Average position: 4</td>
</tr>
<tr>
<td></td>
<td>Highest position from the site in the results</td>
<td>(2 + 5 + 5) / 3 = 4</td>
</tr>
</tbody>
</table>

**Drawbacks in K-means Grouping Algorithm**

- One of the main disadvantages of K-means is the specification of clusters as an input to the algorithm. Due to its design K-means cannot determine the number of clusters and depends upon the user to identify this in advance.

- K-means does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
• It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance.

• Another disadvantage is the requirement for the concept of a mean to be definable which not the case is always.

5.3 PRE-PROCESSING SOCIAL MEDIA TEXT

The data extraction technique uses at least one structured data in the mining process. The information collected gives the minimum accuracy of result when similar data is collected together. Data’s annotation and alignment are the major issues that need proper treatment for accuracy in results. While aligning data some labels may be falsely detected as descriptive, thus leading to incorrect grouping and thus creating a need to refine the process for minimizing unwanted data. SMAGA uses filtered tweets for analysis in a table form of oracle which can be defined as a matrix $M$ from the tweet analysis table values as listed in Table 5.1 below.

<table>
<thead>
<tr>
<th>USERID</th>
<th>PRODUCT</th>
<th>LOCATION</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Samsung Mobile</td>
<td>Delhi</td>
<td>High Price</td>
</tr>
<tr>
<td>1</td>
<td>Samsung TV</td>
<td>Delhi</td>
<td>Like</td>
</tr>
<tr>
<td>1</td>
<td>Samsung Mobile</td>
<td>Delhi</td>
<td>Like</td>
</tr>
<tr>
<td>2</td>
<td>Samsung TV</td>
<td>Chennai</td>
<td>Not Like</td>
</tr>
<tr>
<td>2</td>
<td>Samsung Mobile</td>
<td>Chennai</td>
<td>Like</td>
</tr>
<tr>
<td>3</td>
<td>Samsung TV</td>
<td>Mumbai</td>
<td>Good</td>
</tr>
<tr>
<td>3</td>
<td>Samsung Mobile</td>
<td>Mumbai</td>
<td>Nice</td>
</tr>
<tr>
<td>3</td>
<td>Samsung Mobile</td>
<td>Mumbai</td>
<td>Nice</td>
</tr>
<tr>
<td>3</td>
<td>Samsung Mobile</td>
<td>Mumbai</td>
<td>Fantastic</td>
</tr>
<tr>
<td>4</td>
<td>Samsung TV</td>
<td>Hyderabad</td>
<td>Like</td>
</tr>
<tr>
<td>4</td>
<td>Samsung Mobile</td>
<td>Hyderabad</td>
<td>Not Like</td>
</tr>
<tr>
<td>4</td>
<td>Samsung Mobile</td>
<td>Hyderabad</td>
<td>Okay</td>
</tr>
<tr>
<td>5</td>
<td>Samsung TV</td>
<td>Calcutta</td>
<td>Not Like</td>
</tr>
<tr>
<td>5</td>
<td>Samsung Mobile</td>
<td>Calcutta</td>
<td>Not Like</td>
</tr>
<tr>
<td>5</td>
<td>Samsung Mobile</td>
<td>Calcutta</td>
<td>Not Bad</td>
</tr>
</tbody>
</table>

The above tabular values can be represented as a matrix as shown below
\[ M = \begin{pmatrix}
X_{11} & X_{12} & \ldots & X_{1m} \\
X_{21} & X_{22} & \ldots & X_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\end{pmatrix} \]

Where the value of \( m \) is 4 and the value of \( n \) is 15 and evaluated as

\( X_{11} = 1, \ X_{12} = \text{Samsung Mobile}, \ X_{13} = \text{Delhi}, \ X_{14} = \text{High Price} \)

\( X_{21} = 1, \ X_{22} = \text{Samsung TV}, \ X_{23} = \text{Delhi}, \ X_{24} = \text{Like} \)

\( X_{31} = 1, \ X_{32} = \text{Samsung Mobile}, \ X_{33} = \text{Delhi}, \ X_{34} = \text{Like} \)

\( X_{41} = 2, \ X_{42} = \text{Samsung TV}, \ X_{43} = \text{Chennai}, \ X_{44} = \text{Not like} \)

\( \ldots \)

\( X_{15,1} = 5, \ X_{15,2} = \text{Samsung Mobile}, \ X_{15,3} = \text{Calcutta}, \ X_{15,4} = \text{Not bad} \)

The Matrix \( M \), generated by text tools or exporting geographical area and the tweet gets converted into an Oracle Table or the Dataset OD as shown in Screen shot below. The Data set OD contains users Tweets with geographical area.
5.4 TWEET EXTRACTION

Data extraction in SMAGA is the process of retrieving tweets from unstructured data sources for processing. The process is executed by using pattern matching in tweets for comments of importance and thus identifies records and their associated data. The table based approach is used to find commonality within a limited domain. In text analysis common or required terms for example ‘like’ or ‘dislike’ are extracted from SNS tweets. Though Web content mining is unlike text mining, since Web data is a semi-structured, SMAGA converts the web data into a structured format to analyze and classify.
5.5 LOADING TWEETS INTO ORACLE

The structured format of web data is attached with a description on particular area of interest like a product or travel place, actual comment and a user’s geographical area. Social media text can be loaded into Oracle databases as tables using Oracle’s SQL*Loader facility into the following table structure.

SQL*Loader is also flexible and offers various options to maximize the speed of data loads as indicated in Table 5.2 below. [http://www.remote-dba.net/teas_rem_util18.htm]

<table>
<thead>
<tr>
<th>SQL*Loader Option</th>
<th>Elapsed Time (Seconds)</th>
<th>Time Reduction</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct=false rows=64</td>
<td>135</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>direct=false bindsize=512000 rows=10000</td>
<td>92</td>
<td>32%</td>
<td>The number records in each read</td>
</tr>
<tr>
<td>direct=false bindsize=512000 rows=10000 database in noarchivelog mode</td>
<td>85</td>
<td>37%</td>
<td>Transactions are not backup paralelly increasing the time</td>
</tr>
<tr>
<td>direct=true</td>
<td>47</td>
<td>65%</td>
<td>Operation is between read and load but with recoverable errors</td>
</tr>
<tr>
<td>direct=true unrecoverable</td>
<td>41</td>
<td>70%</td>
<td>Further increased since errors are not checked</td>
</tr>
</tbody>
</table>

Oracle provides many choices for online data loading like inserting or bulk loading where SQL*Loader is the fastest in loading bulk data into oracle compared with other oracle data tool options and is indicated in Figure 5.1 below [http://www.dba-oracle.com/art_orafaq_data_load.htm].
SQL*Loader is also flexible and offers various options to maximize the speed of data loads as indicated in Table 5.2. [http://www.remote-dba.net/teas_rem_util18.htm]

5.6 SOCIAL MEDIA ANALYSIS GROUPING ALGORITHM USING DATABASES

Before grouping, a similarity/distance measure is determined which reflects the degree of closeness of objects to the characteristics that are believed to distinguish the groups embedded in the data. In many cases, these characteristics are dependent on the context and there is very little universal measure to suit all kinds of grouping problems. The basic idea of the proposed algorithm (SMAGA) is to retrieve specific text from user posts and then aim to retrieve groups with maximum correlation between users and classify them based on their geographical regions.
5.6.1 SMAGA Architecture

Social Sites have millions of members and millions of activities like 50 million page views, above 300K invitations, above 1.5 million answers and about 1.8 million email messages every day. Social Networks Architecture can be viewed as a combination of hardware and software. The software used typically have one Solaris server, application servers like Tomcat, Databases servers like Oracle with JDBC. The server Architecture uses one gigantic web application, One Core Database and Web Applications update the Core Database directly. The Communication Services uses JMS heavily and messages are routed via a routing service to the appropriate mailboxes. The mining process on real time data retrieval has to find a feasible solution to the extracted and aligned data. Pattern recognition in these data and data analysis requires that the representation to be processed data has to be of high relevance.

Data grouping can also be helpful in classifying documents for discovery during Information Retrieval. Hence the Web content mining combined with Data grouping makes this proposed research a simple solution to a complicated problem. String similarity for approximate string matching, for example the strings "like" and "dislike" seems to be comparable. The string metric provides the difference between them. Thus the SMAGA algorithm uses the core database for focused retrievals for categorizing the tweets into several areas for the value of x like Technology, Merchandize, Travel and Products with geographical areas y and users who like or dislike. The main concept of SMAGA Grouping is to group like-minded users based on their likes or dislikes shared in tweets of a social network. Thus extracting textual posts based on search criteria is a part of the algorithm. SMAGA implements the grouping of tweets with geographical area and likes or dislikes. The result is split into groups and their aggregate is computed. Aggregation of Geographical area of likes \(G_1, \ldots, G_n\) is written as \(\Gamma_{G_1, \ldots, G_n, OP(G)}(OD)\). Fig.5.2 below illustrates the proposed SMAGA Architecture.
5.6.2 Grouping similar data using SMAGA

Data grouping (or just grouping), is an unsupervised classification for creating groups of objects, or groups, in such a way that similar objects belong to the same group. Group analysis has been widely used in numerous applications of research. In business, grouping can help marketers discover customer interests based on purchasing patterns and characterize groups of customers. Data grouping can also be helpful in classifying similar documents in Information Retrieval. The Accuracy-Extract (AE) ratio has recently received much attention in the research field. Accuracy and Extract are used for measurement where accuracy is the ratio of the number of relevant tweets retrieved to the total number of tweets and expressed as a percentage. In the field of information retrieval, accuracy is the fraction of retrieved documents that are relevant to the search query.
\[
\text{Accuracy} = \frac{\{\text{relevant tweets} \} \cap \{\text{retrieved tweets} \} }{\{\text{retrieved tweets} \} }
\]

As Accuracy value shows the total no of retrieved tweets from the website based on given topic. It computes the highest number of tweets found in different pages and the percentage of overall tweets found in all websites. Extract is the ratio of number of relevant tweets retrieved to the total number of relevant tweets in the database. It is usually expressed as a percentage. Extract in information retrieval is the fraction of documents relevant to a query that are successfully retrieved.

\[
\text{Extract} = \frac{\{\text{relevant tweets} \} \cap \{\text{retrieved tweets} \} }{\{\text{relevant tweets} \} }
\]

Extract represents the highest number of relevant and repeated tweets found in different web pages and the percentage of overall tweets data found in all websites. It gives the total number of percentage result.

### 5.6.3 Mathematical Notation of SMAGA operations

**OD – Oracle Database**

The selection operation of choosing rows in a table may be denoted as

\[\sigma_C(\text{OD}) = \{r \in \text{OD} \mid r \text{ satisfies C} \}\]

Where \(\sigma\) is the Selection operation and \(\sigma_C\) is the operator working on D for the selection of rows in a condition C and C may also have null or no condition then

\[\sigma_C(\text{OD}) = \text{SELECT * FROM OD when C is null}\]

or

\[\sigma_C(\text{OD}) = \text{SELECT * FROM OD WHERE } \{(x, y) \mid x \text{ is a string, and } y \text{ is a string}\}\]

78
The resultant rows can be depicted as

$$\sigma_c(OD) = \{rs \in OD \mid rs \{ (x, y) \mid x \text{ is a string, and } y \text{ is a string} \}$$

Where the set of all elements X satisfying the condition Y

5.6.4 SMAGA Schema Tables

users table for user registration.

Create table users(

userid int(15) not null auto_increment,
username varchar(60),
‘password’ varchar(10),
‘email’ varchar(60),
‘friendcount’ int(15),
‘profilepic’ varchar(120),
primary key(‘userid’));

updates table for status updates data.

Create table ‘userupdates’ (‘updateid’ int(15) auto_increment,
‘update’ varchar(60),
‘useridfk’ varchar(60),
‘created’ int(15),
‘ip’ varchar(60),
Primary key (‘updatedid’),
Foreign key (useridfk) references users(userid));

Friends table for relation data.

Create table ‘userfriends’ (‘friendone’ int(15),
‘friendtwo’ int(15),
‘status’ enum(‘0’, ’1’, ’2’) default ’0’,
--0,1 and 2 refer to pending friend request, confirm friend request and you
Primary key (‘friendone’, ’friendtwo’),
Foreign key (friendone) references users(userid);
Foreign key (friendtwo) references users(userid));
Create table TweetAnalysis(
  Userid int(15) not null auto_increment,
  Area_product varchar2(50),
  Commnts varchar2(140));

5.6.5 SMAGA Algorithm

Input : User Product/Service Search Keyword x and Database OD
Output: User Likes on Product/Service $x_g$ for where $g$ is Geographical
         Area and $g=1...n$

Function filter
Counter =0
Arraylike[1..m]  // array to hold all the like values of the tweets
For counter = Tweet t1 to counter=t_m in OD
   If t_counter = ‘like’ or ‘lik’ then
      Arraylike[counter]=t_counter Geographical Area
   Endif
Next

Function Group
AryCounter =0
For Arycounter = 1 to Arraylike[Max]
   Group by Geographical Area $G_1..G_n$
Next

The input matrix M is regrouped with like-minded users in different geographical
areas and sorted by geographical areas. SMAGA is capable of delivering higher
performance in terms of classification accuracy and speed than other algorithms. The
Oracle Equivalent of this grouping is SELECT $G_1, \ldots, G_n$ OP(OD) FROM OD
GROUP BY $G_1, \ldots, G_n$.
5.6.6 SMAGA Result

Given a group of users the intention was to find the clusters of people with similar interests. The distribution maximizes the correlation between the intra-group users. In SMAGA the grouping is based on the content of the messages with the main goal to find centers of interest around which the input data is concentrated. Different types of operation on the database yielded different results in the database. CPU and Memory utilization was 3% and 7MB. Table 5.3 depicts the performance timings of SMAGA on Tweet data, and the resulting graphs are shown in Figure 5.3 and Figure 5.4.

Table 5.3 Performance Timings of SMAGA

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Operation</th>
<th>The Query</th>
<th>Execution Time in Seconds</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Select All</td>
<td>SELECT * FROM TweetAnalysis</td>
<td>0.115</td>
<td>.99</td>
</tr>
<tr>
<td>2</td>
<td>One Filter on ‘like’</td>
<td>SELECT * FROM TweetAnalysis WHERE UPPER(commnts) LIKE</td>
<td>0.625</td>
<td>.99</td>
</tr>
<tr>
<td>3</td>
<td>Grouping in Geographical Area</td>
<td>SELECT Area_Product,COUNT(Area_product) FROM Tweet Analysis GROUP BY Area_product</td>
<td>4.005</td>
<td>.99</td>
</tr>
<tr>
<td>4</td>
<td>SMAGA filter on likes and</td>
<td>SELECT Area_Product,COUNT(Area_product) FROM Tweet Analysis WHERE Upper(commnts) LIKE GROUP BY Area_product having count(Area_product) &gt; 0</td>
<td>13.5</td>
<td>.98</td>
</tr>
</tbody>
</table>

The classification results of SMAGA shows a high value on precision and low time of processing tweets. As per Table 5.3, from Figure 5.3 and Figure 5.4, it can be inferred that the majority of the values of precision exceed 0.95 which proves SMAGA can assign majority of users to the right group and also group users sharing the same likes or dislikes with a low processing time. SMAGA is also successful in identifying like- minded tweets and distributing them geographically.
Figure 5.3  SMAGA Execution times on Groupings

Figure 5.4  SMAGA Precision