International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 2, Issue 2, March – April 2013 ISSN 2278-6856

Tracing High Quality Content In Social Media For Modeling & Predicting The Flow Of Information - A Case Study On Facebook

Karuna Gull[1], Akshata Angadi[2], Dr. Santoshkumar Gandhi [3] and Santoshkumar B. Shali[4]

[1,2]Department of Computer Science and Engineering
K.L.E. Institute of Technology, Hubli, India

[3]CMJ University, Shillong (Meghalaya), India

[4]Innovative IT Solutions Hubli, India

Abstract: ***Efficiently extracting high-quality content from Social media is a challenging problem due to the heterogeneous and noisy nature of the data. Facebook is currently the second most popular site in the world just after search site Google that provides many features to the user to add up the content in form of posts comments & likes to establish a connection between the producers and the consumers of information. Longer posts, supplemented with a picture, tell compelling stories. Posts are well written and inspire hundreds of comments from fans. Each piece of content is highly relevant to the brand's audience. So, considering Facebook as a case study, in this paper we are presenting a Social Network Mining approach using the Web, to trace the high quality content present in the posts, generated by user, for Modeling and Predicting the flow of information. Our work provides a Linear Influence Model to analyze the social media from data mining view, to trace high quality content which can be used for marketing strategy of products. This enables companies to gain feedback and insight of how to improve and market products better.***

Keywords: DOM, Facebook, Intensity, Intimacy, Social Media, Social Networking Services(SNS).

1. INTRODUCTION

The era of Social Media as we understand it today probably started about 20 years earlier, when Bruce and Susan Abelson founded ''Open Diary,'' an early social networking site that brought together online diary writers into one community. The term ''weblog'' was first used at the same time, and truncated as ''blog'' a year later when one blogger jokingly transformed the noun ''weblog'' into the sentence ''we blog.'' The growing availability of high-speed Internet access further added to the popularity of the concept, leading to the creation of social networking sites such as MySpace (in 2003) and

Facebook (in 2004).This, in turn, coined the term ''Social Media,'' and contributed to the prominence it has today.
The concept of Social Media is top of the agenda for many business executives today. Decision makers, as well as consultants, try to identify ways in which firms can make profitable use of applications such as Wikipedia, YouTube, Facebook, Second Life, and Twitter. Yet despite this interest, there seems to be very limited understanding of what the term ''Social Media'' exactly means:

Social media is defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content [1]. [2]Lietsala et al. (2008) think social media is a term for describing web service including contents sharing among users in sites.

1.1 The challenges and opportunities of Social Media

- Collaborative projects: Collaborative projects enable the joint and simultaneous creation of content by many end-users.
- Blogs: They are the Social Media equivalent of personal web pages and can come in a multitude of different variations, from personal diaries describing the author's life to summaries of all relevant information in one specific content area.
- Content communities: The main objective of content communities is the sharing of media content between users.
- Virtual social worlds: The second group of virtual worlds, often referred to as virtual social worlds, allows inhabitants to choose their behavior more freely and essentially live a virtual life similar to their real life.
- Social networking sites: Social network sites (SNSs) are increasingly attracting the attention of academic and industry researchers intrigued by their affordances and reach.

Social networking sites are applications that enable users to connect by creating personal information profiles, inviting friends and colleagues to have access to those profiles, and sending e-mails and instant messages between each other. These personal profiles can include any type of information, including photos, video, audio

files, and blogs. According to Wikipedia, the largest social networking sites are U.S.-based Facebook (initially founded by Mark Zuckerberg to stay in touch with his fellow students from Harvard University) and MySpace (with 1,500 employees and more than 250 million registered users). Social networking sites are of such high popularity, specifically among younger Internet users, that the term ''Facebook addict'' has been included in the Urban Dictionary, a collaborative project focused on developing a slang dictionary for the English language. Several companies are already using social networking sites to support the creation of brand communities [3] or for marketing research in the context of netnography [4]. To promote the movie ''Fred Claus,'' a 2007 Christmas comedy film, Warner Brothers created a Facebook profile via which visitors could watch trailers, download graphics, and play games.

Facebook is defined as "a social utility that helps people share information an communicate more efficiently with their friends, family and co workers" (facebook. com).The mission of facebook is "Giving people the power to share & make the world open & connected". As of November 2009, with 316 million users, Facebook is the most popular and well known social network throughout the world. Facebook is currently the second most popular site in the world just after search site Google according to Alexa traffic rankings.

Fundamental features to the experience on Facebook are a person's Homepage and Profile. Facebook also has numbers of core features with which users can interact. They include "The Wall", the message board on the user's profile page that allows friends to post messages which other users can see & advertisements can be added up if interested; "Gift", the user can send virtual gifts to friends; "Pokes", users can "poke" each other by sending a poke icon for interaction; "Status", shows friends about the user's whereabouts and actions; "Events", the function that informs users activities will happen online and offline. In 2007, Facebook open its platform to third party application developers. It appeals a great amount of brilliant applications' integration with Facebook's platform, which extends Facebook's core functionalities and features tremendously. Nowadays, Facebook runs over 24,000 applications, there are about 400,000 experts develop applications all over the world, and around 140 applications are published daily.

The paper is organized as follows: Section 2 describes the related work on social media & social networking services. Section 3, presents the market consideration of facebook. The methodology for the extraction & finding the quality of content is defined in Section 4. Experimental results are shown in Section 5. Lastly in Section 6 we concluded the paper with future work.

2. LITERATURE SURVEY

Social network sites constitute an important research area for scholars interested in online technologies and their social impacts, as evinced by recent scholarship in the area [5][6]. Social network sites (SNSs) are "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system". The first social network site was launched in 1997 and currently there are hundreds of SNSs across the globe, supporting a spectrum of practices, interests and users [7].

One of the largest social network sites among the U.S. college student population is Facebook, created in February 2004 by Mark Zuckerberg, then a student at Harvard University. According to Zuckerberg, "The idea for the website was motivated by a social need at Harvard to be able to identify people in other residential houses"[8]. Facebook has become very popular among undergraduates, with usage rates upwards of 90% at most campuses [9]. It has also stimulated much recent research on various aspects of Facebook use, such as the use of Facebook in academic settings and the demographic predictors of Facebook use [10]. One strand of research focuses on the outcomes of Facebook use.

The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance- from very high-quality items to low-quality, sometimes abusive content.

According to Boyd and Ellison, social network sites have three fundamental properties.

- First, they allow individuals to "construct a public or semi-public profile within a bounded system."
- Second, the individual is able to articulate connections to alter in the system.
- Third, these lists of connections can be "viewed and traversed" by others within the system. In these systems, status updates, wall postings, shared links and pictures and other fundamental activities are centralized on the profile [7].

Table.1. Compares the characteristics of four popular social networks

	Main Audience	**End User Feature**	**General Features**
MySpace	Teens, young People	Blog, video, photo, address book, bulletin calendar, email	Browse profiles, search, and invite new people, film/comedy/music forum, favorites, videos, classifieds,

Orkut	Teens, young people	Photo, video, bookmarks, scrapbook,profile, testimonials	Friends (rank, best, good, acquaintances), search, communities, Orkut media, news
LinkedIn	Business professionals	Connections, network data, email list, recommendations	People search, jobs, hiring, categorized service providers recommendations, ask a question from professionals
Facebook	Students Profile,	friend finder, photos, myShares, notes, events, inbox	Browse profiles, search, invite new people

Practically, most of social media services are available to use without charge. The free service attracts users to come and join the network. However, as the user amount increases and the contents enrich, the desire for making business also grows. In order to make profits, companies holding substantial users base like Facebook, Twitter are starting to look for the ways for making money behind the free service. And marketers are already thought about the strategy to monetize the social media.

3. MARKET CONSIDERATIONS

There are numerous SNS sites presently in global internet market, they have rather fierce competitions. eBizMBA Rank (2011) summarizes the top 15 most popular social networking sites base on the average of each website's Alexa Global Traffic Rank, and US Traffic Rank from both Compete and Quantcast by February 2011, in order, they are Facebook, Twitter, Myspace, LinkedIn, Ning, Tagged, classmate.com, hi5, myyearbook, meetup, bebo, mylife, Friendster, myHeritage, and Multiply. However, each site targets distinct user groups and provides different social purposes. Among the most competitive three ones, Twitter serves as a micro-blogging tool, MySpace mainly drives social interaction by providing a highly personalized experience around entertainment and connecting with music, celebrities, TV, movies, and games they love, LinkedIn operates the world's largest professional network, and helps users to maintain the business connections. But Facebook takes the major social networks market.

3.1 Financial aspects

Social media activity now constitutes a substantial fraction of time spent on the Web [12]. Users of social networking technologies create explicit representations of their relationships with other users (their peers) [13], and use those connections as channels for information dissemination. They also establish connections with other entities in order to express their identities and subscribe to content. The widespread adoption of such technologies has led to advertising approaches that differ from existing approaches, such as search-based advertising. For example, rather than inferring consumer intent via search terms, social advertising systems can match ads to consumers who have peers that are affiliated with the brand, product, or organization being advertised.

We regard social advertising as any advertising method that uses information about consumers' social networks to target ads and/or provide personalized social signals. Thus, there are two ways in which the use of social information in advertising can affect consumer responses: social networks encode unobserved consumer characteristics, which allow advertisers to target likely adopters; and the inclusion of social cues creates a new channel for social influence. Recent work on social advertising has recognized these mechanisms, but has been largely unable to identify the extent in which social influence actually plays a role.

According to Mashable [14] there are generally three basic types of advertisement on Facebook platform.

- Firstly, ads places can be bought directly on Facebook's page. It can be simply created by anyone. To make an advertisement, an advertiser needs to address the links you want to be directed to your ads first, and then give the name and description; specify the target customers by age, gender, location, sex, relationship, etc, and state the payment model either by pay per click or pay per view. Finally the ads will be displayed on target users' pages.
- Secondly, Facebook provides social ads pages. Any brands, businesses, organizations, bands and so on can create their own Facebook page with adding the contents they want, including photos, videos, music and Facebook platform applications. Users can directly interact with the business by commenting on business' Wall or giving the feedback by clicking "thumbs up" or "thumbs down" buttons on the top of that page. The actions could appear in users' News Feed which enables to ones' friends can all see this information. Social ads actions are more powerful than conventional ads because they act as trusted referrals. Above two ads approaches are both based on Facebook's self-serve ads platform, according to Mashable, it takes 60% Facebook's ads revenue.
- Thirdly, Facebook and Microsoft Corp. announced in 2007 that the two companies would expand their advertising partnership, and Microsoft will be the exclusive third party advertising platform partner for Facebook. Thus Microsoft is able to post advertising banners on Facebook's page [14].

3.2 User activity and engagement metric

User activity indicates users' actions when using the services, it typically reflects the popularity of the service, and can be measured by the site traffic, or the number of registered or active users. Engagement metric can be understood as how much time people spend on a site and how many pages they look at on average during each visit to more fully understand sites' popularity. A successful service attracts more users' attentions and activities, people are willing to engage more time on it rather than other services. According to Alexa[15], internet company data statistics, the most frequently used metrics are reach

rate, percentage of daily pageview, daily pageview per users, bounce rate, and daily time on site. We collect data of Tencent, Facebook, and Myspace and do the comparison of their popularity as shown in Table.2.

Table .2. Traffic and user engagement

	Reach rate (%)	Daily Pageview (%)	Daily Pageview /User	Bounce rate (%)	Daily time spent
Tencent	7.16	0.64	9.23	29.2	15.18
Facebook.com	40.29	5.1	13.1	15.1	32.76
Myspace.com	1.31	0.053	4.23	44.7	3.97

Reach measures the number of users. It is the percentage of all internet users who visit a given site. Facebook has a reach of 40.29%, which means that all global internet users measured by Alexa, 40.29% of them visit Facebook.com. Tencent's website portal obtains a much lower reach. There are 7.16% global netizens visit qq.com. Myspace has the lowest reach rate 1.31%. Page views measure the number of pages viewed visitors. Daily pageview metric indicates the percentage of global page views per day. Facebook still excels Tencent and Myspace in this metric. Meanwhile, the page views per user are the average numbers of unique pages viewed per user per day by visitors. A Facebook's user visits around 13 pages every day, while a user of qq.com views about 9 pages and a user of MySpace views 4 pages on average. Bounce rate is the percentage of visitors who enter the site and leave rather than continue viewing other pages within the site. There are 29.2% visitors of qq.com views only one page of the site and leave, they are less willing to go deeper into the site than Facebook. And MySpace did the worst then. People prefer to spend more time on Facebook than qq.com and MySpace.

3.3 Facebook has real business value

Many internet activities that once took place on separate, isolated venues are now funneled through Facebook. These include email, instant messaging, blogging, gaming, video-sharing and online shopping. As long as users feel they can use Facebook as a gateway for these and other functions, the social network will remain vital to the internet experience and relevant to marketers.

According to Source-Outsell, December 2009, facebook is considered as effective marketing tool (as shown in Fig.1**)** & most U.S B2B & B2C marketers use facebook for about 80% & 90% respectively.

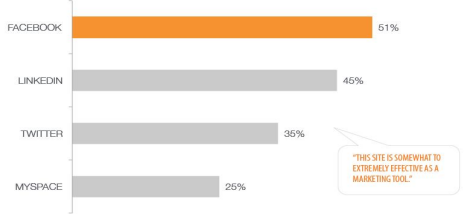

Figure.1.shows which SNS is effective(in %).

The number of marketers who say Facebook is "critical" or "important" to their business has increased 83.In 2009 it was 24% & there was sudden drastic increase to 83% in 2011 as per report given by Hubspot, state of inbound marketing.

4. METHODOLOGY

Data mining is a process that uses a variety of data analysis tools to discover patterns and relations in data that may be used for prediction purposes. Supervised data mining techniques are used to model an output variable based on one or more input variables and these models can be used to predict or forecast future cases.

The purpose of the present study is to discover high quality content of Facebook using data mining techniques. It is believed that social network based application development and advertisement programs can be enhanced by the findings of this study.

4.1 Facebook" As case study

Various data mining techniques are employed during the analysis and their prediction performances are compared. Thus, in this section, brief information is presented about the methodology we followed. Facebook is the largest social network in the world with 500 million active users of whom 50% log in on a daily basis. So, it is no wonder why marketers are interested in figuring out how to reach their target audience using Facebook. Fig.2. provides even more statistics about Facebook's huge user base.

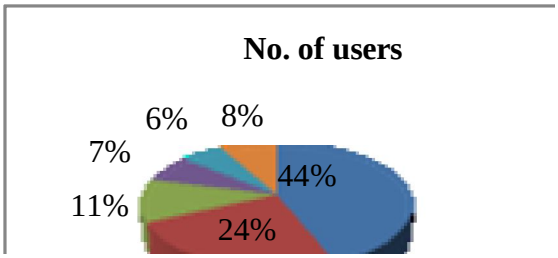

Figure.2. No. of Users (in %) to the popular sites listed above.

4.2 Strength of tie:

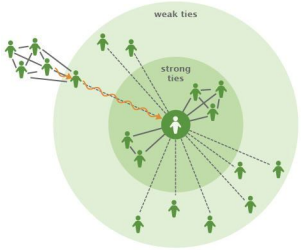

Figure.3. Strength of Tie

According to Granovetter, "The strength of a tie is a combination of the amount of TIME, the emotional INTENSITY, the INTIMACY, and the reciprocal SERVICES which characterize the tie." [16]. Thus these

variables are used as suitable measures to strengthen the predicting model that we are designing.

4.3 Framework

Processes undergone in Extraction of content present in Webpage

- Data Acquisition

  The first step is to obtain the web page for data extraction. The data source for extraction can be a data on local disk or data on the network [2].
- Data Preprocessing

  The web page in the HTML format is not well-formed because it does not conform to HTML specification. That is, in HTML ignoring closing tag will not give any error message. Therefore first it must be converted into well-structured XHTML format.
- Data Conversion

  XSL (eXtensible Style Sheet language) is used to convert the XHTML to XML. The conversion is needed because of poor structural of XHTML documents. Even though it is based on the XML syntax structure, still it contains a lot of HTML vocabulary. So an XSL file has to be designed to convert the XHTML to XML. Next to extract information, it is necessary to find the reference point which contains the actual content.
- Data Extraction

  Our method is used to extract the data present in webpage. Social networking site i.e. Facebook which is a structured model. Based on the kind of data present in social media, extraction method is applied. We write the query that supports to dump the content in the database.

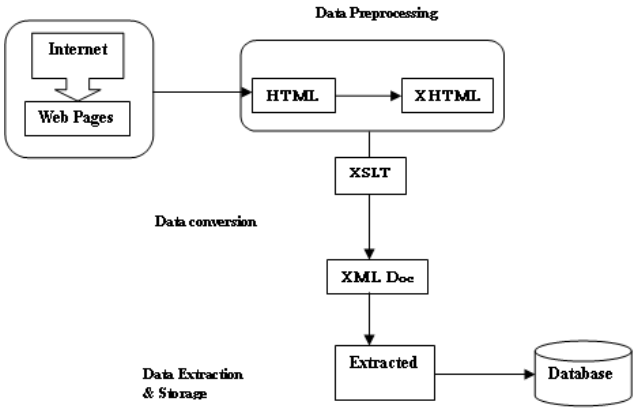

Figure.4. Shows the flow of processes in sequence.

4.4 Steps drawn to extract the high quality content (to best of our knowledge) from the post

- Facebook authentication

  User authentication is required to access the contents of facebook.com
- Extraction of data present in Posts

  Social networking site i.e. Facebook is a structured model. Based on the kind of data present in social media, extraction method is applied. The data present in the posts come under structured whereas related data of the same, such as likes, comments given by user come under unstructured. Web pages are in HTML/XHTML format, which may be in semi-structured or structured language. We write the query that supports to dump the post content in the JSON format. In this paper we are working on posts, which are structured.
- Identify the number of shares, likes, comments and links (Services). Classify positive and negative comments (Emotional support) based on simple NLP process where keywords are picked from the database itself.
- Assign weightage to posts' based on shares, likes and links.
- Extract Creation time and Updated Time from the post data (Duration) and assign the weight based on the duration.
- Assign weight based on size of content or number of words (Intensity).
- Find high quality content based on simple NLP process where keywords are picked from the wordnet dictionary.
- Assign weight based on user interaction (Intimacy).
- Find average weightage for posts. Assign rank for posts based on weightage.

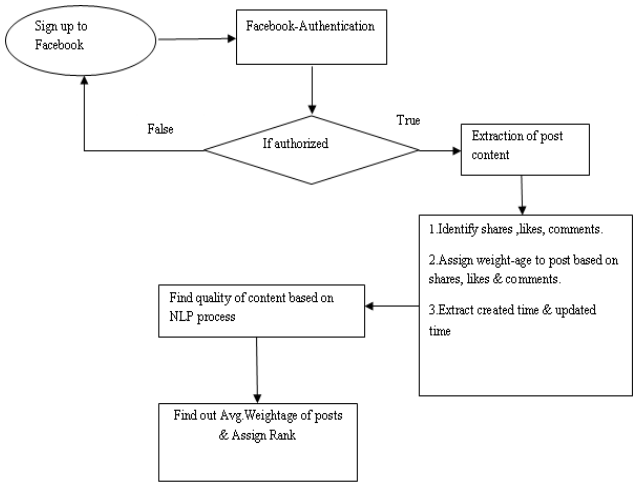

Figure.5. Flowchart to extract the high quality content (to best of our knowledge) from the post.

4.5 Methodology for Information Extraction

In our paper we work on dynamic web pages to exploit the special features of web page, i.e. the organizational structure of the fields in the DOM tree architecture.

As mentioned before in system design, all the pages are transformed into XHTML format before the extraction process can be performed. So the DOM tree should be valid, correct and well formatted. Those requirements are essential and pre-requisite for our system, because the whole extraction algorithm is relied on the HTML tags.

Let p be a web page, and *PTagSet* be the set of tags in p, that is $PTagSet = \{tag_1, tag_2, tag_3, \ldots, tag_n\}$, and each of the tag is either open or close tag. From the result

training process, keywords exist in the schema information. After analyzing, a set of keywords, $KWSet_i=\{KW_1, KW_2, \ldots, KW_l\}$ for extraction field f_j is formed.

For each KW_i in $KWSet_i$, the location of the KW_i will be identified in the page p. Then the first most nearest open tag against KW_i the keyword *KTagSet* will be put into a set. After all KW_i in $KWSet_i$ are applied, $KTagSet=\{KTage_1, KTage_2, \ldots, KTage_n\}$ will be formed.

For each $KTag_i$ in *KTagSet*, the page p will be split up to several parts by using $KTag_i$. For every part of the page, if any KW_i is located in between, then the whole part will be added to the output result set.

```
PTageSet  ={ tag1,tag2,... tag n }? p
KWSeti ={ KW1, KW2,..., KWl }
For each  KWi ∈ KWSeti {
If KW ? i  p {
        ∃ ktagi ∈ PTagSet ,  where  ktagi =nearest (ot )
        KTagSet = KTagSet + ktagi
        }
        }
For each ktagk(k=1..n) ∈ KTagSet {
         WPl = Split( p, ktagk)
        For each  WPl{
                If ∃ KWi ∈ WPl {
                        Output = Output + WPl
                         p = p - WPl
                }
        }
        }
```

Figure.6. Extraction Algorithm

The above methodology will be illustrated in Fig.6 and Fig.8. For example, if keyword *KWi* appears between the first pair of the <LI> as shown in Fig.7, then the *KTag* will be <UL>, since it is the first open tag for the keyword *KWi*. Then the whole page is split by the tag <UL>. After that, for each separated part, if the *KWi* exists, the part will add to the result set.

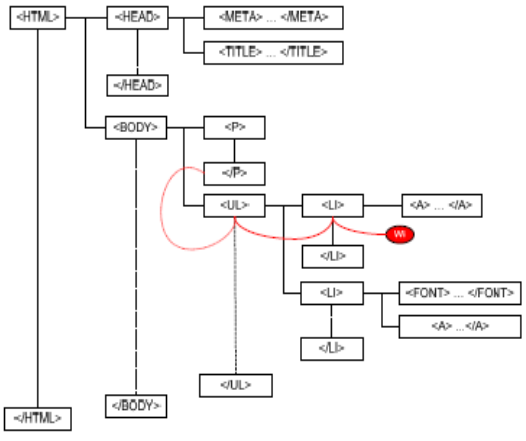

Figure.7. DOM tree structure of sample web page

As shown in Fig.7, keyword *KWi* appears more than one time in the page. In this example, the key tag list is {<UL>, <LI>}. Then the page is split by <UL> first and any part that contains *KWi* will be output to the result set. After that, the remaining parts of the page will be further split by <LI> and the process continues.

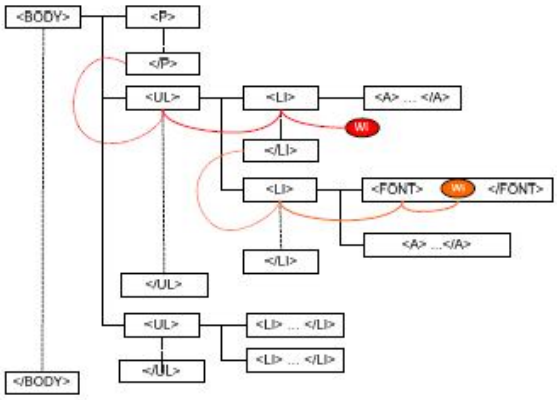

Figure.8. DOM tree structure of sample web page

The methodology mentioned above is only the basic one. In order to enhance the extraction result, instead of collecting the keyword *KW* from different training sample pages, some equivalent and synonymous keywords will be added. In this paper, the additional keywords added to the *KWSet*, are selected from a lexical dictionary called WordNet, which provides semantic relations among words. For example, the "Teach" in WordNet 2.1 has the synonymous words "instruct" and "learn".

Thus extraction of structured data is summarized as follows:

1. Input the Query for the extraction.
2. Extract required elements by Tag Name or By ID from HTML DOM structure.
3. Observe the elements and identify required data carefully.
4. Depending on requirement, process the data element.
5. Extract particular Class Names or Ids.
6. Data/information is stored in database.
7. Repeat steps 4), 5) and 6) for all the Data.

Thus the extracted data from web page which is in structured format is put in the database. Now we work on post message and comments which are in the unstructured format.

Extraction of Unstructured Data

1. Input the Query for the extraction.
2. Extract element by Tag Name ordered list from HTML DOM structure.
3. Extract elements by Tag Name 'li' (List) from the ordered list.
4. Extract inner-Html property from list.
5. Extract text using inner-Text property of ordered list.
6. Repeat steps 3), 4) and 5) for the entire list in the ordered list.
7. Finally the required data is extracted and dumped into database.

4.6 Experimental Setup:

We are implementing using Java Script and VB.net or ASP.net and running it on a Pentium – IV with 1GB RAM and 200 GB Hard disk. The Operating system used is Windows XP. The server side script is written in

VB.net. The MySQL 5.0 is used for creation of database and ODBC connector is used for connection.

5 EXPERIMENTAL RESULTS

4.1. Extraction of posts

In the similar fashion we can extract the content present in each post using the extraction method discussed in [17].

Figure 8 . Sample dataset taken from Facebook.com

Table.3. Data collected from the Figure.8 (post).

Name	Value
Id	444687735552770
From	"name":"Arvind Kumar H", "id": "100000346437504"
To	All
Message	It was in Pune that I met Narayan Murty through my friend Prasanna who is now the Wipro chief, who was also training in Telco(TataMotors) . Most of the books that Prasanna lent me had Murty's name on them which meant that I had a preconceived image of the man. Contrary to expectation, Murty was shy,bespectacled and an introvert. When he invited us for dinner. I was a bit taken aback as I thought the young man was making a very fast move. I refused since I was the only girl in the group. But Murty was relentless and we all decided to meet for dinner the next day at 7.30 p.m .. at Green Fields hotel on the Main Road ,Pune. Continuation ……..
Link	http://www.facebook.com/photo.php?fbid=444687735552770&set=o.104719951487&type=1
Likes	{ { "id": "100000545136960", "name": "Shyama Mohanan" }, { "id": "100002264984054", "name": "Manasi Joshi" }, { "id": "100000179234927", "name": "Augustin Xaxa" }, continuation…

Comments	{
	"id": "444687735552770_1146491",
	"from": {
	"name": "Sumitra Pandey",
	"id": "1142301181"
	},
	"message": "amazingggggggggggggg",
	"can_remove": false,
	"created_time": "2012-07-26T08:23:35+0000",
	"like_count": 0,
	"user_likes": false
	},
	{
	"id": "444687735552770_1147154",
	"from": {
	"name": "Kanjan Sharma",
	"id": "1495433079"
	},
	"message": "Very interesting story.....",
	"can_remove": false,
	"created_time": "2012-07-26T18:53:57+0000",
	"like_count": 0,
	"user_likes": false
	},
	continuation…

Created _time	2012-07-16T03:37:31+0000
Updated _time	2012-09-16T01:15:32+0000

Database Tables containing the actual data extracted from Fig.8.(post) are as follows .

Links Table:

Sl No	Post-ID	URL
1	444687735552770	http://www.facebook.com/photo.php?fbid=444687735552770&set=o.104719951487&type=1

Likes' Table:

Sl. No	Post-ID	Facebook-ID
1	444687735552770	100000545136960
2	444687735552770	100002264984054
3	444687735552770	100000179234927
4	444687735552770	100001493160701

Comments' Table:

Post-ID	Comment-ID	Facebook-ID	MESSAGE	Creat-Time	LIKE-COUNT
444687735552770	444687735552770_1146491	1142301181	Amazingggggggggggggg	2012-07-26T08:23:35+0000	0
444687735552770	444687735552770_1147154	1495433079	Very interesting story....	2012-07-26T18:53:57+0000	0
444687735552770	444687735552770_1162006	100003711798939	simply nice...	2012-08-03T12:56:44+0000	0
444687735552770	444687735552770_1182039	100002392967087	good team work	2012-08-14T09:10:33+0000	1

Data extracted from the post is stored in JSON format. The JSON format is often used for serializing and transmitting structured data over a network connection. It is used primarily to transmit data between a server and web application, serving as an alternative to XML. The

partial data present in JSON format has been added up in the above table for the post in Fig.8.
In the Table.4, we have taken out the values from 5 posts from facebook.com.

Table.4. Extracted values from the .json file 5 posts.

Post No.	Shares	Likes	Comments	Links
1	86	8226	468	25
2	50	4557	268	10
3	10	1011	301	0
4	28	2068	108	5
5	112	6781	278	30

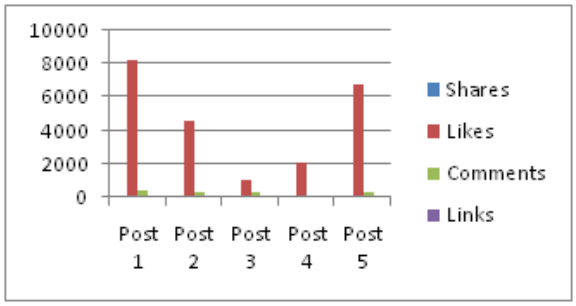

Figure.9. Graph showing information as per Table.5.

The graph shown in Fig.9 gives the information about the number of shares, likes, comments, links each post poses, according to our above Table.4.

5.1 Calculation of weightage

With the best of our knowledge, we have assigned ratings to each Share as 10, Like as 5, Comment as 10 & link as 1.Based on these consideration, we obtained the total weightage for each post as shown in Table.5 & a graph in Fig.10 respectively.

Table.5. Calculation of weightage

Post No.	Shares	Likes	Comments	Links	Total
1	860	41130	4680	25	46695
2	500	22785	2680	10	25975
3	100	5055	3010	0	8165
4	280	10340	1080	5	11705
5	1120	33905	2780	30	37835

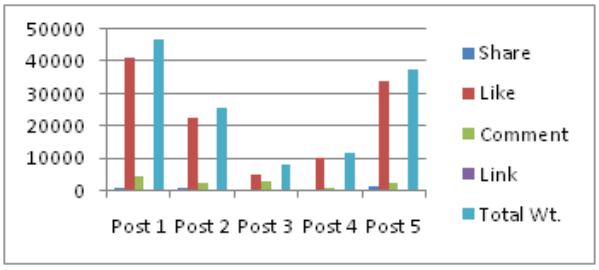

Figure.10. Graph shows information as per Table.5

After analyzing the high quality content of post in the social media, we conclude that the "Highest weightage post has highest rank ". According to above Table., Post 1 has highest weight because of that it has 1st rank ,Post 5 2nd rank, Post 2 3rd ,Post 4 4th and Post 3 is last as in Fig.11.

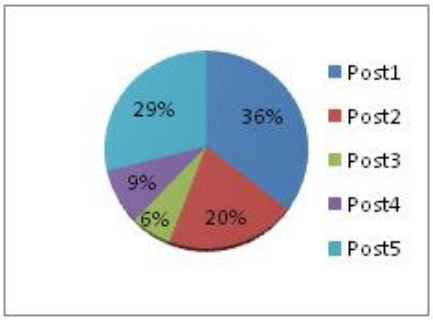

Figure.11.shows the weightage (in percentage) of 5 posts that we analyzed.

6 Conclusion

In social media platform such as Facebook.com, there are many settings in which users can publicly post content. A post's influence depends not just on its time, but also on the content when it is posted and its relation to the users' interests. Considering this as a point in this paper we analyze the post displayed on the Facebook Wall and provided a methodology to trace the content from social media. As part of our approach, we also propose a simple method to model and predict the effectiveness of a post in gaining social influence. The experimental result on the real-world dataset is consistent with our findings.

Our work provides a new method to analyze the social media from data mining view, which contrasts with a number of theories from marketing and sociology. Merging the vast amount of information on the Web and producing higher-level information might contribute to many knowledge-based systems in the future. Efficiently extracting high-quality content from Web is crucial for many Web applications such as information retrieval, automatic text categorization, topic tracking, machine translation, abstract summary, helping end users to access the Web easily over constrained devices like PDAs and cellular phones. The extracted results will be the basic data for the further analysis. So content extraction from social media has attracted many researchers recently.

References

[1] Kaplan, A.M. & Haenlein, M. (2009) "Users of the World, Unite, The Challenges and Opportunites of Social Media", *Business Horizons*, Vol.53, pp. 59-68.
[2] Lietsala, K., Sirkkunen, E. (2008) Social Media: Introduction to the Tools and Processes od Participatory Economy University of Tampere, Tampere 191.
[3] Muniz, Albert, M. and Thomas O'Guinn (2001), "Brand Community, " Journal of Consumer Research, 27 (March), 412-32.
[4] Kozinets, Robert V. (2002), "The Field Behind the Screen: Using Netnography for Marketing Research

in Online Communities,"Journal of Marketing Research, 39 (February), 61-72.

[5] Valkenburg, P. M., & Peter, J. (2007). Online communication and adolescent well-being: Testing the stimulation versus the displacement hypothesis. *Journal of Computer-Mediated Communication*, *12*(4),article2. http://jcmc.indiana.edu/vol12/issue4/valkenburg.html

[6] Wilkinson, D.M., Huberman, B.A.: Assessing the value of cooperation in wikipedia (Feb 2007)

[7] Boyd, D.M. & Ellison, N.B. (2007) "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, Vol. 13, pp. 210-230

[8] Moyle, K. (2004, Dec. 7). Internet helps people connect with past friends. : University Wire.

[9] Lampe, C., Ellison, N., and Steinfield, C. (2006). A Face(book) in the Crowd: Social Searching vs. Social Browsing. CSCW'06, Banff, Canada.

[10] Hargittai, E. (2007). Whose Space? Differences Among Users and Non-Users of Social Network Sites ,Journal of Computer Mediated Communication, 13(1), Article 14.

[11] Miltiadis, D.L., Ernesto, D. & Patricia, O.P. (2009) Web 2.0, The Business Model, Springer, New York.

[12] S. Goel , Quang Duong, Jake Hofman, and Sergei Vassilvitskii Sharding Social Networks Proceedings of the Fifth Conference on Web Search and Data Mining (WSDM 2012).

[13] Sun, E., Rosenn, I., Marlow, C., Lento, T. 2009. Gesundheit! Modeling contagion through Facebook News Feed. *Proc. ICWSM '09*.

[14]O'Dell, J. (2011). Online. Available at: http://mashable.com/2011/01/17/facebooks-ad-revenue-hit-1-86b-for-2010, [1.3.2011].

[15] Alexa.com, retrieved at 3 March, 2011, Web information site , http://www.alexa.com/siteinfo

[16] Granovetter, M. (1973). The Strength of Weak Ties. The American Journal of Sociology 78 (6), 1360 – 1380.

[17] Karuna C. Gull, Akshata Angadi, Dr.Santosh kumar Gandhi , Santoshkumar B. Shali , "A Survey on Tools used in Web Harvesting and a Methodology to Extract Data from Facebook", accepted in IJORCS , 2013.

AUTHOR

Karuna Gull is from Hubli, India. She has born on 7th June 1974. She has received the B.E. degree in Electronics and Communication from Karnataka University, India in the year 1996 and the M.Tech degree in Computer science and Engineering from the Visvesvaraya Technological University, India in the year 2008.
She has been working in the area of data mining and social networking since 2009. She has published 3 papers on Data Mining, 1 paper on Cloud Computing & 2 papers on Image Processing in International Journals. She has also published 5 National and 4 International papers in Conference Proceedings. She has also attended many of the workshops and conferences held in different places on High Impact Teaching Skills, Embedded System Using Microcontroller, Information Storage and Management (ISM), Data Mining, and many more. She worked as a Lecturer and Senior Lecturer for about 10 years. She is currently working as an Assistant Professor in K.L.E.I.T., Hubli, India since 2011.

Akshata B. Angadi received the BE degree in Computer Science from Visvesvaraya Technological University, India in 2011.
She is currently working as a Lecturer in K.L.E.I.T., Hubli since 2011. She has attended many conferences. She has published 2 papers on Data Mining, 1 paper on Cloud computing & 1 paper on Mobile application in International Journals. She has also published 2 National papers in Conference Proceedings.

Santoshkumar B.Shali completed B.E degree in Computer Science Engineering from B.V.B College of Engineering and Technology, Hubli. 5 years experience in Information technology which involves pre-sales, requirements gathering, analysis, project leadership, product delivery and support. Extensive programming background in domains such as education, Health-Care, Accounts/Billing, Retail and Finance focusing on product development, components, customization of legacy applications. 2.5 years experience in teaching, Worked as visiting faculty at SJMVS Business Administration College for Women, Hubli. Conducted various technical workshops and seminars on Mobile application development using Android. Carried out various projects in domains like Image Processing, Data Mining, Mobile Applications, Networking, Cloud Computing and so on.