Survey on Online Social Networks Analysis Concepts and Knowledge Discovery Techniques

Noha Negm
Faculty of Science, Menoufia University,
Shebin El-Kom, Egypt
Faculty of Science and Arts, King Khalid University,
Saudi Arabia

Hany Mahgoub
Faculty of Computers and Information, Menoufia University, Shebin El-Kom, Egypt
Faculty of Science and Arts, King Khalid University,
Saudi Arabia

ABSTRACT
In the recent decade, the Online Social Network (OSN) has gained remarkable attention. Accessing to OSN sites such as Twitter, Facebook, LinkedIn and Google Plus; the most dominant social media in the world, through the internet and the web 2.0 technologies has become more comfortable. These days through these online social networks, it becomes very easy for anyone to meet the people of the same interests for learning and sharing precious information. Online Social Network Analysis (OSNA) is an essential and important technique to understand the social structure, social relationships and social behaviors of OSN. OSNA deals with the interaction between individuals by considering them as nodes of a network whereas their relations are mapped as network edges. Now, it has increased various challenges for the evolution of the web and simultaneously increased the dynamic changes in its structure so it became harder to manually analyze very broad OSN. This survey investigates the current progression in the field of knowledge discovery in OSNA and covers all basic techniques of Data, Text, and Web mining that are widely used for the exploration of the unstructured and structured data available on the OSNA. The targets for OSNA are mainly focused on resources from the web, such as content, structure, and user behaviors. The main goal of this paper is to introduce a roadmap for the researchers who are interesting on the topics of knowledge discovery techniques for discovering totally different trends in OSN data. Discussion of all the challenges that face researchers in OSNA is also included.

Keywords
Social Network, Online Social Network Analysis, Knowledge Discovery, Data Mining, Text Mining, Web Mining

1. INTRODUCTION
Recognizing that "*we all connect, like a net we cannot see*", Online Social Network Analysis (OSNA) signifies one of the most important social and computer science phenomena and gains enormous importance by researchers around the world. This is because of many reasons including the popularity of online social networks, availability of huge amount of Online Social Networks (OSN) data, representation and analysis of OSN data as graphs, the market interests of social networks, and so on. The advent of the most popular OSN sites among the society such as Facebook, Twitter, WhatsApp, and LinkedIn has caused a shift on how people communicate and share knowledge, how politicians contest and influence, and how businesses operate and compete. The term Social Network (SN) is used to describe web-based services that allow people creating a public profile within a domain such that they can communicate with other users within that network as shown in Figure 1.

Fig 1: Sample of Social Network

Some common examples of OSN are given in figure 2. OSN has some inherent properties which make it much more powerful than the traditional networks [1]:

- ***Accessibility:*** the social media network is easily accessible and does not require any special skills, knowledge to use.it is absolutely simple to connect with others and be a part of communities.
- ***Speed:*** the content that you create on the social media network is available to everyone in your network/forum/community as soon as you publish.
- ***Interactivity:*** social media affords multiple communication channels. Users will interact with others; raise questions, discuss products/services, share opinions and anything else they might be interested in doing.
- ***Longevity / Volatility:*** because of the nature of the medium, social media content remains accessible for a long time. In addition to this, the content can be altered /updated anytime.
- ***Reach:*** the internet offers limitless reach to all content available. Anyone will access it from anywhere and anyone will reach everyone.

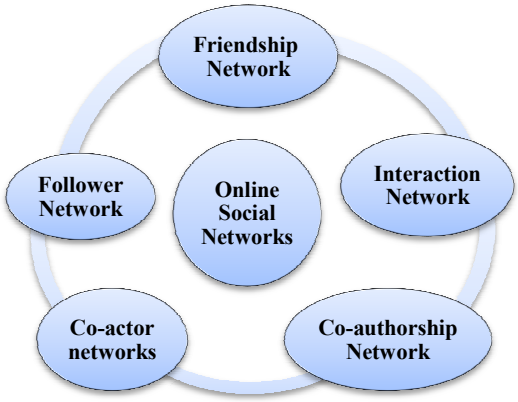
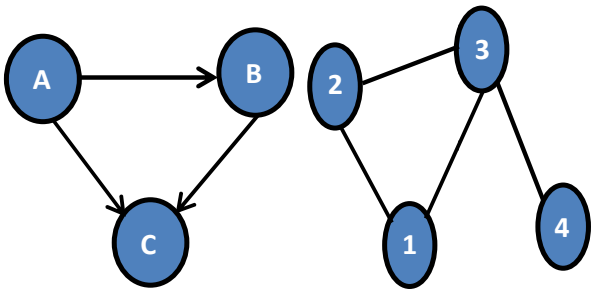


Fig 2: Common examples of online Social Networks

SN is commonly modeled by a graph which consists of users or groups called nodes connected by edges or links which is represented as a graph $G=(V, E)$ where V is a set of nodes or entities and E is a set of edges or relationships that connects the nodes. Edges may be directed or undirected, depending on the type of network that is modeled. The total number of nodes n i.e. $n=|V|$ represents the order of a graph G [2]. In the literature, graphs can be classified according to the direction of their links to:

1) Directed graph: is graph where all the edges are directed from one node to another. It is sometimes called a digraph. The order of nodes in the pairs in the edge set important in this type of graph. In a direct graph, edges are drawn as arrows indicating the direction. Twitter is an example of a directed graph since a person can be followed by others without necessarily follows them.

2) Undirected graph: is graph where all the edges are bidirectional. In this type of graph, the order of nodes in the pair in the edge set doesn't important and the edges are drawn as straight lines. Facebook is an example of undirected graph since the established friendship tie is mutual. Figure 3 shows an example of a directed and undirected graph.



a) Directed graph **b)** Undirected Graph

Fig 3: Directed and Undirected graph

The main benefit of social networked data is that they enable to understand individuals and society, provided the approval and trust, individuals have shown towards them [3]. With the exponential growth of social media sites in the recent years, social media sites provide data which are vast, noisy, distributed and dynamic and it became harder and unsuitable for traditional statistical methods to manually analyze this massive data [4]. The challenging features of online social networks are that it is very large, noisy and dynamic. These challenges are overcome by using different kinds of knowledge discovery techniques for the analysis of online social networks [5]. Hence, Knowledge discovery techniques provide researchers the tools needed to analyze such large, complex, and frequently changing social media data. The analysis targets of online social networks are mainly focused on resources from the web, such as its content, structures, and user behaviors.

To the best of our knowledge, there is no previous survey that systematically concentrates on review the knowledge discovery in OSNA and covers all basic techniques of Data, Text, and Web mining. All these techniques are widely used for mining the unstructured and structured data available on the OSN. Data mining techniques can be used for mining user information from OSN structured data while Text mining techniques are more complex than data mining because it contains irregular and unstructured OSN data patterns. In Web mining, main analysis targets are from OSN in the form of web content mining, web structure mining, and web usage mining [6].

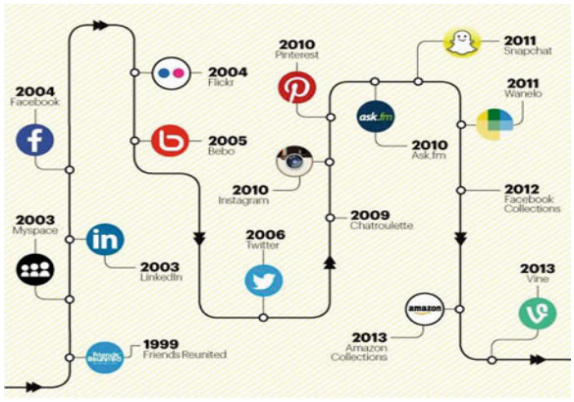
In this survey, we aim to introduce a reference survey for researchers who are working on the topics of Web knowledge discovery and online social network analysis. It systematically presents the theoretical background of online social networks; in addition, some basic concepts and properties of social networks are addressed. Some of the corresponding techniques used to process data collected from OSN are described. Techniques and concepts of knowledge discovery process are introduced and reviewed along with how to use different knowledge discovery techniques for OSNA.

The rest of the survey is organized as follow: theoretical background and some common basic concepts and properties of OSN will be reviewed in section 2. Section 3 introduces some of the social network analysis tools. In section 4, a study of how to use various techniques of Knowledge discovery for OSNA will be included. In section 5, there is a discussion of the challenges that face researchers for OSNA. Section 6 gives the conclusion.

2. ONLINE SOCIAL NETWORKS CONCEPTS AND PROPERTIES

SN is a social structure made up of actors called nodes, which are connected by various types of relationships or edges. A graph may be directed or undirected: for instance, an e-mail may be from one person to another and will have a directed edge, or a mutual e-mailing event may be represented as an undirected edge. To analyze and measure these relationships between people, groups and other knowledge processing entities, OSNA is used to provide both a structural and mathematical analysis. Online social networking has the ability to connect geographically dispersed users and provides social contact using the Internet. The ever-increasing popularity of many online social networks such as e.g. Facebook, Twitter, and MySpace etc. is a good pointer for that online social network has become very popular in recent years. Figure 4 shows the history of OSN sites in terms of when they were created. OSN contain a huge amount of content and linkage data which can be useful for analysis. These data can be divided into two types: unstructured data and structured data. Structured data are usually graph-structured which is represented as a graph $G = (V, E)$.

On the other hand, unstructured data are the content data shared in OSN, they are include text, images, videos, tweets, product reviews, and other multimedia data. Figure 5 shows the two types of data in social networks.



. **Fig 4: History of Online Social Networks**

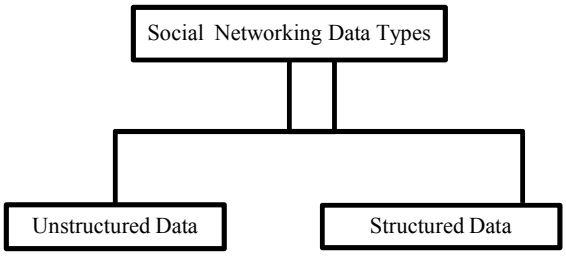


Fig 5: Social Networking Data Types

In the context of OSNA, some basic concepts and properties are well used to analyze the connections or interactions between nodes in the network and provide us with visions about the role of nodes in the network. Even with additional powerful computing paradigms are introduced into social network analysis, they still form a solid foundation for advanced social network analysis. Let's look at some of them [2]:

1. Centrality Gives a rough indication of the social power of nodes in the network based on how well they impact the network. Degree, Betweenness, Closeness, and Density are the most important closely related concepts of centrality.

2. Degree or Ties The degree of graph G is the actual number of edges m i.e. $m=|E|$. The maximum number of edges in undirected graph is $m_{max} = \frac{n(n-1)}{2}$, and for directed graph is $m_{max} = (n-1)$ where n is the number of nodes in the graph. It is an essential and effective measure to assess the importance and influence of a node in a social network. This parameter can give us a general idea of how large the network is.

3. Betweenness Measures the extent to which a node lies between other nodes in the network and can be computed as the percentage of shortest paths that pass through the node. This measure takes under consideration the property of the node's neighbors, giving a higher value for nodes which bridge clusters.

$$b_v = \sum_{s,t \in V(G)\setminus v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} denotes the number of shortest paths between nodes s and t and $\sigma_{st}(v)$ expresses the number of shortest paths passing through node v.

4. Closeness Closeness is a measure of reachability that measures how fast a given node can reach everyone in the network, giving an idea about how long it will take to reach alternative nodes from a given beginning node. Thus the closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network.

$$Cl_v = \frac{n-1}{\sum_{u \in V(G)\setminus v} d(u,v)}$$

Where $d(u,v)$ all the shortest distances between node u and node v in the network.

5. Density In real social networks, the fully connected network is rarely happens instead, the less connected network is more often. Density is a measure of the closeness of a network. The density of a graph is quantitatively defined as the number of links divided by a number of nodes in a complete graph with the same number of nodes. It is an indicator for the general level of connectedness of the graph. Given a number of nodes, the more links between them, the larger the density.

For directed graph $p = \frac{2l}{n*(n-1)}$

and for undirected graph $p = \frac{l}{n*(n-1)}$

where l is the number of links

6. Clique A clique in a graph represents a subset of a network in which nodes are more closely connected to one another than to other nodes of the network. In some extents, the clique is a similar concept to the community, which means the members within the same group have high similarity in some aspects, such as cultural or religious belief, interests or preferences and so on. The clique membership gives us a measure of how likely one node in the network belongs to a specific clique or community.

7. Clustering Coefficient A measure of the probability that two associates of a node are associates them. A higher clustering coefficient indicates a greater cliquishness.

The clustering of the entire network is the average clustering coefficient taken over all nodes in the graph.

$$C_i = \frac{2|e_{jk}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E$$

Where N_i is the neighborhood of node v_i , e_{jk} represents the edge that connects node v_j to node v_k , k_i is the degree of node v_i , and $|e_{jk}|$ indicates the proportion of the links between the nodes within the neighborhood of node v_i

8. Path Length Nodes may be directly connected by a line or they may be indirectly connected through a sequence of lines. A sequence of lines in a graph is a "walk", and a walk in which each point and each line are distinct is called a "path". The length of a path is measured by the number of lines which makes it up. The distance between two nodes is the length of the shortest path which connects them.

3. SOCIAL NETWORK ANALYSIS TOOLS

Since network data are different from the traditional attribute data, SNA uses corresponding tools to process data collected. SNA software is used to identify, represent, analyze, visualize, or simulate nodes and links from various types of input data. Some popular social networks tools are UCINET, PAJEK, STRUCTURE, and NETMINER. Existing techniques seem to be inadequate to handle new types of social network data that are continuous, dynamic, and multilevel [7]

UCINET This type of software can process, read and write a multitude of differently formatted text files, as well as Excel files, and handle a maximum of 32,767 nodes. Centrality measures, subgroup identification, role analysis, elementary graph theory, permutation-based statistical analysis, and other SNA measures can be performed on the software [8].

PAJEK PAJEK is an open source Windows program for analysis and visualization of large networks having some thousands or even millions of nodes. The main goals in the design of PAJEK are to support abstraction by (recursive) factorization of a large network into several smaller to provide the user with some powerful visualization tools and to implement a selection of efficient algorithms for analysis of large networks [9].

STRUCTURE The program STRUCTURE may be a free software package for using multi-locus genotype data to analyze population structure. Its uses embrace inferring the presence of distinct populations, assigning individuals to populations, learning hybrid zones, distinctive migrants and admixed individuals, and estimating population allele frequencies in situations wherever many individuals are migrants or admixed. It is often applied to most of the commonly-used genetic markers, including SNPs, microsatellites, RFLPs and AFLPs. Furthermore, functions that STRUCTURE provides cannot be found in other social network data processing tools [10].

NETMINER NETMINER is an innovative software tool for Exploratory Analysis and Visualization of Network Data. It is often used for general analysis and teaching in social networks. This tool permits researchers to explore their network data visually and interactively, helps them to find underlying patterns and structures of the network. Especially, it can be effectively applied to various business fields where network structural factors have a lot of influences on the performance. Statistically, this program supports many standardized computer methods, including descriptive statistics, ANOVA, correlation, and regression [10].

4. KNOWLEDGE DISCOVERY AND OSNA

With the explosive growth of OSN in recent years, social network sites provide a huge amount of data which are vast, noisy, distributed and dynamic. The structure of social media data is unorganized and is displayed in different forms such as text, images, videos, and other multimedia data [11]. The three main characteristics of the social media data are that it is large, noisy and dynamic. Moreover, it became harder and harder to manually analyze very wide-ranging social networks and it is important to find a computational means to filter, categorize, classify, and analyze the contents of the social network. Various researchers put their insight to overcome these challenges by using automated information processing techniques from Knowledge discovery techniques. The concept of online social media mining is the process of representing, analyzing, and extracting actionable patterns and trends from raw social media data [12]. Knowledge discovery techniques provide the capability to discover new and meaningful knowledge from the large dataset collected from OSN. There are different techniques of knowledge discovery based on the type of collected data from OSN. The process of discovering hidden knowledge from structured data is called Data mining Techniques while the Text mining techniques is the process of discovering meaningful knowledge from unstructured text data. Web mining Techniques can be defined as to discover or extract useful information from the web data.

4.1 Data Mining

Data mining (DM) can be viewed as a result of the natural evolution of information technology [13]. DM is the process of extracting patterns/models and relationships from a huge amount of raw data and transforms it into an understandable representation using different techniques for more use [14]. The data could be expressed in different data types, such as transaction data in E-commerce applications or genetic expressions in bioinformatics research domain. Due to the huge amount of data in social media, social network sites appear to be typical sites to mine with data mining techniques [15]. Various data mining techniques are used for detecting useful knowledge from huge datasets like trends, patterns and rules [16]. Data mining includes many techniques such as clustering (supervised learning), Classification (unsupervised learning), Association rule mining, Sequential pattern mining, feature selection, instance selection, and visual analytics [17] [18]. Moreover, Figure 6 summarizes an overview of online social media mining process that includes different online media sources, social media data for social user, data preprocessing, data analysis, data interpretation, and pattern evaluation. Data Mining Algorithms are mostly classified into:-.

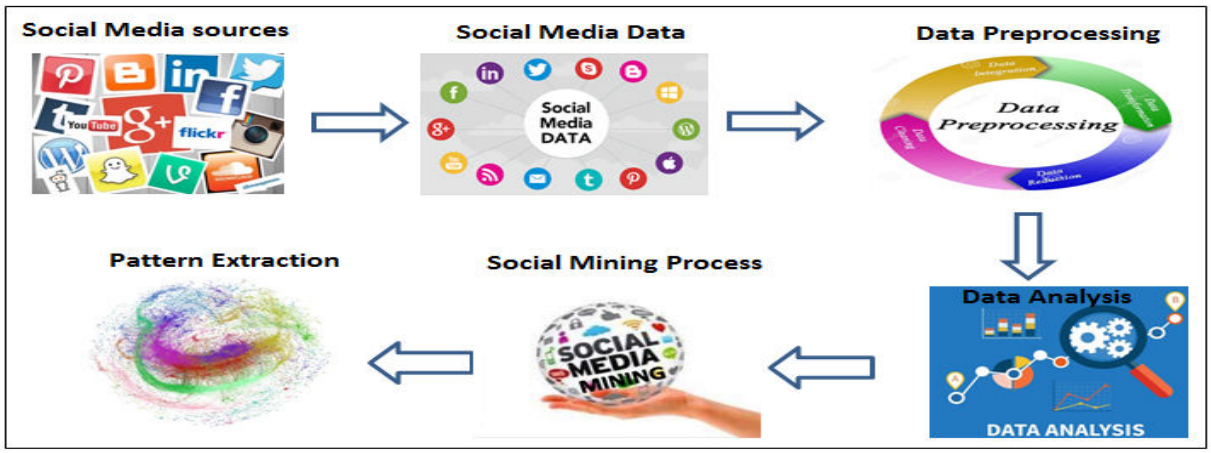


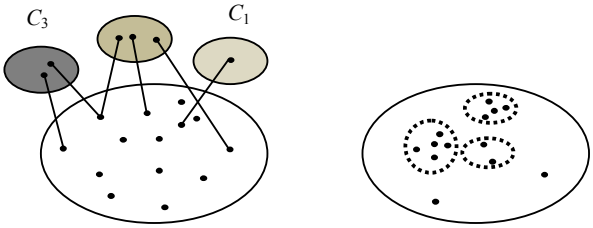
Fig 6: General Diagram of Data mining process in OSNA

4.1.1 Supervised Learning Algorithms

Classification is a common example of a supervised learning approach. The goal of data classification is to organize and categorize the data into distinct classes with known class labels according to a classification model. The main steps of the classification process are: - 1) establish the classifier through a learning process using the training dataset, 2) evaluate accuracy using the testing dataset by counting misclassified cases, and 3) predict the class for a new dataset. There are several classification methods such as the Statistical method, k-Nearest Neighborhood, Decision Tree Induction, Artificial Neural Networks, Bayesian Classification, Case-Based Reasoning, Genetic Algorithms, Rough Set, and Rule Induction.

4.1.2 Unsupervised Learning Algorithms

Clustering is an unsupervised classification where the number of classes, or class label, is not defined in advance. The goal of clustering is to analyze a set of data and to generate a set of classes or clusters that can be used to classify future data. In other words, clustering is a process of partitioning the dataset into a set of meaningful subclasses, called clusters. The clustering process depends on the similarity measure, intra: measure distance among the members within a cluster or inter: measure distance among clusters. The common clustering algorithms are partitioning algorithms, Hierarchical algorithms, Density-Based algorithms and, Model-Based algorithms. The difference among clustering methods has more than one dimension. The common dimensions are Scalability, Preprocessing, Different Attribute, and Cluster Output Format. The distinction between clustering and classification is illustrated in Figure 7.



(a) Classification (b) Clustering

Fig 7: In (a), three classes are known a priori, and documents are assigned to each of them. In (b), an unknown number of groupings must be inferred from the data based on a similarity criterion.

4.1.3 Semi-supervised Learning Algorithms

Semi-supervised learning is midway among supervised and unsupervised learning. In addition to unlabeled data, the algorithm is provided with some supervision information but not necessarily for all patterns. Semi-supervised learning algorithms are most appropriate where there exist small amounts of labeled data and large amounts of unlabeled data. Semi-supervised classification and semi-supervised clustering are the two types of semi-supervised learning.

4.2 Text Mining

In social networking, the nature of data is in the unstructured form and People used unstructured or semi-structured language to communicate. During the life chats, people do not care about the spellings and precise grammatical construction of a sentence that may result in different types of ambiguities, such as lexical, syntactic, and semantic [19]. Therefore, analyzing and extracting information patterns from such unstructured data are more complex. With the increasing number of electronic information such as digital libraries, electronic mail, and blog so it is difficult and costly to categorize it manually, Text mining is the solution for the above problems. Text mining is a knowledge discovery process to extract interesting and non-trivial patterns from natural language in a shortest time period [19]. Text mining techniques become much more complex as compared with data mining because of the unstructured and irregular nature of natural language text, whereas data mining deals with the structured sets of data [20]. Facebook are rich website in texts contents in the form of comments, wall posts, social media,

and blogs. Most of the surveys centered on the applying of various text mining techniques on unstructured data however do not specifically target the datasets in social networking websites [21- 25]. Classification and clustering are the two text mining techniques that are widely used for mining the unstructured text available on the web. An important phase in text mining process is the pre-processing phase since it organizes documents into a fixed number of pre-defined categories. Feature extraction and feature selection are two basic methods of text pre-processing in text mining. In feature selection, the document is treated as a sequence of word strings and split words by removing punctuations and removing all stop words to improve the effectiveness and efficiency of text processing after that return each word to the root form by using stemming algorithm. Eliminate irrelevant and redundant information from the target text is the basic purpose of the feature selection method. In this step, text document is represented as a vector space model and each dimension represented a separate term as a single word, keyword, or a phase. Term frequency and inverse document frequency are the two basic methods used to calculate feature vector. Then apply one of the most commonly text mining techniques for text analysis in social networking. Figure 8 reviews a framework of the Text mining process in analysis OSN. It begins with collected one online social media source or collection of sources, after that text gathering process from all sources, the preprocessing phase includes tokenization, filtration, and steaming steps, document presentation, and then apply one of the text mining techniques for analysis online social network.

4.2.1 Text mining using Classification (Supervised)

Classification is the process that classifies each text to a certain category and can be divided into two categories: a) machine learning-based text classification and b) ontology-based text classification.

4.2.2 Text mining using Clustering (Unsupervised)

In document clustering, the numbers of the classes are not known in advance. Documents are often grouped along based on a particular class. The clustering techniques can be divided into three categories: a) hierarchical clustering b) partitioned clustering, and c) semantic-based clustering that are detailed in the subsequent text.

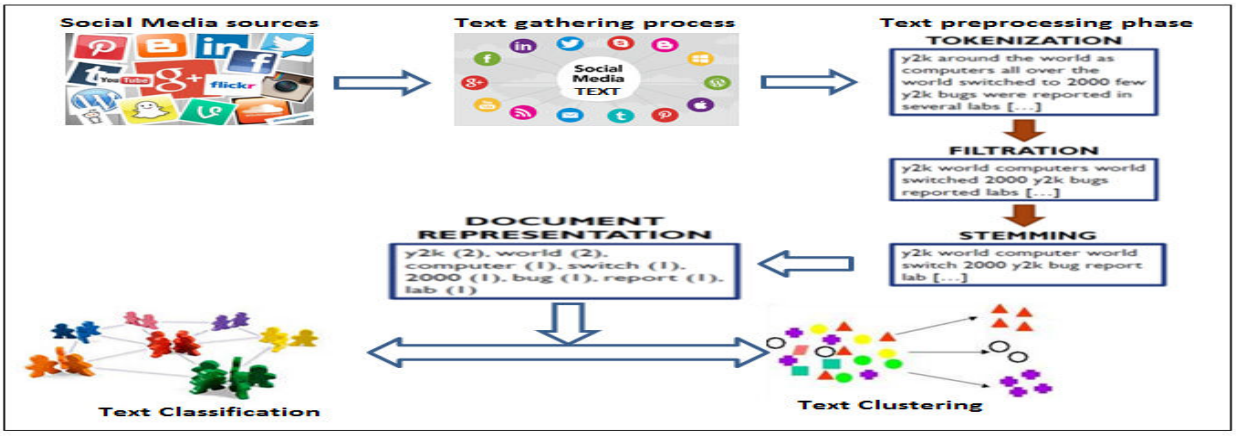


Fig 8: Text Mining process in OSNA

4.3 Web Mining

Web mining is considered an application of data mining techniques to the World Wide Web and focuses on the discovery of potentially meaningful and previously unknown knowledge from data such as online mailing lists, blogs, and social media [26]. In Web mining, the targets of analysis are in the form of web content mining, web structure mining and web usage mining [27]. Data collection in Web mining can be a substantial task especially for Web structure and content mining, which involves crawling for a large number of target Web pages.

4.3.1 Web mining Taxonomy

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined as shown in Figure 9 into: Web Structure Mining, Web Content Mining and Web Usage Mining.

4.3.1.1 Web content mining

Web content mining is a web mining technique aims to extract and analyze the contents in the web, such as HTML (semi structured), plaintext (unstructured), or XML (structured) documents and can categorize or classify documents on an on-line social networking website content by applying techniques from multidisciplinary fields including data mining, machine learning, natural-language processing, information]n retrieval, and statistics [28]. It can also be used in social networks analysis to analyze users reading interests to determine their favorite content. The common applications for web content mining are finding keywords, discovering grammatical rules and collocations, text classification, patterns discovery, text and document clustering etc.

4.3.1.2 Web structure mining

Web structure mining is a technique that can be used to easily understand, analyze, and construct social networks to extract the Web's hyperlink structure, e-mail's links or other sources. Web structure mining usually uses graphs and visualized means to represent the data and information collected from online social networks, enabling the analyst to easily understand and analyze social networks much easily [29]. Researchers have developed different algorithms to extract the structure of web site based on hyperlink analysis [30-33]. The common application of web usage mining are finding out mirrored web sites, discovering the nature of the hierarchy of hyperlinks in the web sites of a particular domain to study how the flow of information affects the design of the Web sites, personalization and recommendation system etc.

4.3.1.3 Web usage mining

Web usage mining refers to the process of discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining plays an essential role in the analysis of online social networks where usage data and users communication over the online social network can be transformed into relational data from which social network structure can be constructed [34]. The sources of data used here are any data related to user interaction to the web site such as Web log sources such as server access logs, proxy server logs, client browsers history files, cookies files, mouse clicks, user sessions [35, 36]. In order to produce the right data for mining in Web usage mining, the pre-processing of click stream data in usage logs is considered one of the key issues to achieve this. The common applications of web usage mining are the classification of user navigation pattern predicting user's future intentions, improvement in customer relationship management by finding out interested users, understanding the patterns and the web structure, important customers, improving the design of this collection of the information, and particular pages.

All the above three categories of web mining can be used combined to analyze OSN data. There are studies where hyperlinks are used to predict Web content [37]. In another study, web usage mining and web content mining techniques were combined to create user content profiles. Several studies also exist who have combined web usage data with semantics and ontologies for improving the Web personalization [38-41]. Table 1 shows a comparison between the three Web mining taxonomy in OSNA

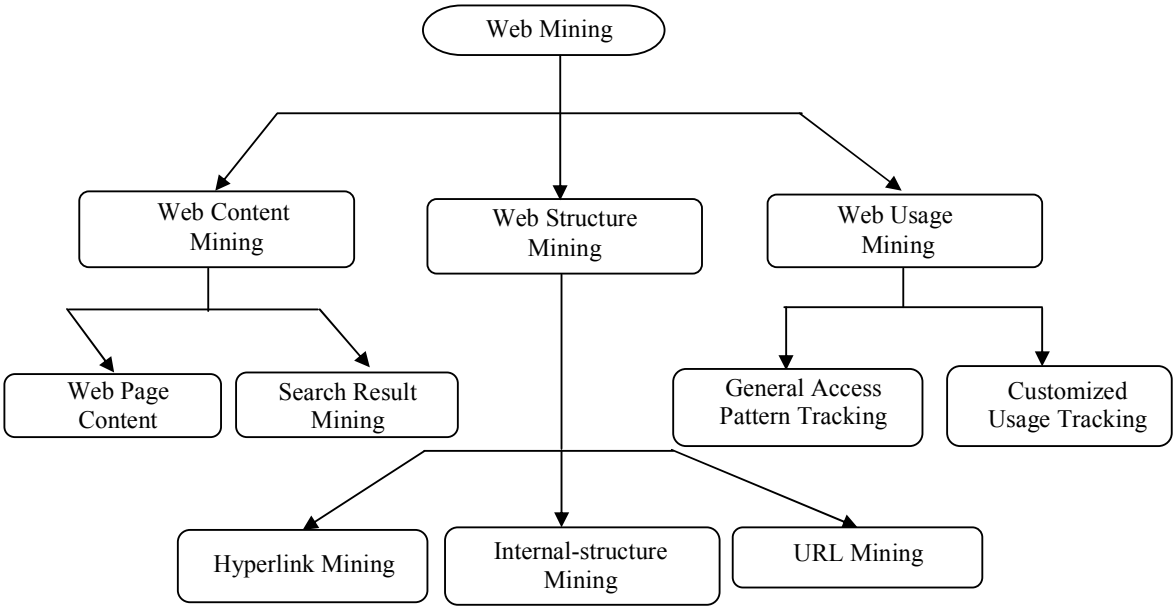


Fig 9: Web mining taxonomy

4.3.2 Web mining Techniques

In Web mining, there are many different kinds of web mining techniques. In this section, examples will be given of the most two web mining techniques used for social networks analysis are Clustering and Association Rule Mining.

4.3.2.1 Clustering

Discovering the group of closest people in the network is usually the main mission in social networks analysis. Clustering is similar to classification but it is an unsupervised learning process. The clustering technique can be used for identifying more clusters and groups in large social networks based on their similarities. A similarity between objects is defined by similarity functions usually; similarities are quantitatively specified as distance. In addition, clustering can provide more information of the members in a group and the relationship between groups in a social network.

4.3.2.2 Association Rule mining

Association rule mining technique in the social network analysis can used to discover the hidden relationships between nodes in a social network or even cross networks. The extracted patterns from the association rules technique can be represented in the form of association rules. The two most interestingness measures of rule are support and confidence. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold [42]. Users or domain experts can set such minimum support and confidence thresholds. For example, an association rule for social networks analysis may be "person A knows person B and also knows person C, the support is s=0.9 and the confidence is c=0.5". In addition, association rule mining is helpful for the application after social networks analysis, such as recommendation systems or information filtering systems.

The overall process of using Web mining for social networks analysis is shown in Figure 10. The first step includes the selection of analysis targets, such as Web, Email, Facebook etc. and sometimes more than one target will be selected. After that, the selection of online social networks analysis methodology will be completed. In the data preparation step, data will be collected for analysis, cleaned and formed in the format to store. The next step is selecting the suitable web mining techniques to analyze the data collected and sometimes a combination of techniques is required. The results of the analysis process are presented and interpreted, recommendation and action. Visualization techniques are sometimes used to assist the presentation of the results of the analysis, such as the extracted social networks.

Table 1. The obtained overall F-measure comparison for four clustering algorithms on the four datasets

	Web Structure mining	Web Content mining	Web Usage mining
The main aim	Focuses on the hyperlinks at the inter web site level	Focuses on the structure of the inner web site	Focuses on the secondary data extracted from the interaction of the user with the web
The Purpose	Extract the Web's hyperlink structure from the social networks	Extract and analyze the contents in the social media	Analyze the navigation behavior of users in the social networks
Data Form	Link structure	Semi structured, Unstructured, or Structured	User communications on online social networks websites
Source of Data	Web's hyperlink e-mail's links other sources	HTML ,plaintext, XML discussion topics, opinions, positive or negative sentiments expressed	Server access logs, Proxy server logs, Client browsers history files, Cookies files, and Mouse clicks
Representation	Graph Visualized means	Graph Relational Table	Graph Relational Table
Common Applications	Clustering Categorizations	Clustering Categorizations Association Rules	Classification Association Rules
Used Combined	Hyperlinks are used to predict Web content		
		web usage mining and web content mining techniques were combined to create user content profiles	

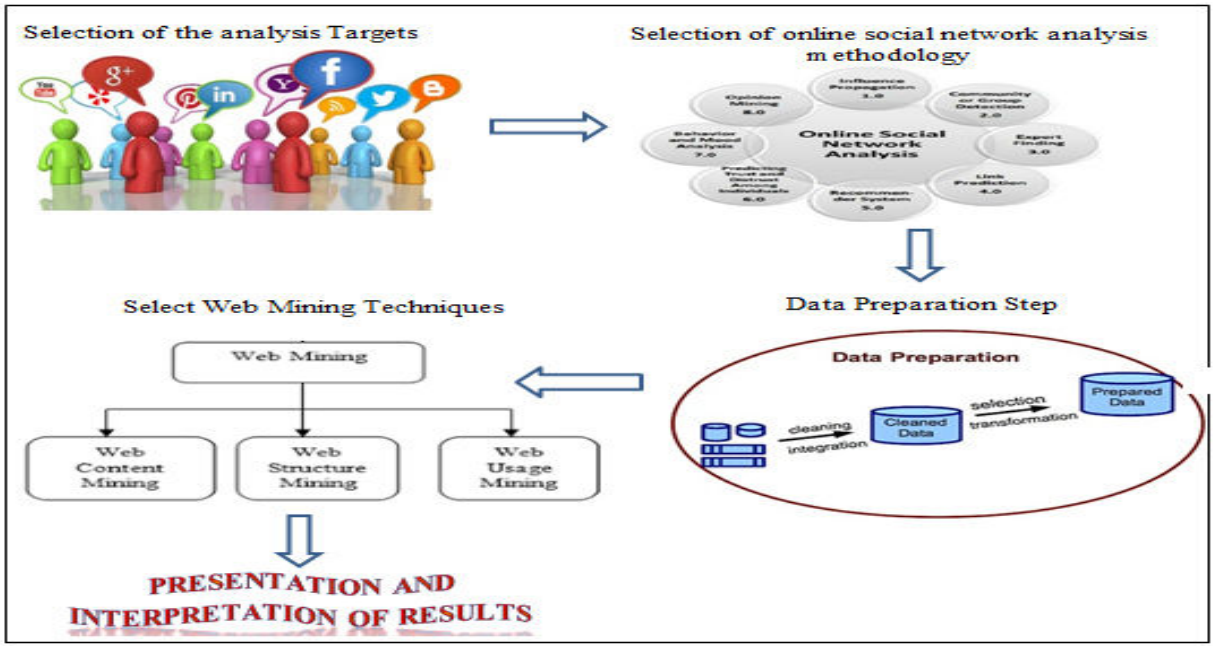


Fig 10: Web mining process in OSNA

5. CURRENT CHALLENGES IN OSNA

With the explosive growth of social networks, online social network analysis faces a number of challenges concerning the nature of social network data, data collection, data quality, and existing mining and analysis methods such as:

1. Massive and High Dimensionality of social data

Social network data are generated in very large amounts and highly complex in their nature. In social networks, different data types such as structured, semi-structured or unstructured are found in the form of texts, videos, images, hyperlinks or metadata. This high dimensionality data cannot be processed easily using traditional algorithms or database management tools which are a challenge to the researchers. Selecting suitable samples of data from the dimensional data is also a challenge. Moreover, information overload represents a significant challenge that requires large computing capacities and advanced sampling, extraction, analysis, and mining methods.

2. Heterogeneous Content

Due to the diverse writing, the same information is presented in different ways, using a completely different set of words, data types and other personal styles (i.e. grammar, abbreviations, mixed languages, etc.). This makes integration of information from multiple sources in social networks a

challenge to the researchers during content mining of information. Advanced algorithms are essential for mining knowledge from this heterogeneous content.

3. Noisy Data

Normally, in online social networks, websites contain many sections of information such as the main content, navigation links, advertisements, copyright notices, privacy policies, hyperlinks etc. For a specific application, only part of the information is useful and the rest is considered noise which is a research challenge. To perform web analysis and mining, the noise should be removed without losing too much.

4. Dynamic Data

The nature of information on the online social networks evolves continuously. For many applications, keeping up with the change, monitoring and analyzing the change, robust Web mining techniques and algorithms are required that is another challenge for researchers. More efficient methods must be improved for analyzing constantly changing data.

5. Low Quality Data

Due to the fact that there is no quality control of data in online social network, a large amount of information on online social network sites suffer from the low of quality, erroneous, and sometimes misleading data. Researchers face challenge during the data selection and pre-processing phase of web mining process to a large volume of inconsistently presented data. For example, consider the suicidal tweet below, presented exactly as it was posted:

"Tonight I want to dei, sorry every1, bye"

Because of typographical errors/deliberate misspellings ("dei") or shorthand (every1) used, this sort of statement could easily confound any detection efforts.

6. Efficient Mining Algorithms

Efficient algorithms are required to study social interactions, finding community structures, and analyzing overlapping communities in online social networks. Other challenges include finding communities in social networks, finding patterns in social networks and analyzing overlapping communities.

6. CONCLUSION

Online social networks consider a perfect reflection of the structure and dynamics of the society since they are platforms to share media, ideas, news, and links. Due to the significant and rapid growth in the amount of information and the number of users in OSN, Web users always suffer the problems of information overload. The eminent challenge lies in mining this great amount of data to extract, represent and exploit meaningful knowledge from OSN.

This survey provided a thorough understanding of different Knowledge discovery techniques as well as the application of these techniques in online social network analysis. The survey provided a comprehensive overview of all the existing Data, Text, and Web mining techniques that can be used for the extraction of meaningful knowledge from various types of OSN data. This survey will definitely provide a roadmap for researchers to proceed with knowledge discovery techniques that will be useful for online social network analysis

7. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their help to improve this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

8. REFERENCES

[1] Taprial V. and Kanwar P. 2012. Understanding Social Media. Published in Book Boon first edition. https://bookboon.com/en/understanding-social-media-ebook

[2] S. Tabassum, F. Pereira, S. Fernandes, J. Gama, "Social network analysis: An overview", Journal of WIREs Data Mining Knowlge Discovery 2018.

[3] P. Chaudhary, S. Gupta, B.B. Gupta, V.S. Chandra, S. Selvakumar, M. Fire, R. Goldschmidt, Y. Elovici, S. Gangwar, M. Kumar, P.K. Meena, L. Sharma, "Auditing defense against XSS worms in online social network-based web applications", in: Handbook of research on Modern Cryptographic Solutions for Computer and Cyber Security, Vol. 36, IGI Global, 2016, pp.216-245. No. 5, 1AD.

[4] H. Chen, R.H.L. Chiang, V.C. Storey, "Business intelligence and analytics: from big data to big impact", Mis Q 36 (2012) 1165–1188.

[5] Chakrabarti, S. 2003. Mining the web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publisher, USA.

[6] E. Raju, K. Sravanthi, "Analysis of social networks using techniques of web mining", Journal of advanced research in computer science and software engineering, Vol. 2, Issue 10, 2012. pp.: 443-450.

[7] Zhu, J. J. H. 2007. Opportunities and Challenges for Network Analysis of Social and Behavioral Data. Seminar Series on Chaos, Control and Complex Networks City University of Hong Kong, Poly U University of Hong Kong & IEEE Hong Kong R&A/CS Joint Chapter.

[8] Borgatti, S., Everett, M. and Freeman, L. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

[9] It is freely available, for noncommercial use, at its homepage: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[10] B. Furh, "Handbook of Social Network Technologies and Applications", Springer New York Dordrecht Heidelberg London, Springer Science+Business Media, LLC 2010.

[11] A.L. Kavanaugh, E. a. Fox, S.D. Sheetz, S. Yang, L.T. Li, D.J. Shoemaker, et al., "Social media use by government: From the routine to the critical", Gov. Inf. Q. 29 (2012) 480–491.

[12] "Social Network Marketing: The basics", http://www.labroots.com/Social_Networking_the_Basics.pdf, Aug 01, 2013.

[13] Jiawei H. and Kamber, M. 2001. Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, New York, USA.

[14] Umadevi B., Sundar D., and Alli Dr.P. 2013. An Optimized Approach to Predict the Stock Market Behavior and Investment Decision Making using Benchmark Algorithms for Naive Investors. In proceedings of Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on IEEE Explore Digital Library, pg1 -5.

[15] Cortizo, J., Carrero, F., Gomez, J., Monsalve, B., Puertas, E. 2009. Introduction to Mining SM. In Proceedings of the 1st International Workshop on Mining SM , 1 – 3.

[16] Kagdi, H., Collard, M. L., Maletic, J. I. 2007. A survey and taxonomy of approaches for mining software repositories in the context of software evolution. J. Softw. Maint. Evol.: Res. Pract, 19, 77-131.

[17] Richardson, M. and Domingos, P. 2001. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In NIPS, pages 1441–1448.

[18] B. Umadevi, P. Surya, "A Review on Various Data Mining Techniques in Social Media", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2017.

[19] Sorensen, L. 2009. User managed trust in social networking comparing Facebook, MySpace and LinkedIn. In Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology, (Wireless VITAE 09), 427–431.

[20] Y. Kano, W. A. Baumgartner, L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunter, and T. Tsujii, "Data mining: concept and techniques", Oxford Journal of Bioinformatics 25(15), 2009.

[21] Yin, S.,Wang, G., Qiu, Y and Zhang,W. 2007. Research and implement of classification algorithm on web text mining. In Proceedings of 3rd International Conference on Semantics, Knowledge and Grid, 446–449.

[22] Tekiner, F., Aanaiadou, S., Tsuruoka, Y. and Tsuji, J. 2009. Highly scalable text mining parallel tagging application. In Proceedings of IEEE 5th International Conference on Soft Computing, Computing with Words and Perception in System Analysis, Decision and Control (ICSCCW), 1–4.

[23] T. Jo, "NTC (Neural Text Categorizer): neural network for text categorization", International Journal of Information Science 2(2), 83–96, 2010.

[24] Ringel, M. M., Teevan, J. and Panovich, K. 2010.What do people ask their social networks, and why: a survey study of status message question & answer behavior. In Proceedings of International Conference on Human Factors in Computing Systems (CHI 10), 56–62.

[25] Li, J., Khan, S. U., Li, Q., Ghani, N., Bouvry, P. & Zhang, W. 2011a. Efficient data sharing over large-scale distributed communities. In Intelligent Decision Systems in Large-Scale Distributed Environments, Bouvry, P., Gonzalez-Velez, H. & Kolodziej, J. (eds). Springer, New York, NY, USA, 2011, pp. 110–128, ISBN: 978-3-642-21270-3.

[26] Chakrabarti, S. 2003. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, USA.

[27] Cooley, R., Mobasher, B. and Srivastave, J. 1997. Web Mining: Information and Pattern Discovery on the World Wide Web. In Proceedings of the 9th IEEE InternationalConferenc onToowithArtificial Intelligence, pp. 558-567, Newport Beach, CA, USA.

[28] Liu B. Carey MJ, Ceri S, eds. 2006. Web data mining: exploring hyperlinks, contents and usage data. In carey MJ, Ceri S, eds. Berlin: Springer.

[29] S. M. Goodreau, "Advances in exponential random graph (p*) models applied to a large social network", Social Networks, 29(2), 231–248. 2007.

[30] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM (JACM), 46(5), 604–632. 1999.

[31] Biswal, B. 2008. Web site optimization through mining user navigational patterns, web engineering and application. New Delhi: Narosa Publishing House.

[32] Li, F. 2008. Extracting structure of web site based on hyperlink analysis. In proceedings of fourth international conference on wireless communication. Networking and Mobile Computing, 1–4.

[33] X. Fang, and O. Sheng, "LinkSelector: Web mining approach to hyperlink selection for web portals", Journal of ACM Transactions on Internet Technology, 4(2), 209–237.2004.

[34] Lento, T., Welser, H. T., Gu, L., and Smith, M. 2006. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In proceedings of 3rd Annual Workshop on the Weblogging ecosystem (Vol. 12).

[35] Nina, S. P., Rahaman, M., Bhuiyan, K., and Khandakar E. 2009. Pattern Discovery Of Web Usage Mining. In proceedings of International Conference on Computer Technology and Development, Vol. 1.

[36] R.Kosala, and H. Blockeel, "Web mining research: A survey", Journal of ACM SIGKDD Explorations Newsletter, 2(1), 1–15. 2000

[37] Mladenic, D., Grobelnik, M. 1999. Predicting content from hyperlinks. In Proceedings of the 16th International ICML99 Workshop on Machine Learning in Text Data Analysis (pp. 109–113).

[38] B. Berendt, "Using site semantic to analyze, visualize and support navigation", Journal of Data Mining and Knowledge Discovery, 6, 37–59. 2002.

[39] Dai, H. Mobasher, B. 2003. A road map to more effective Web personalization; Integrating domain knowledge with Web usage mining. In Proceedings of the International Conference on Internet Computing (IC 2003), Las Vegas, Nevada.

[40] Oberle, D., Berendt, B., Hotho, A., Gonzalez, J. 2003. Conceptual user tracking. Lecture notes on artificial intelligence, Vol. 2663, pp. 155–164.

[41] M. Spiliopoulou, and C. Pohle, "Data mining for measuring and improving the success of Web sites", Journal of Data Mining and Knowledge Discover, 5(1–2), 85–114. 2001

[42] Agrawal, R. and Srikant, R. 2017. Fast algorithms for mining association rules. In proceedings of International Conference of VLDB, pp. 487– 499, 1994. V. Bhatia, R. Rani, "A parallel fuzzy clustering algorithm for large graphs using Pregel", Journal of Expert System with Applications, 78(c):135-144.