A Data Analytic Framework for Unstructured Text

Hassanin M. Al-Barhamtoshy and Fathy E. Eassa
Computing and Information Technology, King Abdulaziz University (KAU), Saudi Arabia
hassanin@kau.edu.sa and feassa@kau.edu.sa

Abstract. This paper describes a systematic flow of the unstructured data in industry, collected data, stored data, and the amount of data. Big data uses salable storage index and distributed approach to retrieve required information. Therefore, the paper introduces an unstructured data framework for managing and discovering using the 3Vs of big data: variety, velocity, and volume. Different approaches for managing, collecting, and classification of twitter data, e-mail data and free text are required to manage resources more efficiently, and building software platform around scalable analytics. The development processes in this paper is implemented in Python, build up lexicon and calculated sentiment score. Analyzing twitter data and e-mail data answered many of questions; what are people talking about?, what is the most important? … etc. The accuracy of the proposed classifier was 77.78, without stop words and was 78.76 and 79.94 with stop words (25 and 174) respectively. If the stop words are increased, the accuracy will be 87.69.It has been 10% better accuracy between Naïve Bayes and Maximum Entropy classifier.
[Hassanin M. Al-Barhamtoshy and Fathy E. Eassa. **A Data Analytic Framework for Unstructured Text.** *Life Sci. J* 2014; 11(10):339-350] (ISSN: 1097-8135). http://www.lifesciencesite.com. 48

Keywords: Sentiment analysis, Cloud computing; big data; twitters, Social network analysis, and unstructured data.

1. Introduction

"Big Data" is defined as the amount of data to store, manage, and process in effective manner [1]. Such process includes robust analysis of the data itself, and capability of tools used to analyze it. Accordingly, to the technology capabilities; tens hundreds of terabytes storage are needed to handle big data. Therefore, big data and related analysis are very essential in modern science and business [2]. Such big data are generated from audio, videos, images, event streams, logs, posts, search queries, health records, social networking interactions, online transactions, emails, science data, sensors and mobile phones and their applications [3, 4].

Therefore, growth rate of data collected is national challenging. This growth rate is very fast exceeding design capability to handle data effectively and also extract relevant meaning for decision making.

Many of governmental sectors emphasize on how big data create "value" in different domains and across discipline fields [1]. Structured, semi structure and un-structure data will continue to grow. Consequently, different organizations are challenged to manage the big data they have.

For example, from the medical field; Fox [5] illustrates how patients medical data record and current health situation are used to plan and predict patient participation in wellness and disease management systems. Therefore, doctors should collect and analyze patient's data using such systems. Accordingly, many of cloud services require users to share their data like health records for data mining and analytical process, taken into consideration privacy and security [6].

Internet Data Center (IDC) announced that the global big data will increase by 50 times next decade [7]. How to accumulate these fast-increasing, massive amounts of information? How to analyze this amount of data? Therefore, today many researches focused on this era. So, IDC defines big data as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis" [8]. This definition includes "three main characteristics of big data": volume, velocity, and variety. Two other characteristics seem relevant: value and complexity. Summarize as "4V+C" [7].

Chamber Commerce in a recent report of the U.S., stated that the market's data is expected to grow to $16.9 billion in 2015 [9]. IBM's statement "Every day, we create 2.5 quintillion bytes of data" – so greater than90 % of the data in the world has been created in the last two years only.

A social media data mining system is implemented [10] to be used in forecasting events related to Latin American community. The related method extracts a small number of tweets from the available data on "twitter.com" using public API as well as a commercial data streaming.

Another research focuses on the second parameter of the Big Data (Velocity); therefore, a popular open-source stream processing engine (Storm) is used to perform real integration and trend detection on Twitter and Bitly streams [11]. Also, Clowd Flows platform with the real-time data streams is used to create specialized type of streaming and a stream mining, [12].

One of the first studies on Twitter was published in 2010 [13]. The study investigated Twitter's topological characteristics, and its influence as a new standard of information sharing [13,14]. Accordingly, OLAP and data mining tools are used for exploring the data and for additional complicated analysis [14].

The goal of the paper is to provide general view of unstructured data, challenges, technology, and data manager implementation. Therefore, the paper is organized as follows. Section2 introduces to web scaling with text processing taken into consideration the cloud service layers. The proposed high level framework, architecture and related definition of the big data will be illustrated in Section 3. Section 4 illustrates the proposed frame work of the knowledge data discover (KDD) including: service-based, meta-data storage and data preparation. The implementation manager of data annotation for free text and overall evaluation will present in Section 5.Where, Section 6 discusses conclusion and future work.

Gartner Laney describes Big Data as 3 V: Volume, Variety and Velocity [1,4] as shown in figure 1. Two additional features seem relevant: Value and Complexity [1]. Another suggestion is proposed that includes three fundamental areas to be addressed: storage, management and processing issues.

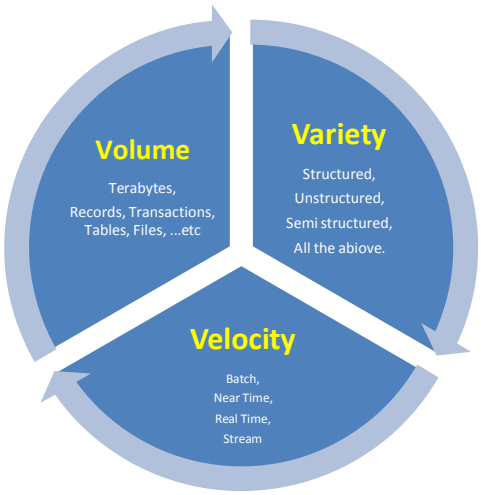

Figure 1: Big Data Definition (3 Vs)

Currently, high energy physics experiments generate more than one terabyte of data per day, such as DZero. Face book serves more than 570 billion pages per month, and captures 3 billion new photos every month. Therefore, American government announced the "Big Data Research and Development Initiative", i.e., the unstructured data will be national policy for the first time.

Hey T. and et. al., at Microsoft research center, announced in the "The Fourth Paradigm Data Intensive Scientific Discovery" a new scientific computing architecture, it composed from experiments, simulations, archives, literatures and instruments. They mentioned that the scientific computing storage is doubling every two years [15].

Recently, there is a new storage medium, moreover, social media includes more than 12 terabytes of tweets every day, and average retweets are 144 per tweet [1].

Now, the disk space is around 4 terabytes; consequently, 12 terabytes would require 3 disks (per day). Therefore, collection of data should be transmitted along with integrity and stream management process. The sources of such unstructured data are varied in format, and method of collection (texts, images, pictures, drawings, sounds, videos, user interface designs, etc.).

User's requirements, technologies that are needed, and the design of the proposed system are challenges required to work with big data discovery. Consequently, Stonebraker and Hong [6] support the previous idea. They said that users' need understanding and technologies that can be used to solve the problems should be investigated.

In another way, small to large volumes of data (produced via transaction processing) are not capable to extract such huge volumes and therefore cannot be executed in minimum time.

There is no known tool to access large quantities of unstructured or semi-structured big data [1, 16]. The online analytical processing (OLAP) may be show or have limited reading all the data into memory or have limited supporting discovery functions.

According to the previous discussion, unstructured data collection and processing of very large data need a new paradigm to handle, manipulate and discover such data. However, due to the multiple data types categorizations (processed formats): structure, unstructured and semi structured. In added to difficulty of analyzing such large datasets, the new paradigm accomplishes architecture challenges. Therefore, knowledge discovery from data (KDD) became strategically important enterprises, business, research enterprises, governmental sectors, educational institutes, and scientific activities [17].

Recent survey of big data reported that 63% of databases were smaller than 10 TB (TDWI Research Web Site. Therefore, many of servers harness to perform map/reduce analysis by using datasets to be held entirely in memory.

2. Web Scale Text Processing Structure

Most of web text processing, internet tweets, e-mail streaming, and therefore their requirements of

dataset annotator relates to database; and fits into an n-tier application. Within the web-scale text processing, the cloud computing services; Software as a services (SaaS), Platform as a Services (PaaS), Infrastructure as a Services (IaaS) [18] and Data as a Services (DaaS) be supposed to be employed, as illustrated in figure 2 [19].

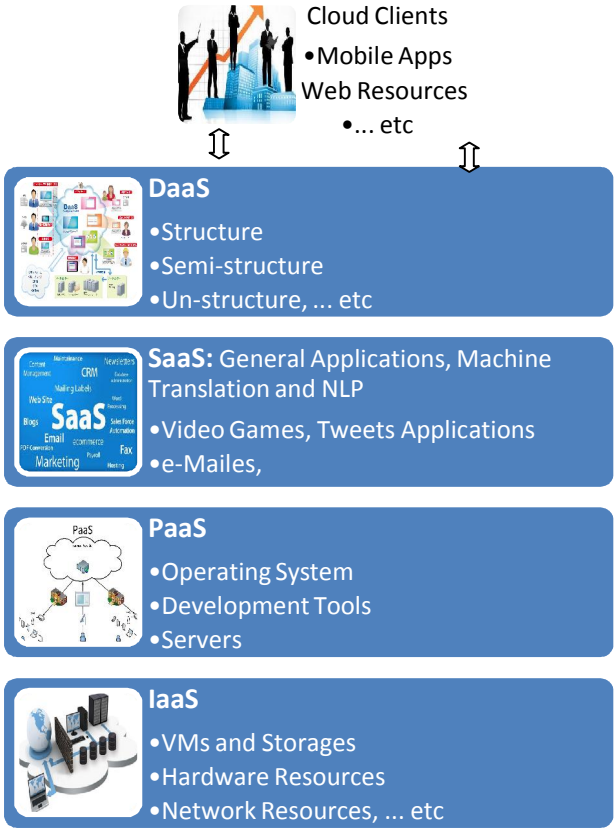

Figure 2: Proposed Cloud Service Layers [19]

The SaaS layer introduces software applications, such as general editors, word processors, spread sheets, etc. over the cloud networks. PaaS presents a host operating system, cloud development tools, hardware servers being used to support media streaming using much more demanding Quality of Service (QoS) restrictions. While, IaaS delivers virtual machines or processors, supports storage memory or auxiliary space and uses network resources to be introduced to the clients. Finally, DaaS includes large quantity of available data in significant volumes (Peta bytes or more). Such data may have online activities like social media, mobile computing, scientific activities and the collation of language sources (surveys, forms, etc.).

Therefore, cloud clients can access any of the previous web browsers or a thin client with the ability to remotely access any services from the cloud. Consequently, Amazon's EC2 and apple's iCloud introduced their products with a way of smaller, lighter and portable devices [18]. So, classification of unstructured data (big data varieties) should be considered as services. Such services are used to process data into information(as shown in figure 3 [20]), process information to get knowledge, and process knowledge to result intelligence, using the following steps: 1. Knowledge Discovery; 2. Web Scale Data mining; 3. Employer people with knowledge; 4.Employer Apps and services with intelligence. In other direction, location based service can be exemplified to use cloud computing services: SaaS, PaaS, IaaS, and DaaS. Consequently, the data as a service (DaaS) includes:

1. Data classification and varieties that are needed.
a. Structured; b. Semi-structured and; c. Un-structured.
2. Location Based Service (LBS) provider, that generates different data that are needed.
a. Maps; b. Road Networks; c. Search Results; d. … others
3. User generated data, that can be classified into:
a. Location history; b. Photos and videos; c. Search queries; and d. … Others.
4. Interface that includes third party services to do something; e.g.;
a. To educate; b. To navigate; and c. … Others

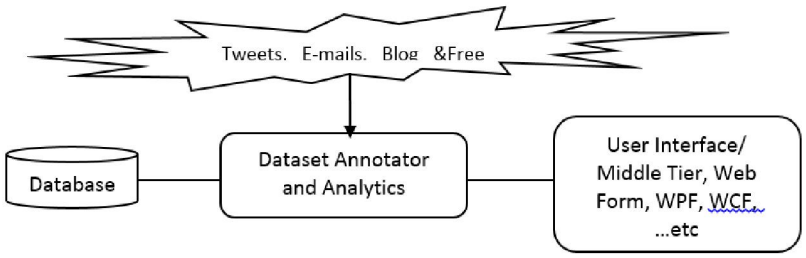

Figure 3: Web-Scale Text Processing Structure [20]

3. Data and Text Classification

The following section recommends data collection, dataset description, and rates unstructured text using numerical values, taken into consideration people's behavior. The practical case uses numbers, frequency and voting in particular case, at the classification level. Therefore, in output case of the dataset includes numerical attributes in a representation form (table), at it is called "structured data". But unstructured dataset may include twitter messages, emails messages, newspaper, blogs, and free text books. These types of data do not represented in tables.

Opinion mining and sentiment analysis are two empirical research areas to analyze opinions over the world web [23, 24]. Therefore, classification by polarity uses positive and negative classes to satisfy opinion. Accordingly, the positive and negative classes should be balanced and equalized to have same number of documents for each [24]. On the other direction, Abdul-Majeed *et al.* [25] have used unbalanced Arabic datasets and they have published results on those datasets. The second approach [26] performed a comparative study between balanced and unbalanced opinion for customer review.

3.1 Data Collection

Big data analysis still needs hundreds of servers running in parallel using diverse software applications. Examples of literature are available in Arabic culture (free textbooks). Biological, scientific, government, astronomy, financial services, medical services, web blogs, chatting, text and document analysis, photography, video and audio streams, … etc. Therefore, many peoples and organizations have begged data and can benefit from its analysis to solve real problems [32].

Four Arabic datasets will be used in different two domains. The four datasets have been collected from online web: DS_1, DS_2, DS_3 and DS_4. Table 1 illustrates numbers and percentage of such datasets. Accordingly, web contains many of Arabic reference's books and Hadith's Sciences, many of works such as analysis, verification and clustering are needed. In addition, predictive analytics on social networks, text document classification, POS information, and machine translation will be employed.

Table 1. Classification of Arabic Datasets - Hadith

	Positive	Negative	Total Document
DS_1:Sahih	25 (50 %)	25 (50%)	50
DS_2: Hasan	15 (50 %)	15 (50 %)	30
DS_3: Da'eef	13 (50 %)	13 (50%)	26
DS_4: Mawdoo'	12 (50 %)	12 (50 %)	24

Suppose that, we want to create a model that can tell whether Hadith texts are Sahih, Hasan, Da'eef, or Mawdoo'. In other system, the model decides the decision to be like or dislike various thing products. Therefore, a list of words that includes evidence that a system like such things and another list of words that provides such system does not like. So, number of "like" words, number of "dislike" words are needed in addition to classify which number is higher.

Naïve Bayes with training dataset will be used. Accordingly, the formula:

$$P(h|D)=P(D|h)P(h)/P(D); \quad (1)$$

and;

$$H=\text{argmax}P(D|h)P(h) \quad (2)$$

Will be applied and used, and can be defined as the probability of seeing evidence to some data D given the hypothesis h [1].

Big data of Arabic texts and free textbooks were created and streamed in the web. The proposed approach here is to build management system to categorize and control of the Arabic text according to its value and available. The proposed architecture should support download text streaming, text preprocessing, text annotation, text analysis, text clustering / categorizing, and therefore text discovery, as illustrated in figure 4.

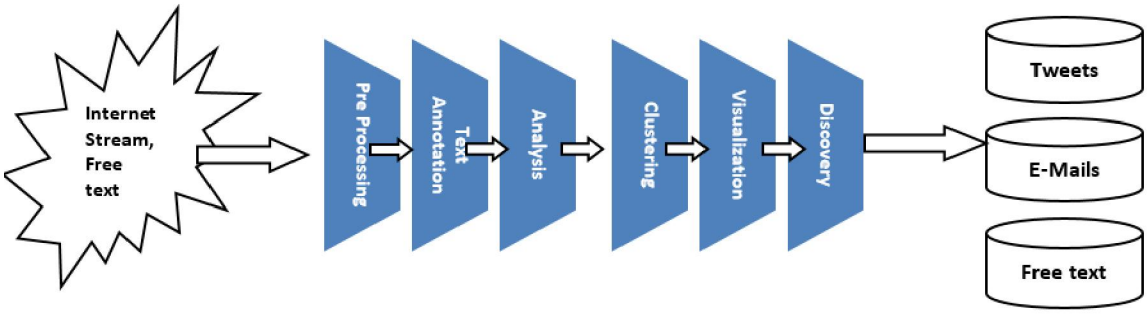

Figure 4: Proposed Approach for Text Processing

3.2 Training Datasets

Number of unique words (or vocabulary) will be used to compute probability for each word by given the hypothesis illustrates how to compute.

Step 1. Aggregate the documents tagged that highly deducted text stream.

Step 2. Count the number of occurrences for each word in the stream.

Step 3. Count the number of occurrences for each word in the vocabulary (n_v) exist in the stream.

Step 4. Compute $P(D_v| h_i) = (n_v + 1) / (n +|vocabulary|)$

Step 5. Compute: P(Hypothesis) x P ($Word_1$ | Hypothesis) x P($Word_2$ | Hypothesis) … x P($Word_n$ | Hypothesis) e.g.;

P(Like) x P ($Word_1$ | Like) x P($Word_2$ | Like) … x P($Word_n$ | Like)

and;

P (Dislike) x P ($Word_1$ | Dislike) x P($Word_2$ | Dislike) … x P($Word_n$ | Dislike)

Step 6. Compute the summation of the two hypothesis

$\prod = \sum P(W_i)$

$\prod' = \sum P(W'_i)$

Step 7. Find and select the hypothesis associated with the highest probability.

Step 8. IF the probability of the hypothesis (Like) is greater than for hypothesis (Dislike)

Then the decision will be Positive or Like.

Else the decision is negative or dislike

3.3 Dataset Description

Over 400 million tweets are generated everyday [21]. So, twitter's data is very important. One objectives of this paper is extracting twitters information, such as user information (twitter user Id), Tweets published by the user (textual description, no of tweets published by the user), user's network connections (geographic regions, and URL), and tweets creation date. Also, the dataset includes 4 types of Arabic text books (as mentioned in section 3.2).

This section proposes machine learning methodology for Arabic texts in sentiment analysis or data opinion. Such machine learning uses Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. Figure 4 illustrates the processes of the Arabic texts analysis. The processes start by collecting and reading texts from internet's datasets. The second step is used to decide if each text will be positive (like, favorite) or negative (dislike, un-favorite).The third process is feature extraction process, and it is employed to extract features and represent it in feature vector. Consequently, such feature vector will be used in the training process of the classifier.

There are three main components, within unstructured text. Such components are tweets, E-mail and free text services, respectively. All the three components expose a set of well-defined APIs in addition to two classifiers to analyze and classify of the unstructured data.

The preprocessing phase is used to annotate different data, besides to employ the stop words list [33]. Such stop words is useful to organize unstructured text with mining in good manner. So, preprocessing method should take place to remove such stop words that affect the performance of text mining tasks [33]. Subsequently, many words that occur frequently are eliminated to speed processing and save storage.

As mentioned before, there is an explosion in the size of social media over the internet, accordingly, a new data storage paradigm has been created, NoSQL is published to store big data. Mongo DB is one example of NoSQL implementation [21]. Document oriented storage; index support, queries, and speed are principles to support NoSQL implementation. Twits consist of connected nodes, between nodes there are connections or edges. There are two types of connections; one direction and both directions (→ and ↔).

The location can identified by two different methods: 1. Geotagging information; if the tweet was

published by smart mobile with GPS capability, and 2. User profile. Location field in user's profile is extracted from user location using 'API's translate function' to identify coordinates (longitude and latitude). Text's location is necessary information to be used in different visualizations. The IaaS manager provides management tool to support the network file stream (NFS) using GUI and virtualized environment. It interacts with the three different clusters: Tweets, e-mail and free text to collect physical and virtual machines within the clusters. In addition, it calls the APIs exposed by some public IaaS sites to employ more VMs for supporting.

Figure 5 illustrates the proposed multimodal for sentiment analysis, which integrates tweets, emails, and free text to make knowledge discovery. The three modules are based on the fact that they have been used routinely in text mining and text classification. Further, these classification methods complement one another in their advantaged and strengths.

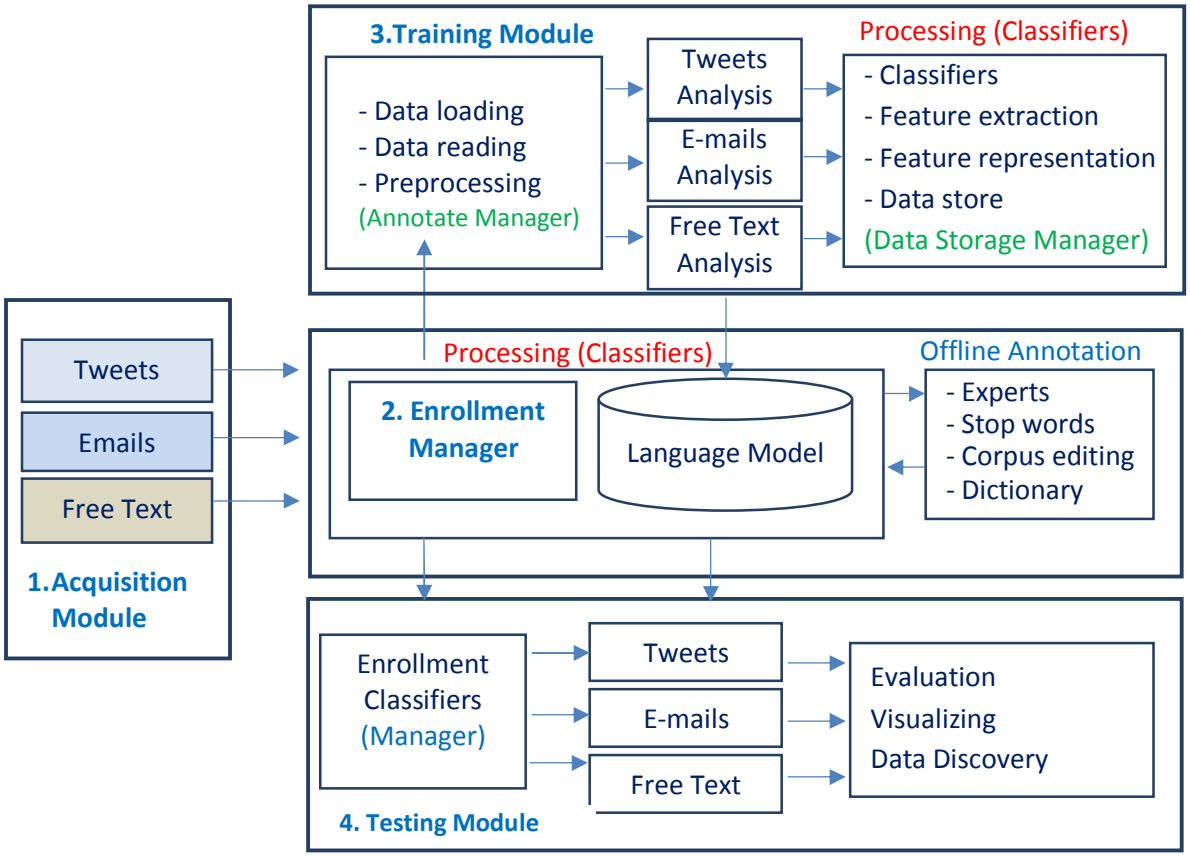

Figure 5: Multimodal Sentiment Analysis System

While tweets processing provides high accuracy due to tweets length (140 character), during training and testing, Emails and free text -on the other hand- are routinely uses extra processing routines (Tokenizer, Lemmatizer, Parser, Semantic analyzer …etc.). Our proposed system targets to implements most processes of natural language toolkit (NLTK). The block diagram consists from four modules: (1) Acquisition module, (2) Enrollment Manager, (3) Training module, and (4) Testing and verification module. The Acquisition module is responsible for acquiring tweets, emails and free text of a user who intend to use the system. The process of second module is used to manage decides to enroll the system inside training phase or testing phase, as well as to enroll offline operations (add, delete and update experts). The training module consists of three sub-modules to analyze tweets, emails, and free text using tokenizer, lemmatizer parser, and semantic analyzer. One method that is used for annotation by tokenizing the text input and giving each token a classified value. The tokenization process usually based on whitespace and punctuation- each word and punctuation mark has been pulled apart. It also classifies tweets, emails and text using feature extraction processes via standard classifiers (SVM and NB). The last module used to classify, evaluate and visualize the output and accuracy of the proposed solution.

4. The Architecture of Unstructured Data/Text

As mentioned in previous paper, that the data manager consists of two managers: metadata extraction manager and knowledge discovery manager [20].The manager is used to extract and retrieve all metadata of the unstructured data over the network and store them in the storage. The improved architecture of the unstructured data manager consists of many agents: stationary and mobile agents; see figure 6.

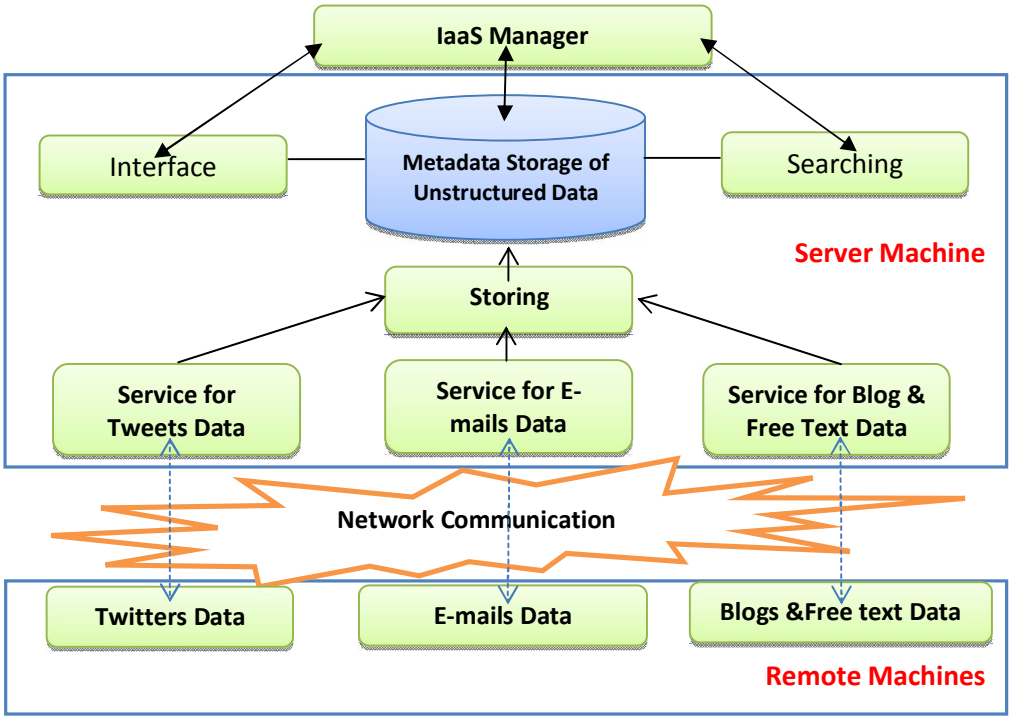

Figure 6: Modified Data Manager Architecture

Each service is migrated from the unstructured data server machine to a remote machine to collect the metadata from the existing unstructured data. There are services for tweets data, services for E-mails data, and services for blog and free text data. The mechanism of extracting the metadata from the existing data depends on the type of the data. The details of the mechanisms, algorithms, and techniques of extracting the metadata are employed. The services modules return with the collected (extracted) metadata to the unstructured data server and store it in the storage by the "storing module".

The interface receives the requests from the user of the unstructured-data manager and displays the outputs. Based on the received requests, the searching module joins the metadata and selects the information about existing data all over the distributed system and sends the information to the interface module to be displayed to the user.

The algorithms and techniques of extracting the metadata of each module depend on the varieties of unstructured-data (Tweets, E-mails, Blogs or Free Text). The natural language processing algorithms and techniques, data mining algorithms, or statistical techniques can be used for extraction. Also, the lexical analysis, data mining and parsing techniques can be used especially for extracting metadata. After the extraction activity, all modules return to the metadata server for giving the metadata to the storing module for registering them in the metadata storage.

4.1 The Metadata Storage Design

Different detail classification of each element in the figure 7will be described in the research.

4.2 Data Preparation

Data preparation is important during the analytics phase. At this level the data may be contain unusable format, missing values, errors, and compressed format. Therefore, additional tool may be used at this level (e.g. Hadoop is a good tool for this stage [27] with Google's Big Table, Map Reduce [28] and Google File System [29]). Additional data management solutions are needed such as Hive and HBase to store data sets, [28].

Distributed analytic databases (e.g. EMC Greenplum and HP Vertica) can be used to store and

analyze such structured data [28]. Such storing is employed using column-oriented fashion and distributed data over large scale multiple machines. Therefore, SQL and business intelligence software can be used for querying and backend for data streaming [20].

In other direction, natural language processing, semantic web analysis, semantic annotation, and data mining are extremely useful for social networks, e-mails, blogs and free text. Consequently, HBase (Apache HBase Web Site: http://hbase.apache.org) and Cassandra [30] can be used to prepare data as a key value data organization. Also, semantic ontologies will be used in information architecture to help system manager the diversity of data types generated and used by multi-disciplinary groups [20].

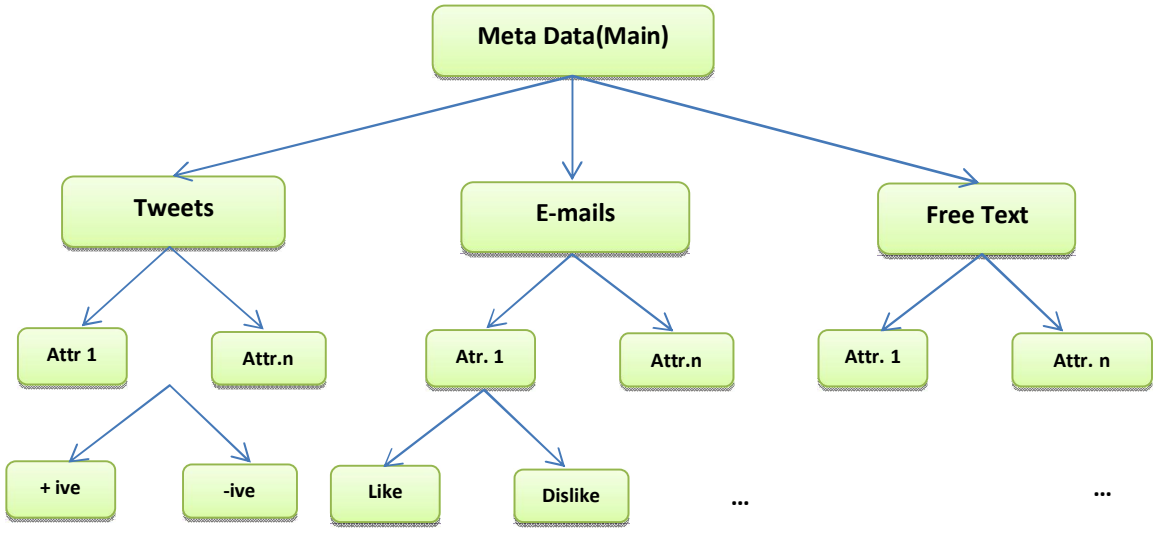

Figure 7: The Classification of the Metadata Storage

5. Data Management and Annotation

Annotation tasks range from simple data document labeling to text extent tagging and tag linking. Therefore, more information is needed to be contained within the annotation; category labeling. However, multiple categories use labels for a document, e.g. positive, negative, or neutral. Other labeling uses likes, dislikes and indifferent, or rates everything on a scale from 0 to 100, or binary classification. Classification and identification tasks are two annotation labels that refer to the entirety of a document. Therefore, many annotation tasks required in this approach, where tags are applied to specific areas of the text, rather than all of it at once. There are other types of task: part-of-speech tagging, named entity recognition, the time and event identification parts, opinion analysis and so on. Principally, any annotation that requires sections of the text to be given different labels falls into this class [31].

The metadata-type tags used for the document classification task could contain start and end indicators or could leave them out. However, with stand-off annotation it is required that locational indicators are present in each tag.

5.1 NLP Classifier Implementation

Text clustering is a methodology to organize collections of documents, and it is used with other fields such as information retrieval (IR) and topic identification [34]. Also, it goals to partition a given documents into meaningful classes. The quality of clusters can be evaluated using entropy [34]. Therefore, the entropy of a cluster C_r using size $n_r$is calculated as:

$$E(Cr) = -\frac{1}{Log\ q}\sum_{i=1}^{q}\left(\frac{Nri}{Nr}Log\frac{Nri}{Nr}\right)$$

Where q is the number of clusters that are considered correct for evaluation, and Nri is the number of documents for cluster i found in cluster r. The overall entropy is defined as follows:

$$Entropy = \sum_{r=1}^{k}\left(\frac{Nr}{N}E(Cr)\right)$$

The better clustering classification is the smaller values in the range of 0 to 1.

The manager consists of three manager modules named as **Loader** module, **Feature Extraction** module, and **Classifier** module**,** as shown in figure 8. The loader loads subsequent features from opinion mining and placed in the **warehouse**. Feature Extraction module defines a set of properties, which

are ranked based upon the role and importance of title/subject and determines such properties for each subject. Classifier Module gets the result from Feature Extraction module and classifies it by using classifier (Naïve Bayes or Support Vector Machine (SVM)).

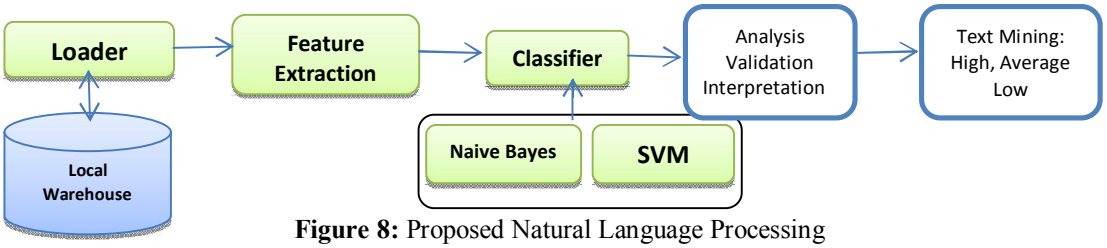

Figure 8: Proposed Natural Language Processing

At last, the proposed manager works with the analysis, validation and interpretation of collected data through an access to the classified documents, and decides the behavior or the attitude of the text to be high or average or low (text mining). Consequently, the practical applications of each of these methods use named part of speech (POS) as a case study. POS aims at marking up what probably think of as proper nouns- objects in the real world that have specific designators, not just generic labels. Therefore, the text can assigned by numbers to the tokens. Simply number every token in order, starting at 1 and going until there are no more tokens left. Another way is to assign numbers to each sentence, and identify each token by sentence number and its place in that sentence [31]. Figure (9-a) shows what it might look like to create this annotation for holy Quran using annotation tool (MAE).

Certainly, more identifying features could be added, such as paragraph number, document number, and so on. The advantage of having additional information used to identify tokens is the information that can be used later to help define features for the machine learning algorithms. Figure (9-b) illustrates screen shot of the Holy Quran after annotation model taking into consideration Part of Speech Tagging (POS).

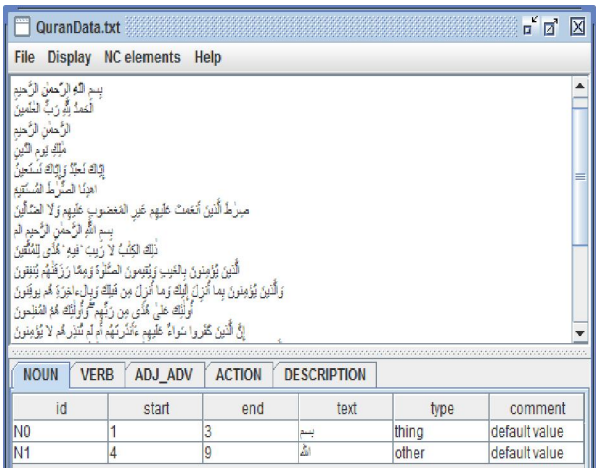

Figure 9-a: Named Entity Annotation of Quran

Figure 9-b: POS Annotation of Holy Quran

5.2. Data Model Manager Analysis

In this section basic analytic framework will be discussed, which has consideration of various elements of manager service, and the social network approaches. One example of the services is Vedio-on-Demand (VoD) services with opinion observations. This service will be applied in the proposed approach. An analysis of the VoD and associated methods are used to define these elements with respect to time analysis. Suppose that the VoD includes c clusters (blocks or groups) of data; therefore, the time is given by:

$T = N(\text{Cloud Network}_{l}) + K(\text{Cloud Knowledge}_{j}) + T(\text{Cloud Cluster}_{k}) + A(\text{Cloud Agent}_{l})$

where N is the network time of the i^{th} network, K represents knowledge time discovery, T is the average time per cluster, and A is equivalent to the service time to search (mobility time), group, discover, etc. Thus the distributed services scheme is adapted in the proposed architecture. Multiple

domains as clusters will be managed with a total number of resources of N virtual machines. Each virtual machine includes domain as a cluster.

If the time of i network is larger than the time it takes for the virtual machine (network device) to consume, then jitter is assumed, therefore;

$$T(i) <= T_{\text{Virtual machine}} \text{ x } i$$

where $T_{\text{Virtual machine}}$ represents the time it takes for virtual machine (network device) to consume a service.

The proposed model includes data model based on big table suggested in this paper. Such table contains Big Table description by cloud providers; like Google to manage and store social network data contents. The table has n rows, and each row has unique key in the row field. Each row includes several columns associated with column key and column value. Table 2 illustrates the suggested data model description.

Table 2: Proposed Unstructured Data Model Description

Row Key	Cluster Info	Client Comments			
		Client-ID_1	Client-ID_2	Client-ID_3	Client-ID_n
Cluster ID/ Group ID	Subject: Date: Content: Classification:	User ID: Date: Comment: Rating:	… … … …	… … … …	User ID: Date: Comment: Rating:

The proposed framework will be deployed and executed on a private IaaS cloud at KAU. The choice of a private cloud allows collocation with 8 nodes with 8 CPU cores connected through gigabit Ethernet. The manger deploys the variety of data up and down on a VM to bulk data loads that occur every day. The manager runs round the clock and loads the semantic data repository that supports mobile apps and smart grid portal at KAU. Lexical ontology used in the proposed framework helps manage the diversity of data sources and data types generated. Data items from different sources are mapped to concepts in the ontology by the lexical and semantic annotation. The lexical and semantic annotation accept input resources semi and un-structured data, interpret such input and create semantic RDF that relate key and value events with the ontology domain, store the RDF to semantic database.

5.3. Overall Evaluation

Finally, let's have a method that classifies every document in the test directory and prints out the percent accuracy of this method. Accordingly, the proposed classifier decides the value of the accuracy; such value is taking into consideration: (1) Correct classification; and (2) Wrong classification. Table (3) illustrates the evaluation results for the proposed classifier relative to the human experts' judgment for the 10 datasets.

Table 3: Overall Accuracy Evaluation of Classifiers

Classifier	Stop Words	Accuracy
Naïve Bayes	0 words	77.78
Naïve Bayes	25 words	78.76
Naïve Bayes	174 words	79.94
Naïve Bayes (Java)	500 words	86.43
Max. Entropy	500 words	87.69

Figure 10 shows the comparison of classification accuracy on dataset that is in the English language. As shown in table 2, for Naïve Bayes classifier, the accuracy was 77.78, 78.76 and 79.94 without stop words and with stop words (25 and 174) respectively. The last two rows of the table include Naïve Bayes and Max Entropy with 500 stop words. Such figure shows the accuracy plot obtained by the experimental testing of proposed test. It has been 10% better accuracy between Naïve Bayes (without stop words) and Maximum Entropy classifier.

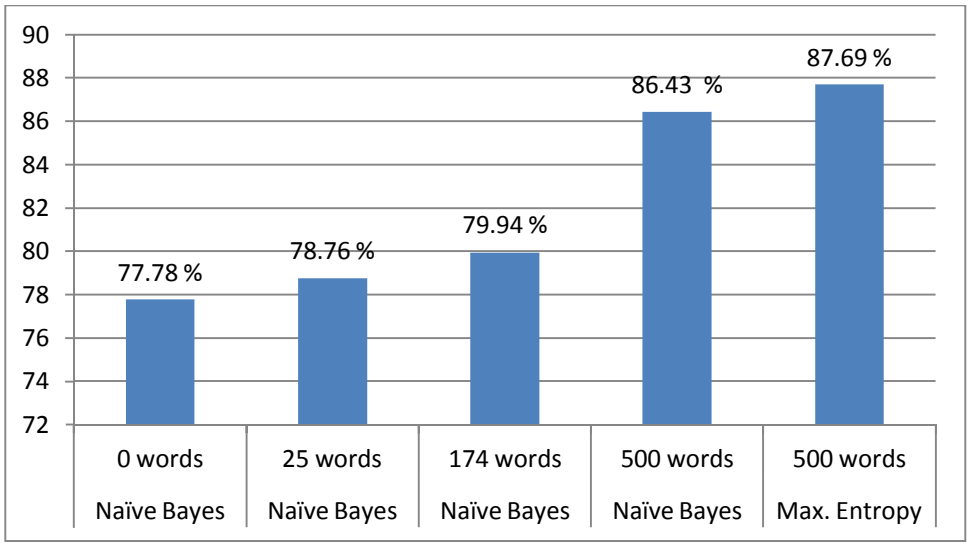

Figure 10: Accuracy Evaluation of Proposed Classifiers

The data stream of Arabic text in Hadith been increased and therefore, the analysis of the Arabic dataset become more competitive and challenging. The challenge is the management and classification of such data with vast size and different categories of this data. Consequently, data growth, policy, infrastructure, integration velocity, variety and data visualization are several challenges [32, 33]. Now we are preparing such dataset to work with.

6. Conclusion and Feature Work

This paper introduced unstructured data, collected data and stored data issues and challenges. Some of the major issues are identified. In addition, many of technical points have been introduced; such as, big data contents, samples and challenges.

We also mentioned to additional major challenges- that must be addressed with next few years and will be establish a framework for management big data in future.

The feature research of our collaborative research will focus on developing unstructured data analysis, annotating and designing methodologies, taken into consideration language processing, machine translation and the issues that we have raised in this paper.

There are major challenges for unstructured data processing. These challenges include analytics challenges, owner of the data, and scaling challenges. One such challenge is related to scaling issue, as the data set increases linearly with increasing computational resources. Solving this issue requires new algorithm. Another challenge is the time-varying nature of large graphs. Therefore, this challenge requires more intensive computation and new algorithms [1].

Consequently, the results of the related case study have shown that the effectiveness of many additional services should be done. Besides, the classification and clustering are two critical issues of the big data will be discussed in more detail in future.

References

1. Kaisler S., F. Armour, J. Espinosa, J. and G. Washington,"Big Data: Issues and Challenges Moving Forward". 46th Hawaii Conference on System Sciences, IEEE Computer Society, pp 995 – 1004, 2013.
2. Sagiroglu S. and D. Sinanc, Big Data: A Review, Collaboration Technologies and Systems (CTS)Onference, IEEE, 2013.
3. Eaton C., D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012.
4. Schneider R. D., Hadoop for Dummies Special Edition, John Wiley&Sons Canada, 978-1-118-25051-8, 2012.
5. Fox B., "Leveraging Big Data for Big Impact", Health Management Technology, 2011. http://healthmgttech.com/.
6. Hsiao-Ying L. and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, 2012.
7. Zhang X., F. Xu, (2013). Survey of Research on Big Data Storage, 2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science, (DCABES), IEEE, 2-4 Sept. 2013, pp. 76-80.
8. Gantz J. A. E. R., "Extracting Value from Chaos," IDC's Digital Universe Study 2011.
9. Rats J., G. Ernestsons, (2013). Using of Cloud Computing, Clustering and Document-oriented

Database for Enterprise Content Management, Informatics and Applications (ICIA) Second International Conference, IEEE, pp. 72-76.
10. Compton R., L. De Silva, and M. Macy, (2013). Detecting future social unrest in unprocessed Twitter data "Emerging Phenomena and Big Data", ISI 2013, June 4-7, 2013, Seattle, Washington, USA. IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 56-60.
11. Chardonnens T., B.Perroud, (2013). Big Data Analytics on High Velocity Streams: A Case Study, 2013 IEEE International Conference on Big Data, pp. 784 – 787.
12. Kranjc J., V. Podpecan, N. Lavrac, (2013). Real-time data analysis in Clowd Flows, 2013 IEEE International Conference on Big Data, pp. 15 – 22.
13. Kwak H., C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in Proceedings of the 19th international conference on World wide web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.
14. Rehman N. U., S. Mansmann, A. Weiler, M. Scholl, (2012). Building a Data Warehouse for Twitter Stream Exploration, Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference on 2012, pp. 1341 – 1348.
15. Hey T., S. Tansly and K. Tolle, "The fourth Paradigm: Data Intensive Scientific Discovery". Microsoft Research, October 2009.
16. Stonebraker M., and J. Gong, "Researchers' Big Data Crisis: Understanding Design and Functionality", Communication of the ACM, 2012, 55(2): 10-11.
17. Belgoli E., S. Horey, "Design Principles for effective Knowledge Discovery from Big Data", Software Architecture and 6th European Conference on Software Architecture, IEEE Conference, 2012.
18. Sadis F., G. Mapp, J. Loo, M. Aiash, and A. Vinel, On the Investigation of Cloud-based Mobile Media Environments with Service-Populating and QoS-aware Mechanisms, IEEE Transactions on Multimedia, Issue 99, 2013.
19. Al-Barhamtoshy H. M., and F. Mujallid. Building Mobile Dictionary System, The International Conference on Digital Information Processing, E-Business and Cloud Computing (DIPECC 2013), http://sdiwc.net/digital-library/ building-mobile-dictionary. The Society of Digital Information and Wireless Communica-tion (SDIWC), October 23-25, 2013.
20. Eassa F., H. M. Al-Barhamtoshy, K. Jambi, An Architecture for Metadata Extractor of Big Data in Cloud Systems, International Journal of Scientific and Engineering Research (IJSER) -(ISSN 2229-5518) http://www.ijser.org, January 2014, IJSER Volume 5, Issue 1.
21. Kumar S., F. Morstatter, and H. Liu. Twitter Data Analytics, Springer, August 19, 2013.
22. Zacharski R.. A Programmer's Guide to Data Mining, ,www.guidetodatamining
23. Shoukry A., and A. Rafea, Sentence-Level Arabic Sentiment Analysis. 2012 IEEE Xplore.
24. Mountassir A., H. Benbrahim and I. Berrada, An Emperical Study to Address the Problem of Unbalnced Data Set in Sentiment Classification. International Conference on Systems, Man, and Cybernetics, October 14-17, 2012, COEX, Seoul, Korea, pp. 3298-3303.
25. Abdul-Mageed M., S. Kubler, and M. Diab, SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. The Association for Computational Linguistics, WASSA 2012, 3rdWorkshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'12). Republic of Korea.
26. Bums N., Y. Bi, H. Wang, and T. Anderson, "Sentiment Analysis of Customer Review: Balanced versus Unbalanced Datasets". KES 2011, Part I, LANAI6881, 2011, pp. 161-170.
27. Manyika J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers, "Big Data: The Next Frontier for Innovation, Competition and Productivity," McKinsey Global Institute, May, 2011.
28. Begoli E., J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Conference of Software Architecture and 6th European Conference on Software Architecture, IEEE Computer Society 2012.
29. Dean J. and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of ACM, vol. 51, no. 1, pp.107-113, 2008.
30. Lakshman A. and P. Malik, " Cassandra: A Decentralized Structured Storage System, "ACM SIGOPS Operating Systems Review, vol. 44, no. 2, pp. 35-40, Apr. 2010.
31. White T., Hadoop: The Definitive Guide, O'Reilly, Yahoo Press, 3rd Edition, 2012.
32. Sagiroglu S. and D. Sinanc. Big Data: A Review, Collaboration Technologies and Systems (CTS) International Conference, pp. 42-47, 20-24 MayIEEE 2013.
33. Alajami A., E. Saad and R. Darwish, Towards an Arabic Stopwords List Generation, International Journal of Computer Applications, Vol. 46- No 8, May 2012.
34. Mihalcea R., D. Radev, Graph-Based Natural Language Processing and Information Retrieval, Cambridge Publisher, 2011.