Retrieving Electronic Data Interchange (EDI) Dataset using Text Mining Methods

Zakaria Suliman Zubi
Sirt University, Faculty of Science, Computer Science Department
Sirte, P.O Box 727, Libya, zszubi@yahoo.com

Abstract: - The internet is a huge source of documents, containing a massive number of texts presented in multilingual languages on a wide range of topics. These texts are demonstrating in an electronic documents format hosted on the web. The documents exchanged using special forms in an Electronic Data Interchange (EDI) environment. Using web text mining approaches to mine documents in EDI environment could be new challenging guidelines in web text mining. Applying text-mining approaches to discover knowledge previously unknown patters retrieved from the web documents by using partitioned cluster analysis methods such as k- means methods using Euclidean distance measure algorithm for EDI text document datasets is unique area of research these days. Our experiments employ the standard K-means algorithm on EDI text documents dataset that most commonly used in electronic interchange. We also report some results using text mining clustering application solution called WEKA. This study will provide high quality services to any organization that is willing to use the system.

INTRODUCTION

The growth of the stored electronic documents is increasing day by day on the web. These documents contain an electronic media for a particular end user represented in texts, pictures, audios and videos format. The electronic texts in these documents characterized in multilingual languages and classified into two catalogs such as Latin and non-Latin languages. These languages correspond to the electronic text contents of the documents stored on the web. The text contents became the most important item in the document and the most frequently distributed in that document as well. Electronic documents on the internet had a tremendous number of electronic texts defined in million of topics. Internet users actively are exchanging documents with each other asking about subjects of interest or sending requests to Web-based expert forums, or any other services in electronic text forms. Organizations such as governments and companies institutions are exchanging electronics documents in different form throughout the internet media in a security environment called electronic documents interchange (EDI). The electronic documents interchange (EDI) is a computer-to-computer exchange environment of electronic documents between organizations. EDI replaces the faxing and mailing of paper documents. EDI documents uses a precise computer record formats based on a commonly accepted standards. However, each organization will use the flexibility allowed by the standards in a unique way that fits their daily inquires needs. The data in the EDI documents mainly represented in text formats translated from one host to another through a network media. EDI usually transfers text data between different originations using internet or network environments. Such environment could be a VANs or the Internet. As more and more organizations connected to the Internet, EDI is becoming progressively more significant as an easy mechanism for organizations to manage, buy, sell, and trade information. ANSI has approved a set of EDI standards known as the X12 standards. Theses standards play a necessary condition for organizations to join EDI community. Moreover, the X12 standards developed uniform standards for inter industry electronic exchange of business and managements transactions electronic data interchange (EDI). EDI standards used as a national format based on the organization location and activity. Each international format is an international EDI standard designed to meet the needs of both government and private industry. These standards sets many types of transactions in the organization for different purposes such as product/pricing transactions, ordering transactions, materials management transactions, shipping/receiving

transactions, inventory management transactions, financial transactions and control transactions were each transaction type had several sub-types. On the other hand, these transactions fashioned to insure the daily work in any organization. Moreover, each transaction forms is a document consists of text data and unique data such as the transaction number, type, date and more. This text data or dataset could be storied in a database after the transaction process and the translation procedure are completed [1].

The similarities between EDI files and databases give us the courage to unify them to improve new concepts. Even thought an EDI file, and a database are similar in many ways, there really is not a one to one correlation between them. Some times a lot of people for the purpose of simplicity compare data elements of an EDI file to fields of a database, and a data segment of an EDI file to records of a database. This comparison may be acceptable to map the EDI data segment located in the EDI file after parsing and translating into a database.

The database contains high-quality information in text forms. High quality of information typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining, sometimes alternately referred to as text data mining, almost equivalent to text analytics, refers to the process of deriving high-quality information from text. Figure 1, shows how EDI format conducted to the database after that text mining approaches could be used to retrieve the databases were hosted on the server side.

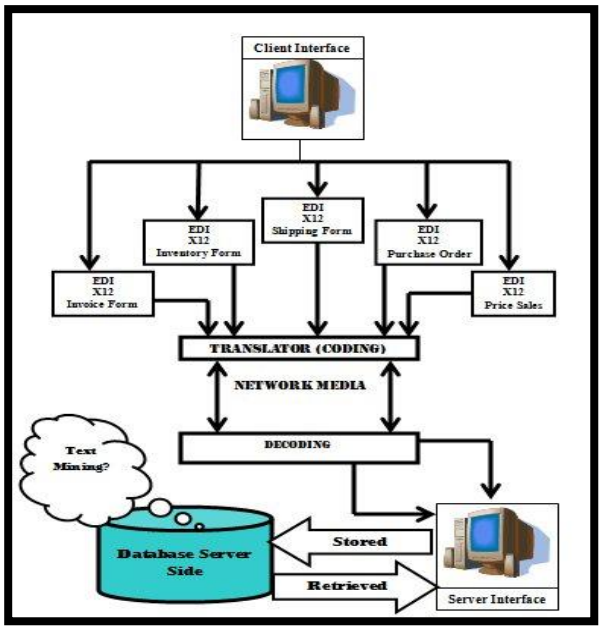


Figure 1. Shows the EDI documents- to- database – to- text mining life cycle.

The principle of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information enclosed in the text accessible to the various data mining (statistical and machine learning) algorithms. Information extracted to develop summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, etc., or you could analyze documents and conclude similarities between them or how they are related to other variables of interest in the data mining issue.

In the most general expressions, text mining will "turn text into numbers" (meaningful indices), which can then be included in other analyses such as predictive data mining, the application of unsupervised learning methods (clustering), etc. Text mining [4] is a technique for the automatic clustering of large volumes of documents, which applied to the problem using some common clustering algorithms such as k-means. Text mining can use cluster analysis methods to identify

groups of documents (e.g., vehicle owners who described their new cars), to identify groups of similar input texts. This type of analysis also could be exceedingly useful in the context of market research studies, for example of new car owners. A k-means clustering analysis, we would observe the means for each cluster on each dimension to evaluate how distinct our k clusters are. In an ideal world, we would obtain very different means for most, if not all dimensions, used in the analysis. The magnitude of the F values (frequent value) from the analysis of variance performed on each dimension is another indication of how well the respective dimension distinguishes between clusters.

The k-means method will produce exactly k different clusters of greatest possible distinction. It should mentioned that the best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data [5, 6], moreover, text mining is an approach were we can apply several methods such as clustering , classification, sequence and association data. Figure 2 summarizes these methods.

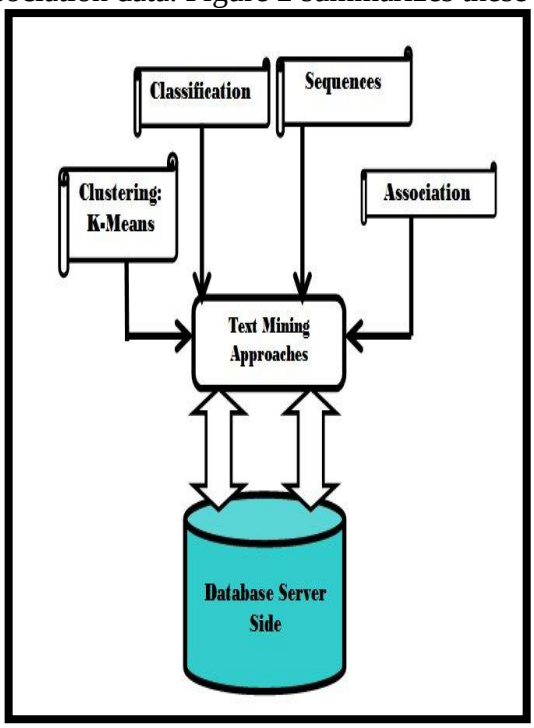


Figure 2. Information value and information collection methods in text mining

The figure illustrate that there are variety of methods could be apply in text mining when retrieving EDI databases .In this paper we are focusing on mining text documents using a common clustering approaches called k-means.

2 TYPES OF TEXT MINING

Any data mining approaches usually used for four main purposes: (1) to improve customer achievement and maintenance; (2) to reduce fraud; (3) to identify internal inefficiencies and then revamp operations, and (4) to map the unexplored environment of the Internet. The major types of tools used in text mining are:

- Artificial Neural Networks;
- Decision trees;
- Genetic algorithms;
- Rule induction;
- Nearest Neighbor Method;

- Data Visualization;

Text mining uses discovery-based approaches in which pattern-matching and other algorithms used to discover key relationships in the data, formerly unknown to the user.

The discovery model is different because the system automatically discovers information hidden in the data. The data examined in a search of frequently occurring patterns, trends, and generalizations about the data without intervention or supervision from the use. An example of such a model is a bank database, which mined to discover the many groups of customers to aim for a mailing campaign. The data searched with no hypothesis in mind other than for the system to group the customers according to the common characteristic found.

3 TYPES OF INFORMATION AND METHODS

Text mining usually produces five types of information, theses information illustrated also in figure 2:

- Associations;
- Sequences;
- Classifications;
- Forecasting
- Clusters;

Associations: turn out when occurrences linked in a single occasion. For example, a study of supermarket baskets might expose that when corn chips purchased, 65% of the time cola also purchased, unless there is a support, in which case cola purchased 85% of the time.

Sequences: procedures linked over time based on the event that happen. For example, if we bought a house, then 45% of the time a new oven will be bought within one month and 60% of the time a new refrigerator will be bought within two weeks as well.

Classification: is possibly the most common text mining motion today. Classification can assist you to discover the personality of customers who are likely to leave and provides a model that used to expect who they are. It can also help you decide which types of promotions have been useful in keeping which kinds of customers, so that you expend only as much money as required to retain a customer [5, 20].

Forecasting: Most of the applications that uses expectation may involve predictions, such as whether a customer will renew a subscription forecasting, a dissimilar form of prediction. In other hand, it guesses the future value of continuous variables like sales figures based on patterns within the data.

Clustering: is one of the essential methods used in text mining approaches to discover different groupings with the data. This can be applied to problems as dissimilar as detecting defects in manufacturing or finding affinity groups for bankcards or electronic documents. Data clustering is one of the most effective techniques that can improve performance of text in electronic documents. In this paper we will use clustering methods to group electronic documents which inherently an unsupervised learning process that organizes document (or text) data into distinct groups without depending on pre-specified knowledge [2,7,8].

In our case of using clustering is to give a source of a text documents with distance measures e.g., how many words are common in these documents. It helps us to find the several clusters that are relevant to each other. Figure 3, illustrates the needs of the clustering distance measure.

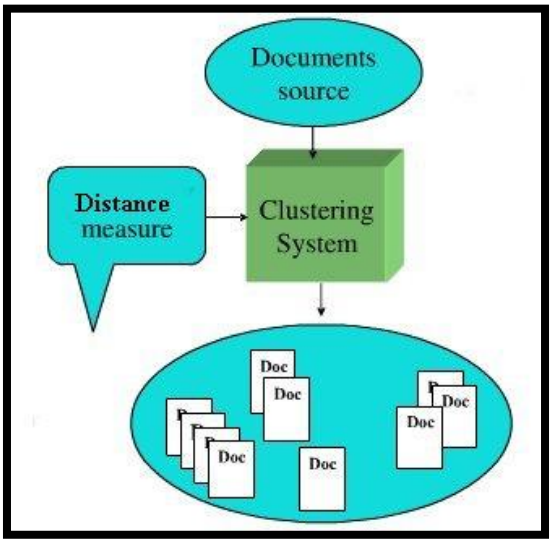


Figure 3. Clustering system mechanisms

In clustering it, find out whether the documents in one cluster are more close to one another or documents in separate clusters are more far to one another to find a correct set of documents. Moreover, clustering is an unsupervised learning method works with unstructured data (documents), which deals with unknown class labels of training data. It also gives a set of measurements with the aim of establishing the existences of clusters in the data. We can evaluate clustering methods to produces high quality clusters with a high intra-class or a low inter-class distances to figure out the quality of the clustering method which measured by the ability to discover some or all the hidden patterns. There are two main methods in clustering such as partitioning methods and hierarchical methods. In this work we are going to focus on partitioning methods since we are working with k-mean algorithm as a partitioning method in clustering. The main feature of partitioning method is to construct a partition of n documents into a set of k cluster. In this paper, we are going to use partitioned clustering analysis using k- mean algorithm [3].

4 Methods and Algorithms used

4.1 Clustering using k- means Algorithm

Since 1950s, people have proposed many kinds of clustering algorithms. They roughly separated into two brands, of which one based on division and the other based on level. At the same time, a third type, namely the combination of these two methods emerged. Among those based on division-clustering algorithms, the most famous is the k- means type algorithm. The basic members of k-mean type algorithm family include K-Means, K-Modes [1] and K-Prototypes [2]. K-Means algorithm used in value data, K-Modes algorithm used in attribute data, and K-Prototypes algorithm used in mixed data of value and attributes [2].

The k- means type algorithm has such advantages as fast speed, easy realization and suitable for those kinds of data clustering analysis as in text, picture characteristic but the iterative process of this algorithm is likely to terminate it soon [4].

Therefore, an excellent result achieved by, owing to its random selection of initial centers, unstable results often gotten. Because clustering often applied in data, which the final user is also unable to judge clustering quality, these kinds of unstable results is difficult to accept. Therefore, it is significant to improve the quality and stability of clustering result in text clustering analysis [5].

The k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Figure 4, shows how k-mean algorithm works [6].

Example: The data set has three dimensions and the cluster has two points: X = (x1, x2, x3) and Y = (y1, y2, y3). Then the centroid Z becomes Z = (z1, z2, z3), where z1 = (x1 + y1)/2 and z2 = (x2 + y2)/2 and z3 = (x3 + y3)/2.

The algorithm steps are:

0. Input D:= {d1,d2,….,dn}; k:= the cluster number;
1. Select k document vectors as initial centriods of k cluster;
2. Repeat;
3. Select one vector d in remaining documents;
4. Compute similarities between d and k centriods;
5. Put d in the closest cluster and recomputed the centriods;
6. Until the centriods don't change;
7. Output: k clusters of documents.

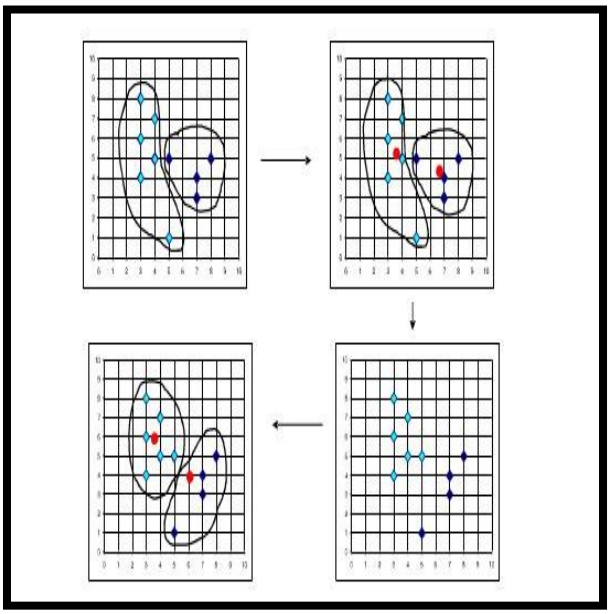


Figure 4. K-mean clustering algorithm.

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets, the k-medoids variant is appropriate. Other popular variants of K-means include the Fast Genetic K-means Algorithm (FGKA) and the Incremental Genetic K-means Algorithm (IGKA) [14].

The characteristic of k-mean algorithm summarized as follow:

1. Works with numeric data only.
2. Pick a number (K) of cluster centers (at random).
3. Assign every item to its nearest cluster center (e.g. using Euclidean distance).
4. Move each cluster center to the mean of its assigned items.

Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

4.2 BAG-OF-WORDS DOCUMENT

The generation of electronic documents as a bag of words in EDI database will leads to the following features:

- Text document is represented by the words it contains (and their occurrences) e.g., "Lord of the rings" → {"the", "lord", "rings", "of"}. This representation has a high efficient which makes learning far simpler and easer. The order of words in this case is not important for certain application.
- Stemming to identify a word by it's root is also conducted e.g., flying, flew → fly, it's used to reduce dimensionality.
- Stop words are also used whereas, the most common words are unlikely to help text mining e.g., "the", "a", "an", "you" ..etc.

Text document representation based on the bag of words model is a subfield of natural language processing (NLP). The model can solve problems such as the following two simple text documents:

- John likes to watch movies. Mary likes too.
- John also likes to watch football games.

Based on these two text documents, a dictionary constructed as:

Dictionary={1:"John", 2:"likes", 3:"to", 4:"watch", 5:"movies", 6:"also", 7:"football", 8:"games", 9:"Mary", 10:"too"},

Which has 10 distinct words? In addition, using the indexes of the dictionary, each document represented by a 10-entry vector:

- [1, 1, 1, 1, 1, 0, 0, 0, 1, 1]
- [1, 1, 1, 1, 0, 1, 1, 1, 0, 0],

Where each entry of the vectors refers to count of the corresponding entry in the dictionary, (This is also the histogram representation). As we can see, this vector representation does not preserve the order of the words in the original sentences. This kind of representation has several successful applications.

Therefore, we propose a bag of words format practically will uses a file extension, which includes documents in a full processed form supporting the bag-of-words representation. Each document represented by the set of its word frequencies and categories that it belongs too. This format corresponds to the commonly used representation of a text document with a word-vector ignoring position of words in the document.

The purpose of the format is to enable efficient execution of algorithms working with the bag-of-words representation such as, clustering, learning, classification, visualization, etc [7].

4.3 TEXT IN EDI DOCUMENT REPRESENTATION

There are several ways to model an EDI text document. For example, it can be representing as a bag of words, where words are assumed to appear independently and the order is immaterial.

The bag of word model is widely used in information retrieval and text mining [20]. Words counted in the bag, which differs from the mathematical definition of set. Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension. Here we use the frequency of each term as its weight, which means terms that appear more frequently are more important and descriptive for the document.

Let $D = \{d_1, \ldots, d_n\}$ be a set of documents and $T = \{t_1, \ldots, t_m\}$ the set of distinct terms occurring in D.

→

A document represented as a vector t_d. Let *tf(d, t)* signify the frequency of term $t \varepsilon T$ in document $d \varepsilon D$. Then the vector representation of a document *d* is as follow:

$\overrightarrow{t_d} = (tf(d, t_1), \ldots, tf(d, t_m))$ ……(1)

Although words are more frequent assumed more significant as mentioned above, this is not frequently the case in practice. For example, words like a "and" the "are" probably the most frequent words that appear in English text, but "neither" are "neither" descriptive nor significant for the document's subject. With documents presented as vectors, we measure the degree of distance of two documents as the correlation between their corresponding vectors. For instance, "terms" are words. However, we applied several standard transformations on the basic term vector representation as well [8].

First, we have to remove stop words. There are words that are non-descriptive for the topic of any document even EDI documents, such as "a"," and", "are" and "do". Following common practices, we used the one implemented in the Weka machine-learning workbench system, which contains 527 stop words.

Second, we must stemmed words using Porter's suffix-stripping algorithm [14], so that words with different endings will be mapped into a single word. For example, "production" , "produce", "produces" and product will be mapped to the stem product.

The principal assumption is that different morphological variations of words with the same root/stem are thematically similar and should treat as a single word.

Third, we measured the effect of including infrequent terms in the document representation on the overall clustering performance and decided to discard words that appear with less than a given threshold frequency. The rationale by discarding infrequent terms is that in many cases they are not very descriptive about the document's subject and make little contribution to the distance between two documents [9].

Meanwhile, including unusual terms can also introduce noise into the clustering process and make distance computation more expensive. Therefore, we choose the top 2000 words ranked by their weights and use them in our experiments [10].

In the clustering process, we also need to measure the distance between two clusters or between a cluster and an object. In hierarchical clustering this is normally computed as the complete-link, single-link or average-link distance [8, 11]. However, in partitional clustering algorithms, which we choose one of their algorithm to be applied, a cluster is usually represented with a centroid object. For example, in the K-means algorithm the centroid of a cluster is the average of all the objects (documents) in the cluster that is, the centroid's value in each dimension is the arithmetic mean of that dimension over all the objects in the cluster. Let C be a set of documents. Its centroid is defined as [12]:

$$\overrightarrow{t_C} = \frac{1}{|C|} \sum_{\overrightarrow{t_d} \in C} \overrightarrow{t_d},$$

……(2)

Which is the mean value of all term vectors in the set? Moreover, we normalize the vectors to a unified length to avoid long documents dominating the cluster.

4.4 DISTANCE MEASURES

Before clustering, a distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects (documents) [13]. It should correspond to the

characteristics that are supposed to decide the clusters embedded in the data. In many cases, these characters are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems.

Moreover, selecting a suitable distance measure is also critical for cluster analysis, especially for a particular type of clustering algorithms. The closeness of the corresponding text data object to its neighboring documents by recalling that closeness quantified as the distance value [14]. We can see that large numbers of distance calculations are required for discovering dense areas and approximate cluster assignment of new text data objects. Therefore, understanding the usefulness of different measures is of great importance in helping to choose the best one.

In general, distance measures map the distance between the representative description of two objects into a single numeric value, which depends on two factors the properties of the two objects and the measure it [15]. Based on that we decided to use one common distance measure called Euclidean distance measure.

To qualify a distance measure as a metric, a measure *d* must satisfy the following four conditions.

Let *x* and *y* be any two objects (electronic document) in a data set and *d(x, y)* be the distance between *x* and *y* [16].

1. The distance between any two points must be nonnegative, that is, $d(x, y) \geq 0$.
2. The distance between two objects must be zero if and only if the two objects are identical, that is, $d(x, y) = 0$ if and only if $x = y$.
3. Distance must be symmetric, that is, distance from *x* to *y* is the same as the distance from *y* to *x*, i.e. $d(x, y) = d(y, x)$.

4. The measure must satisfy the triangle inequality, which is $d(x, z) \leq d(x, y) + d(y, z)$.

Euclidean Distance: Euclidean distance is a standard metric for geometrical problems [17, 20]. It is the ordinary distance between two points and can be easily measured with a ruler in two or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. It satisfies all the above four conditions and therefore is a true metric. It is also the default distance measure used with the K-means algorithm. Measuring distance between EDI text documents, given two documents d_a and d_b represented by their term vectors t_a and t_b respectively, the Euclidean distance of the two documents defined as:

$$D_E(\vec{t_a}, \vec{t_b}) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{1/2}, \quad (3)$$

Where the term set is $T = \{t_1, \ldots, t_m\}$. As mentioned previously, we use the *tfidf* value as term weights, that is $\boldsymbol{w_{t,a} = tfidf(d_a, t)}$,

Alternatively, it can be denoted also as:

$distance(x,y) = \{\sum i\ (xi - yi)2\ \}½$ ……(4)

Another Euclidean algorithm called squared Euclidean distances usually computed from raw dataset, and not from standardized data. This method has positive advantages (e.g., the distance between any two objects affected by the addition of new objects to the analysis, which may be outliers). However, the distances greatly affected by differences in scale among the dimensions from which the distances computed. For example; if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be

greatly affected (i.e., biased by those dimensions which have a larger scale), and therefore, the results of cluster analyses may be very different. Commonly, it is good practice to transform the dimensions so they have similar scales.

Squared Euclidean distance: You may want to square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance computed as the following:

$$distance(x,y) = \Sigma i\ (xi - yi)^2 \quad \ldots\ldots(5)$$

4.5 DATASET

In the dataset we propose a collection of a banking transaction of EDI electronic text data that been gathered from EDI databases. EDI text data collected and aggregated in seven main categories. These categories create what we called EDI corpus [18, 20]. This corpus represent the datasets that consist of 2000 EDI electronic documents of different lengths that belongs to seven categories, the categories are transactions divisions in X12 standard EDI format. These seven categories presented as follow: product/pricing transactions, ordering transactions, materials management transactions, shipping/receiving transactions, inventory management transactions, financial transactions and control transactions. Table 1 represents the number of documents for each category.

Table 1 - number of EDI documents in the dataset

Clusters	Category Name	No. Of EDI Documents
1	product/pricing transactions	290
2	ordering transactions	280
3	materials management transactions	300
4	shipping/receiving transactions	320
5	inventory management transactions	296
6	financial transactions	289
7	control transactions	225
	Total	**2000**

4.6 TRANSLATING EDI TO DATABASES

Translating EDI to database formats is essential for storing and accessing your transaction information. While an increasing number of databases support storing EDI messages in their native formats, it is necessary to translate this data into a valid database structure in order for it to be available for search and retrieval [19].

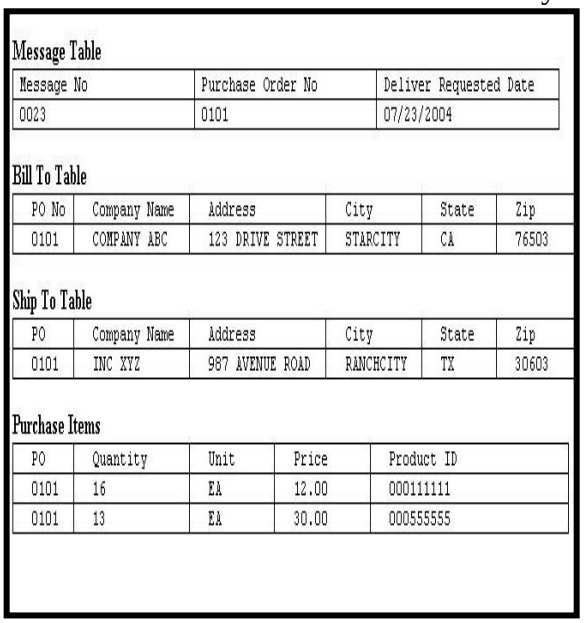
The system should supports all major databases, including:

- Microsoft® SQL Server® 2000, 2005, 2008
- IBM DB2® 8, 9
- IBM DB2 for iSeries® v5.4, 6.1
- Oracle® 9i, 10g, 11g
- Sybase® 12
- MySQL® 4, 5
- PostgreSQL 8
- Microsoft Access™ 2003, 2007

To insert a decoded EDI standard format form to the database. We will translate simply an EDI message EDI X12 standards formats into a variety of transactions. We will then prompt to specify a sample EDI file that can used to view the output of your mapping [20].

The system enables us easily to transform our data into EDI formats by visually mapping it from a wide variety of other usable file formats, including XML, databases, flat files and other EDI standards.

EDI formats are standards for electronic data exchange and are specifically suited for exchanging information between disparate systems. Mapping proprietary data to EDI for transmission to partner companies via extranets, Web services, or value-added networks (VANs), is a very common business requirement in EDI environments [21]. Mapping the translated EDI message into the database will constricts a database more likely as illustrated in figure 5.



Message Table

Message No	Purchase Order No	Deliver Requested Date
0023	0101	07/23/2004

Bill To Table

PO No	Company Name	Address	City	State	Zip
0101	COMPANY ABC	123 DRIVE STREET	STARCITY	CA	76503

Ship To Table

PO	Company Name	Address	City	State	Zip
0101	INC XYZ	987 AVENUE ROAD	RANCHCITY	TX	30603

Purchase Items

PO	Quantity	Unit	Price	Product ID
0101	16	EA	12.00	000111111
0101	13	EA	30.00	000555555

Figure 5. Shows some EDI mapped database tables.

These database tables are stored temporary in database. Moreover, this database contains redundant data and cannot be manipulated as well.

In this case, we need to normalize the data in a flat file. This flat file can be in any common form for instance in comma-separated format or any common format. The redundancy of data in the flat table can be clearly seen from a small portion of an EDI file [22]. Having more loops in the EDI file, the size of the table would have been exponentially bigger. The size as well as the redundancy of data in the flat table would make managing and translating the information over into an existing production database easer. Figure 6. Viewed the flat file of the database tables.

Msg No	PO NO	PO DATE	Qlfr	Company Name	Address	City	State	Zip	Qty	Unit	Price	Product ID
00023	0101	20040723	BT	COMPANY ABC	123 DRIVE STREET	STARCITY	CA	76503	16	EA	12.00	000111111
00023	0101	20040723	ST	INC XYZ	987 AVENUE ROAD	RANCHCITY	TX	30603	16	EA	12.00	000111111
00023	0101	20040723	BT	COMPANY ABC	123 DRIVE STREET	STARCITY	CA	76503	13	EA	30.00	000555555
00023	0101	20040723	ST	INC XYZ	987 AVENUE ROAD	RANCHCITY	TX	30603	13	EA	30.00	000555555

Figure 6. illustrates the shape of flat database table.

On the other hand, if the EDI file were to be translated directly into a normalized relational database, the transfer of data over to the production database would be a one-to-one correlation between the fields of the two databases [23].

5 Types of Outputs

Generally then, applications of text mining can generate outputs Such as in retail, every time merchandise is handled it costs the merchant. By incorporating text-mining techniques, retailers can improve their inventory logistics and thereby reduce their cost in handling inventory. Through text mining, using EDI data a retailer can identify the demographics of its customers such as gender, martial status, number of children, etc. and the products that they buy [24]. This information can be extremely beneficial in stocking merchandise in new store locations as well as identifying "hot" selling products in one demographic market that should also be displayed in stores with similar demographic characteristics. For nationwide retailers, this information can have a tremendous positive impact on their operations by decreasing inventory movement as well as placing inventory in locations where it is likely to sell [25].

Buying patterns of customers; associations among customer demographic characteristics; predictions on which customers will respond to which mailings.

Patterns of fraudulent credit card usage; identities of "loyal" customers; credit card spending by customer groups; predictions of customers who are likely to change their credit card affiliation [13]

Predictions on which customers will buy new insurance policies; behavior patterns of risky customers; expectations of fraudulent behavior;

Characterizations of patient behavior to predict frequency of office visits.

6 Applications of Text Mining in EDI databases

From the above examples, we can say that text-mining applications can be used in a variety of sectors: consumer product sales, finance, manufacturing, health, bank, insurance, and utilities these sectors are EDI sectors as well. Thus if a business has data about its customers, suppliers, products, or sales, it can benefit from text mining. As the rapid technological advancement in the computer industry is making the data acquisition, dissemination, storage, and usage cheaper by megabytes, so we can easily predict that data mining will be one of the greatest tools to be used by the business community in the next century [14].

The types of data that are needed to perform text-mining applications for customer-based businesses which is available also in EDI databases are:

1) demographics, such as age, gender and marital status;
2) banking and economic status, such as salary, profession and household income; and,
3) geographic details, such as city, state or regions.

Other demographics like education, hobbies or marital status can also be used [2].

All of these data types can be used to group data according to a particular datasets or segments of customers that share similar interests and have common product requirements. The benefits of these applications can be seen in the experimental part of this paper were we illustrate a banking system using the k-mean clustering distance measure of the EDI bank text dataset.

In this paper, we are going to use WEKA as a text mining solution application. WEKA is a collection of machine learning algorithms for data mining and text mining tasks. The algorithms can either be applied directly to a dataset or called from any Java code. WEKA also contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also a well suited for developing new machine learning schemes such as in text mining fields.

7 Experimental Results

In this paper, we are going to use WEKA as a text mining solution application. WEKA is a collection .

The dataset presented in section 4, generated by using Euclidean distance measures in k-mean algorithms to assign every item to its nearest cluster center using a common text mining application called WEKA. The WEKA as a data mining or text mining solutions illustrates the use of k-means clustering with WEKA.

The EDI banking text dataset normalized in a flat file and represented in a comma-separated format. This document supposes that suitable data preprocessing achieved.

In this case a version of the primary dataset has been formed in which the ID field has been removed and the "children" attribute has been converted to categorical (This, however, is not essential for clustering but it's necessary when dealing with EDI format).

The resulting data file consists of 600 instances.

As an illustration of performing clustering in WEKA, we will use its implementation of the K-means algorithm to cluster the customers in this bank dataset, and to characterize the resulting customer data segments.

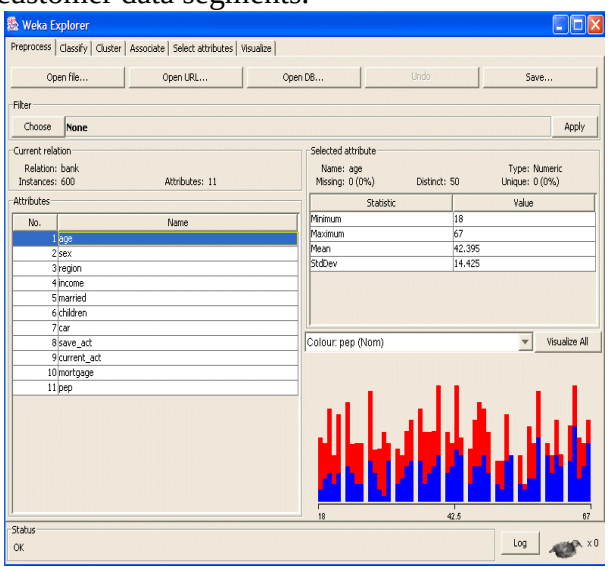


Figure 7, Shows flat dataset file loaded in WEKA.

Some implementations of k-means only permit numerical values for attributes. Therefore, it may be obligatory to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be obligatory to normalize the values of attributes that measured on significantly different balance (e.g., "age" and "income"). Even as WEKA offers filters to achieve all of these preprocessing tasks, which are not necessary for clustering in WEKA. This is because WEKA k-means algorithm automatically handles a mixture of categorical and numerical attributes. In addition, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA k-means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

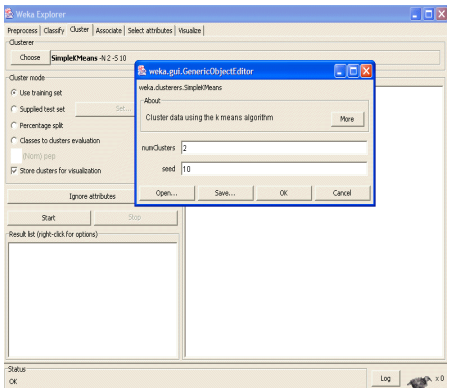


Figure 8, Shows editing the clustering parameter.

Entering seven clusters and seed values as well to generate a random number for making the initial assignment of instances to clusters. In general, k-means is quite sensitive to how clusters initially assigned. Thus, it is often necessary to try different values and evaluate the results when generating a training dataset. Figure 9. Shows the results.

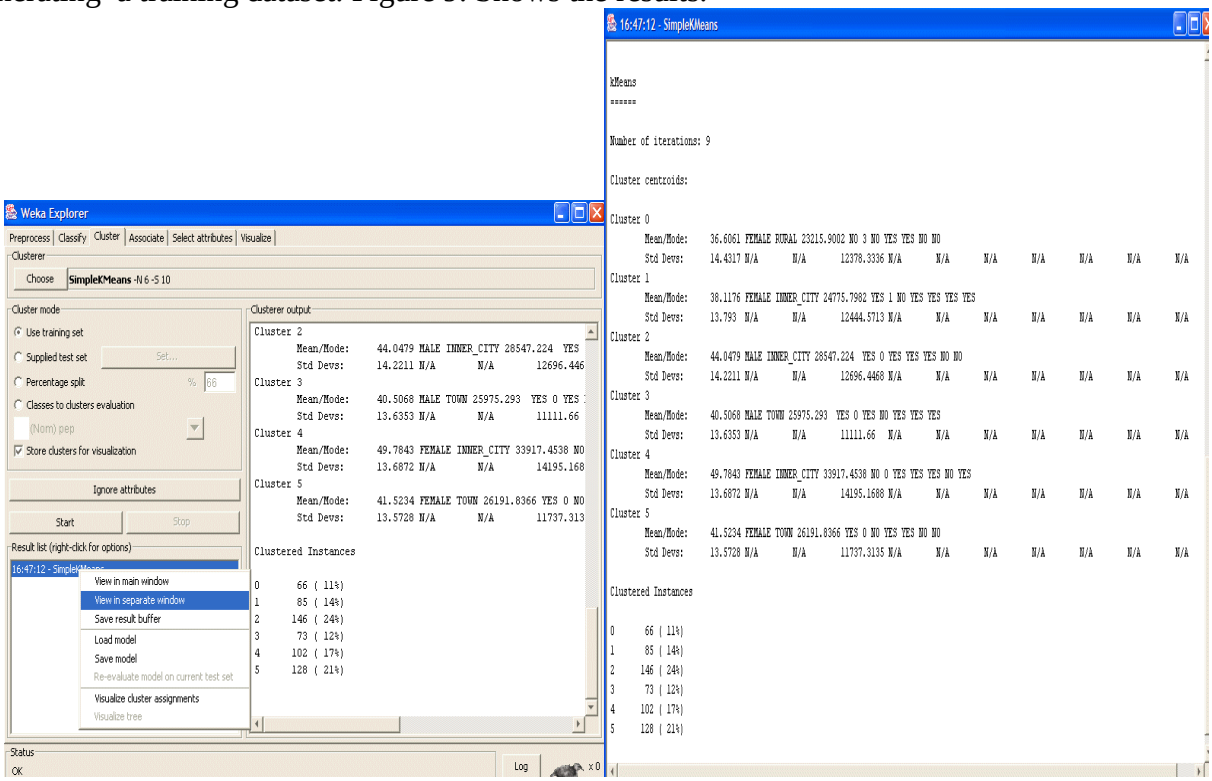


Figure 9, Shows the results of the clustering algorithm.

In figure 9, the WEKA technology illustrates the centroid of every cluster as well as statistics on the number and percentage of instances assigned to dissimilar clusters. Cluster centroids are the mean vectors for each cluster (so, each dimension value in the centroid corresponds to the mean value for that dimension in the cluster). Thus, centroids used to distinguish the clusters. For example, the centroid for cluster 1 shows that this is a segment of cases instead of middle aged to young (approx. 38) females living in inner city with an average income of approx. $28,500, who are married with one child, etc. Moreover, this group has on average said YES to the PEP product. Another way to recognize the character of each cluster is through visualization. Figure 10, demonstrate the visualization window for WEKA.

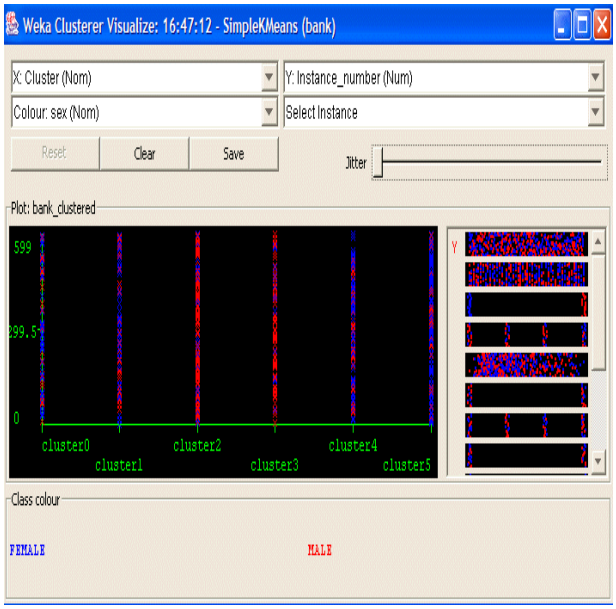


Figure 10, Shows a clustering visualization representation in WEKA system.

The cluster number and the other attributes in a three dimensions are represented in (x-axis, y-axis, and color) forms. Different combinations of choices will result in a visual rendering of different relationships within each cluster. In the example, we have chosen the cluster number as the x-axis, the instance number (assigned by WEKA) as the y-axis, and the "sex" attribute as the color dimension. This will result in a visualization of the distribution of males and females in each cluster. For instance, you can note that males dominate clusters 2 and 3, while females dominate clusters 4 and 5. In this case, by changing the color dimension to other attributes, we can see their distribution within each of the clusters.

Finally, we may be interested in saving the resulting dataset, which included each instance along with its assigned cluster figure 11 shows, the resulting dataset.

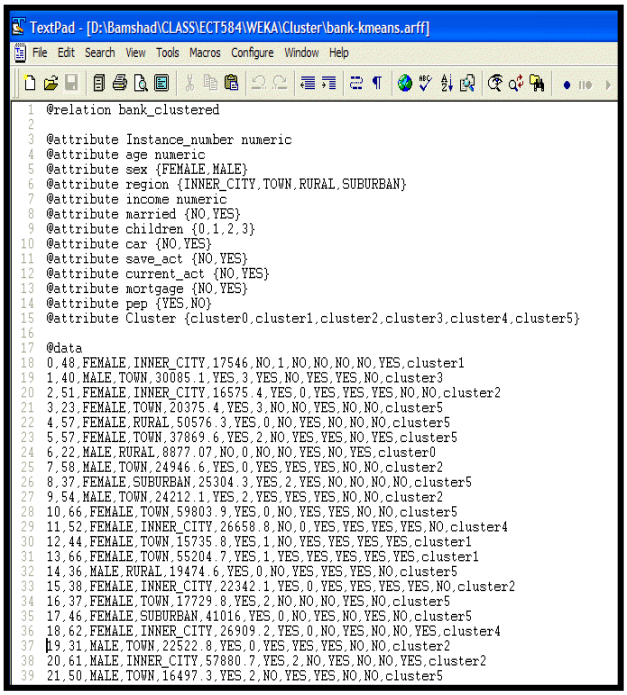


Figure 11 shows, the resulting dataset, which identify the clusters.

The attribute in WEKA are added to the original datasets. In the data portion, each instance has its assigned cluster as the last attribute value. By doing some simple manipulation to this dataset, we can easily convert it to a more usable form for additional analysis or processing. For example, here we have reconverted this dataset in a comma-separated format and sorted the result by clusters. Additionally, we have added the ID field from the original dataset (before sorting).

8 CONCLUSION

In this paper, we have used a homogenous mixture of two common technologies such as EDI and Text mining. EDI with a transformation process represented the database storage and on the other hand, text mining is the technology that extracts useful hidden and previously unknown patterns or information from EDI text databases. Both approaches present the challenges that go well beyond the technical. Many data management challenges remain, both technical and societal. Large online databases raise serious societal issues. To cite a few of the societal issues: Electronic data interchange and text mining software make it relatively easy for a large organization to track all of our financial transactions. We underline all the common principles between both technologies. Based on that, we circled only the most interesting intersection point that correlates between EDI and text mining.

In the case of EDI format, the file translated into a normalized flat file in a comma-separated format. The flat file represented the EDI database where we propose a dataset collected from a banking transaction of EDI electronic text data which been gathered from EDI databases. In text mining, we suggest to use k-mean algorithm in clustering method. We also calculate the Euclidean distance measures in k-mean algorithms to assign every item to its nearest cluster center. In the experimental section, we used a text mining application program solution called WEKA to represent our results in a visual fashion.

REFERENCES

[1] Anna Nick Wreden, Communications Week Interactive, February 17, 1997.

Arsitk Karen Watterson, Datamining poised to go mainstream October 1999.

Barbaros A.,Information and Privacy Commissioner/Ontario, Data Mining: Staking a Claim on Your Privacy, January 1998.

Bran Dick, Author unknown Data Mining: What is Data Mining?

[Chatterjee, P., Hoffman, D. L. and NOVAK, T. (2003). Modeling the clickstream: Implications for web-based advertising efforts. Marketing Sci. 22 520–541.

G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.

Pilot Software, Data Mining White Paper — Profitable Applications found at,www.pilotsw.com/dmpaper/dmindex.htm#dmapp.

Jim Gray, Data Management: Past, Present, and Future,at www.research.microsoft.com/~gray/DB_History.htm

J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2):638–645, Jun. 1998.

Jonathan Wu, Business Intelligence: The Value in Mining Data, DM Review online, February, 2002.

Kim, H. and Lee, S. 2002. An effective document clustering method using user-adaptable distance metrics. In Proceedings of the 2002 ACM Symposium on Applied Computing (Madrid, Spain, March 11 - 14, 2002). SAC '02. ACM, New York, NY, 16-20. DOI= http://doi.acm.org/10.1145/508791.508796.

M. F. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.

M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

Omran, M., Salman, A. and Engelbrecht, A. P., 2002. Image classification using particle swarm optimization. Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning 2002 (SEAL 2002), Singapore. pp. 370-374.

Richard A. Spinello, Case Studies in Information and Computer Ethics. New Jersey: Prentice Hall, 1997, p. 73.

Van D. M., Engelbrecht, A. P., 2003. Data clustering using particle swarm optimization. Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia. pp. 215-220.

Wisely Lasker RD. Strategies for addressing priority information problems in health policy and public health. J Urban Health 1998 Dec;75(4):888-895.

[Wexy Feldman R, Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing UnstructuredData. Cambridge: Cambridge University Press; 2007.

Zak Pines, Data Mining – A universal Tool, from www.hpcwire.com/dsstar .

Zakaria Suliman Zubi, 2009. Using some web content mining techniques for Arabic text classification. In Proceedings of the 8th WSEAS international Conference on Data Networks, Communications, Computers (Baltimore, MD, USA, November 07 - 09, 2009). M. Jha, C. Long, N. Mastorakis, and C. A. Bulucea, Eds. Recent Advances In Computer Engineering. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 73-84.

Zakaria Suliman Zubi, 2008. Knowledge discovery query language (KDQL). In Proceedings of the 12th WSEAS international Conference on Computers (Heraklion, Greece, July 23 - 25, 2008). N. E. Mastorakis, V. Mladenov, Z. Bojkovic, D. Simian, S. Kartalopoulos, A. Varonides, C. Udriste, E. Kindler, S. Narayanan, J. L. Mauri, H. Parsiani, and K. L. Man, Eds. Recent Advances In Computer Engineering. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 497-519.

Zakaria Suliman Zubi ,2008. I-extended databases. In Proceedings of the 10th WSEAS international Conference on Mathematical Methods and Computational Techniques in Electrical Engineering (Sofia, Bulgaria, May 02 - 04, 2008). D. P. Dimitrov, D. Simian, V. Mladenov, S. Jordanova, and N. Mastorakis, Eds. Electrical And Computer Engineering Series. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 126-137.

Zhang Ling-ling, and Lin Jian, "Study on information technology and corporate strategy, business processes and organizational structure of the conformity relationship model," Systems Engineering, March 2002.

Zhao Y. and Karypis G., 2004. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, Machine Learning, 55 (3): pp. 311-331.

Zhong, H. Zhang, and S. Chang. Clustering methods for video browsing and annotation. In Proc. IS&T/SPIE Symposium on Storage and Retrieval for Image and Video Databases, 1996.

ISBN 978-5-8114-1068-2