Knowledge Sources and Automatic Classification: A Literature Review

Mario Barcelo-Valenzuela
Miguel Romero-Ochoa
Alonso Perez-Soltero
Gerardo Sanchez-Schmitz
Universidad de Sonora
Departamento de Ingeniería Industrial
Rosales y Blvd. Luis Encinas S/N CP. 83000
Hermosillo, Sonora, México

Abstract

In organizations, a very important task is to identify and exploit knowledge assets in order to be competitive. Managing Knowledge alone has been a difficult task, therefore, the use of information systems to help manage organizational knowledge is strongly recommended.This article contains the introduction of knowledge possessed by organizations and the use of knowledge management systems, which are composed of different sources of knowledge such as those related to text, and multimedia communication systems. Because organizations have the need of knowledge classification, a review of the literature based on classification techniques that do not use information technology is presented. In the other hand, automatic techniques that work with different types of algorithms are mentioned. Finally, we propose a synthesis of knowledge sources and its classification.

Key Words: Knowledge, Knowledge management system, Data mining, Classification

1. Introduction

Knowledge is a fluid mix of framed experience, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences, new information and it is originated and applied in the minds of connoisseurs (Davenport and Prusak, 1998). Generally, knowledge always involves a user or an agent who are used to perform the necessary actions to achieve a goal. Knowledge can and should be evaluated by decisions or actions to which it leads (Davenport, 1998). The ability of organizations to create sustainable competitive advantages in time, determines the success of these to their competitors. The knowledge found within organizations, in addition to being a resource, is becoming an important strategic asset for them (Sîrbu et al. 2009).

The capacity in which organizations have to effectively utilize this knowledge and distribute it to all its departments, is becoming ever more crucial and important for competitive advantage and this will be increasingly critical to the success and survival of these (Ichijo and Kohlbacher, 2006). Knowledge is a concept that differs radically from what is information, but that information becomes knowledge once it is processed in the mind of the individual tacit knowledge (Polanyi, 1962). Again this knowledge will become information: once explicit knowledge is communicated to others in the form of text, screen output, audio, or other means (Nonaka, 1994).
One of the strategies to take advantage of the knowledge possessed by organizationsis using knowledge management (KM). KM is the construction of systematic, deliberate and explicit knowledge that increase organizational effectiveness in order to use their knowledge assets (Wiig, 1997).

According Quintas et al. (1997) KM is the process of manage critical knowledge to meet existing needs as identify and better use of existing knowledge assets, acquired and develop new opportunities. Uriarte (2008) defines KM as the process by which organizations add value to their assets using knowledge, understanding that the KM is directly related to the process of identification, acquisition and maintenance of knowledge that are essential for organization.

Alavi et al. (1999) mentions that what is new and exciting in the area of the KM is the ability to use information technology: Internet,intranets, browsers, data warehouse, data filtering and software agents. Knowledge has been considered a key factor in an organization; therefor its effective management is critical to develop new competitive advantages. To carry out the above, a large number of organizations have begun to participate in KM initiatives and investing in knowledge management systems (KMS) (O'Brien et al. 2006, Hahn et al. 2009). The high demand for KMS includes the ability of organizations to be flexible and respond quickly to changing market conditions, and their potential for better decisions and productivity (Harris,1996).

The implementation of a KM system usually requires tools that support thatprocess, facilitating the flow of information between the different elements in a workgroup, therefore, the selection and implementation of these tools technologies becomes an important factor to support the KM(Ramirez and Martin,2003). With afunctional KMS, administrations can increment their resources related to organizational knowledge because they would be able to use, build, share and create very important knowledge (Kuo et al. 2011). A KMS must provide the information / knowledge to meet the needs of different tasks and a clarity way of the different types of users(Kuo et al. 2009, Liu et al. 2008).

While methodologies and KM tools are constantly evolving, the fundamentals of access to knowledge and information retrieval are unchanged. No matter if what is sought is a table of records, a finite element model of a neural network,a text document,an image,etc., Basic operation required to obtain the information and / or knowledge of every source can be modeled as a function that converts a set of input arguments (request for information) into a set of output arguments (informationrequested). Any of the functions analyzed can be defined as a source of knowledge (Bless et al., 2012).

To manage and store knowledge inside of organizations, it could use different sources of knowledge. Table 1 shows various sources to manage and store, a brief description and the authors are mentioned.

Table 1: Some Media Available to Store / Classify Knowledge

Knowledgesource	Description	Author
Documents (text, spreadsheets, slides, etc...)	Text classification, involves assigning a set of words in a document to a number of pre-defined classes automatically, using a machine learning technique. Labelisneededtoclassify data.	Dharmadhikari (2012) Dalal et al. (2011)
Use cases	It help to make the analysis, design, implementation and testing process. They appear to structure UML development concepts and application user requirements.	Dolques (2012) Jacobson (2004)
E-mail	It occupies a central place primarily for its low cost, wide spread (customers, suppliers, public companies, etc.), ease and variety of uses. It is widely used in business and commerce. It can be classified using various algorithms.	Park (2010) Zhu (2008)
Social networks (Facebook, Twitter, Google+, chat, Linked in, etc.)	Allow and favor to publish and share information, self-learning, teamwork, communication, feedback, access to other sources of information to support on a certain topic and even facilitate constructivist and collaborative learning as well as keep in touch with experts.	Imbernon et al. (2011)
Images	Visual representation, mental or verbal of a real or imaginary object. Go through three processes: production, circulation and reception. Its meaning becomes of the result of multiple social, moral, religious, among others.	Soto (2012) Paklone (2011)
Audio	It is any recording, or self-made magnetic or digital form presented in different media (CD, MP3, etc.). It can be "pure" or contain a series of mixed sounds.	Yu (2009) Lin (2005)
Video	They can be defined as the electric light, that generate containing visual information, it would be possible to say that if a text is show in a screen through electrical signals that contain visual information, it will be a form of video. Mayapplydifferentmethodstoperformclassification.	Dimitrova (2011) Pérez (2010)

The repositories of knowledge or organizational memories can even be considered a strategic product of the KM, which are supported by information technologies for their potential use. Nevo et al. (2005) mention that a repository can be described as the way like organizations store knowledge of the past to support the activities of the present. Walsh (1991) refers repositories as all information that is taken through the history of the organization, which can be retrieved to support the decisions of the present. For Stein (1995), the repository is the way in which knowledge of the past is transferred to the activities of the present, obtaining results of higher organizational effectiveness. Finally, Ackerman et al. (2000) considered repositories as a tool that contains information of some type for the entire organization.

When we are considering the processes of capture and retrieval of knowledge, the repository of knowledge, must have a structure that facilitates these processes.This paper is a review of the literature of sources of knowledge and the way in which they can be qualified. Organizations always have the need to classify information.It is presentedone of the most used techniques that do not use information technologies. Next, the relation of structured knowledge and its relationship to data mining is exposed. On the other hand it gives an overview of data mining, continuing with a series of algorithms used to classify information:supervised, semi- supervised and unsupervisedalgorithms. Then,there is a final classification analysis with the respective sources of knowledge:text, media andmultimedia. Later a summary of the sources of knowledge and its techniques / algorithms is shown. Finally we present the discussion and conclusions of this review.

2. Manual Classification Techniques

Classification is a procedure used to identify categories and assign entities to thosecategories, considering their attributes. For example a particular event,we can classify in KMas: data, information or knowledge. A classification strategy is widely used in KM taxonomies. Those are basic classification systems to describe concepts and their dependencies in a hierarchical manner.

Among the organizations there are different concepts that can be seen as "snapshots" of building knowledge and experience. Once the key concepts have been identified,captured and organized it can be categorized in a hierarchy form that is known as structured knowledge taxonomy. It allowsthe representation of knowledge graphically in a way that reflects the available information. If the concept is on the top, it is considered more general. If the concept is placed in a lower position, it will be more specific with respect to an instance of top-level categories. Building a taxonomy involves identifying, defining, comparing, and grouping items manually,so all organizational actors must agree on the classification system to be used to obtain the desired categories (Dalkir, 2011).

3. StructuredKnowledge

Once knowledge is explicit (documents, e- mail, audio, etc.),it should be organized in a structured way to allow better analysis and reuse it. A variety of techniques can be utilized to capture and encode structured knowledge. To structureknowledge within an organization is necessary to analyze the information in their databases, documents produced as a group, electronic meeting systems, organizational manuals, among others (Dalkir,2011). With the emergence of the need to analyze large amounts of information within a knowledge repository, the problem arises of how to reach and quickly analyze existing information. The term data mining (DM), also known as "knowledge discovery database" refers to the removal of potentially useful information which was previously unknown (Han et al. 2000).

To carry out the DM, the use of information systems is needed, which must have the ability to "self-learning", and the ability to work with series of algorithms and techniques for understanding the information they process. Here, users perform searches in a graphical interface to locate desired data(Tunstall,2010).

4. Data Mining(DM)

There are a number of applications for machine learning (ML), the most significant is DM(Vaithiyanathan et al. 2013).The DM involves the use of various tools that are sophisticatedfor data analysis to discover "unknown", valid patterns and relationships between large sets of data. These tools are just methods of machine learning, statistical models and mathematical algorithms. Not only is composed of a collection and data management, but also includes the analysis and prediction of them. Fayad (1996) mentions that it is a non-trivial process of identifying valid,novel, potentially useful and understandable patterns that are hidden in the data. Classification techniques are capable of processing a wide variety of data.

According Chauhan (2010) there are different techniques to carry out the DM: preprocessing,association, classification, pattern recognition and clustering. Rajeswari et al. (2011,2012)mention that the most popular techniques are classification and association, which serve to predict the interest of users and the relationship among those data items that have been used by them.

Nowadays, large amounts of information are manage, it causes the use of robust tools to ensure the reliability of the system, in spite of this, within this vast amount of data there is still a huge mass of hidden information from large importance, which is not accessible by conventional techniques of information retrieval. The discovery of this hidden information is possible by the DM, which enables appropriate processing results, because the correct algorithm is applied automatically to extract knowledge of inherent data (Solarte et al. 2009).
The application of mining algorithms has allowed to detect patterns on the data and, thus create models that support decision-making,which contributesin improvement of competitiveness indices or particular problem (Roddick and Lees, 2001).

4.1 Classification

Classification is a DM technique with wide applications, which classifies the data of different types. It can be used in all areas of our life, which helps categorize each data element within a predefined set of classes or groups (Vaithiyanathan et al. 2013).

With the arrival of "the information age", the growth of obtaining infinite resources of information has presented a major challenge when trying to process the data. In one hand, the information resources rise too fast and in the other, persons asking for valuable information also increase (Wei et al. 2012).

Hernandez (2004) defines classification as a process of information categorization. To set the categories, is required to prepare a pre- training base, which contains information on the different classes. If you have an object and want to know if it belongs to a class, identify the most significant attributes and evaluates how similar are the attributes of the training base.
In DM exists, a number of techniques to carry out the classification of different knowledge sources, among them are:

A) SupervisedClassificationAlgorithms

Dharmadhikari et al. (2011)supervised algorithms are those that use training data, where each document is labeled in zero or more categories in order to learn to classify, based on a categorizer for new texts. Here, a document is considered as a positive example for all the categories in which it's tagged and as a negative example for all those where it is not. Ozgur (2004)mentioned the task of a training algorithm for classification is to find "a weight vector" which categorizes a best new textual content.

Some of the supervised classification algorithms used are:

- **K – NearestNeighborClassifier**

Is known to be a good pattern recognition algorithm. Taking a test document,this algorithm is a "k - neighbors" closer to each training document and use the categories of the k- neighbors in order to assign a class (Narayana, 2003). For Dharmadhikari et al. (2011), this algorithm is also based on the assumption that the characteristics of the members of the same class should be similar.

Advantages: It is an effective, simple, non-parametric and easy to implement.

Disadvantages: The main disadvantage is that it becomes slow when the training set size grows, so that the presence of irrelevant features severely degrades its accuracy.

- **Artificial Neural Network (ANN) Classifier**

The method consists of multi - layer feed- in which there is at least one input layer, one hidden and one output which are based on standard back propagation learning. Learning occurs by adjusting all those "weights" in the nodes to minimize all differences between the output node activation with normal output (Atkinson et al., 1997).

Advantages: High level pattern recognition.

Disadvantages: High -level learning and computational.

- **ClassifierParallelepiped**

Peruma et al. (2010) mention that this classifier consists of two image bands that are used to determine the "training area" of the pixels based on maximum and minimum values. Many pixels are often left without classify because there may be overlap between each of the pixels. Another detail about this type of classification is that the data values of each pixel are compared with the upper and lower limits set.

Advantages: Within supervised classifiers usually one of the simplest.

Disadvantages: Although it is considered the most accurate classification techniques, it´s not the most used,because can leave off a lot of pixels from classification.

- **MinimumDistance**

It is based on the minimum distance decision rule that calculates the spectral distance between the measurement vectors for the candidate pixel class that has the minimum distance of the standard comparison. The texture descriptors are useful for distinguishing between different types of surface, here the most common analysis methods are based on analyzing any tone "gray space" or occurrence analysis (Parinello et al., 2006).

Advantages: Easy implementation.

Disadvantages: Classes may not become well generated because they arise based on the pixels that sometimes are not all analyzed.

- **NaiveBayesClassifier**

This method is one that makes assumptions of data independence (Narayana,2003). Furthermore Khan et al. (2010) mention that this classifier makes predictions by reading a series of attributes taken as an example for the representation and then apply Bayes theorem to estimate the probabilities. The assumption of independent characteristics that makes the whole presence of irrelevant features does not affect the time to perform the classification.

Advantages: Requires a small amount of training data to estimate the parameters necessary for classification. Classifiers based on this algorithm, shows a high accuracy and speed when applied to databases.

Disadvantages: Works well only if the features are assumed independent, so when dependence exists it doesn´t.

- **DecisionTreeClassifier**

In this method, the classification of the training documents is done by constructing true / false definitions that has to be well defined within the tree structure. The leaves represent the category for sorting and branches the conjunctions of features that lead to the various categories (Khan et al. 2010).

Advantages: Fits any type of data. Usually fast, even in large amounts of attributes.

Disadvantages:Its greatest risk is that only fits the training data by the occurrence of an alternative tree.

- **ClassifierDecision Rules**

This method uses inference based on rules to put the documents in their categories (Khan et al. 2010). These classifiers are useful for non-standardized data analysis. Another feature is that it builds a set of rules that describe the profile of each category. The rules have the structure: "If condition then conclusion", where conditions will be those that correspond the characteristics of the class and the conclusion by the name of the categories or some other rule to be tested (Dharmadhikari et al. 2011).

Advantages: Capacity of a good systematic performance analysis.

Disadvantages: The main disadvantage is the need for the involvement of experts to build the set of rules.
Table 2 shows a list of other algorithms that use techniques based on supervised classifications. As can be seen each algorithm has a different base for its development, which depends on the type of information that is intended to be classify.

Table 2: Comparative Analysis of Algorithms for Supervised Classification Type

Supervisedclassificationmethods				
Algorithm	**Advantages**	**Disadvantages**	**Base**	**Source**
BR	Fitscalculation diagonal matrices.	No tag correlations performed explicitly.	GaussianMatrix	Geist (1999)
Adaboost	Excellent for sorting. Better accuracy.	Generalizing results in decreased performance.	Algebraic	La et al. (2012)
Back Propagation	Learning iteratively. More capacity of generalization.	Computationally complex presented by the algorithm.	Algebraic	Pradeep et al. (2011)
C4.5	Based on decision trees, improving accuracy and prediction. Easy to understand, popular and powerful.	Not takes correlations between classes.	Algebraic	Hantan et al. (2010)

As can be seen each algorithm has a different base for its development, which depends on the type of information that is intended to be classify.

B) UnsupervisedClassificationAlgorithms

Dharmadhikari et al. (2011) mention that unsupervised algorithms do not have a collection of labeled documents. The goal is to group documents without knowledge or additional intervention such that documents within a group have the same similarities. This type of methodology is classified into two major groups: partitioned and hierarchical. Hierarchical algorithms are those that produce nested partitions by dividing data (divisiveapproach) or by fusion (agglomeratesapproach) similarities based groups. In the other hand, partitioned algorithms grouped data into non-overlapping partitions making grouping criteria optimized.
Some of the non-supervised classification algorithms used are:

- **HierarchicalClustering**

This type of algorithm produces a cluster hierarchy as tree structure called "dendrogram". The root of the tree consists of a single group that contains all the observations, and the leaves correspond to individual observations (Berkhin,2002).
Advantages: Simplicity and ability to capture the information correctly.
Disadvantages: Not discreet groups and groups have different categories for different type of groups.

- **DivisiveHierarchicalClustering**

This type of algorithms start with one group. During each iteration, it is divided in different groups until the most relevant is found and certain criteria, such as a requisition number or some "k" value, is reached. At each step of this technique, the group with the biggest size is divided. As document similarity is used rather than alienation of patterns, the group is divided in one that contains fewer similarities to the base documents. Then, all documents with less significant similarities are removed from the group to form a new group focused in the requirements(Dharmadhikari et al.,2011).
Advantages: It has a higher average similarity within the documents of the new cluster.
Disadvantages: Whilethe size of group increments, it becomes complex.

- **HierarchicalAgglomerativeClustering**

In this type of algorithms documents start in a separate group. Each iteration merges all those groups that have similarities until it meets the search criteria. They are classified primarily as single link, complete link and average link depending on the method defined in similarity groups (Mitchell, 2006).
- Single link: Defines the similarity of two sets Ci and Cj as the two most similar groups.
- Complete link: Defines the similarity of two sets Ci and Cj as less similar the two groups.
- Average link: Defines the similarity of two sets Ci and Cj as the average of similarities between the two groups.

- **ClusteringPartitional**

In this type of grouping, classes are mutually exclusive. Each object is a member of one that has the highest similarity, so that each object can be classified into a representative group. Once given the number of "k" clusters, initial partition will be built, then the solution group is refined to move documents from one group to another (Ozgur,2004).

Advantages: It only takes a look through the data set and is relatively fast.
Disadvantages: The resulting clusters are not independent related with the order in which the documents are processed.

- **Kohonen'sSelfOrganizing Network**

Dharmadhikari et al. (2011) mentioned that this type of algorithm uses a special type of neural network which is called "self-organizing Kohonenwork". The novelty of this method is that it automatically detects the number of classes present in a set of documents given and then places it in its appropriate class. Initially uses self-organizing network to explore the location of possible groups within the feature space. After that, they will check if any of the groups can be merged on any basis that is appropriate grouping, which represent the different classes present in a given set of documents.
Advantages: They tend to be easier for humans to see the relationships between large amounts of data.
Disadvantages: Computationally complex and extensive.

C) SemisupervisedClassificationAlgorithms

It uses unlabeled data along with some that are labeled to get categories for new documents that have not been tagged yet. In text classification most of the times there are limitations on data showing labels,making it costly for organizations. Based on the above, semi- supervised algorithms are a good solution to this situation because their application framework provides grouping (Pise, 2008).
Some of the semi- supervised classification algorithms used are:

- **Co-Training**

This type of algorithm is usually applied when a data set has a natural division of their characteristics and requires two data views (Subramanya,2008). It is assumed that each set provides a set of features that are conditionally independent (Dharmadhikari et al. 2011). First it works with a classifier for each view using examples that are labeled. Confidence predictions of each classifier are then used to carry out the labeling of the data that are not labeled, and thus, build additional training data (Nigam et al. 2006).

Advantages: Simplicity and applicability to almost all existing classifiers.

Disadvantages:Naturally, in some division applications, features may not be available as their performance in such cases is usually low.

- **ExpectationMaximization (EM) Based**

Also known as expectation maximization algorithm, often used to train classifiers by estimating parameters of a generative model of expectation maximization (Nigam et al. 2006).

Advantages: It is a very simple representation with respect to the complexity of a text, another advantage is that using established classes and large amounts of unlabeled data,it can find a more likely model than if you use only labeled data.

Disadvantages: No guarantee maximum global in the probability model covering space.

- **Graphbased**

Dharmadhikari et al. (2011) mentioned that this type of algorithms work with a set of test data which is displayed when "training",and assumes that the data are inserted into a low-dimensional manifold expressed by a graph. Each simple data is represented by a vertex in a graph that is weighted by weights providing a measure of similarity between the vertices.

Advantages: not only provides a binary solution but also can be measure multiclass and uncertainty.

Disadvantages: Require excessive calculations.

Since there are a variety of classification methods in Table 3 are illustrated some methods often used for semi-supervised algorithms.
There are a lot of algorithms used in semi-supervised methods, some of them are mentioned in table 3:

Table 3: Some Semi-Supervised Classification Algorithms

Semi-supervisedclassificationmethods				
Algorithm	**Advantages**	**Disadvantages**	**Base**	**Source**
Multi-labelclassificationbyconstrained non-negative matrix factorization	Adaptable to semi-supervised environments along with the representation of documents in rank matrix factorization using the non-negative. Is adapted to an environment with a small number of training data and a large number of class labels.	There is a strong influence from two parameters on the performance: latent variables and tuning parameters. Theincorrectvaluechosenwillsignificantly reduce performance.	Featurematrixanalysis	Y. Liu, R. Jin, L. Yang (2006)
Graph-based SSL with multi-label	Effective use of large amounts of unlabeled data and the ability to exploit the relationships between labels.	Most of the time is used for video files. It does not adapt well to texts	Algebraic	Z. Zha, T. Mie, Z. Wang, X. Hua (2008)
Multi-label learning by using dependency among labels	Improving accuracy by configuring SSL	Time increment for large data sets	Analysis of relationships, cross-validation method	Wei, Yang, Zhu y Wang (2011)
Semi supervised multi-label learning by solving a Sylvester Eg (SIAM)	Use of large amounts of unlabeled data as well as the ability to exploit the relationship between labels. Significantimprovement in theprecision	May become slow when using large data sets	Graphics for data entry, analysis of nodes	Chen,Song, Wang y Zhang (2008)
Semi-supervised non-negative matrixfactorization	Using NMF in conjunction with SSL allows the extraction of the most discriminating than if MFN were used only	ComputationalComplexity	Data matrix, factorization joint data matrix characteristics	Lee ,Yoo y Choi (2009)

Since there are a variety of classification methods in Table 3 are illustrated some methods often used for semi-supervised algorithms.

5. Source of KnowledgeClassification

This section presents sources used to store knowledge in an electronic organizational repository organizational and its relationship with automatic classification techniques.

5.1 Classification of Text (Documents,Spreadsheets,Slides, Use Cases)

The wide availability of text documents in digital form has provided a wealth of information for people whouse it,and it has must be organized systematically. Perform the above, facilitates better information storage, search and retrieval of relevant text content to the needs of whoever seeks (Dharmadhikari et al. 2011).
Automatic text classification is a major problem seen from different areas which has many applications such as automatic indexing of scientific articles, spam filtering, gender identification of specific document, author attributions,automated classification tests,survey coding, classification of news, among others (Fabrizio et al.,2002).

Machine learning techniques have the great advantage of beingunderstood in a better way from the theoretical view, allowing users to rely on the performance and configuration of certain parameters to carry it out. All techniques are divided into three groups, which are supervised, unsupervised and semi- supervised learning (Wang et al.,2006).

Wei (2012) mentions that text classification focuson all areas devoted to process information, where a large set of files are preset and assigned to predefined some kind of attributes, in which,the main point will be to split hypertext files in several categories based on predefined content,so in this case, what is studied is the content of a text file. In the classification system the text classifier is a key, so that the quality of performance of the classifier is directly related to the effect of classifier and effectiveness of the text.
Some of the algorithms used for this type of classification are: K - nearest neighbor, Naive Bayes, Decision tree, Decision rules,SVM,ANN,partitionalcluster, based Graph:Dharmadhikari et al. (2011), Khan (2010), Wang et al. (2006),Narayana (2003).

5.2 Media Communication Classification(E - Mail, Social Networking, Chat)

Taiwo et al. (2010)mention that the use of (email) has become one of the fastest and most effective forms of communication that significantly impacts on the transfer of knowledge within a particular group of people. However, the increase of users using it and the large volumes of messages can result inunstructured mails,congestion,overload, priority loss, among others. This results in a direct impact for wanting to have tools that help to manage and sort the contents of this media.

Massive mail sending, also known as spam, has generated a need for reliable anti-spam filters. Therefore, Awad et al. (2011) mention that the use of a classifier based on machine learning techniques could automatically filter mail spam or even any kind of information you want.

Social networking (Facebook,Twitter, Google+, Linked in,etc.), makes that organizational learning can be more interactive and meaningful as they facilitate greater interaction with the staff that is working there. On this basis it is important to analyze in detail and if possible to classify the information handled in order to facilitate and improve job performance. Make appropriate use of the above will generate a more dynamic organizational environment (Imbernon et al., 2011).

Some of the algorithms used for this type of classification are: Graph based, Naive Bayes, K - nearest neighbor,ANN,SVM,non-negative matrix factorization,partitionalcluster, Back Propagation:Awad et al. (2011),Pradeep et al. (2011), Park (2010),Taiwo et al. (2010).

5.3 Media Classification (Images,Audio,Video)

For Palaniswani et al. (2006) image classification is a process to “assemble” identical groups of pixels that can eventually be found in data obtained by detecting features within an image based on predefined classes, which correspond to a certain type requested by user categories based on comparison of pixels with each other and the identity that is already known (classes). The identification of images has come a significantly impact because ofprovidingimportant information and meanings for organizations as it is a common capture medium that is used today. This type of classification is performed by identifying pixels and comparing them with previously classified pixels in different types of classes in order to classify the unknown pixels. An example of known identity pixel is anyone that is located within the “training areas” of the class. The statistical characteristics of the classes of pixels estimated from training, depend on the algorithm that will be used for classification (Parinello et al.,2006). Only a small number of image classification algorithms have been shown to have good accuracy with respect to the classification and detection of images. Today we need to classify to be effective when trying to categorize the existing contents of any type of image(Perumal et al., 2010).

To retrieve the information (knowledge) of users in large multimedia data content, the automatic classification of audio -video plays an important role to perform it. Both can be classified and store in a well-organized database for later use. To classify the audio data is necessary to use sound wave coefficients represented by images and video features color histograms that are extracted from the images of the video clips which are used to obtain visual features (Subashini et al. 2012).

In order to have an automatic recovery video it´s necessary that it will be classify. Brezeale (2008) has studied and regarding surveys related with automatic classification techniques as text, visual range, color histogram, textures movements, among others.

For audio classification much effort has been devoted to investigating the relevant characteristics of the various types of sound that are available, as well as algorithms that can be applied for classification, which is done by studying the waves generated in the time domain (Yu,2009).

Some of the algorithms used for this type of classification are: ANN Parallepiped, Minimum distance: Atkinson et al. (1997),Perumal et al. (2010), Parinello et al. (2006),SVM, Audio segmentation:Subashini et al. (2012), Yu (2009),Dhanalakshmi et al. (2008).

6. Synthesis of Knowledge Sources and Their Classification

In table 4 can be seen various sources of knowledge that have been mentioned in this article and classification techniques used in each (table 4).

Table 4: Sources of Knowledge and Techniques Used for Classification

Knowledgesource	Techniquesused
Text (documents, spreadsheets, slides, use cases)	K-nearest neighbor, Naive Bayes, Decision tree, Decision rules, ANN, Partitional cluster, Graph based.
Media (E-mail, social networking, chat)	Graph based, Naive Bayes, K- nearest neighbor, ANN, SVM, non-negative matrix factorization, Partitional cluster, Back Propagation.
Multimedia (Images, Audio, Video)	SVM, Audio segmentation, ANN, Parallepiped, Minimum distance.

7. Discussion

Dalal et al. (2011) note that some researchers have reported improved accuracy of classifiers by combining techniques and automatic learning methods. To Goyal (2007) the performance of text classification based on neural networks was improved by assigning probabilities from the Naive Bayes method. Furthermore, Isa et al. (2008) used the Naive Bayes method as a size reduction processor in combination with the method of obtaining a better SVM detailed classification in their study. Dharmadhikari et al. (2011) presents another example of application, jointly applying the EM algorithm and Naive Bayes classifier resulting in much more efficient text. Also set a decision tree with neural network techniques so you can build networks through direct allocation or decision nodes by generating neural units rules eliminating redundant data between the connections.
As noted in Table 4, depending source of knowledge that is wanted to classify, there exists various techniques to do that. The selection of the most appropriate technique will depend on exactly what you will want to find and in such case the need to implement hybrid techniques should be analyze what would be the best options to suit the data type to be studied.

Finally we can assume that there is a great need to experiment with hybrid techniques based on the different methods presented, in order to get better benefits and results to classify information (Dalal et al.,2011). With respect to the different sources of knowledge mentioned throughout this, the methods of classification and clustering are fundamental in decision-making processes, helping to identify, distinguish, and even to establish criteria for evaluating different alternatives for dealing with a specific situation or problem (Galindo,2010).

8. Conclusions

We have presented a review of the knowledge within organizations and the use of KMsystems. These are made out of records containingorganizational knowledge coming from different sources, which is important to identify for a better management and an efficient retrieval.It is importantto mention that there is no order or properly manner on how to classify the information that is reflected in the different sources,that is the reason of this literature review about the various techniques used to classify information available in different knowledge sources and how they impact substantially on the improvingof organizational performance.

There have been supervised, semi - supervised and unsupervised classification methods in which a series of algorithms were use and the way they work. Some techniques are usually more complicated than others and depend on the type of information or knowledge that organizationrequire to be classified in order to determine which would be the most appropriate algorithm to implement.It is important to know that classification requires previousinformation, which is not always available in an organization. In this case the grouping process is more suitable because of no previous knowledge is requireand it is generated base on different machine learning patterns. In some cases a combination of two or more techniques is proving to be a point of interest for researchers, because doingit properlywill impact directly by satisfactory results when making categorization.

9. References

Ackerman, M. and Halverson, C. (2000). Re-examining organizational memory.Communications of the ACM.

Alavi, M. andLeidner D. (1999). Knowledge management systems: Issues, challenges, and benefits. Smith School of Business.Volume 1, Article 7.

Atkinson, P. andTatnall, A. (1997). Introduction to neural networks in remote sensing, International Journal of Remote sensing, volume 11, pp. 699-709.

Awad, W. andELseuofi, S. (2011). Machine Learning methods for E-mail Classification. Math.&Comp.Sci.Dept., Science faculty, Port Said University Information. System Dept., Ras El Bar High inst

Berkhin, P. (2002). Survey of Clustering Data Mining Techniques. Research paper, Accrue Software, http://www.accrue.com/products/researchpapers.html

Bless, P., Klabjan, D. and Chang, S. (2012). Automated knowledge source selection and service composition. Computational optimization and applications.

Brezeale, D. and Diane, J. (2008). Automatic video classification: A Survey of the literature. IEEE Transactions on systems, man, and cybernetics-part c: application and reviews, vol. 38, no. 3, pp. 416-430.

Britos, P., Hossian, A., García-Martinez, R. and Sierra, E. (2005). Minería de Datos Basada en Sistemas Inteligentes. Nueva Librería.

Chen, G., Song, Y. and Zhang, C. (2008). Semi-supervised Multi-label Learning by Solving a Sylvester Equation.In SDM.

Chauhan, R., Kaur, H andAlam, M. (2010). Data Clustering Method for Discovering Clusters in Spatial Cancer Databases. International Journal of Computer Applications Volume 10– No.6

Dalal, M. andZaveri, M. (2011). Automatic text classification: A technical review. International Journal of Computer Applications, volume 28- No. 2.

Davenport, T. andPrusak L. (1998). *Working Knowledge: How Organizations Manage What They Know,* Boston: Harvard Business School Press.

Dhanalakshmi, P., Palanivel. S. andRamaligam. V. (2008). Classification of audio signals using SVM and RBFNN.In Elsevier, Expert systems with application, Vol. 36, pp. 6069–6075.

Dharmadhikari, S., Ingle, M. and Kulkarni, P. (2011).Empirical Studies on Machine Learning Based Text Classification Algorithms. Advanced Computing: An International Journal (ACIJ), Vol.2, No.6.

Dharmadhikari, S., Ingle, M., and Kulkarni, P. (2012). Analysis of semi supervised learning methods towards multi label text classification. International Journal of Computer Applications.Volume 42- No. 16.

Dimitrova, N. andAgnihotri, L. (2011). Video classification using object tracking.

Dolques, X. (2012). Fixing generalization defects in UML use case diagrams. Campus universitaire de Beaulieu, 35042 Rennes, France.

Fabrizio, S. (2002).Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1.

Fayyad, U. (1996). Data mining and Knowledge discovery in databases: Applications in Astronomy and Planetary Science. Microsoft Research.

Galindo, K., Juganaru, M., Áviles, C. and Vázquez, H. (2010). Desarrollo de una aplicación destinada a la clasificación de información textual y su evaluación por simulación. Dpto. DeSistemas, DCBI. México, DF.

Geist, G., Howell, G. and Watkins, D. (1999). The Br eigenvalue algorithm.Society for industrial and applied mathematics. Vol. 20, No. 4, pp. 1083-1098.

Goyal R. (2007). Knowledge based neural network for text classification. In proceedings of the IEEE international conference on Granular Computing, pp. 542 – 547.

Hahn, J. and Wang, T. (2009). Knowledge management systems and organizational knowledge processing challenges: a field experiment. Decision Support Systems, Vol. 47 No. 4, pp. 332-42.

Han, J., and Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Harris, D. (1996). Creating a Knowledge Centric Information Technology Environment, http://www.htca.com/ckc.htm

Hernández, J., Ramírez, M., andRamírez F. (2004). Introducción a la Minería de Datos.

Ichijo, K. andKohlbacher, F., (2006). Global Knowledge Creation – The Toyota Way. Int. J. Automotive Technology and Management, 7, pp. 116-134.

Isa, D., Lee, L., Kallimani, V. andRajKumar, R. (2008).Text document pre-processing with the Bayes formula for classification using the support vector machine.

Imbernón, F.; Silva, P. andGuzmán, C.(2011). Competencias en los procesos de enseñanza-aprendizaje virtual y semipresencial. Co -municar, 36; 107-114.

Jantan, H., Handan, A. and Othman, Z. (2010). Human talent prediction in HRM using C4.5 classification algorithm.International Journal on Computer Science and Engineering.

Khan, A., Baharudin, B., and Lee, L. (2010).A Review of Machine Learning Algorithms for Text- Documents Classification.Journal of Advances in Information Technology, Vol. 1, No. 1.

Kuo, R., Lai, M. and Lee, G. (2011).The impact of empowering leadership for KMS adoption. Management Decision.

Kuo, R. and Lee, G. (2009). KMS adoption: the effects of information quality. Management Decision, Vol. 47 No. 10, pp. 1633-51.

La, L., Guo, Q., Yang, D. and Cao, Q. (2012). Multiclass boosting with adaptive group-based kNN and its application in text categorization. Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2012, Article ID 793490, 24 pages

Lee, Yoo and Choi. (2009). Semi-supervised Non-negative Matrix factorization. IEEE.

Lin, R. and Chen, L. (2005). A new approach for classification of generic audio data.Department of computer and Information Science. National Chiao Tung University.

Liu, D. and Wu, I. (2008). Collaborative relevance assessment for task-based knowledge support. Decision Support Systems, Vol. 44 No. 2, pp. 524-43.

Liu, Y., Jin, R. and Yang, L. (2006). Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization .In: AAAI.

Mitchell, T. (2006). The Discipline of Machine Learning, CMU-ML-06-108, July 2006.

Narayana, K. (2003). Advances in Automatic text categorization.DRTC Workshop on Semantic Web, Bangalore, India, 8-10 December, 2003.

Nevo, D. and Wand, Y. (2005). Organizational memory information systems: a transactive memory approach. Decision Support Systems 39, 549-562.

Nigam, K., McCallum, A., and Mitchell. T. (2006). Semi-supervised Text Classification Using EM. MIT Press.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39, 103–134.

Nonaka, I. (1994). "A Dynamic Theory of Organizational Knowledge Creation" Organization Science, (5) 1, pp. 14-37.

O'Brien, J. andMarakas, G. (2006). Management Information Systems, 7th ed., McGraw-Hill International, New York, NY.

Ozgur, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization. Thesis submitted in Department of Computer Science, Bogaziki University.

Paklone, I. (2011). Conceptualization of visual representation in urban planning. ISSN 2029-7475.

Palaniswami, C., Upadhyay, A. K., andMaheswarappa, H. P. (2006). Spectral mixture analysis for sub pixel classification performance of multispectral images.Journal of Computing, 2(2), 124-129.

Park, S. and Dong, U. (2010). Automatic E-mail classification using Dynamic category hierarchy and semantic features.School of information technology, Chonbuk national university, South Korea.

Parrinelo, T., Vaughan, R. (2006). On comparing multifractal and classical features in minimum distance classification of AVHRR imagery. International Journal of Remote Sensing Vol. 27, No. 18, 3943-3959.

Pise, N andKulkarni, P. (2008). A survey of Semi-Supervised Learning Methods".IEEE International conference on Computational Intelligence and Security. 2008.30-34.

Perez, J. (2010). Internet's Information as a Video Signal and its Editor Vol. 29 Issue 57, p178-191, 14p.

Perumal, K. andBhaskaran, R. (2010). Supervised Classification Performance of Multispectral Images.Journal of Computing, Volume 2, Issue 2, February 2010, ISSN 2151-9617.

Polanyi, M. (1962). Personal Knowledge: Toward a Post-Critical Philosophy, New York, NY: Harper Torchbooks.

Pradeep, T., Srinivasu, P., Avadhani, P.andMurthy, Y. (2011). Comparison of variable learning rate and Levenberg-Marquardt back-propagation training algorithms for detecting attacks in Intrusion Detection Systems. TummalaPradeep et al. / International Journal on Computer Science and Engineering (IJCSE).

Quaglini, S., Panzarasa, S., Cavallini, A., Micieli, G., Pernice, C. andStefanelli, M. (2005). Smooth Integration of Decision Support into an Existing Electronic Patient Record. Artificial Intelligence in Medicine, pp. 89 – 93.

Quintas, P., Lefrere, P. and Jones, G. (1997). Knowledge management: a strategic agenda. Journal of Long Range Planning, Vol. 30, No. 3, pp. 385-91.

Rajeswari, K andVaithiyannathan, V. (2012). Mining Association Rules Using Hash Table. International Journal of Computer Applications.

Rajeswari, K. andVaithiyannathan, V. (2011). Heart Disease Diagnosis: An Efficient Decision Support System Based on Fuzzy Logic and Genetic Algorithm. International Journal of Decision Sciences, Risk and Management by Inderscience Publications.

Ramírez, P. N. and Martín, M. A., (2003). Herramientas para la gestión del conocimiento. Categoría Administración, Gestión del Conocimiento, pp. 1-34.

Roddick, J. and Lees, B. (2001). Paradigms for spatial and spatio-temporal data mining. In MILLER, H. y HAN, J. Geographic data mining and knowledge discovery. London: Taylor & Francis.

Sîrbu, M., Doinea, O. andGoirgiana, M., (2009). Knowledge based economy - the basis for insuring a sustainable development. Annals of the University of Petroşani, Economics, 9, pp. 227-232.

Solarte, G andOcampo, C., (2009). Técnicas de Clasificación y análisis de representación del conocimiento para problemas de diagnóstico. Scientia et Technica Año XV, No 42 Agosto de 2009. Universidad Tecnológica de Pereira.

Soto, J. (2012). Images, society and its decoding. ISSN: 1578-8946.

Stein E. (1995). Organizational Memory: Review of Concepts and Recommendations for Management: International Journal of Information Management, Vol. 15, No 2, pp. 17-32.

Subashini, K., Palanivel, S. andRamalingam, V. (2012). Audio-video based classification using SVM and AANN. International Journal of Computer Applications (0975 – 8887).

Subramanya, A. andBilmes, J. (2008). Soft-Supervised Learning for text classification.Proceedings of the Conference on Empirical Methods in Natural Language Processing. Pages: 1090-1099.2008.

Taiwo, A., Shikun, Z. andRinat, K. (2010). Email Classification Using Back Propagation Technique. Department of Electronics and Computer Engineering University of Portsmouth, United Kingdom.

Tunstall, W. (2010). True Knowledge: Open-Domain Question Answering Using Structured Knowledge and Inference. Association for the Advancement of Artificial Intelligence (www.aaai.org).

Uriarte, F. A., (2008). Introduction to Knowledge Management. Jakarta, Indonesia: ASEAN Foundation.

Vaithiyanathan, V., Rajeswari, K., Tajane, K. andPitale, R. (2013).Comparison of different classification techniques using different datasets.International Journal of Advances in Engineering & Technology.

Walsh, P. (1991). Organizational Memory, Academic of Management Review. 1, pp. 57-91.

Wang, Z., Sun, X. and Zhang, D. (2006). An optimal Text categorization algorithm based on SVM.Communications, Circuits and Systems Proceedings,

Wei, L., Wei, B., and Wang B. (2012). Text Classification Using Support Vector Machine with Mixture of Kernel. A Journal of Software Engineering and Applications, 2012, 5, 55-58

Wei, O. Yang, J. and Wang (2011). Semi-supervised Multi- label Learning Algorithm using dependency among labels.In IPCSIT vol. 3.

Wiig, K. (1993). Knowledge management foundations: Thinking about thinking. How people and organizations create, represent and use knowledge. Arlington, TX: Schema Press.

Wiig, K., Towe B. andPizziconi V. (1997). Knowledge Management: Where Did It Come From and Where Will It Go? Expert Systems with Applications, Vol. 13, pp 1.14.

Yu, G. andSlotine, J. (2009). Audio classification from time-frequency texture.

Zha, Z. Mie, T., Wang, Z. and Hua, X. (2008). Graph-Based Semi-Supervised Learning with Multi-label. In ICME.pp 1321-1324.

Zhu, Z. (2008). An email classification model based on rough set and support vector machine. In the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5, pp. 236-40.