The Knowledge Engineering Review, Vol. 30:2, 157–170.
doi:10.1017/S0269888914000277

A survey on text mining in social networks

RIZWANA IRFAN[1], CHRISTINE K. KING[1], DANIEL GRAGES[1], SAM EWEN[1], SAMEE U. KHAN[1], SAJJAD A. MADANI[2], JOANNA KOLODZIEJ[3], LIZHE WANG[4], DAN CHEN[5], AMMAR RAYES[6], NIKOLAOS TZIRITAS[4], CHENG-ZHONG XU[4], ALBERT Y. ZOMAYA[7], AHMED SAEED ALZAHRANI[8] and HONGXIANG LI[9]

[1]*North Dakota State University, Fargo, 58102 ND, USA;*
e-mail: rizwana.irfan@ndsu.edu, christine.k.king@ndsu.edu, daniel.grages@ndsu.edu, sam.ewen.2@ndsu.edu, samee.khan@ndsu.edu;
[2]*COMSATS Institute of Information Technology, 44000 Islamabad, Pakistan;*
e-mail: madani@ciit.net.pk;
[3]*Cracow University of Technology, 30001 Cracow, Poland;*
e-mail: jkolodziej@uck.pk.edu.pl;
[4]*Chinese Academy of Sciences, 100864 China;*
e-mail: lzwang@ceode.ac.cn, cz.xu@siat.ac.cn, nikolaos@siat.ac.cn;
[5]*China University of Geosciences, 430000 Wuhan, China;*
e-mail: chendan@pmail.ntu.edu.sg;
[6]*CISCO Systems, San Jose, 94089 CA, USA;*
e-mail: rayes@cisco.com;
[7]*University of Sydney, 2006 NSW, Australia;*
e-mail: albert.zomaya@sydney.edu.au;
[8]*King Abdulaziz University, 21589 Saudi Arabia;*
e-mail: asalzahrani@kau.edu.sa;
[9]*University of Louisville, 40292 KY, USA;*
e-mail: h.li@louisville.edu

Abstract

In this survey, we review different text mining techniques to discover various textual patterns from the social networking sites. Social network applications create opportunities to establish interaction among people leading to mutual learning and sharing of valuable knowledge, such as chat, comments, and discussion boards. Data in social networking websites is inherently unstructured and fuzzy in nature. In everyday life conversations, people do not care about the spellings and accurate grammatical construction of a sentence that may lead to different types of ambiguities, such as lexical, syntactic, and semantic. Therefore, analyzing and extracting information patterns from such data sets are more complex. Several surveys have been conducted to analyze different methods for the information extraction. Most of the surveys emphasized on the application of different text mining techniques for unstructured data sets reside in the form of text documents, but do not specifically target the data sets in social networking website. This survey attempts to provide a thorough understanding of different text mining techniques as well as the application of these techniques in the social networking websites. This survey investigates the recent advancement in the field of text analysis and covers two basic approaches of text mining, such as classification and clustering that are widely used for the exploration of the unstructured text available on the Web.

1 Introduction

Social networking websites create new ways for engaging people belonging to different communities (Baumer *et al.*, 2010). Social networks allow users to communicate with people exhibiting different moral and social values. The websites provide a very powerful medium for communication among individuals

that leads to mutual learning and sharing of valuable knowledge (Sorensen, 2009). The most popular social networking websites are Facebook, LinkedIn, and MySpace where people can communicate with each other by joining different communities and discussion groups. Social networking can solve coordination problems among people that may arise because of geographical distance (Evans *et al.*, 2010; Li *et al.*, 2011b) and can increase the effectiveness of social campaigns (Li & Khan, 2009a, 2009b; Baumer *et al.*, 2010) by disseminating the required information anywhere and anytime. However, in social networking websites, people generally use unstructured or semi-structured language for communication. In everyday life conversation, people do not care about the spellings and accurate grammatical construction of a sentence that may lead to different types of ambiguities, such as lexical, syntactic, and semantic (Sorensen, 2009). Therefore, extracting logical patterns with accurate information from such unstructured form is a critical task to perform.

Text mining can be a solution of above-mentioned problems. Owing to the increasing number of readily available electronic information (digital libraries, electronic mail, and blogs), text mining is gaining more importance. Text mining is a knowledge discovery process used to extract interesting and non-trivial patterns from natural language (Sorensen, 2009). The technique comprises of multidisciplinary fields, such as information retrieval, text analysis, natural language processing (NLP), information classification, and database technology. In Liu and Lu (2011), the authors defined text mining as an extension of data mining technique. The data mining techniques are mainly used for the extraction of logical patterns from structured database. Text mining techniques become more complex as compared with data mining owing to unstructured and fuzzy nature of natural language text (Kano *et al.*, 2009).

Social networking websites such as Facebook are rich in texts that enable user to create various text contents in the form of comments, wall posts, social media, and blogs. Owing to ubiquitous use of social networks in recent years, an enormous amount of data are available via the Web. Application of text mining techniques on social networking websites can reveal significant results related to person-to-person interaction behaviours. Moreover, text mining techniques in conjunction with social networks can be used for finding general opinion about any specific subject, human thinking patterns, and group identification (Aggarwal, 2011). Recently, researchers used decision trees (DTs) and hierarchical clustering (text mining techniques) for group recommendation in Facebook where user can join the group based on similar patterns in user profiles (Baatarjav *et al.*, 2008).

For the past few years there has been a lot of research in the area of text mining. In the scientific literature (Yin *et al.*, 2007; Tekiner *et al.*, 2009; Jo, 2010; Ringel *et al.*, 2010), various text mining techniques are suggested to discover textual patterns from online sources. In Baharum *et al.* (2010), the authors restrict the analysis to techniques that are specifically associated with text document classification. Brucher stated various clustering-based approaches for document retrieval and compared different clustering techniques for logical pattern extraction from unstructured text, but most of the techniques presented in the papers are not recent (Brucher *et al.*, 2002). In Durga and Govardhan (2011), the authors proposed a new model for textual categorization to capture the relations between words by using WordNet ontology (Xu *et al.*, 2008). The proposed approach maps the words comprised of same concepts into one dimension and present better efficiency for text classification. In Xu *et al.* (2008), the authors indicated a best practice in information extraction process based on semantic reasoning capabilities and highlighted various advantages in terms of intelligent information extraction. The author explained the suggested methods, such as query expansion and extraction for semantic-based document retrieval, but did not mention any results associated with the experiments. In Tekiner *et al.* (2009), the author introduced general text mining framework to extract relevant abstract from large text data of research papers. However, the proposed approach neglected the semantic relations between words in sentences.

Most of the scientific literature (Xu *et al.*, 2008; Tekiner *et al.*, 2009; Li *et al.*, 2011a) focuses on specific techniques of text mining for information extraction from text documents. However, a thorough discussion is lacking on the actual analysis of different text mining approaches. Most of the surveys emphasize on the application of different text mining techniques on unstructured data but do not specifically target the datasets in social networking websites. Moreover, the existing research papers cover the text mining techniques without mentioning the pre-processing phase (Yin *et al.*, 2007; Xu *et al.*, 2008) that is an important phase for the simplification of text mining process. In contrast, this survey attempts to

address all the above-mentioned deficiencies by providing a focused study on the application of all (classification and clustering) text mining techniques in social networks where data is unstructured.

The rest of the survey is organized as follows. Section 2 presents different pre-processing techniques. Section 3 describes and different classification-based algorithms for text mining in social networks. In Section 4, the clustering techniques used for text mining are described. Section 5 presents current challenges and future directions. Finally, Section 6 concludes this survey.

2 Pre-processing in text mining

During the text gathering process, the text may be loosely organized and can be interpreted as irrational text integration or missing information. If the text has not been scanned carefully to identify the problems (as reported in Section 1), then text mining might lead to the 'garbage in garbage out' phenomena (Dai *et al.*, 2011). Unstructured text may lead to poor text analysis that affects the accuracy of an output (Forman & Kirshenbaum, 2008). The pre-processing phase organizes documents into a fixed number of pre-defined categories. Pre-processing guarantees successful implementation of text analysis, but may consume considerable processing time (Forman & Kirshenbaum, 2008). There are two basic methods of text pre-processing: (a) feature extraction (FE) and (b) feature selection (FS), which are detailed in the subsequent sections.

2.1 Feature extraction

The process of FE can be further categorized as: (a) morphological analysis (MA), (b) syntactical analysis (SA), and (c) semantic analysis. MA deals with individual words represented in a text document and mainly consists of tokenization, remove-stop-word, and stemming-word (Forman & Kirshenbaum, 2008). In tokenization the document is treated as a sequence of word strings and splits word by removing punctuations (Negi *et al.*, 2010). In remove-stop-word phase, stop words, such as 'the', 'a', and 'or' are removed. Remove-stop-word phase improves the effectiveness and efficiency of text processing because the number of words in the document are reduced (Shekar & Shoba, 2009). Stemming-word is the linguistic normalization technique generally used to reduce a word to the root form, such as the word 'honesty' can be reduced to root form of 'honest' or the word 'walking' can be reduced to the root form of 'walk'. Different stemming algorithms are available in the literature, such as brute-force, suffix-stripping, affix-removal, successor variety, and n-gram (Forman & Kirshenbaum, 2008; Shekar & Shoba, 2009).

To interpret a logical meaning from a sentence, a grammatically correct sentence is required (Yuan, 2010). SA provides knowledge about the grammatical structure of a language that is often termed as syntax. For instance, the English language comprises of noun, verb, adverb, punctuation, and other parts of speech. The SA technique comprises of: (a) part-of-speech tagging (POS tagging) and (b) parsing.

The POS tagging process is commonly used to add contextually related grammatical knowledge of a single word in a sentence. If the lexical class of the word is known, then performing linguistic analysis becomes much easier (Yoshida *et al.*, 2007). Various approaches are mentioned in the scientific literature for implementing POS tagging based on dictionaries (Yuan, 2010). The most promising approaches used are rule-based MA and stochastic model, such as Hidden Markov Model (HMM). In a rule-based approach, the text is decomposed into tokens that can be further used for analysis. Moreover, HMM is a stochastic tagging technique mainly used to discover the most similar POS tagging from sequence of input tokens (Yuan, 2010). Parsing is a technique used for examining the grammatical structure of a sentence. The sentence is represented in a tree-like structure, termed as parse tree, that is mainly used for analysis of correct grammatical order of a sentence. A parse tree can be constructed by using a top-down or bottom-up approach (Ling *et al.*, 2006).

To fulfil the needs of a distributed knowledge society, available natural communication tools must understand the meaning of a sentence (Strapparava & Ozbal, 2010). Keyword spotting technique is used to determine the useful contents from the textual message (Ling *et al.*, 2006). The keyword spotting technique is completely based on the WordNet-Affect, which is a semantic lexicon commonly used for the categorization of words that express similar emotions (Strapparava & Ozbal, 2010). Another example is

SentiWordNet that generally uses WordNet synonyms for measuring the emotions on the basis of two scales, such as positive emotions (happiness) and negative emotions (hate) (Esuli & Sibastiani, 2006). A state-of-the-art comparison between keyword spotting and semantic analysis has been presented in Ling *et al.* (2006). Ling analyzed the sentence syntactically and identified the basic emotions by analyzing words with respect to context and structure patterns (Ling *et al.*, 2006).

The keyword spotting technique is based on keywords specifically used for the description of certain emotions in the text (Wollmer *et al.*, 2009). For instance, in English language verb, noun, and adjective can be used as the keywords for emotion detection. However, the basic disadvantage of keyword spotting technique is the dependency on the presence of obvious affective words in the text. For instance, the emotion 'sadness' cannot be derived from the sentence 'I lost my money', as the sentence does not specifically mention the word 'sad'.

To overcome the limitations of keyword spotting and to achieve true understanding the authors Ling *et al.* (2006) introduced a new paradigm named semantic networks. Semantic networks are used to represent concepts, events, and relationships between them. The authors Ling *et al.* (2006) concluded the paper by declaring better performance in detecting human emotions using semantic networks versus keyword spotting because the semantic networks do not depend on detecting emotions based on keywords. In semantic networks the emotions are detected based on the contextual information. The authors Ling *et al.* (2006) and Li and Khan (2009a, 2009b) explained the method specifically but did not mention any results associated with the experiments. Moreover, a very large database is required, such as a combination of WordNet-Affect and SentiWordNet for increasing the accuracy of the results.

2.2 *Feature selection*

The basic purpose of FS is to eliminate irrelevant and redundant information from the target text. FS selects important features by scoring the words. The importance of the word in the document is represented by the assigned score (Hua *et al.*, 2009).

The text document is represented as a vector space model. In a vector space model, each dimension represents a separate term as a single word, keyword, or a phrase. Document matrix can be represented with n documents and m terms where any non-zero entry in the matrix indicates the presence of a term in the document (Hua *et al.*, 2009). Feature vectors represents document feature. Two basic methods have been proposed to calculate feature vectors: (a) term frequency (TF) and (b) inverse document frequency (IDF) (Shekar & Shoba, 2009).

TF determines how often a term is found in a collection of documents. The information about the topic of the document can be identified by the number of occurrences of a term associated with the topic. IDF considers the least frequent words in the document that have information about the topic. Whereas, TFIDF technique is the combination of term TF and IDF and is mainly used for calculating the frequency and relevancy of a given word in a document (Yoshida *et al.*, 2007).

Ma *et al.* (2005) and Li *et al.* (2011b) used similarity measuring techniques for text pre-processing. Similarity measuring is a technique used to measure the affinity of any group of words or phrases to occur together frequently. Similarity measuring can be further categorized on the basis of grammatical construct, such as 'because of' and semantic relations, such as 'student' and 'teacher'. The terms have a high probability to occur in close proximity and exhibit affinity for each other. In Ma *et al.* (2005), emotional weights are assigned to each word to determine specific emotions. However, similarity measure technique performs poorly when handling complicated sentence structures. For instance, the sentence 'This was not a failure for our company' most likely represents success. However, the word 'failure' has 75% probability of expressing negative emotion. For perfect implementation of the similarity measuring techniques, the user needs to have a sufficient corpus. Moreover, scaling an algorithm to such large data sets, particularly available on the Web, still needs to be addressed (Zhao *et al.*, 2009).

In scientific literature (Yoshida *et al.*, 2007), different FS techniques such as (a) latent semantic indexing (LSI) and (b) random mapping (RM) are discussed. LSI tends to improve the lexical matching by adopting a semantic approach, as in the case of semantic analysis, while RM creates a map through the

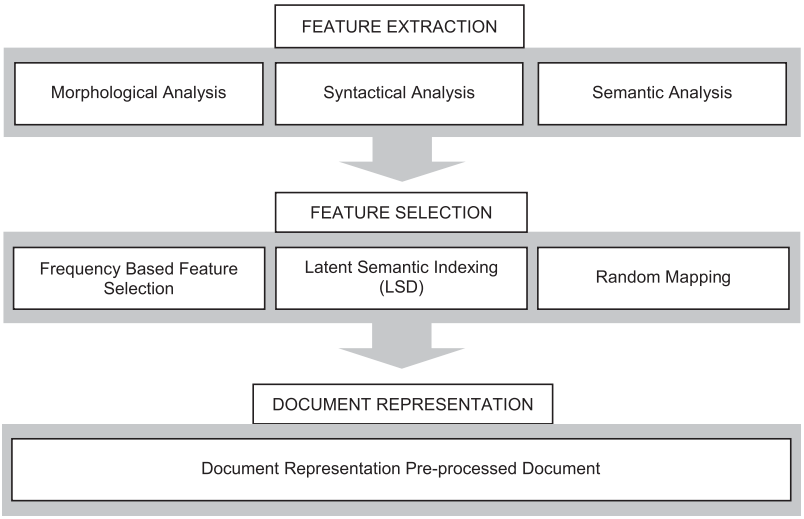

Figure 1 Pre-processing

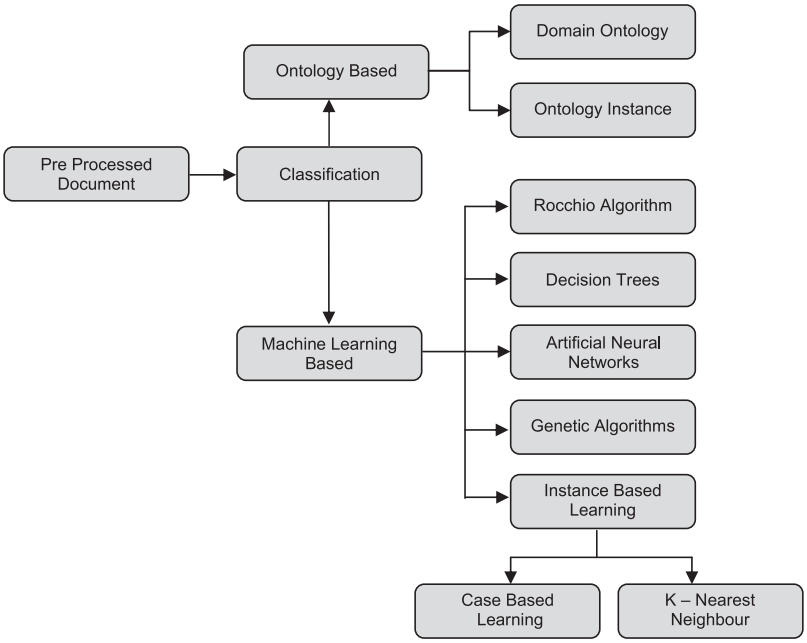

Figure 2 Text mining using classification

contents of a large document set. Any selected region in a map can further be used for the extraction of new documents on similar topics. A pre-processed document can be represented as in Figure 1. Durga and Govardhan (2011) present the two most commonly used text mining techniques for text analysis in social networking: (a) text mining using classification (supervised) and (b) text mining using clustering (unsupervised).

3 Text mining using classification

Supervised learning or classification is the process of learning a set of rules from a set of examples in a training set. Text classification is a mining method that classifies each text to a certain category (Yin *et al.*, 2007). Classification can be further divided into two categories: (a) machine learning-based text classification (MLTC) and (b) ontology-based text classification (Xu *et al.*, 2008) and is illustrated in Figure 2.

3.1 Machine learning-based text classification

MLTC comprises of quantitative approaches to automate NLP that uses machine learning algorithms. Preferred supervised learning techniques for text classification are described in the subsequent text.

3.1.1 Rocchio algorithm

Different words with similar meanings in a natural language are termed as Synonymy. Synonymy can be addressed by refining the query or document using the relevance feedback method. In the relevance feedback method, the user provides feedback that indicates relevant material regarding the specific domain area. The user asks a simple query and the system generates initial results in response to the query. The user marks the retrieved results as either relevant or irrelevant. Based on the user-marked results the algorithm may perform better. The relevance feedback method is an iterative process and plays a vital role by providing relevant material that tracks user information needs (Liu & Lu, 2011).

Rocchio algorithm is an implementation of the relevance feedback method and is mainly used for document refinement. However, the relevance feedback algorithms have drawbacks as illustrated in Liu and Lu (2011). The user must have sufficient knowledge to indicate relevance feedback (Luger, 2008). Moreover, the relevance feedback algorithm may not work efficiently when the user spells a word in a different way. Various spelling correction techniques can be used at the cost of computation and response time, such as hashing-based and context-sensitive spelling correction techniques (Udupa & Kumar, 2010).

3.1.2 Instance-based learning algorithm

Instance-based learning algorithms (also known as lazy algorithms) are based on the comparison between new problem instances and instances already stored during training (Chang & Poon, 2009). On arrival of a new instance, sets of related instances are retrieved from the memory and further processed so the new instance can be classified accordingly. Algorithms exhibiting instance-based learning approaches are described in the subsequent text.

K-nearest neighbour algorithm is a form of instant-based learning. The algorithm categorizes similar objects based on the closest feature space in the training set. The closest feature space may be determined by measuring the angle between the two feature vectors or by calculating the Euclidean distance between the vectors. For more details, we encourage the readers to browse Chang and Poon (2009).

Case-based reasoning comprises of three basic steps: (1) classification of a new case by retrieving appropriate cases from data sets, (2) modification of the extracted case, and (3) transformation of an existing case (Forman & Kirshenbaum, 2008). Textual case-based reasoning (TCBR) primarily deals with textual knowledge sources in making decisions. A novel TCBR system, named SOPHIA-TCBR has been detailed in Patterson *et al.* (2008) for organizing semantically related textual data into a group. Patterson *et al.* (2008) stated better results of knowledge discovery in the SOPHIA-TCBR system. However, in the TCBR approach, extracting similar cases and representing knowledge without losing key concepts with low knowledge engineering overhead are still challenging issues for researchers (Patterson *et al.*, 2008).

3.1.3 Decision trees and Support Vector Machine

Relationships, attributes, and classes in ontology can be structured hierarchically as taxonomies (Forman & Kirshenbaum, 2008). The process of constructing lexical ontology by analyzing unstructured text is termed as ontology refinement. DT is a method to semantically describe the concepts and the similarities between the concepts (Forman & Kirshenbaum, 2008). Different algorithms of DT are used for classification in many application areas, such as financial analysis, astronomy, molecular biology, and text mining. As text classification depends on a large number of relevant features, an insufficient number of relevant features in a DT may lead to poor performance in text classification (Forman & Kirshenbaum, 2008).

Support Vector Machine (SVM) algorithm is used to analyze data in classification analysis. In contrast to other classification methods, SVM algorithm uses both negative and positive training data sets to construct a hyper plane that separates the positive and negative data. The document that is closest to decision surface is called support vector.

3.1.4 Artificial neural networks

Artificial neural networks (ANN) are parallel distributed processing systems specifically inspired by the biological neural systems (Jo, 2010).The network comprises of a large number of highly interconnected processing elements (neurons) working together to solve any specific problem. Owing to their tremendous ability to extract meaningful information from a huge set of data, neurons have been configured for specific application areas, such as pattern recognition, FE, and noise reduction. In the neural network, connection between two neurons determines the influence of one neuron on another, while the weight on the connection determines the strength of the influence between the two neurons (Jo, 2010).

There are two basic categories of learning methods used in neural networks: (a) supervised learning and (b) unsupervised learning. In supervised learning, the ANN gets trained with the help of a set of inputs and required output patterns provided by an external expert or an intelligent system. Different types of supervised learning ANNs include: (a) back propagation and (b) modified back propagation neural networks (Luger, 2008). Major application areas of supervised learning are pattern recognition and text classification (Jo, 2010; Kolodziej *et al.*, 2012). In unsupervised learning (clustering), the neural network tends to perform clustering by adjusting the weights based on similar inputs and distributing the task among interconnected processing elements (Luger, 2008).

The field of text mining is gaining popularity among researchers because of enormous amount of text available via Web in the form of blogs, comments, communities, digital libraries, and chat rooms. ANN can be used for the logical management of text available on Web. Jo proposed a new neural network architecture for text categorization with document presentation called Neural Text Categorizer (NTC) (Jo, 2010). NTC comprises of three layers: (a) input layer, (b) output layer, and (c) learning layer. Input layer is directly connected with output layer, whereas learning layers determine the weights between input and output layer. The proposed approach can also be used for organizing the text in social networks (Jo, 2010).

3.1.5 Genetic algorithms

A genetic algorithm (GA) is a heuristic search that simulates the natural environment of biological and genetic evolution (Luger, 2008; Kolodziej *et al.*, 2011). Multiple solutions of a problem are presented in the form of a genome. The algorithm creates multiple solutions and applies genetic operators to determine the best offspring. GAs are widely used to solve optimization problems. Therefore, researchers are trying to use the utility of GAs in social networking websites (Luger, 2008; Guzek *et al.*, 2010).

A GA was used for FS and termed weight method in Khalessizadeh *et al.* (2006) for assigning weights to each concept in the document on the basis of relevant topics. Weighted topic standard deviation was the proposed formula used to present the concentration of a topic in a document as a fitness function. As the process is recursive, an end function needs to be specified based on monitoring the improvement of results in the consecutive generations. In Khalessizadeh *et al.* (2006), the authors revealed better results by using a GA for text classification.

3.2 Ontology-based text classification

Statistical techniques for document representation (as described in Section 3.1) are not sufficient because the statistical approach neglects the semantic relations between words (Luger, 2008). Consequently, the learning algorithm cannot identify the conceptual patterns in the text (Luger, 2008). Ontology can be the solution of the problems by introducing explicit specification of conceptualization based on concepts, descriptions, and the semantic relationships between the concepts (Zhao *et al.*, 2009; Li *et al.*, 2012a). Ontology represents semantics of information and is categorized as: (a) domain ontology consists of concepts and relationship of the concepts about a particular domain area, such as biological ontology or industrial ontology and (b) ontology instance related with automatic generation of web pages (Luger, 2008).

Basic components of ontology include (a) classes, (b) attributes, (c) relations, (d) function terms, and (e) rules (Wimalasuriya & Dou, 2010). Ontology needs to be specified formally (Luger, 2008). Formal relation can be represented as (a) classes and (b) instances (Zhao *et al.*, 2009). Ontology-based languages

Table 1 Comparison of hybrid approaches

Authors	Hybrid approaches						Success rate (%)
	ANN	RA	DT	SVM	K-NN	GA	
Miao *et al.* (2009)	No	Yes	No	No	Yes	No	83.8
Wu (2009)	Yes	No	Yes	No	No	No	93.67
Aci *et al.* (2010)	No	No	No	No	Yes	Yes	75.52
Gazzah and Ammara (2008)	Yes	No	No	Yes	No	No	91.5
Meesad *et al.* (2011)	No	No	Yes	Yes	No	No	92.20
Mitra *et al.* (2005)	Yes	No	No	Yes	No	No	99.66
Lee *et al.* (2010)	No	No	No	Yes	Yes	No	97
Remeikis *et al.* (2005)	Yes	No	Yes	No	No	No	90.9

ANN, artificial neural networks; RA, Rocchio algorithm; DT, decision tree; SVM, Support Vector Machine; K-NN, k-nearest neighbour; GA, genetic algorithm.

are declarative languages and generally express the logic of computation based on either first-order logic or description logic. For instance, the W3C organization introduced standardized Ontology Web Language that supports interpretability of language by providing additional vocabulary with formal semantics (Xu *et al.*, 2008). Common Logic and Semantic Application Design Language (Wimalasuriya & Dou, 2010) are the popular ontology-based languages commonly used for semantic evaluation of data sets available in social networking websites.

Online information usually resides in digital libraries in the form of online books, conference, and journal papers. In digital libraries, searching techniques are based on a traditional keyword matching approach that may not satisfy requirements of users owing to lack of semantic reasoning capabilities. Xu recommended an ontology-based digital library system that analyzed the query with respect to semantic meanings and revealed better results when compared with traditional keyword-based searching approach (Xu *et al.*, 2008). However, semantic analysis is computationally expensive and challenging for researchers, especially for large text corpora such as text data in social networking websites (Xu *et al.*, 2008).

3.3 Hybrid approach

Different classification algorithms have been used for text classification and analysis. However, literature (Miao *et al.*, 2009; Aci *et al.*, 2010; Li *et al.*, 2011b; Meesad *et al.*, 2011) shows that the combination of different classification algorithms (hybrid approach) provides better results and increased text categorization performance instead of applying a single pure method. The result of applying hybrid approach to large text corpora heavily depends on the test data sets. Therefore, there is no guarantee that a high level of accuracy acquired by one test set will also be obtained in another test set. Moreover, for better performance of the hybrid approach, several parameters need to be defined or initialized in advance. Table 1 provides an overview of different hybrid approaches used for text classification that can be further used for the text analysis in social networking. However, selecting the classification approach for text analysis in social networks totally depends on the data set and nature of the problem being investigated (Miao *et al.*, 2009).

The result of the analysis shows that SVM and ANN performed well in several comparisons. The main purpose of the comparison of hybrid approach is to highlight the applicability of different classification algorithms and complement their limitations (Aci *et al.*, 2010).

4 Text mining using clustering

Document clustering includes specific techniques and algorithms based on unsupervised document management (Jain, 2010). In clustering the numbers, properties, and memberships of the classes are not

known in advance. Documents can be grouped together based on a specific category, such as medical, financial, and legal.

In scientific literature (Sathiyakumari & Manimekalai, 2011), different clustering techniques are comprised of different strategies for identifying similar groups in the data. The clustering techniques can be divided into three broad categories: (a) hierarchical clustering, (b) partitional clustering, and (c) semantic-based clustering that are detailed in the subsequent text.

4.1 Hierarchical clustering

Hierarchical clustering organizes the group of documents into a tree-like structure (dendrogram) where parent/child relationships can be viewed as a topic/subtopic relationship. Hierarchical clustering can be performed either by using (a) agglomerative or (b) divisive methods, which are detailed in the subsequent text (Kavitha & Punithavalli, 2010).

An agglomerative method uses a bottom-up approach by successively combining closest pairs of clusters together until the entire objects form one large cluster (Kavitha & Punithavalli, 2010). The closest cluster can be determined by calculating the distance between the objects of n-dimensional space. Agglomerative algorithms are generally classified on the basis of inter-cluster similarity measurements. The most popular inter-cluster similarity measures are single-link, complete-link, and average-link (Sathiyakumari & Manimekalai, 2011). Several algorithms are proposed based on the above-mentioned approach, such as Slink, Clink, and Voortices use single-link, complete-link, and average-link, respectively. The Ward algorithm uses both the agglomerative as well as divisive approach as illustrated in Figure 3. The only difference between the aforementioned algorithms is the method of computing the similarity between the clusters.

In Yonghong and Wenyang (2010), the authors suggested agglomerative hierarchical clustering techniques for text clustering. First, GA was applied to achieve the FS phase in the text document. Second, similar

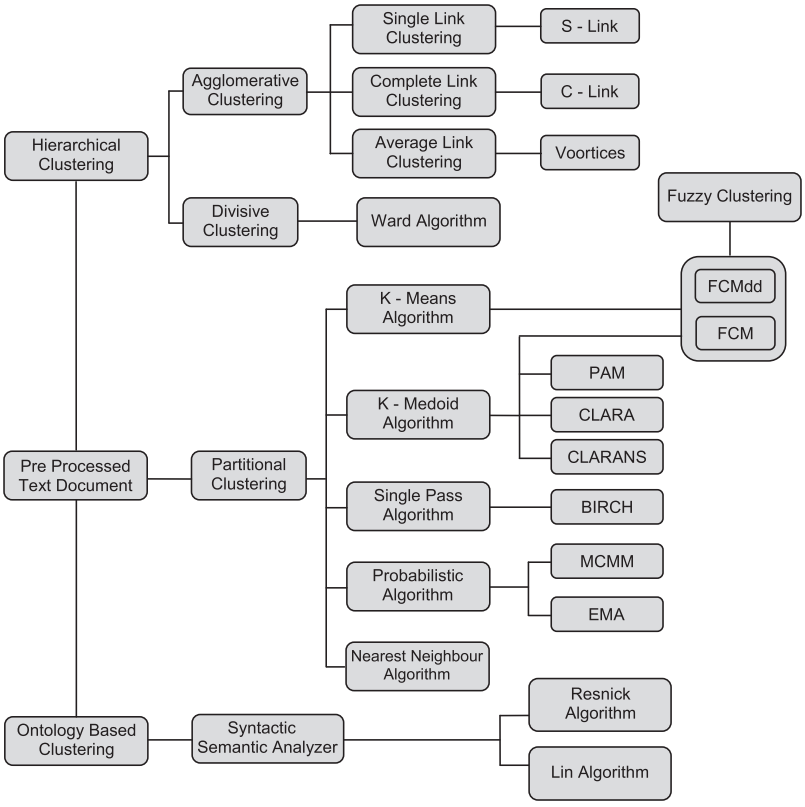

Figure 3 Text mining using clustering

document sets were grouped together into small clusters. Finally, the authors proposed text clustering algorithm to merge all clusters into final text cluster (Yonghong & Wenyang, 2010). The proposed approach can be used for grouping the similar text from social networking websites, such as blogs, communities, and social media.

The divisive method uses a top-down approach by starting with the same cluster and recursively splitting the cluster into smaller clusters until each document is in a classified cluster (Sathiyakumari & Manimekalai, 2011). The computations required by divisive clustering are more complex as compared with the agglomerative method. Therefore, the agglomerative approach is the more commonly used methodology.

Hierarchical clustering is very useful because of the structural hierarchical format. However, the approach may suffer from a poor performance adjustment once the merge or split operations are performed that generally leads to lower clustering accuracy (Sathiyakumari & Manimekalai, 2011). Moreover, the clustering approach is not reversible and the derived results can be influenced by noise.

4.2 Partitional clustering

Partitional clusters are also known as non-hierarchical clusters (Kavitha & Punithavalli, 2010). To determine the relationship between objects, partitional clustering uses a feature called vector matrix. Features of every object are compared and objects comprised of similar patterns are placed in a cluster (Liu & Lu, 2011). The partitional clustering can be further categorized as iterative partitional clustering, where the algorithm repeats itself until a member object of the cluster stabilizes and becomes constant throughout the iterations. However, the number of clusters should be defined in advance (Liu & Lu, 2011). Different forms of the iterative partitional cluster-based approaches are described as follows.

4.2.1 K-mean, k-medoid, c-mean, and c-medoid

In the *k*-mean approach the data set is divided into *k* clusters (Jain, 2010). Each cluster can be represented by the mean of points termed as the centroid. The algorithm performs in a two-step iterative process: (1) assign all the points to the nearest centroid and (2) calculate the centroids for a newly updated group (Jain, 2010). The iterative process continues until the cluster centroid becomes stabilized and remains constant (Liu & Lu, 2011).

The *k*-mean algorithm is widely used because of the straightforward parallelization (Jain, 2010). Moreover, *k*-mean algorithm is insensitive to data ordering and works conveniently only with numerical attributes. However, the optimum value of *k* needs to be defined in advance (Liu & Lu, 2011).

The *k*-medoid algorithm selects the object closest to the centre of the cluster to represent the cluster (Jain, 2010). In the algorithm, the *k* object is selected randomly. Based on the selected object, distance is computed. The nearest object with respect to *k* will form a cluster. Remaining objects take the place of *k* recursively until the quality of the cluster is improved (Liu & Lu, 2011). The *k*-medoid algorithm has many improved versions, such as PAM (Partitioning Around Medoid), CLARA (Clustering Large Applications), and CLARANS (Clustering Large Applications Based Upon Randomized Search). *K*-medoid algorithms work well for small data sets, but give compromised results for large data sets (Liu & Lu, 2011).

C-mean is a variation of *k*-mean that exhibits a fuzzy clustering concept that generates a given number of clusters with fuzzy boundaries and allows overlapping of clusters (Chen & Wang, 2009). In overlapping clusters process, the boundaries of clusters are not clearly specified. Therefore, each object belongs to more than one cluster. Fuzzy *c*-mean (Chen & Wang, 2009), and fuzzy *c*-medoids (Hang *et al.*, 2008) algorithms are widely used examples of *c*-mean algorithm (Li *et al.*, 2012), as illustrated in Figure 3.

4.2.2 Single-pass algorithm

The single-pass algorithm is the simplest form of partitional clustering (Mehmed, 2011). The algorithm starts with empty clusters and randomly selects a document as a new cluster with only one member (Mehmed, 2011). Single-pass algorithm calculates a similarity coefficient by considering a second object. If the calculated similarity coefficient is greater than the specified threshold value, then the object will be

added to the existing cluster, otherwise a new cluster will be created for the object. The BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) algorithm is an example of the single-pass clustering algorithm (Sathiyakumari & Manimekalai, 2011). The algorithm uses hierarchical data structure called CF (Clustering Feature) tree for partitioning the data sets (Mehmed, 2011). Nearest neighbour clustering is iterative and similar to the hierarchical single-link method (Mehmed, 2011).

4.2.3 Probabilistic algorithm

Probabilistic clustering is an iterative method that calculates and assigns probabilities for the membership of an object (Sathiyakumari & Manimekalai, 2011). Based on the probability measurements, an object can be a part of any specific cluster. Probabilistic clustering technique is popular because of the ability to handle records of a complex structure in a flexible manner. As probabilistic clustering has clear probabilistic foundations, finding out the most suitable number of clusters becomes relatively easy (Liu & Lu, 2011). Examples of probabilistic clustering are the Exception Maximizing Algorithm and Multiple Cause Mixture Model. However, these approaches are computationally expensive (Sathiyakumari & Manimekalai, 2011).

4.3 Semantic-based clustering

Meaningful sentences are composed of logical connections to meaningful words (Liu & Lu, 2011). A logical construction of words is generally provided by machine readable dictionaries, such as WordNet. In semantic-based clustering, the structured patterns are extracted from an unstructured natural language. Moreover, the approach emphasizes meaningful analysis of contents for information retrieval.

Researchers have proposed several algorithms for computing semantic similarities between text, such as Resnik and Lin algorithms (Liu & Lu, 2011) are proposed to measure the semantic similarity of text in a specific taxonomy. Detailed descriptions of these algorithms are presented in Chen and Wang (2009).

Yu and Hsu (2011) introduced a novel approach to automate the ontology construction process based on data clustering and pattern tree mining. The study comprises of two phases: (1) document clustering phase creates a group of related documents using *k*-mean clustering technique and (2) ontology construction phase creates inter-concept relation from the clustered documents, whereas inter-concept relation is termed as similar concept relationship. The author implemented the proposed approach on weather news collected form e-paper and revealed remarkable results by extracting the regions with high temperature.

5 Current challenges and future directions

Implementing text mining techniques in social networking have several challenges for researchers.

Text in social networks: in social networks, textual data may be large, noisy, and dynamic. Moreover, interpreting emoticons (smile, sad) for expressing any specific concept or emotion is still a challenging issue for researchers. Privacy and trust in online communication is also a major issue. Application of ethical values, such as integrity, veracity, in online communication is the only effective way to build trust online.

Text mining using cloud computing: another challenge of the current era is to implement text mining techniques in cloud-based infrastructure that allow people to access technology-enabled and scalable services via Internet (Yoo, 2012). However, in cloud computing, user may have difficulty in the process of storing and retrieving the document (Yoo, 2012). Automatic document archiving can be performed using the text mining techniques. Moreover, text processing and text aggregation in cloud would be the issues for the researchers.

To overcome the challenges, researchers need to apply different text mining techniques in social networks that can filter out relevant information from the large text corpora. However, determining whether to use clustering or classification approach for text analysis in social networks is still a challenging task that totally depends on the data set and the nature of the problem being investigated. In future, text mining tools can also be used as intelligent agent that can mine user's personal profiles from social networks and forward relevant information to the users without requiring an explicit request.

6 Concluding remarks

Electronic textual documents are extensively available owing to the emergence of the Web. Many technologies are developed for the extraction of information from huge collections of textual data using different text mining techniques. However, information extraction becomes more challenging when the textual information is not structured according to the grammatical convention. People do not care about the spellings and accurate grammatical construction of a sentence while communicating with each other using different social networking websites (Facebook, LinkedIn, MySpace). Extracting logical patterns with accurate information from such unstructured form is a critical task to perform.

This survey attempts to provide a thorough understanding of different text mining techniques as well as the application of these techniques in the social networking websites. The survey investigates the recent advancement in the field of text analysis and provides a comprehensive overview of all the exiting text mining techniques that can be used for the extraction of logical patterns from the unstructured and grammatically incorrect textual data. This survey will definitely provide new ways for researchers to proceed and develop novel classification or clustering techniques that will be useful for analysis of text in social networks.

Acknowledgements

We are grateful to Juan Li, Matthew Warner, and Daniel Grages for their feedback on draft of this survey report. Samee U. Khan's work was partly supported by the Young International Scientist Fellowship of the Chinese Academy of Sciences, (Grant No. 2011Y2GA01).

References

Aci, M., Inan, C. & Avci, M. 2010. A hybrid classification method of k-nearest neighbour, Bayesian method and genetic algorithm. *Expert Systems with Applications* **37**(7), 5061–5067.

Aggarwal, C. 2011. Text mining in social networks. In *Social Network Data Analytics*, Charu, A. C. (ed.), 2nd edition. Springer, 353–374.

Baatarjav, E., Phithakkitnukoon, S. & Dantu, R. 2008. *Group Recommendation System for Facebook*, 2nd edition. Springer.

Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like metamucil: fostering critical and creative thinking about metaphor in political blogs. In *Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010)*. ACM, 34–45.

Brucher, H., Knolmayer, G. & Mittermayer, M. 2002. Document classification methods for organizing explicit knowledge. In *Proceedings of 3rd European Conference on Organizational Knowledge, Learning and Capabilities*, 1–25.

Chang, M. & Poon, C. K. 2009. Using phrases as features in e-mail classification. *Journal of System and Software* **82**(6), 1036–1945.

Chen, W. & Wang, M. 2009. A fuzzy c-means clustering-based fragile watermarking scheme for image authentication. *Expert Systems with Applications* **36**(2), 1300–1307.

Dai, Y., Kakkonen, T. & Sutinen, E. 2011. MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis method. *International Journal of Computer Information System and Industrial Management Applications* **3**, 165–173.

Durga, A. K. & Govardhan, A. 2011. Ontology based text categorization-telugu document. *International Journal of Scientific and Engineering Research* **2**(9), 1–4.

Esuli, A. & Sibastiani, F. 2006. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 417–422.

Evans, B. M., Kairam, S. & Pirolli, P. 2010. Do your friends make you smarter: an analysis of social strategies in online information seeking. *Information Processing and Management* **46**(6), 679–692.

Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In *Proceedings of 17th ACM Conference on Information and Knowledge Management*, 26–30.

Gazzah, S. & Ammara, N. B. 2008. Neural network and support vector machines classifiers for writer identification using Arabic script. *International Arab Journal of Information Technology* **5**(1), 92–101.

Guzek, M., Pecero, J. E., Dorronsoro, B., Bouvry, P. & Khan, S. U. 2010. A cellular genetic algorithm for scheduling applications and energy-aware communication optimization. In *Proceedings of PACM/IEEE/IFIP International Conference on High Performance Computing and Simulation (HPCS)*, 241–248.

Hang, N., Honda, K., Ichihashi, H. & Notsu, A. 2008. Linear fuzzy clustering of relational databased on extended fuzzy c-medoids. In *Proceedings of IEEE International Conference on Fuzzy Systems*, 366–371.

Hua, J., Tembe, W. D., Dougherty, E. R. & Edward, R. D. 2009. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition* **42**(3), 409–424.

Jain, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition* **31**(8), 651–666.

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

Kano, Y., Baumgartner, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. & Tsujii, T. 2009. Data mining: concept and techniques. *Oxford Journal of Bioinformatics* **25**(15), 1997–1998.

Kavitha, V. & Punithavalli, M. 2010. Clustering time series data stream – a literature survey. *International Journal of Computer Science and Information Security* **8**(1), 289–294.

Khalessizadeh, S. M., Zaefarian, R., Nasseri, S. H. & Ardil, E. 2006. Genetic mining: using genetic algorithm for topic based on concept distribution. *Journal of Word Academy of Science, Engineering and Technology* **13**(2), 144–147.

Kolodziej, J., Burczynski, B. & Khan, S. U. 2012. *Advances in Intelligent Modelling and Simulation: Artificial Intelligence-Based Models and Techniques in Scalable Computing*, Springer-Verlag.

Kolodziej, J., Khan, S. U. & Xhafa, F. 2011. Genetic algorithms for energy-aware scheduling in computational grids. In *Proceedings of 6th IEEE International Conference on P2P, Parallel, Grid, Cloud, and Internet Computing (3PGCIC)*, 17–24.

Lee, L. H., Wan, C. H., Yong, T. F. & Kok, H. M. 2010. A review of nearest neighbour-support vector machine hybrid classification model. *Journal of Applied Science* **10**(17), 1841–1858.

Li, J. & Khan, S. U. 2009a. MobiSN: semantics-based mobile ad hoc social network framework. In *Proceedings of IEEE Global Communications Conference (Globecom)*, Zomaya, A. Y. & Sarbazi-Azad, H. (eds). John Wiley & Sons, Hoboken, NJ, USA, 2013, ISBN: 978-0-470-93688-7.

Li, J. & Khan, S. U. 2009b. *On How to Construct a Social Network from a Mobile Ad Hoc Network*. Technical report, NDSU-CS-TR-09-009, North Dakota State University.

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

Li, J., Li, Q., Khan, S. U. & Ghani, N. 2011b. Community-based cloud for emergency management. In *Proceedings of the 6th IEEE International Conference on System of Systems Engineering (SoSE)*, 55–60.

Li, J., Wang, H. & Khan, S. U. 2012. A fully distributed scheme for discovery of semantic relationships. *IEEE Transactions on Services Computing* **6**(4), 257–469.

Ling, H. S., Bali, R. & Salam, R. 2006. Emotion detection using keywords spotting and semantic network. In *Proceedings of International Conference on Computing and Informatics IEEE (ICOCI)*, 1–5.

Liu, F. & Lu, X. 2011. Survey on text clustering algorithm. In *Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS)*, 901–904.

Luger, G. F. 2008. *Artificial Intelligence: Structure and Strategies for Complex Problem Solving*, 6th edition. Addison Wesley.

Ma, C., Helmut, P. & Mitsuru, l. 2005. *Emotion Estimation and Reasoning Based on Affective Textual Interaction*, 3rd edition. Springer.

Meesad, P., Boonrawd, P. & Nuipian, V. 2011. A chi-square-test for word importance differentiation in text classification. In *Proceedings of International Conference on Information and Electronics Engineering*, 110–114.

Mehmed, K. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd edition. John Wiley & Sons.

Miao, D., Duan, Q., Zhang, H. & Jiao, N. 2009. Rough set based hybrid algorithm for text classification. *Journal of Expert Systems with Applications* **36**(5), 9168–9174.

Mitra, V., Wang, C. & Banerjee, S. 2005. A neuro-SVM model for text classification using latent semantic indexing. In *Proceedings of International Joint Conference on Neural Networks*, 564–569.

Negi, P. S., Rauthan, M. M. S. & Dhami, H. S. 2010. Language model for information retrieval. *International Journal of Computer Applications* **12**(7), 13–17.

Patterson, D., Rooney, N., Galushka, M., Dobrynin, V. & Smirnova, E. 2008. SOPHIA-TCBR: a knowledge discovery framework for textual case-based reasoning. *Knowledge-Based Systems* **21**(5), 404–414.

Remeikis, N., Skucas, I. & Melninkaite, V. 2005. Hybrid machine learning approach for text categorization. *International Journal of Computational Intelligence* **1**(1), 63–67.

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

Sathiyakumari, K. & Manimekalai, G. 2011. A survey on various approaches in document clustering. *International Journal of Computer Technology and Application (IJCTA)* **2**(5), 1534–1539.

Shekar, C. B. H. & Shoba, G. 2009. Classification of documents using Kohonens self organizing map. *International Journal of Computer Theory and Engineering (IACSIT)* **1**(5), 610–613.

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

Strapparava, C. & Ozbal, G. 2010. The color of emotion in text. In *Proceedings of 2nd Workshop on Cognitive Aspects of the Lexicon*, 28–32.

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

Udupa, R. & Kumar, S. 2010. Hashing-based approaches to spelling correction of personal names. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 1256–1265.

Wimalasuriya, D. C. & Dou, D. 2010. Ontology-based information extraction: an introduction and a survey of current approach. *Journal of Information Science* **36**(5), 306–323.

Wollmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B. & Rigool, G. 2009. Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 3949–3952.

Wu, C. 2009. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications* **36**(3), 4321–4330.

Xu, X., Zhang, F. & Niu, Z. 2008. An ontology-based query system for digital libraries. In *Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, 222–226.

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

Yuan, L. 2010. Improvement for the automatic part-of-speech tagging based on Hidden Markov Model. In *Proceedings of 2nd International Conference on Signal Processing System IEEE (ICSPS)*, 744–747.

Yonghong, Y. & Wenyang, B. 2010. Text clustering based on term weights automatic partition. In *Proceedings of 2nd International Conference on Computer and Automation Engineering (ICCAE)*, 373–377.

Yoo, K. 2012. Automatic document archiving for cloud storage using text mining-based topic identification technique. In *Proceedings of International Conference on Information and Computer Application*, 189–192.

Yoshida, K., Tsuruoka, Y., Miyao, Y. & Tsujii, J. 2007. Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. In *Proceedings of 20th International Conference on Artificial Intelligence*, 1783–1788.

Yu, Y. & Hsu, C. 2011. A structured ontology construction by using data clustering and pattern tree mining. In *Proceedings of International Conference on Machine Learning and Cybernetics*, 45–49.

Zhao, P., Han, J. & Sun, Y. 2009. P-Rank: a comprehensive structural similarity measure over information networks. In *Proceedings of 18th ACM Conference on Information and Knowledge Management*, 233–238.