Text data mining: a proposed framework and future perspectives

Article *in* International Journal of Business Information Systems · February 2015

DOI: 10.1504/IJBIS.2015.067261

3 authors, including:

Sanaa Alwidian
University of Ottawa
7 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Hani Bani-Salameh
Hashemite University
25 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Text data mining: a proposed framework and future perspectives

Sana'a A. Alwidian*

Department of Computer Science and Applications,
Faculty of Prince Al-Hussein Bin Abdallah II for IT,
The Hashemite University,
Zarqa 13115, Jordan
Fax: 962(05) 3826625
Email: sanaaa@hu.edu.jo
*Corresponding author

Hani A. Bani-Salameh

Department of Software Engineering,
Faculty of Prince Al-Hussein Bin Abdallah II for IT,
The Hashemite University,
Zarqa 13115, Jordan
Fax: 962(05) 3826625
Email: hani@hu.edu.jo

Ala'a N. Alslaity

Department of Computer Science,
Faculty of Computer and Information Technology,
Jordan University of Science and Technology,
Irbid 22110, Jordan
Email: analslaity@just.edu.jo

Abstract: With the increased advancements in technology and the emergence of different kinds of applications, the amount of available data becomes enormous, and the large proliferation of such data becomes evident. Therefore, there is an essential need for some techniques or methods to interact with data and extract useful information and patterns from them. Text data mining (TDM) is the process of extracting desired information out of mountains of textual data that are inherently unstructured, without the need to read them all. In this paper, we shed the light on the-state-of-the-art in text mining as an interdisciplinary field of several related areas. To facilitate the understanding of text data mining, this paper proposes a framework that visualises this field in a step-wise manner, taking into consideration the semantic of the extracted text. In addition, this paper surveys a number of useful applications and proposes a new approach for spam detection based on the proposed TDM framework.

Keywords: text mining; clustering; categorisation; spam filtering; semantic; information retrieval; IR; text data mining; TDM; natural language processing; NLP; knowledge discovery from databases; KDD; knowledge discovery from text; KDT; semantic analysis.

Reference to this paper should be made as follows: Alwidian, S.A., Bani-Salameh, H.A. and Alslaity, A.N. (2015) 'Text data mining: a proposed framework and future perspectives', *Int. J. Business Information Systems*, Vol. 18, No. 2, pp.127–140.

Biographical notes: Sana'a A. Alwidian is a full time Lecturer at the Department of Computer Science and Applications (CSA) in the Prince Al-Hussein Bin Abdallah II Faculty for Information Technology at the Hashemite University. She received her BSc in Computer Information Systems from Jordan University of Science and Technology in 2007 and her MSc in Computer Science from Jordan University of Science and Technology in 2009. Her research interest includes: data mining, Arabic language processing, evolutionary computing, distributed systems and computer networks.

Hani A. Bani-Salameh is an Assistant Professor of Software Engineering, in the Prince Al-Hussein Bin Abdallah II for Information Technology College at the Hashemite University. His research interests include computer supported cooperative work (CSCW), software development environments, collaborative software development in virtual environments, and social networking. He studies social interactions in social networks and online environments, including Facebook and SourceForge.

Ala'a N. Alslaity is a Programmer at the Public Security Directorate/Jordan and a part time Instructor at the Jordan University of Science and Technology. He received his Bachelor degree in Computer Science and Computer Information Systems from the Jordan University of Science and Technology in 2006, and he completed his Master degree in Computer Science from Jordan University of Science and Technology in 2012. Through his study, he worked under the supervision of Dr. Ismail Ababneh and Dr. Muneer Bani Yassein in the area of mobile networks. He is interested in programming and computer science researches in the fields of wireless networks, distributed computing, and cloud computing. He aims to develop himself especially in education, teaching and research areas.

1 Introduction

Through out the current and the previous five decades, the scientific, technical, and scholarly literatures witnessed an unprecedented revolution (Aggarwal et. al., 2012). As a result, information space proliferates in a dramatic way to the level that it became very difficult, or even impossible, to retrieve this information and read it all at once. Therefore, the introduction of novel methods for information and knowledge extraction becomes a major promising theme in the field of information science.

The process of extracting information from mountains of data is not modern. It is as old as the information retrieval (IR) itself (Saracevic, 2001a). IR had been matured enough since the time of its introduction (in the 1970s) until recently. However, it is still vital since it contributes to nearly all fields that deal with information processing and analysis. The popularity of IR is supported by the widespread usage of internet and the birth of information linguistic known as natural language processing (NLP) (Xie et al., 2012).

Although the traditional IR systems perform a fundamental role in retrieving tremendous amount of data, it requires a massive effort from users to obtain useful

information from text data sources. Traditional IR often simultaneously retrieves both 'too little' information and 'too much' text (Adebola and Adeyemo, 2013). Therefore, non-traditional strategies that broaden the concept of IR are needed, such that the text data sources and databases are viewed more holistically, and the view that 'a truly useful system must go beyond simple retrieval' (Liddy, 2000) is supported. Text mining stands to be that promising, non-traditional strategy of text extraction.

The concept of text mining which is concerned with getting all the desired information out of 'mountains' of textual data is nearly as old as IR itself. However, text mining possesses essential features that distinguish it from IR as well as other related fields. Text mining aims at obtaining useful information from textual data that are inherently unstructured, unorganised, and erratic (Jusoh and Alfawareh, 2012).

The rest of the paper is organised as follows. Section 2 discusses text data mining (TDM), its definition, importance, purposes and tasks. Section 3 proposes a new framework that illustrates text data mining tasks. Section 4 proposes a new approach for e-mail spam detection based on text clustering. A list of useful applications of text mining is provided in Section 5. Section 6 discusses some open issues and future directions, and Section 7 concludes the paper.

2 Text data mining

This section defines TDM, its importance and its purposes. In addition, it presents the other related fields that lie under the TDM umbrella.

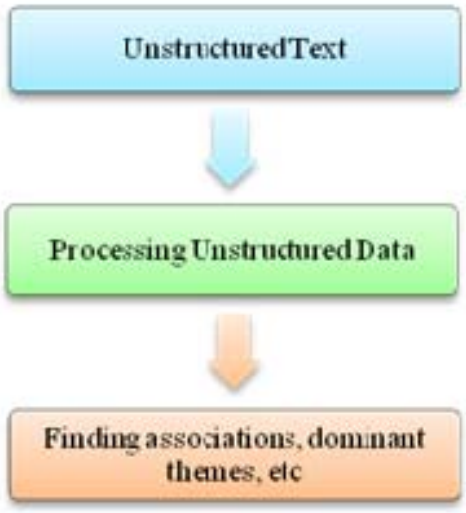
2.1 What is TDM?

TDM, or alternatively text mining or knowledge discovery from text (KDT) (Karanikas and Manchester, 2001), refers to the process of deriving a high quality, interesting, and non-trivial patterns, information or knowledge from *unstructured* text documents (Michael, 2007). TDM is different from data mining (Elmasri and Navathe, 1999) in the sense that TDM explores through textual data, which usually have unstructured nature, in an attempt to extract useful information, whereas data mining attempts to discover knowledge from the *structured* databases (that is why it is referred to as knowledge discovery from databases – KDD).

Hearst (1999), one of the pioneers in the field of text mining, provided a complete definition of TDM, wherein she made a clear distinction between TDM and the traditional IR. According to Hearst, traditional IR is concerned with the retrieval of documents that are relevant to a user's information needs (not the retrieval of the information itself), and then, selecting the desired information is left up to the user. On the other hand, TDM does not only deal with the direct retrieval of information from documents, but also it attempts to discover new patterns of information from documents, such information are useful, non-trivial and unknown previously (even by the author of the documents).

According to Havre et al. (2001), TDM is viewed as an iterative process (see Figure 1) of getting some unstructured text that is related to the domain of interest, processing this unstructured text and represent it in some medium format and finally applying specific activities on these formats such as finding associations, dominant themes, and so on.

Figure 1 The iterative process of TDM (see online version for colours)



Fan et al. (2005) discussed a generic model for TDM. In their proposed model, given a set of documents, TDM tools retrieve particular documents that are related to user's queries and pre-process them by checking character sets and formats. After that, the tools repeatedly go through text analysis phases until information are extracted. The extracted information then is placed in a management information system to produce an abundant amount of knowledge for the user of that system.

2.2 Why TDM?

The continuous advancement in technology causes a dramatic increase in information space, especially in the technical and scientific literatures. As a result, tracking information is a challenging task for scientists that have always complained about keeping up with their literature (Li et al., 2011). Therefore, TDM strategies are required to help scientists and researchers to obtain useful information from the large information spaces.

Recent statistical studies showed that 80%–85% of companies store the huge amount of their business information in the form of text (http://www.thearling.com/text/dmwhite/dmwhite.htm). Textual data share some common characteristics. They are fuzzy, unstructured, and contain ambiguous relationships between text documents (Berry, 2004). Therefore, TDM is believed to have a high potential to expose the concealed information by providing means of techniques and algorithms that are able to:

1 cope with the tremendous amount of text (that are expressed by natural language expressions)

2 handle the fuzziness and ambiguity.

Saracevic (2001b) proposed specific criteria that serve to be obvious illustrations of TDM goals and distinguish TDM from other IR concepts. He mentioned that TDM primary goals include:

1 working on natural language, large text collections

2 discovering new knowledge

3 applying principal algorithms more than heuristics and manual filtering

4 extracting phenomenological units of as patterns rather than, or in addition to documents.

2.3 TDM: the interdisciplinary field

In the literature, text mining problem proved to involve the tackling of several problems including: text representation, clustering, classification (or categorisation), text extraction, and the modelling of hidden information (Hotho et al., 2005). Therefore, it is obvious that TDM is not a standalone field per se, rather, it is an interdisciplinary field that lies in the intersection of several related fields including: IR, information extraction (IE), machine learning (ML), NLP, statistics and data mining (KDD) (Vishal and Gurpreet, 2009). This section discusses the above mentioned fields briefly.

IR, as discussed earlier, concerned with finding the documents that contain the related information and not finding the information themselves (Hearst, 1999). In a broader perspective, IR is considered to deal with information processing in its entire range from data retrieval to knowledge retrieval (Smith, 2002).

In IE, the main goal is to navigate through the huge amounts of text and to sieve useful information from them. Such information varies from a single word to a large passage. However, in IE, it must be known in advance what kind of information we are looking for, then the extracted documents are stored in a DB-like patterns for future use (Nahm and Mooney, 2003).

ML is an artificial intelligence (AI) area, wherein, the main concern is to develop machines (computers) that are able to learn by the analysis of data. Another concern of ML is the measurement and the analysis of the algorithm computational complexities (Sebastiani, 2002).

NLP, or alternatively known as linguistic analysis, refers to the deployment of techniques that facilitate the understanding and processing of natural languages through the use of computers (Jusoh and Alfawareh, 2009). NLP has several levels of analysis (Jurafsky and Martin, 2009): the phonological (speech) level, the morphological (word structure) level, the syntactic (grammar) level, the semantic (meaning of multiword structures, especially sentences) level, the pragmatic (sentence interpretation) level, the discourse (meaning of multi-sentence structures) level, and world (how general knowledge affects language usage) level.

Statistics is a branch of applied mathematics that deals with the analysis of empirical data. With statistics theory, uncertainty is modelled by probability theory. A good overview is provided in (Un Yong and Raymond, 2004).

Data mining or KDD is absolutely considered as the most important field from which TDM extracts its existence. According to Han and Kamber (2001), data mining is the process of finding interesting patterns in data that are not explicitly part of the data.

In his definition of KDD, Benoit (2002, p.265) defined data mining as: “multistage process of extracting previously unanticipated knowledge from large databases, and applying the results to decision making. Data mining tools detect patterns from the data and infer associations and rules from them. The extracted information may then be

applied to prediction or classification models by identifying relations within the data records or between databases. Those patterns and rules can then guide decision making and forecast the effects of those decisions". Data mining can be viewed as a multi-steps process that involves:

1 selection of data from the database and analysing them

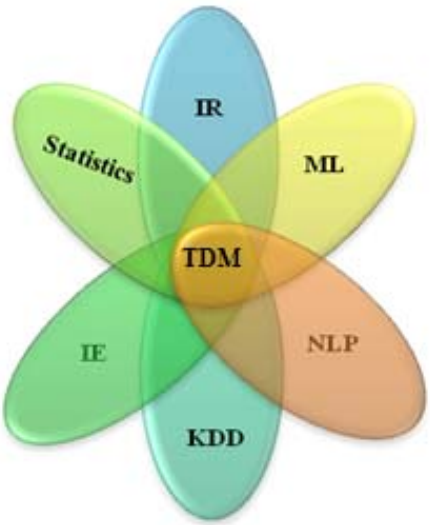
2 preparing the data, by apply data cleaning activities

3 applying the data mining algorithms

4 interpreting and evaluating the results.

The main difference between data mining (KDD) and TDM (KDT) is that the former applies its algorithms on structured data stored in the database, whereas the later applies its algorithms on the unstructured data (Benoit, 2002). Figure 2 provides our own representation of the above mentioned fields in relation with TDM.

Figure 2 TDM, the interdisciplinary field (see online version for colours)



2.4 TDM tasks

TDM involves a set of tasks including: text classification or categorisation, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarisation, and entity relation modelling (i.e., learning relations between named entities) (Hotho et al., 2005).

This paper focuses mainly on the first two tasks of TDM text classification and text clustering since these tasks will be deployed in our proposed framework (see Section 3) and our proposed spam detection approach (see Section 4).

Text classification refers to the process of assigning a particular document to one or more categories (class labels). Classification algorithms are known to be supervised learners.

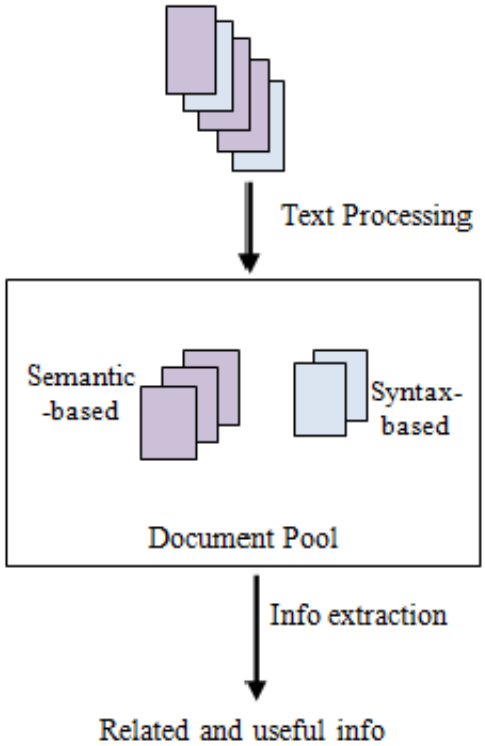
Since the class label is known in advance, and the mission is to assign the new data (with unknown class) to one (or more) class label (Aixin, 2012). In the classification task, knowledge about domain is very necessary to improve the efficiency of learning as well as to increase the quality of the learn model (Goncalves et al., 2012).

Text clustering is a more specific technique for unsupervised document organisation, in which documents are automatically grouped into meaningful categories (called clusters), wherein, the documents are similar within the same cluster and dissimilar between different clusters (Steinbach et al., 2003). The efficiency and quality of clustering is considered high if the documents that belong to the same cluster are more similar and between the other clusters more dissimilar.

3 Proposed framework

This section proposes a simple framework (see Figure 3) that facilitates the understanding of TDM activities, and visualises them as a two-steps process: text processing and IE.

Figure 3 The proposed TDM framework (see online version for colours)

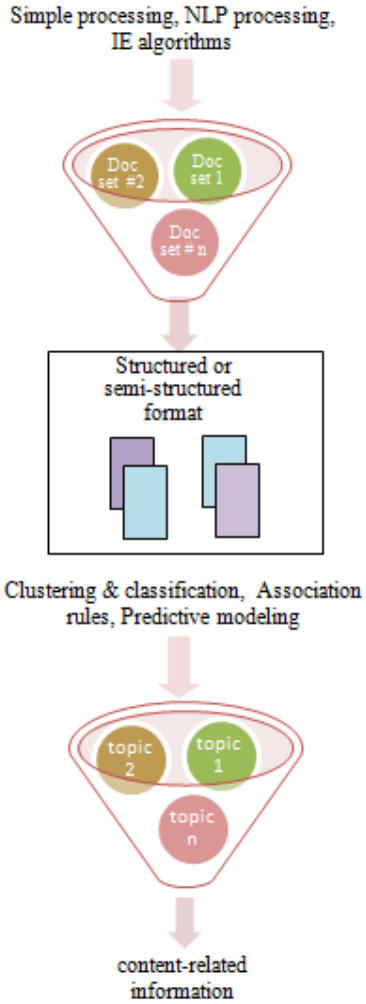


In the *Text processing* step, the text is converted from its natural, structure-free format into a structured (or semi-structured) intermediate format. The resulting format either represents a document of interest (syntax-based format) or a concept inside a document (semantic-based format). Then the set of resulting documents are stored in a large repository (for the purpose of further processing) called documents pool.

In the *Information extraction* step, a set of activities are applied to the documents stored in the document pool. For instance, in the case of the syntax-based format, clustering, classification and visualisation tasks are applied for the purpose of deducing relationships across documents, and in the case of semantic-based formats, predictive modelling and association rule mining are applied to deduce relationships across concepts within documents.

Further details of our proposed approach are illustrated in Figure 4. Applying our proposed framework on TDM process has an extra advantage over the traditional TDM frameworks. This advantage comes from taking the semantic features within text into consideration, therefore, the task of extracting and retrieving becomes more vital, objective, and meaningful. In addition, applying the proposed framework can assist in aligning the applications and future directions of TDM-based on the text processing and the IE tasks.

Figure 4 The details of the TDM framework (see online version for colours)



4 A proposed approach for spam detection

E-mails are one of the essential means of communication. They are prone to the threat of spam, which becomes an annoying problem for all types of users.

Many spam-filtering techniques were proposed to combat spammers based on the assumption that in any spam (junk) e-mail, there are a specific words/patterns that provide indications of spam (Ahmed, 2007). There are predefined barriers that distinguish a spam message from a legitimate one. The most commercially available filters use black-lists and hand-crafted rules (Gordon and David, 2006). However, these methods are proved to be inefficient for detecting all types of spam.

To alleviate this problem, content-based recognition techniques must be applied on e-mails to detect spam. TDM-based techniques, especially text classification-based techniques offer the possibility to be implemented as spam filters that can quickly adapt to new types of spam. A good example is the naïve Bayes classifiers which perform its job through learning. For further details about text mining-based classifiers, Michelakis et al. (2004) is a good reference.

This section proposes a new spam detection approach based on applying the *text clustering* task of TDM. The proposed approach for spam detection uses one of the popular text clustering algorithms called *K-means* (Tapas et al., 2002) algorithm to facilitate the grouping of e-mails into two clusters (or buckets). One cluster corresponds to the spam e-mails, and the other corresponds to the non-spam (ham) e-mails.

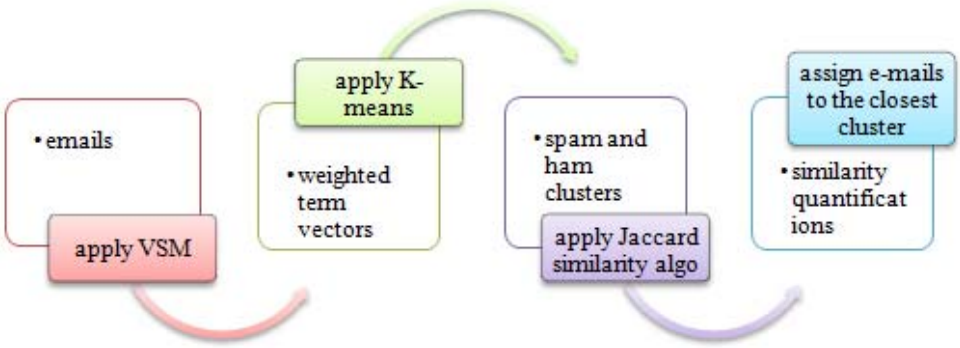
Since each e-mail contains a large number of words, then it becomes tedious to cluster the e-mail based on its terms. For this reason, the suggested approach uses *vector space model* (Pablo et al., 2007) to represent each e-mail document as a vector containing the weights of the most frequent terms. After the e-mails (represented as a weighted terms vectors) being clustered, a cluster representative, *centroid*, is selected for each cluster. The main goal of having a centroid as a cluster's representative is to facilitate the process of clustering. For instance, instead of comparing each new e-mail (which is represented as weighted- terms vectors) with all vectors in the cluster; the algorithm compares it only with the centroid, this will increase the performance of the spam classifier by decreasing the classifying time and the number of comparisons associated with the classifying process. For instance, if the cluster contains V vectors and one centroid, C, then the newly added vector will be compared only once with C instead of being compared V times with the entire vectors.

The similarity between each new e-mail and the centroid is then computed using one of the similarity measures; our proposed approach uses the *Jaccard* similarity measure (Marios and Divesh, 2010). Finally, the result of similarity will decide where to cluster the new e-mail, either to spam or to ham cluster. Figures 5 and 6 illustrate the steps of the proposed approach.

Figure 5 The steps of the spam detection proposed approach

1. Using vector space model, represent each document (e-mail document) as a weighted-term vector.
2. Using K-means algorithm, cluster the resulting vectors into two clusters, spam and ham.
3. For each cluster, select one vector as a centroid, to be the representative of this cluster.
4. Using Jaccard similarity measure, compute the similarity between the new e-mail vector and the centroids.
5. Assign the new e-mail to the cluster whose centroid is much closer than the other centroid.

Figure 6 The flow of the proposed spam detection approach (see online version for colours)



5 Applications

In the era of information revolution, where computer becomes the backbone for both the scientific and the economical sectors, TDM proved to assist these sectors efficiently and provide them with valuable contributions. This section briefly discusses three of the successful applications of TDM in different fields other than computer science field.

The high level of interest in biotechnology has made the area of biology as one of the most active application domains for text mining. Most of the text mining research efforts were devoted to assist the so called MEDLINE (Srinivasan, 2001).

MEDLINE is an application in Bioinformatics domain that obtains records that consist of a title, an abstract, a set of manually assigned metadata terms (known as mesh terms). A particular problem associated with MEDLINE is the huge and growing size of records, thus, it becomes almost impossible for someone to keep track of all the literature reviews in their domain. Therefore, TDM tools that navigate through the literature and help in discovering new relationships and suggesting hypothesis are greatly applied. Various text mining approaches such as co-occurrence based mining, IR-based metadata profiling (Andrea et al., 2005), and speculative sentence annotation have been proposed almost exclusively for MEDLINE data.

A major innovation of TDM in the biology domain is in the field of protein/gene analysis. Besides genes and proteins, researchers also wish to extract several other entities including organs, cells, methods and more broadly, biological pathways.

As the bioinformatics field has different aspects, TDM methods are used quit differently, each suit the particular features of the area they applied to. In Hong and Eugene (2003), ML approaches have been used to disambiguate gene and protein names and assign appropriate class labels to them. Hidden Markov model based approaches (Paaß and deVries, 2005) have also been used for the same problem. Mining of gene and protein functions are observed in Aaron and William (2005), functional relationships between genes, protein-protein interactions, and interactions between genes and gene products.

The sectors of *media* and publishing have witnessed a great revolution. As a consequence, a large number of news stories and information arrive to publishing houses each day. It is clear that users like to retrieve their desired information as easy and fast as possible. Moreover, users like to have these stories tagged with categories and the names of important persons, organisations and places.

Deutsche Presse-Agentur (DPA) was one of the pressing agents that made use of TDM techniques to automate the process of fetching and processing data.

For this purpose, seven systems were tested with a two given test corpora of about half a million news stories and different categorical hierarchies of about 800 and 2,300 categories (Paaß and deVries, 2005). For the corpus with 2,300 categories the best system achieved at an F1-value of 39%, while for the corpus with 800 categories an F1-value of 79% was reached. Especially good are the results for recovering persons and geographic locations with about 80% F1-value. In general there were great variations between the performances of the systems.

In usability experiment with human annotators, the formal evaluation results were confirmed leading to faster and more consistent annotation. It turned out, that with respect to categories the human annotators exhibit a relative large disagreement and a lower consistency than text mining systems. Hence the support of human annotators by text mining systems offers more consistent annotations in addition to faster annotation.

The techniques of TDM are also applied to manage human resources, especially to analyse employees' opinions, to read and store CVs for the selection of new personnel and to track the level of personnel satisfaction. In the context of human resources management, the TDM techniques are often utilised to monitor the state of health of a company by means of the systematic analysis of informal documents (Vishal and Gurpreet, 2009).

6 Future directions

This section discusses some of the promising future directions of TDM.

6.1 Processing multilingual text

As discussed earlier, TDM deals solely with the unstructured text documents, therefore, TDM tasks are dependent on the language in which the text is written and on the way that is used to express these documents. Therefore, it is necessary to adapt the text processing task that is suggested in our framework to process multilingual textual documents and produce a language-independent text format.

6.2 Semantic analysis

In our proposed framework we illustrated the need for semantic analysis to capture a better understanding of the textual documents representations and to gain a good realisation of the relationships between the concepts described in the documents. However, performing a semantic analysis is not that easy task; it is slow (operates in order of few words per second) and also computational expensive. So it is left as an open issue to find an efficient way to perform semantic analysis, especially with large and structured set of texts.

7 Conclusions

This paper proposes a framework that facilitates the understanding of TDM and relates the main tasks (clustering and classification) to TDM procedural steps. A new promising approach for spam detection that is based on e-mail‘s text clustering is proposed. In addition, a set of open problems and their future directions are provided. Moreover, the paper provided as a survey of the state-of-the-art in TDM field. This paper sheds the light to the most related areas to TDM, namely: IR, IE, NLP, ML, data mining (KDD) and statistics.

Acknowledgements

We would like to express our deep gratitude for the reviewers for their efforts, time and constructive comments that enrich the paper and enhance its quality.

References

Aaron, M. and William, R. (2005) ‘A survey of current work in biomedical text mining’, *Briefings in Bioinformatics*, Vol. 6, No. 1, pp.57–71.

Adebola, K.O. and Adeyemo, A.B. (2013) ‘Knowledge discovery in academic electronic devices using text mining’, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 11, No. 2, pp.10–20.

Aggarwal, C.C., Yuchen, Z. and Yu, P.S. (2012) ‘On the use of side information for mining text data’, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 99, p.1.

Ahmed, K. (2007) ‘An overview of content-based spam filtering techniques’, *Informatica*, Vol. 31, No. 3, pp.269–278.

Aixin, S. (2012) ‘Short text classification using very few words’, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ‘12)*, pp.1145–1146.

Andrea, H., Andreas, N. and Gerhard, P. (2005) ‘A brief survey of text mining’, *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology*, Vol. 20, No. 1. pp.19–62.

Benoit, G. (2002) ‘Data mining’, in Chronin, B. (Ed.): *Annual Review of Information Science and Technology*, Vol. 36, pp.265–310, American Society for Information Science and Technology, Silver Spring.

Berry, M.W. (Ed.) (2004) *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer-Verlag, 2003, 2007.

Elmasri, R.A. and Navathe, S.B. (1999) *Fundamentals of Database Systems*, 3rd ed., in Shanklin, C. (Ed.): Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Fan, W., Wallace, L., Rich, S.m. and Zhang, Z. (2005) 'Tapping into the power of text mining', *Communications of the ACM-Privacy and Security in Highly Dynamic Systems*, Vol. 49, No. 9, pp.76–82.

Goncalves, P.M., Barros, R.S.M. and Vieira, D.C.L. (2012) 'On the use of data mining tools for data preparation in classification problems', *IEEE/ACIS 11th International Conference on Computer and Information Science*, Centro de Inf., Univ. Fed. de Pernambuco, Recife, Brazil, pp.173–178.

Gordon, V. and David, R. (2006) 'Email spam filtering: a systematic review', *Foundations and Trends in Information Retrieval*, Vol. 1, No. 4, pp.335–455.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques (Morgan-Kaufman Series of Data Management Systems)*, Academic Press, San Diego.

Havre, S., Hetzler, E., Perrine, K., Jurrus, E. and Miller, N. (2001) 'Interactive visualization of multiple query result', in *Proc. of IEEE Symposium on Information Visualization*, pp.105–112.

Hearst, M.A. (1999) 'Untangling text data mining', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp.3–10.

Hong, Y. and Eugene, A. (2003) 'Extracting synonymous gene and protein terms from biological literature', *Bioinformatics*, Vol. 19, Supplement 1, pp.1340–1349.

Hotho, A., Nürnberger, A. and Paass, G. (2005) 'A brief survey of text mining', in *Proceedings of LDV Forum*, pp.19–62.

Jurafsky, D. and Martin, J.H. (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall, USA.

Jusoh, S. and Alfawareh, H.M. (2009) 'Agent-based knowledge mining architecture', in *Proceedings of the 2009 International Conference on Computer Engineering and Applications, IACSIT*, World Academic Union, Manila, Philippines, pp.602–606.

Jusoh, S. and Alfawareh, H.M. (2012) 'Techniques, applications and challenging issue in text mining', *IJCSI International Journal of Computer Science Issues*, Vol. 9, No. 2, pp.431–436.

Karanikas, H. and Manchester, B.T. (2001) *Knowledge Discovery in Text and Text Mining Software*, Centre for Research in Information Management, UK.

Li, Q., Sam. A., Wen-Pin, L., Li, X. and Ji, H. (2011) 'Joint inference for cross-document information extraction', *Proc. 20th ACM Conference on Information and Knowledge Management (CIKM2011)*, pp.2225–2228.

Liddy, E.D. (2000) 'Text mining', *Bulletin of the American Society for Information Science*, Vol. 27 [online] http://www.asis.org/Bulletin/Oct-00/liddy.html (accessed October 2012)

Marios, H. and Divesh, S. (2010) 'Weighted set-based string similarity', *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol. 33, No. 1, pp.25–36.

Michael, B.W. (2007) *Survey of Text Mining: Clustering, Classification and Retrieval*, 2nd ed., Springer, New York.

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G. and Stamatopoulos, P. (2004) 'Filtron: a learning-based anti-spam filter', *Proceedings of the 1st Conference on Email and Anti-Spam*, Mountain View, CA, USA.

Nahm, U. and Mooney, R. (2003) 'Text mining with information extraction', *Proceedings of the 4th International MIDP Colloquium*, pp.141–160.

Paaß, G. and deVries, H. (2005) 'Evaluating the performance of text mining system', *Proc. 29th Annual Conference of the German Classification Society*, Son real-world press archives, Springer.

Pablo, C., Miriam, F. and David V. (2007) 'An adaptation of the vector-space model for ontology-based information retrieval', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 2, pp.261–272.

Saracevic, T. (2001a) *Personal Communication and Class Discussions, Seminar in Information Studies*, Rutgers University, School of Communication, Information and Library Studies, New Brunswick, NJ.

Saracevic, T. (2001b) *Text Mining Seminar in Information Studies*, Rutgers University, School of Communication, Information and Library Studies, New Brunswick, NJ.

Sebastiani, F. (2002) 'Machine learning', *ACM Computing Surveys*, Vol. 1, No. 34. pp.1–47.

Smith, D. (2002) 'Detecting and browsing events in unstructured text', in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.73–80.

Srinivasan, P. (2001) 'MeSHmap: a text mining tool for medline', in *Proceedings of the American Medical Informatics Annual Symposium*, pp.642–646.

Steinbach, M., Ertoz, L. and Kumar, V. (2003) 'Challenges of clustering high dimensional data', in Wille, L.T. (Ed.): *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition*, Springer-Verlag.

Tapas, K., David, M., Nathan, S., Christine, D., Ruth, S. and Wu, A.Y. (2002) 'An efficient k-means clustering algorithm: analysis and implementation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, pp.881–892.

Un Yong, N. and Raymond, J. (2004) *Text Mining with Information Extraction*, Doctoral Dissertation, The University of Texas at Austin.

Vishal, G. and Gurpreet, S. (2009) 'A survey of text mining techniques and applications', *Journal of Emerging Technologies in Web Intelligence*, Vol. 1, No. 1, pp.10–20.

Xie, J., Chen, X., Ji, J. and Sui, Z. (2012) 'Multi-mode natural language processing for extracting open knowledge', *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp.154–161.