IJCSI International Journal of Computer Science Issues, Volume 13, Issue 3, May 2016
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org http://dx.doi.org/10.20943/01201603.5461

Text Mining and Knowledge Discovery from Big Data: Challenges and Promise

Amal Mahmoud Yehia[1] Lamiaa Fattouh Ibrahim[1,2] Maysoon Fouad Abulkhair[2]
[1]Department of Computer Science and Information
Institute of Statistical Studies and Research
Cairo University, Cairo, Egypt

[2]Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University
B.P. 42808, 21551- Girls Section, Jeddah, Saudi Arabia

Abstract
With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents text mining and the ways used to categorize document structure techniques in big data. This subject poses a big challenge when it comes to guaranteeing the quality of extracted features in text documents to describe user interests or preferences due to large amounts of noise. This subject has many models and algorithms but still needs more to achieve best results for users, making this an open issue that needs more research.
Keywords: Text mining - Big Data - Knowledge Discovery.

1. Introduction

We live in a flood of data that is too big, too fast, or too hard for existing tools to process. "Too big" means that organizations increasingly must deal with petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere. "Too fast" means that is not only the volume of data is big, but it must be processed quickly. "Too hard" is a catchall for data that does not fit neatly into existing processing tools or that needs some kind of analysis that existing tools cannot readily provide. A Big Data problem has three distinct characteristics: the data volume is huge, the data-producing velocity is very high, and the data type is diverse (a mixture of structured data, semi-structured data, and unstructured data). These characteristics pose great challenges to traditional data processing systems since these systems either cannot scale to the huge data volume in a cost-effective way or they fail to handle data with a variety of types [1].

The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data. Exploring large volumes of data to extract information or knowledge for future action is a principal task for Big Data application [2].

The term data mining has been stretched beyond its limits to apply to any form of data analysis. One of the numerous definitions of Data Mining or Knowledge Discovery in Databases is: Extraction of interesting information or patterns from data in large databases is known as data mining. According to Cheng, H et.al. "Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [3]. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies. According to Prabhu, "Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis" [4].

Text mining is a technique which extracts information from unstructured data and finds patterns. It is also known as knowledge discovery from the text (KDT), it deals with the machine supported analysis of text [5]. Text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files etc. The problem

of Knowledge Discovery from the Text (KDT) is extracting detailed and implied concepts and semantic relations between concepts. Its aim is to get insights into large quantities of text data.

As we can see, Big Data resource comes from a collection of multiple, heterogeneous and autonomous information sources with complex and evolving relationships that keeps growing on social sites like Twitter and Facebook. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today was produced within the past two years [6].

This paper is structured as follows: section 2 talks about Text Mining. Section 3 describes Big Data. Section 4 presents the Text mining Challenges with Big Data. Section 5 concludes this study.

2. Text Mining

Text mining (TM) is the automatic discovery of new, previously unknown information by automatic analysis of various textual resources. TM starts by taking facts and events from textual sources and then forming new hypotheses that are further investigated by traditional Data Mining and data analysis methods.

Text mining, also known as Knowledge discovery from text (KDT), is the process of removing interesting patterns from very large text databases for discovering knowledge. TM applies the same analytical functions of data mining but also applies analytic functions from natural language (NL) and information retrieval (IR) techniques [7].

TM tools are used for:
Extracting relevant information from a document – extracting the features (entities) from a document by using NL, IR, and association metrics algorithms [8] or pattern matching [9].

Finding relationships between people, places, organizations, etc. by aggregating and comparing information obtained from the documents.

Classifying and organizing documents according to their content [10].

Retrieving documents based on the various sorts of information about the document content.

Clustering documents according to their content [11].

A TM system is composed of 3 major components (See Figure 1):

Information Feeders enable the connection between various textual collections and the tagging modules. This component is used to connect to any website, (such a news feed), internal document collections and any other kinds of textual collections.

Intelligent Tagging A component responsible for reading the text and tagging the related information. This component can perform any type of document tagging such as **statistical tagging** (categorization and term extraction), **semantic tagging** (information extraction) and **structural tagging** (extraction from the visual layout of the documents).

Business Intelligence Suite A component responsible for combining the information from disparate sources, allowing for simultaneous analysis of the entire information landscape.

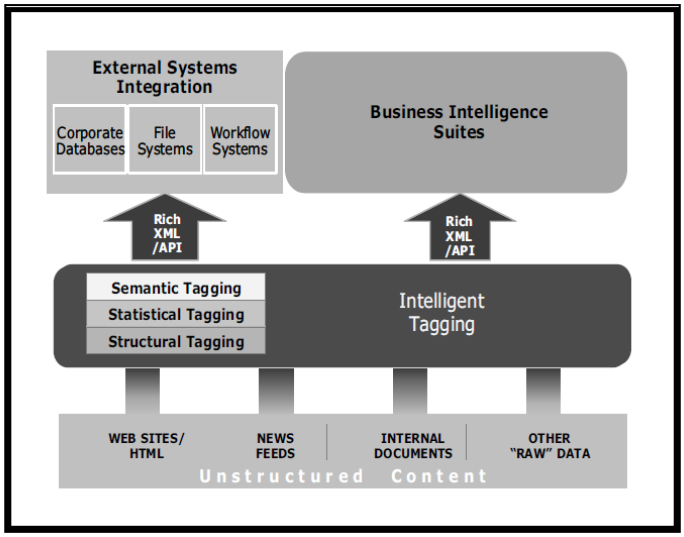

Fig. 1 Architecture of Text Mining system

There are two clear ways to categorize the totality of preparatory document structuring techniques: according to their task and the used algorithms, and to the formal frameworks that they use.

2.1 Task-Oriented Approaches

A document is an abstract that has a variety of possible actual representations. The task of the document-structuring process is to take the most "raw" representation and convert it to the representation where the meaning of the document is understandable.

In order to cope with this extremely difficult problem, a "divide-and-conquer" strategy is typically employed. The

problem is separated into a set of smaller subtasks, each of which is solved separately. The subtasks can broadly be divided into three classes (see Figure 2) – preparatory processing, general-purpose natural language processing ("NLP") tasks, and problem-dependent tasks.

1) ***Preparatory Processing***:

This process converts the raw representation into a structure suitable for further linguistic processing. For example, the raw input may be a PDF document, a scanned page, or even recorded speech. The task of the preparatory processing is to convert the raw input into a stream of text, possibly labeling the internal text zones, such as paragraphs, columns, or tables. It is also sometimes possible for the preparatory processing to extract some document-level fields, such as" Author" or "Title" in cases where the visual position of the fields allows their identification.

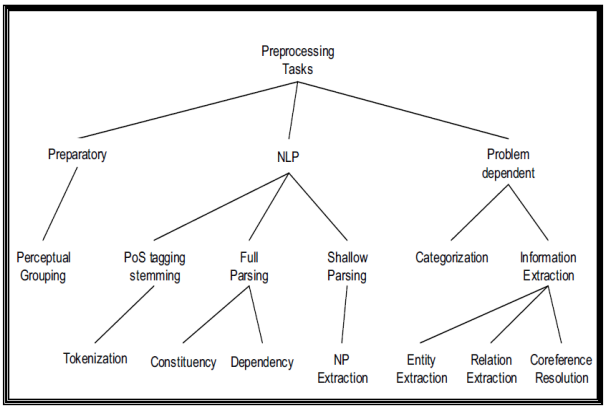

Fig. 2 Text Pre-processing Tasks

2) ***NLP Process***

The general-purpose NLP tasks process text documents using the general knowledge about natural language. The tasks may include tokenization, morphological analysis, part-of-speech tagging, and syntactic parsing whether shallow or deep.

– Tokenization is the process of identifying the words in the sentences within the text and is considered the first step in information extraction from text. For example, the tokenizer in an English text might split the text into white spaces and punctuation marks to separate words. In other languages (e.g., Chinese, Japanese), the spaces do not separate words, making this process more complex. Also, in some languages, hyphens are used to compound words (e.g., German, Dutch), making this is a crucial step.

– Part-of-speech Tagging (PoS) is the process of recognizing a word's part of speech in a sentence (e.g., noun, verb, etc.) by its context. Tagging assists as the basis for the information retrieval system to perform syntax-sensitive filtering and analysis. Usually, PoS taggers at some stage of their processing perform the morphological analysis of words. Thus, an additional output of a PoS tagger is a sequence of stems (also known as "lemmas") of the input words [12].

– Syntactical parsing components perform a full syntactical analysis of sentences, according to a certain grammar theory. The basic division is between the constituency and dependency grammars [13]- [16].

- Shallow Parsing [17] is the task of splitting documents into no overlapping word sequences or phrases such that syntactically related words are categorizing together. Each phrase is then tagged by one of a set of predefined tags such as Noun Phrase, Verb Phrase, Prepositional Phrase, Adverb Phrase, Subordinated clause, Adjective Phrase, Conjunction Phrase, and List Marker. Shallow parsing is generally useful as a pre-processing step, either for bootstrapping, extracting information from corpora for use by more sophisticated parsers, or for end-user applications such as information extraction. Shallow parsing permits morphological analysis and the recognition of relationships between the object, subject and/or spatial/temporal location within a sentence.

3) ***Problem Dependant Task - Information Extraction in Text Mining***

Information Extraction (IE) is the most recent technique currently used in TM pre-processing operations. Without IE techniques, TM systems would have much more limited knowledge discovery capabilities.

From a text we can extract four basic elements:

Entities are the basic building blocks that are used in text documents. Examples include people, companies, locations, genes, drugs, etc.

Attributes are features of the rest entities. Examples of attributes include the title of a person, the age of a person, the type of an organization, etc.

Facts are the relations that exist between entities. Examples could include an employment relationship between a person and a company, Phosphorylation of two proteins, etc.

An Event is an activity or occurrence of interest in which entities participate. Examples could include a terrorist act, a merger between two companies, a birthday, etc.

2.2 Formal Frameworks and Algorithm-Based Techniques

Two steps are used in this technique: Text Categorization and Probabilistic Models for Information Extraction. The following sections discuss these two steps:

1) Text Categorization

There are two main approaches to reveal the **categorization** problem. The first approach is the *knowledge engineering* where the user is manually defining a set of rules encoding human expertise and knowledge about how to classify documents under given categories. The other approach is the *machine learning* that builds an automatic text classifier by learning from a set of pre-classified documents. Additionally, it will be processed automatically through utilizing several methods such as rule induction, naïve-Bayes, neural networks, decision trees, nearest neighbor, and support vector machine. The following two sections will give more detail about these approaches.

a) Knowledge Engineering Approach

An example of knowledge engineering approach is the CONSTRUE system [18]- [21] built by the Carnegie group for Reuters. A typical rule in the CONSTRUE system:

1 **if** DNF (disjunction of conjunctive clauses) formula **then** category **else** ¬ category

An example of this rule being applied might look like the following:

1 **If** ((wheat & farm) or (wheat & commodity) or (bushels & export) or (wheat & tonnes) or (wheat & winter & ¬ soft)) **then** Wheat **else** ¬ Wheat

The main drawback of this approach is that the rules must be manually set by interviewing experts from the domain to define an engineered knowledge. If the set of categories is modified, then more professionals must be interviewed again.

b) The Machine Learning Approach

The machine learning approach is based on the existence of a training set of documents that are already pre-tagged using the predefined set of categories.

There are two main methods for performing machine learning based categorization. One method is to perform (fully automated) or "Hard" classification. For each pair of categories and documents we assign a truth value (either TRUE if the document belongs to the category or FALSE otherwise). The other approach is to perform a ranking (semi-automated) based classification. In this approach, the classifier returns a Categorization Status Value (CSV) instead of the truth value; i.e. a number between 0 and 1 that represents the evidence for the fact that the document belongs to a category. Documents are then ranked according to their CSV value.

(2) Probabilistic Models for Information Extraction

Probabilistic models often show enhancement in accuracy and robustness with less noise level than the categorical models. The ultimate reason for this is not quite clear, and can be an excellent subject for a philosophical debate.

Furthermore, several probabilistic models have proven to be specifically useful for the different tasks in extracting meaning from natural language texts. Most prominent among these probabilistic approaches are Hidden Markov Models (HMMs), Stochastic Context-Free Grammars (SCFG), and Maximal Entropy (ME).

2.3 Hybrid Approaches – TEG

Traditionally, the knowledge engineering (rule-based) systems were top performers in most IE benchmarks, like a series of Message Understanding Conferences (MUC) [22] – [24] that support the evaluation. Recently, machine-learning systems became state-of-the-art, especially for simpler tagging problems, such as named entity recognition [25], [26], or field extraction [27].

The knowledge engineering approach retains some of its advantages. It is focused on writing patterns to extract the entities and relations. The patterns are naturally accessible to human understanding, and can be improved, whereas, improving the results of a pure machine learning system, would require providing it with additional training data. However, the impact of adding more data soon becomes infinitesimal while the cost of manual explanation of the data grows linearly.

TEG [28] is a hybrid entities and relations wresting system which combine the power of knowledge-based and statistical machine learning approaches. The system is based upon SCFGs. The rules for the wresting grammar are written manually while the probabilities are trained from an annotated corpus.

The powerful disambiguation ability of PCFGs allows the knowledge engineer to write very simple and naive rules while retaining their power, thus greatly reducing the required labor [29].

3. Big Data

Big data is generated from many increasing sources, including Internet clicks, mobile transactions, user-generated content, social media, data generated through sensor networks or business transactions like sales queries, and purchase transactions as well as genomics, health care, engineering, operations management, the industrial Internet, and finance [30]. This data requires the use of powerful computational techniques to show trends and patterns within and between these extremely large socioeconomic datasets.

a- Sources of Big Data

Big data is a wrapper for different types of granular data. Some examples of key sources of high volume data are:

(1) public data, (2) private data, (3) data exhaust, (4) community data, and (5) self-quantification data.

"Public data" is the data that is produced by government, organizations, and local communities and that can possibly be harnessed for wide-ranging business and management applications. Examples of such data include those concerning transportation, energy use, and health care that are accessed under certain restrictions in order to guard individual privacy.

"Private data" is the data held by private firms, non-profit organizations, and individuals that reflect private information that cannot readily be imputed from public sources. For example, private data include consumer transactions, radio-frequency identification tags used by organizational supply chains, movement of company goods and resources, website browsing, and mobile phone usage, among several others.

"Data exhaust" refers to a collected different-purpose data, which can be recombined with other data sources to create new value sources, for example, when people make purchases, even at informal markets; when they access basic health care; or when they interact with others. Another source of data exhaust is information-seeking behavior, which can be used to infer people's needs, desires, or intentions. This includes Internet searches, telephone hotlines, or other types of private call centers.

"Community data" is a distillation of unstructured data—especially text—into dynamic networks that capture social trends. Typical community data includes consumer reviews on products, voting buttons (such as, "I find this review useful"), and Twitter feeds, among many others. This community data can then be distilled for meaning to infer patterns in the social structure.

"Self-quantification data" is the data types that are revealed by the individual through quantifying personal actions and behavior.

b- Data Sharing, Privacy, and Ethics

Data sharing agreements need to be linked to the mechanisms for data protection and privacy, including anonymization for open data, access control, rights management, and data usage control.

Issues such as imputed identity, where individual identity can be inferred through data triangulation from multiple sources, will need to be carefully considered and explicitly acknowledged and permitted.

Management scholars will be invited to embed themselves into social issues based on defining research questions that integrate data sharing and privacy as part of their research methodology. Doing so will likely allow us to refine the model for data sharing and data rights, which could be universally beneficial, and define the Big Data collaborations in the future.

c- Analyzing Big Data

There are a number of different techniques for analyzing Big Data. These techniques draw from several disciplines, including statistics, computer science, applied mathematics, and economics. They include cluster analysis, data fusion and integration, data mining, genetic algorithms, machine learning, natural language processing, neural networks, network analysis, signal processing, spatial analysis, simulation, time series analysis, A/B testing, and visualization [31].

The challenge, though, is to shift away from focusing on the observed sample results to focusing, rather, on effect sizes based on variance explained. The next challenge of analyzing big data is to move beyond identifying correlational patterns to exploring causality. Given the unstructured nature of Big Data, causality is not built into their design and the patterns observed are often open to a wide range of possible causal explanations. There are two main ways to approach this issue of causality. The first way is to recognize the central importance of theory. An intuition about the causal processes that generated the data and can be used to guide the development of theoretical arguments, grounded in prior research and pushing beyond it. The second, complementary way is to test these theoretical arguments in subsequent research. Laboratory experiments offer the advantage of greater control, but they usually focus on a very limited number of variables. The nature of Big Data research is that there may be many factors driving the observed correlational patterns. In an

experiment field, a wider net can be cast, as a richer set of data about behaviors and beliefs that can be collected over an extended period of time.

4. Text Mining Challenges with Big Data

Big Data poses contents for text analysis and natural language processing due to its characteristics of volume, veracity, and velocity of the data. The complete volume in terms of numbers of documents contests traditional local repository and index systems for large-scale analysis and mining. Computation, storage, and data representation must work together to provide rapid access, search, and mining of the deep knowledge in the large text collection. [33]

Data preprocessing-remains a daunting task for big textual data, particularly data veracity which is questionable due to the age of original materials. Data velocity is the rate of change of the data but can also be the rate at which changes and corrections are made. [32]

5. CONCLUSION

In this paper, we have dealt with the quality issue of discovered relevant features in text documents for describing user interests or preferences information. We present several kinds of mining approaches which show clear ways of categorizing document-structuring techniques that are based on their task and the algorithms and the used formal frameworks.

We explored the problem that facing the Big Data when performing its analysis. This point of research is still open to discovering new algorithms for text mining for the information.

References

[1] Dawei Jiang, Gang Chen, Beng Chin Ooi, KianLee Tan, Sai Wu, (2014). “epiC: An Extensible and Scalable System for Processing Big Data”, Proceedings of the VLDB Endowment, Volume 7, No. 7.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, (2014). “Data Mining with Big Data” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1.

[3] Cheng, H. Yan, X. Han, J. and Hsu, C. (2007). “Discriminative Frequent Pattern Analysis for Effective Classification,” in 23rd IEEE ICED International Conf. on Data Engineering.

[4] Prabhu, S. (2014) “Data Mining and Warehousing”, paritosh82

[5] Haralampos Karanikas and Babis Theodoulidis Manchester, (2001). "Knowledge Discovery in Text and TM Software", Centre for Research in Information Management, UK.

[6] IBM, (2012). “What Is Big Data: Bring Big Data to the Enterprise,”

[7] Dorre, J., Gerstl, P., and Seiffert, R. (1999). “TM: Finding Nuggets in Mountains of Textual Data,” Proceedings of KDD-99, 5th ACM International Conference on Knowledge Discovery and Data Mining: 398-401. San Diego, US: ACM Press, New York, US.

[8] Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstar, O., Rajman, M., Schler, Y., and Zamir, O. (1998). “TM at the Term Level.” Paper presented at the Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, Nantes, France.

[9] Averbuch, M., Karson, T., Ben-Ami, B., Maimon, O., and Rokach, L. (2004). “Context Sensitive Medical Information Retrieval,” MEDINFO-2004, San Francisco, CA, September. IOS Press, pp. 282-262.

[10] Tkach, D. (1998). “Turning Information into Knowledge.” a white paper from IBM.

[11] Wai-chiu, W. and A. W.-c. Fu (2000). “Incremental Document Clustering for Web Page Classification.” In Proceedings of 2000 International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000), Aizu-Wakameatsu City, Fukushima, Japan.

[12] Brill, E. (1992). “A Simple Rule-Based Part of Speech Tagger.” Third Annual Conference on Applied Natural Language Processing, ACL.

[13] Keller, B. (1992). “A Logic for Representing Grammatical Knowledge.” European Conference on Artificial Intelligence: 538-542.

[14] Pollard, C. and I. A. Sag (1994). “Head-Driven Phrase Structure Grammar.” Chicago, Illinois, University of Chicago Press and CSLI Publications.

[15] Rambow, O. and A. K. Joshi (1994). “A Formal Look at Dependency Grammars and Phrase-Structure Grammars, with Special Consideration of Word-Order Phenomena.” Current Issues in Meaning-Text Theory. L. Wanner. London, UK, Pinter.

[16] Neuhaus, P. and N. Broker (1997). "The Complexity of Recognition of Linguistically Adequate Dependency Grammars." Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics., New Jersey.

[17] Hammerton, J., Miles Osborne, Susan Armstrong, and Daelemans, W. (2002). "Introduction to the Special Issue on Machine Learning Approaches to Shallow Parsing." Journal of Machine Learning Research, 2 (Special Issue Website): 551-558.

[18] Hayes, P. J., Knecht, L. E., and Cellio, M. J. (1988). "A News Story Categorization System," Proceedings of ANLP-88, 2nd Conference on Applied Natural Language Processing: 9-17. Austin, US: Association for Computational Linguistics, Morristown, US.

[19] Hayes, P. J., Andersen, P. M., Nirenburg, I. B., and Schmandt, L. M. (1990). "Tcs: A Shell for Content-Based Text Categorization," Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications: 320-326. Santa Barbara, US: IEEE Computer Society Press, Los Alamitos, US.

[20] Hayes, P. (1992). Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques. Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval: 227-242.

[21] Hayes, P. J. and S. P. Weinstein (1990). "Construe/Tis: A System for Content-Based Indexing of a Database of News Stories." Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence. AAAI Press, Menlo Park, US: 49-66.

[22] Chinchor, N., Hirschman, L., and Lewis, D. (1993). "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)". Computational Linguistics, 3(19): 409-449.

[23] Automatic Content Extraction ACE (2002). http://www.itl.nist.gov/iad/894.01/tests/ace/.

[24] Yeh, A., Hirschman, L., and Morgan, A. (2002). "Background and Overview for KDD Cup 2002 Task 1: Information Extraction from Biomedical Articles." KDD Explorations, 4(2): 87-89.

[25] Bikel, D. M., S. Miller, *et al*. (1997). "Nymble: A High-Performance Learning Name-Finder." Proceedings of ANLP-97: 194-201.

[26] Chieu, H. L. and H. T. Ng (2002). "Named Entity Recognition: A Maximum Entropy Approach Using Global Information." Proceedings of the 17th International Conference on Computational Linguistics.

[27] McCallum, A., Freitag, D., and Pereira, F. (2000). "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proc. 17th International Conf. on Machine Learning: 591-598: Morgan Kaufmann, San Francisco, CA.

[28] Rosenfeld B., Feldman R., et al. (2004). "TEG: A Hybrid Approach to Information Extraction." Conference on Information and Knowledge Management, Washington, D.C., USA.

[29] Gerard George, Martine R. Haas, and Alex Pentland, (2014) "Big Data and Management", Academy of Management Journal, Vol. 57 issue 2, pp. 321-326

[30] Oded Maimo, Lior Rokach, (2010) "Data Mining and Knowledge Discovery Handbook", DOI 10.1007/978-0-387-09823-4, Springer New York Dordrecht Heidelberg London.

[31] McKinsey Global Institute. 2011. "Big data: The Next Frontier for Innovation, Competition, and Productivity**".** June 2011. Lexington, KY: McKinsey & Company.

[32] Beth Plale, (2013). "Big Data Opportunities and Challenges for IR, TM and NLP". ACM New York, NY, USA.

[33] Fatima EL Jamiy, Abderrahmane Daif, Mohamed Azouazi, and Abdelaziz Marza (2014). "The potential and challenges of Big data - Recommendation systems next level application", International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2.

Amal Mahmoud Yehia is a student Ph.D. in Department of Computer Science and Information Institute of Statistical Studies and Research Cairo University, Cairo, Egypt.

Lamiaa Fattouh Ibrahim is a Professor in the Faculty of Computing and Information Technology at King Abdulaziz University in Jeddah and in the Institute of Statistical Studies and Research at Cairo University. She obtained a Ph.D. from the Faculty of Engineering, Cairo University in 1999, Master's degree from the Computer and Systems Engineering Department in the Faculty of Engineering, Ain Shams University in 1993, Master's degree from the Ecole National Superieur de Telecommunication, ENST Paris in 1987, and a BSc

from the Computer and Automatic Control Department, Faculty of Engineering, Ain Shams University in 1984. She has over 32 years of experience in the fields of network engineering and artificial intelligence, focusing on applying knowledge base and data mining techniques in wired and wireless network planning. She has published papers in many journals and international conferences in the areas of networks, data mining, and wired and mobile network planning.

Maysoon Abulkhair is an Assistant Professor and Supervisor of IT department at King Abdulaziz University, Jeddah, KSA. She is a member of ACM. Her main interest in the research field is HCI, associating it with different knowledge areas such as artificial intelligent, machine learning, and data mining.