Survey on Existing Text Mining Frameworks and A Proposed Idealistic Framework for Text Mining by Integrating IE and KDD

Prakhyath Rai[1], Vijaya Murari T[2]

1 PG Scholar, Dept. of Computer Science & Engineering, NMAM Institute of Technology, Nitte, India
2 Asst. Professor, Dept. of Computer Science & Engineering, NMAM Institute of Technology, Nitte, India

Abstract:
Text Mining can be defined as a technique which is used to extract interesting information or knowledge from text documents which are usually in unstructured form. Information Extraction (IE) and Knowledge Discovery from Databases (KDD) are both useful approaches for discovering information in textual corpora. Information Extraction concerns locating of specific terms in natural-language documents. Knowledge Discovery in Databases is the process of discovering useful knowledge from a collection of data. This paper provides an analysis on various emerging text mining framework and methodologies. This paper examines text mining frameworks such as DiscoTEX (Discovery from Text Extraction), RAPIER (Robust Automatic Production of Information Extraction Rules), EPD (Effective Pattern Discovery) and BWI (Boosted Wrapper Induction. This paper provides an analysis of various text mining frameworks and defines their benefits and deficiencies which are then taken into consideration for proposing a novel framework for text mining referred as RDIET. The proposed framework uses an automatically learned IE system to extract a structured database from unstructured textual corpus, and then mines the database for deducing interesting relationships using KDD tools. The proposed technique concentrates in achieving mutual benefits of IE and KDD. IE enables the application of KDD to unstructured textual corpora and KDD can discover predictive rules useful for enhancing IE performance.

Keywords: *BWI, Categorization, Clustering, DiscoTEX, Effective Pattern Discovery, Information Extraction, Knowledge Discovery from Databases, Text Mining, Knowledge Discovery from Text, RAPIER, Standard Rule Induction.*

I. INTRODUCTION

Modern information system allows firms to capture vast amounts of data. Much of this data is structured data that can be analyzed using traditional database software. Increasingly, however, large amounts of data such as textual data are unstructured. Manual analysis of this unstructured textual data is increasingly impractical, and as a result, text mining methods are developed to automate the process of analyzing textual data. Text mining is used to extract the relevant information or knowledge or patterns from different sources that are in unstructured form. It typically consists of (i) information retrieval (IR), which gathers and filters documents, (ii) information extraction (IE), and (iii) data mining for discovering unexpected associations between known facts. IE and KDD have some deficiencies. Information extraction can identify relevant sub-sequences of text, but is usually unaware of emerging, previously unknown knowledge and regularities in a text and thus cannot form new facts or new hypothesis. Knowledge Discovery in Databases limits to deduce explicit relationships from the collection of data. Complementary to information extraction, emerging text mining methods and techniques promise to overcome the deficiencies of information extraction. Developing a knowledge-based or a machine learned IE systems is time and labour intensive. The challenge in this iterative engineering process is that extraction rules must be (i) sufficient background knowledge to extract the full extent of available information and (ii) accuracy in extraction of relevant according to a giving specification. Additional deficiencies of IE approaches are that they (i) extract only explicit knowledge but not new, previously unknown knowledge, such as new relationships between entities, (ii) are better at simple extraction tasks than complex relation or event extraction, and (iii) do not offer information on novelty, reliability, and level of interest of extracted information. Common approaches no longer seem to be appropriate for handling the large amounts of existing information and do not meet the demands of an effective and accurate IE system. Emerging methods and techniques of text and data mining promise to overcome the shortcomings of IE and concentrate in improving the quality of IE. In summary, the open issues in state-of-the-art IE approaches make further developments necessary. There has been much discussion about combining IE and data mining [1] [2] [3], and these first initiatives have been successful, although they address relatively small problems.

II. LITERATURE SURVEY

Text mining is used to extract the relevant information or knowledge or patterns from different sources that are in unstructured form. Text mining mainly concentrates on text refinement and knowledge distillation. Text refinement is an approach of transforming free-form text or document to intermediate form and knowledge distillation is used to deduce patterns or knowledge from intermediate form. Several techniques have been proposed for text mining including conceptual structure, association rule mining, episode rule mining, decision trees, and rule induction methods. In addition, Information Retrieval (IR) techniques have widely used the "bag-of-words" model [1] for tasks such as document matching, ranking, and clustering. Information extraction (IE) and knowledge discovery from databases (KDD) are both useful approaches for discovering information in textual corpora, but they have some deficiencies. Information extraction can identify relevant sub-sequences of text, but is usually unaware of emerging, previously unknown knowledge and regularities in a text and thus cannot form new facts or new hypothesis. Complementary to information extraction, emerging text mining methods and techniques promise to overcome the deficiencies of information extraction. KDD limits itself to deducing relationships implicitly from collection of data. Text mining approaches and applications use IE as a pre-processing task in the text mining process and implement IE and data mining tasks sequentially, making integration of the two techniques impossible. Mooney discussed two approaches: the first one extracts general knowledge directly from a text, and the second one first extracts structured data from text documents or web pages and then applies traditional data mining techniques to discover new knowledge from extracted data. The DiscoTEX system [2] is an example of the second approach, which uses the previously discovered rules to predict information overlooked in the extraction step. In summary, due to the many open issues of state-of-art IE approaches further development is necessary. Many text mining techniques have been proposed in the last decade. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). Even not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. Effective pattern discovery technique overcome the low-frequency and misinterpretation problems of text mining, the technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents [4].The (unheralded) preliminary step in most of the applications of automated text analysis involves keywords to choose documents from large corpus of text data. Some of the computer-assisted statistical approach suggests keywords from available text, without needing any structured data as inputs. The framework suggested by Gary King, Patrick Lam and Margaret E Roberts poses the statistical problem in a new way, which leads to a widely applicable algorithm. This approach is based on training classifiers, extracting information from their mistakes, and then summarizing results with Boolean search strings [6].

III. ANALYSIS OF TEXT MINING FRAMEWORKS

DiscoTEX Framework: DiscoTEX (Discovery from Text Extraction) uses a learned information extraction system to transform text into more structured data which is then mined for interesting relationships. The initial version of DiscoTEX integrates an IE module acquired by an IE learning system, and a standard rule induction module. In addition rules mined from a database extracted from a corpus of texts are used to predict additional information to extract from future documents. DiscoTEX concentrates in improving recall factor of extraction mechanism, thereby enhancing F-measure by a moderate amount. DiscoTEX has a shortcoming in obtaining the precision of underlying documents to the expected mark [2].

RAPIER Framework: RAPIER (Robust Automated Production of Information Extraction Rules) uses relational learning to construct unbounded pattern-match rules for information extraction given a database of texts and filled templates. The learned patterns employ limited syntactic and semantic information to identify potential slot fillers and their surrounding context. RAPIER is bottom-up learning algorithm that incorporates techniques from several inductive logic programming systems and allows patterns to have constraints on the words, parts-of-speech tags, and semantic classes present in the filler and the surrounding text. RAPIER can achieve a good extraction precision. RAPIER lacks itself in achieving effective recall on underlying documents [3].

EPD Framework: EPD (Effective Pattern Discovery) is an innovative approach which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. This approach overcomes the low-frequency and misinterpretation problems which are frequent obstacles in text mining. Effective Pattern Discovery by employing pattern deploying and pattern evolving processes refines the discovered patterns in text documents, thereby enhancing the future mining process. Effective Pattern Discovery approach overcomes major shortcomings found in text mining approaches but involves huge complexities in its approaches [4].

BWI Framework: BWI (Boosted Wrapper Induction) is an approach to build a trainable information extraction system. Like wrapper induction techniques BWI learns relatively simple contextual patterns identifying the beginning and end of relevant text fields. BWI uses AdaBoost algorithm in repeated fashion for learning boundaries. BWI concentrates in repeated execution so that patterns missed by previous rules can be extracted. BWI provides high precision on underling documents. BWI limits itself in acquiring effective recall [5].

IV. PROPOSED TEXT MINING FRAMEWORK

The amount of textual data that is available for researchers and business to analyze is increasing at a dramatic rate. Information extraction (IE) and knowledge discovery from databases (KDD) are both useful approaches for discovering information in textual corpora, but they have some deficiencies. Information extraction can identify relevant sub-sequences of text, but is usually unaware of emerging, previously unknown knowledge and regularities in a text and thus cannot form new facts or new hypothesis. Complementary to information extraction, emerging text mining methods and techniques promise to overcome the deficiencies of information extraction. This paper proposes a framework for text mining that combines the benefits of both the approaches by integrating information extraction and knowledge discovery from databases using an information extraction system for transforming natural-language documents into structured data which can be then used for discovering relevant information and interesting relationships. For Example, suppose we discovered that computer-science jobs requiring “MySQL” skills are “database” jobs in many cases. If Information Extraction system manages to locate “MySQL” in language slot but failed to extract “database” in the area slot, in such cases relationships can be derived. The framework proposed in this paper aims to develop software that allows extracting specific information from unstructured data such as html-tagged text, text documents or documents with .pdf, .doc or .docx extensions and create text databases. The proposed framework in designing a new IE methodology, is referred as RDIET (Recognition and Discovery of Information from Extracted Text), which is based on various statistical and machine learning techniques. Integration of IE and KDD must concentrate on following areas:

Requirement Specification and Analysis: At the preliminary stage all the requirements of each phase have to be specified and analyzed sensibly.

Selection of Techniques: Suitable IE and KDD methodology have to be chosen to meet the strategized solution.

Interface Design: A suitable interface between IE and KDD has to be designed. The interface must facilitate bi-directional communication, so that IE produces accurate and significant hypothesis for the subsequent mining process.

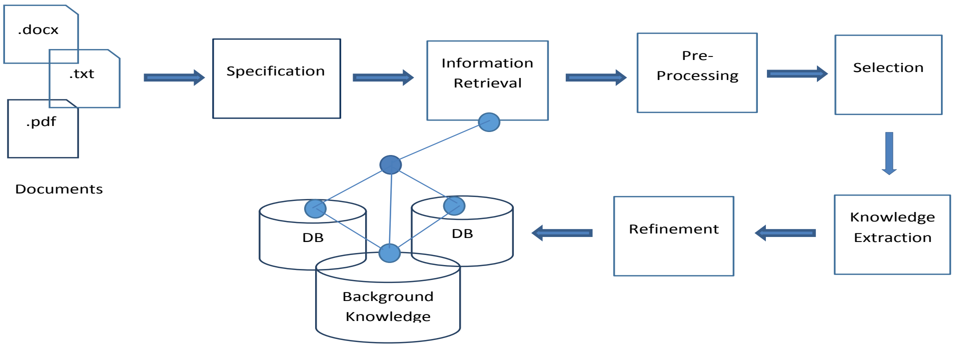

Figure 1: Architecture of RDIET

Figure 1 depicts individual phases of RDIET methodology. The phases are outlined in the following sub-section:

Specification: Specification concentrates in providing information regarding the purpose of text mining analysis and provides description on specification of templates.

Information Retrieval: Information Retrieval aims to collect/Fetch documents either online or offline and also defines certain crawling techniques to be used in functioning of this phase.

Pre-Processing: Several method exist that exploit the syntactic structure and their semantics, using different representations (such as characters, words, terms or concepts) of the documents. Tokenization and text-processing methods such as filtering are applied in RDIET to reduce the size of the data set.

Selection: Functioning of this phase is being strategized into either categorization or clustering based on the requirement, Categorization is a supervised technique based upon the set of input and output, In order to classify the document the set of input and output examples are used to train the classifier on the basis of known examples then unknown examples are categorized automatically. Clustering is a technique used to group similar documents but it differs from categorization, in this documents are clustered on the fly instead of through the use of predefined topics, this is unsupervised technique in which no inputs or patterns are predefined, it is based on the concept of dividing similar text into same cluster with each of it consisting certain number of documents.

Knowledge Extraction: For efficient and easy integration of IE and KDD, it is necessary to evaluate the methods and techniques of data mining in terms of the requirements of the novel IE methodology. The predictive relationships between different slot fillers discovered by data mining methods are the basis for integrating IE and KDD. These provide additional evidence of what information should be extracted from text resources. For example, suppose that the rule "VoiceXML ∈ language" → "Mobile ∈ area". If the IE system extracted "VoiceXML ∈ language" but failed to extract "Mobile ∈ area", we may want to assume there was an extraction error and add "Mobile" to the area slot, potentially improving recall. Therefore, after applying extraction rules to a document, RDIET applies its mined rules to the resulting initial data to predict additional potential extractions. This phase is quiet useful in validation, to mine missing data in order to complete specified templates, to identify inconsistent information, or to de-duplicate information.

Refinement: Because text mining can result in a huge number of templates and slots, which cannot be fixed in the specification phase, the performance measures recall, precision and F-measure are generally more informative than an analysis of the accuracy of extracted novel facts. Measures are used to discard uninteresting extracted information and patterns in the mining process, hence improving mining efficiency. They rank patterns and extracted information to enable a kind of filtering in the early phase of IE. Moreover, measures are applied in the refinement phase to select and present interesting patterns to the user.

Background Knowledge: Semantic and reasoning aspects are used in various points of RDIET methodology. Background knowledge enriches knowledge discovery operations on processed documents and is able to enhance concept extraction and validation. Consequently, background knowledge is important in text mining and IE because it allows pattern abundance to be limited, and is used in pre-processing to provide a consistent lexical representation of documents.

V. CHALLENGES IN PROPOSED FRAMEWORK

Some emerging challenges have been identified for future research:

Choice of appropriate IE and KDD methods: The selection of appropriate variables, data mining algorithms, model assessment and refinement are key components of this project. Automatic feature selection and extraction should support this process.

Performance: Integrating IE and KDD promises to increase performance (in terms of recall and precision). For instance, integrated KDD methods enable more precise feature selection for IE, which in turn reduces the feature space to the most significant information for mining new knowledge.

Extending information extraction measures: Identifying interesting relationships in textual documents is becoming a resource-intensive task because there are many weak links between various entities. A user cannot decide which are really interesting ones – in the volume of information – which is a critical aspect to be taken into consideration. KDD techniques and their measures reduce the amount of information, which in turn increases system performance.

Validating the applicability of Framework: In order to validate the proposed framework, RDIET suitable performance metrics have to be selected. The evaluation of RDIET will be effected by (i) measuring the performance of a current IE process, which will be compared to (ii) an IE process resulting from RDIET, which additionally enables a general benchmark of the interest level of mined Information.

VI. CONCLUSION

Text mining is the discovery, which discovers the previously unknown information by extracting it automatically from different written sources. Text Mining is a new research area which draws on information retrieval, data mining, machine learning, and natural language processing. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text corpora can be developed. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Text mining is similar to data mining expect that data mining approaches are designed for handling the structured data but text mining can work with the unstructured or semi structured data sets such as e-mails, full text documents, HTML files etc. Various applications of text mining are spam filtering, monitoring public opinions, automatic labelling of documents in business libraries, analysis of junk mails etc.

Merits of Text Mining:

- Databases can store less amount of information and this problem has been resolved through text mining as it can extract relevant useful information from large text and put them in appropriate slots of databases.
- IE and KDD approaches support extraction of information from textual corpora accurately and efficiently.

Demerits of Text Mining:

- No programs can be made in order to analyze the unstructured text directly, to mine the text for information or knowledge.
- The information which is initially needed is nowhere written.

Based on the analysis made on various text mining frameworks in the initial sections of the paper the following inferences can be noted;

RAPIER framework and BWI framework provide high precision on underlying documents whereas these frameworks limit themselves in achieving required recall. DiscoTEX framework provides high recall but limits itself to moderate precision on underlying documents. Effective Pattern Discovery framework overcomes the shortcomings of various text mining approaches but involves complex computations.

The benefits and deficiencies of the analyzed frameworks are taken into consideration for proposing a novel framework for text mining referred as RDIET. The proposed framework, RDIET is expected to overcome the deficiencies of the text mining approaches as its deigned in an order to grab the benefits of IE and KDD by taking into account all the various limitations of analyzed frameworks. The proposed framework uses an automatically learned IE system to extract a structured database from unstructured textual corpus, and then mines the database for deducing interesting relationships using KDD tools. IE enables the application of KDD to unstructured textual corpora and KDD can discover predictive rules useful for enhancing IE performance.

REFERENCES

[1] R Baeza-Yates and B Ribeiro-Neto. “Modern Information Retrieval”, ACM Press, New York, 1999.

[2] Raymond J Mooney and Un Yong Nahm, “ Text Mining with Information Extraction”, Proceedings of the 4th International MIDP Colloquium, pages 141-160, Van Schaik Pub., South Africa, 2005.

[3] M E Califf and R J Mooney, “Relational Learning of Pattern-Match Rules for Information Extraction”, Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), pages 328-334, Orlando, FL, July 1999.

[4] Ning Zhong, Yuefeng Li and T. Grance, “Effective Pattern Discovery for Text Mining,” IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

[5] D Freitag and N Kushmerick, “Boosted Wrapper Induction”, Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), pages 577-583, Austin, TX, July 2000.

[6] Gary King, Patrick Lam and Margaret E Aroberts, “Computer-Assisted Keyword and Documents from Unstructured Text”, 2014.