Transforming Digital Unstructured and Semi-structured Data into Structured Data with the Aid of IE and KDT

Prakhyath Rai¹, Vijaya Murari T²

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

¹famous991991@gmail.com, ²vijayamurari.t@nitte.edu.in

Abstract—Data growth has seen an exponential acceleration with the advent of computer and network, which have imparted the digital form to data. Data can be classified into three categories: Unstructured data, Semi-structured data and structured data. Text Mining concerns extraction of relevant information, knowledge or patterns from sources that are in Unstructured or Semi-structured form. This project entitled “Transforming Digital Unstructured and Semi-structured Data into Structured Data with the Aid of IE and KDT” demonstrates a framework for text mining using a learned information extraction system aided with KDT (Knowledge Discovery from Text) principles. The functionality of this project is concentrated over the integrated result of IE (Information Extraction) module, KDT (Knowledge Discovery from Text) module and Standard Protocols module. Pre-processing is employed for transforming unstructured data or Semi-structured data such as HTML documents, text documents, and documents with .doc, .docx or .pdf extensions into a feasible format of data which is then mined for interesting relationships. Standard Protocols are defined for discovery of additional information’s from input sources. For Example, consider if information extraction system has managed to extract skills like “HTML” and “DHTML” from a computer job posting but could not find “XML” in the document, in such cases relationships can be mined through predefined derivations which are framed in the standard protocols module. In addition, rules mined from the database extracted from a corpus of texts are used to predict additional information that could be extracted from future input documents, thereby improving the recall of the underlying extraction system. Results are presented by applying these techniques to a corpus of computer job announcement from an internet news group.

Keywords—Information Extraction, Text Mining, Knowledge Discovery from Text, Standard Rule Induction, DiscoTEX, RAPIER, EPD, BWI.

I. INTRODUCTION

The amount of digital data that is available for researchers and businesses to analyse is increasing at a dramatic rate. Data in the unstructured format makes extraction of information from it difficult. Text mining concerns discovering useful information from unstructured or semi-structured text. Information Extraction (IE) and Knowledge Discovery from Text (KDT) are both useful approaches for discovering information in textual corpora. Information Extraction concerns locating of specific items in natural-language documents. Knowledge Discovery from Text (KDT) is an approach to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. IE and KDT 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 from Text 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. This project demonstrates a framework for text mining by integrating Information Extraction (IE) and Knowledge Discovery from Text (KDT), called “Transforming Digital Unstructured and Semi-structured Data into Structured Data with the Aid of IE and KDT”. A Pre-processing approach is employed for transforming natural-language documents into feasible data format which can be then used for extraction of relevant information. Extracted useful pieces of information’s are stored in structured format, which can be used for discovering relevant information and interesting relationships. The functionality of this project is concentrated over the integrated result of IE (Information Extraction) module, KDT (Knowledge Discovery form Text) module and Standard Protocols
module. Information Extraction (IE) concentrates in identifying relevant pieces of information, phrases and relationships within the text documents. Knowledge Discovery from Text (KDT) is an approach to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Standard Protocols are defined for discovery of additional information’s from input sources. The rules mined from a database extracted from natural-language documents are used for discovering additional information to extract from future unstructured data.

Traditional data mining assumes that the information to be mined is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by integrating IE module and KDT module can be provided to the Standard Protocols module for further mining of knowledge. Standard Protocols are defined for discovery of information from extracted information, for Example information extraction system has managed to extract skills like “HTML” and “DHTML” from a computer job posting but could not find “XML” in the document, and we may assume that there was an extraction error. Since typically the recall (percentage of correct slot fillers extracted) of an IE system is significantly lower than its precision (percentage of extracted slot fillers which are correct), such predictive relationships might be productively used to improve recall by suggesting additional information to extract.

The project aims to develop software that allows extracting specific information from unstructured data and creating text databases, the product facilitates to analyse and search a selected document for specified data and convert the unstructured data into structured data which can be then used for further processing. The product will help get a handle on large external information flows and make them an integral part of mainstream database. The product can serve as a powerful marketing tool that can be used to serve several operations such as to set up recruitment and job resume database, to process raw and random data downloaded from internet, to systemize catalogues: lists and price lists, to process stock quotations, to analyse statistical data on websites and to set up customized databases. This project reports experiments in the computer-related job posting domain demonstrating that predictive rules acquired by applying KDT to an extracted database can be used to improve the recall of information extraction.

II. LITERATURE REVIEW

Text mining is used to extract the relevant information or knowledge or pattern 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. 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 Extraction (IE) 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 text (KDT) 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. KDT 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.

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.

Text mining frameworks such as DiscoTEX, RAPIER, EPD and BWI are designed to withstand certain deficiencies encountered in text mining.

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 (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 (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 (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 underlying documents. BWI limits itself in acquiring effective recall [5].

III. METHODOLOGY

![Architecture Diagram](image)

Figure 1 Architecture Diagram

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

**Information Retrieval:** Information Retrieval aims to collectFETCH documents either online or offline. The documents and files collected are then placed into a specific folder in order to undergo further processing.

**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. Pre-Processing has mainly two piece of functionality Screening and Conversion. Screening concentrates in elimination of non-textual files such as image files, video files or audio files from the folder containing files for processing and places selected textual files such as pdf files, word files, text files or html files in a folder for further processing. Conversion process translates the textual files which are obtained as a results of screening process into text documents i.e. into txt files and places them in folder for future procedures in the processing function.

**Standard Protocols:** 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. Standard Protocols in brief concentrates in defining protocols or rules to be applied.
on text documents for extraction of useful information from text documents.

Knowledge Extraction: For efficient and easy integration of IE and KDT, 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 KDT. 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, the mined rules are applied 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.

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.

Database: Results of extraction process are placed in a suitable database which in turn act as background knowledge. Background knowledge enriches knowledge discovery operations on processed documents and is able to enhance concept extraction and validation. Background knowledge allows pattern abundance to be limited, and are used in pre-processing to provide a consistent lexical representation of documents.

IV. IMPLEMENTATION

Implementation is the realization of an application or execution of a plan, idea, model, design, specification, standards, algorithms or policy. In computer science, an implementation is realization of a technical specification or algorithm as a program, component or system. Many implementations may exist for a given specification or standards.

The implementation is the final and important phase of the software development. It refers to the conversations of a new system design to operations. An implementation plan is made before starting the actual implementation of the system.

Implementation includes refining the system through processing and testing, implementations aims:-
- To define the organization code in terms of implementation organized in layers.
- To implement and objects in terms of components.
- To test the developed components as units.
- To integrate the results produced by implementers into a system.

The project is implemented in a modular fashion. The various modules that constitute the software developed in the project are as follows:

1. Admin Authentication Module: This module is used to authorize admin so as to maintain the integrity of software and secure it from unauthorized users.

2. Documents Retrieval Module: This module is the primary phase of the software it concentrates in information retrieval and pre-processing of documents, here documents of various types such as HTML documents, text documents, images, word documents, video files, audio files and other documents are screened as per the specification and only textual files such as text documents, word documents, HTML files and documents with .pdf extensions are selected which can then undergo conversion process where it is transformed to a feasible data format.

3. Standard Protocols Module: This module mainly concentrates in defining extraction and derivation rules for extracting information from the pre-processed documents. The module is divided into two phases: Specification and Rule Induction.

Specification: This module facilitates the admin to define extraction protocols by providing efficient user interfaces and then stores the described extraction protocols in the database. Extraction protocols mainly include synonyms and patterns to be extracted for input documents.

<table>
<thead>
<tr>
<th>Standard Extraction Term</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActionScript</td>
<td>Action Script, Flash Action Script</td>
</tr>
<tr>
<td>C Language</td>
<td>C, ProC, Objective C</td>
</tr>
<tr>
<td>C# Language</td>
<td>C#, C Sharp</td>
</tr>
<tr>
<td>Microsoft Word</td>
<td>Word, MS Word</td>
</tr>
</tbody>
</table>

Table. 1 Specification Protocols

Rule Induction: This module provides interface to define association rules that aids automatic discovery of useful information from corpus of input text documents. Rules defined are stored in the database.
Table. 2 Rule Induction Protocols

Knowledge Extraction Module: Here the extraction protocols and association rules defined by admin in the respective interface are taken from their respective databases and processing is done using appropriate algorithms to extract relevant information and also to derive new hypothesis from extracted information.

Database Module: Here a database is framed to store and display the result, which could be used to facilitate future extraction process.

<table>
<thead>
<tr>
<th>Extracted Terms</th>
<th>Derived Terms</th>
<th>Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>C Language, C++ Language, Java Language, HTML, DHTML, B.E</td>
<td>XML, UG Scholar</td>
<td>Resume 1</td>
</tr>
<tr>
<td>Windows XP, C Language, C++ Language, HTML, SQL, Microsoft Word, BCA</td>
<td>Windows OS, UG Scholar</td>
<td>Resume 2</td>
</tr>
</tbody>
</table>

Table. 3 Layout of Developed Database

V. RESULTS AND ANALYSIS

The project “Transforming Digital Unstructured and Semi-structured Data into Structured Data with the Aid of IE and KDT” mainly concentrates in representation of huge unstructured and semi-structured digital documents in structured format i.e. in the table’s format of databases.

The analysis procedures illustrates the performance of this project by comparing the speed of transformation with respect to speed acquired in terms of manual processing. The graphs illustrate the measure of performance parameter of this project. These graphs are plotted by considering the total number of documents being processed for extraction of relevant information with respect to time taken.

The Figure. 2 below shows the graph which illustrates how this project performs transformation of unstructured and semi-structured data into structured data more quickly than if it is done manually.
VI. CONCLUSIONS

This project presents an approach that uses an automatically learned IE system to extract a structured database from a text corpus, and then mines this database with existing KDT tools. This project concentrates in utilizing the mutual benefit obtained from integrated result of information extraction and knowledge discovery from text. IE enables the application of KDT to unstructured text corpora and standard protocols can discover predictive rules useful for improving IE performance.

At the preliminary stage documents of various types are screened to separate audio, video and image files from textual documents such as HTML documents, text documents, images, word documents, video files, audio files and other documents, as per the specification and only textual files such as text documents, word documents, HTML files and documents with .pdf extensions are selected and the screened files are transformed to a feasible data format usually into text file format, which can then undergo knowledge extraction process. User Interfaces are provided to the admin to describe the specification and extraction protocols. Suitable databases are framed to store data, which are fed to interfaces by admin and a suitable database is framed to display result in structured format.

VII. FUTURE ENHANCEMENT

Text mining is relatively new research area at the intersection of natural-language processing, machine learning, data mining and information retrieval. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text corpora can be developed. In particular, the growing interaction between computational linguistics and machine learning is critical to the development of text-mining systems.

Some emerging challenges have been identified for future research:

Choice of appropriate IE and KDT Module: 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 KDT promises to increase performance (in terms of recall and precision). For instance, integrated KDT 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. KDT techniques and their measures reduce the amount of information, which in turn increases system performance.

Validating the applicability of the Software: In order to validate the methodology incorporated in the project, a corpus of job applications available on the web and curriculum vitae (eRecruitment) are used. The evaluation of project will be effected by (i) measuring the performance of a current IE process, which will be compared to (ii) an IE process resulting from the project, which additionally enables a general benchmark of the interest level of mined Information.

REFERENCES