IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue, 5, March 2019 E-ISSN: 2347-2693

Information Extraction from Unstructured Documents

R.Jayanthi[1*], D.Nirmala[2]

[1,2]Dept. of Computer Science, Quaid-e-Millath Government College for Women(A), Chennai-600 002, TamilNadu, India

Corresponding Author: srinirmala1419@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si5.146151 | Available online at: www.ijcseonline.org

Abstract -In todays scenario the organization of textual information has become a necessity due to the availability of various digital information. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, makes the information contained in the text accessible to the various data mining (statistical and machine learning) algorithm. Information Extraction is the technique of automatically extracting information from unstructured and/or semi-structured machine-readable documents. An Information Extraction system target a specific topic or domain based on the user's interest and searches for information that has more reliance to the domain. Information Extraction tools make it possible to pull information from text document, database, websites or multiple sources. Information Extraction depends on named entity recognition, a sub-tool used to find targets information to extract. This paper presents the review of various Information Extraction techniques such as Supervised, Unsupervised and Semi-supervised Information Extraction and its application.

Keywords: Text Mining, Information Extraction, Machine Learning, Supervised, Unsupervised, Semi-supervised.

I.INTRODUCTION

Information Extraction (IE) is the automated retrieval of specific information related to a selected topic from a body text document, database, websites or multiple sources. IE may extract information from unstructured, semi-structure or structured machine-readable text. Usually, however, IE is used in Natural Language Processing (NLP) to extract structured from unstructured text. The information extraction software infers the relations between all the identified people places and time to deliver the user with significant information. This technology can be very applicable when dealing techniques. Traditional data mining assumes that the information to be mined is already in the form of a relational database [1]. Unfortunately, for many applications electronic information is only obtainable in the form of free natural language documents rather than structured database [2]. The main objectives of information extraction is to find specific data or information extraction is to find specific data or information in natural language text.

This information is stored in database like patterns and it can be available for further use. IE convert a quality of textual documents into more structured database [3]. The major task of information extraction is Term Analysis, Named Entity, Fact Extraction. Term Analysis identifies the term; the term may contain one or more words. It can be helpful for extracting information for document. Named Entity Recognition identifies the textual information in a document relating the names of person, place, organization or product. Fact Extraction identifies extract the complex fact from the document [4].

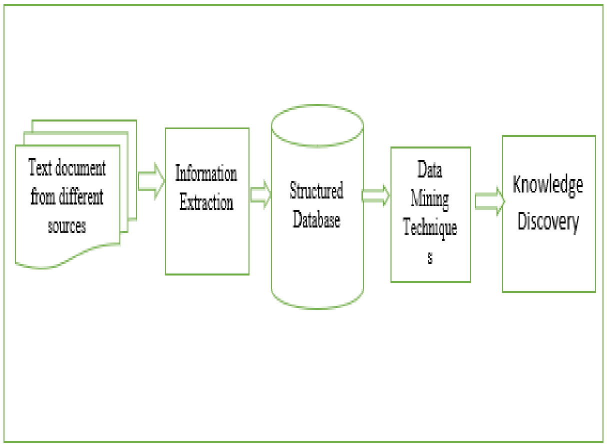

Fig 1: Information Extraction

The unstructured data in the different sources are collected, Information Extraction techniques are used to extract the structured data stored in the structured data base. And then use the data mining techniques are getting the knowledge data as shown in Fig-1.

This paper presents the specific techniques and application for Information Extraction.

II.TECHNIQUES OF INFORMATION EXTRACTION

The techniques in information extraction from machine learning areas such as supervised information extraction, unsupervised information extraction, semi-supervised information extraction.

2.1. Supervised Learning

Supervised learning is the Artificial Intelligence (AI) of data mining task of inferring a function from training data. The training data consist of a set of pre-classified training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value. In supervised learning, example X, Y (pre-classified training examples). Given an observation X, what is the best level for Y. They have been used for a wide variety of tasks, such as textual entailment question answering and knowledge base papulation [5].

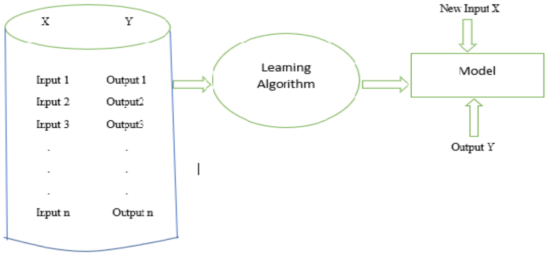

Fig 2: Supervised Learning

In supervised learning have a number of training instances. Each instance comprises of the input and output. So, this is the first training instances, second training instances, third training instances and nth training instances given all the training instances the learning algorithm will comes with the model and this model can be used to classify and to find the output are corresponding Y value for a new appropriate X. So, given a new input X and then use the model to discover the output Y as shown in Fig-2.

This paper analyzes various information extraction techniques. Table1 shows the comparative analysis of supervised information extraction and also discuss about the research challenges in those areas.

Table1: Comparative Analysis of Supervised Information Extraction

S. No	Research Paper Name	Techniques Used	Best Techniques	Research Challenges
1	A Supervised Learning Algorithm for Information Extraction from Textual Data	Self-supervised learning, WOE’s extractors.	Self-supervised learning.	This paper produce extends WorldNet using facts extracted from Wikipedia categories. It only targets a limited number of predefined relations.
2	Supervised Open Information Extraction	bi-LSTM transducer, Rule based algorithm, (QA-SRL) apply QAMR techniques.	(QASRL) apply QAMR techniques.	Released Question-Answer Meaning Representation dataset can be automatically converted into an Open IE corpus which significantly increases the amount of available training data.
3	Identifying Relations for Open Information Extraction	Extraction algorithm.	Relation and argument extraction.	Easy-to-enforce constraints on binary, verb-based relation phrases in English that ameliorate these problems and yield richer and more informative relations.
4	Web-Scale Information Extraction in KnowItAll (Preliminary Result)	Hidden Markov Models, Rule Learning, Ontology, Generic rule	Ontology, Generic rule	KNOWITALL is an autonomous system that extracts facts, concepts, and relationships from the web.
5	A Machine Learning Approach to Information Extraction	Several Machine Learning, regular expression.	Supervised Machine Learning.	information extraction is done by a combination of regular expressions and text classifiers.it is not possible to extract the information expressed in an implicit way

2.2. Unsupervised Learning

Unsupervised Information Extraction (UIE) is the task of extracting knowledge from text without the use of hand-labeled training examples. Because UIE systems do not require human intervention, they can recursively discover new relations, attributes, and instances in a scalable manner. IE without hand-labeled examples is referred to as Unsupervised Information Extraction (UIE) [6]. UIE system such as KnowItAll [7,8,9] and TextRunner [3,4] have demonstrated textual patterns can perform UIE for millions of divers. UIE system such as KnowItAll [7,8,9] and TextRunner [10,11] have demonstrated textual patterns can perform UIE for millions of diver.

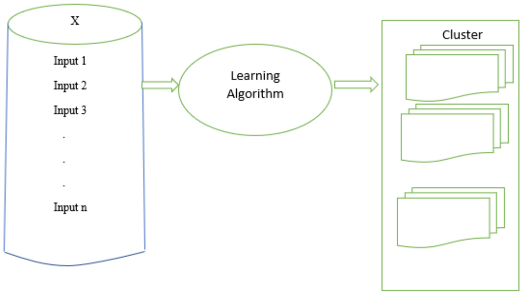

Fig:3 Unsupervised Learning

In Unsupervised learning only have X's. we have different X's that are X1, X2, X3, Xn this is the data and the learning algorithm will produce clusters will group this data. So, it's based on the similarity of the data items each other. We can find out the certain groups among the data as shown in Fig-3.

Table 2 shows the comparative analysis of unsupervised information extraction and also discuss about the research challenges in those area.

Table3: Comparative Analysis of Unsupervised Information Extraction

S. No	Research Paper Name	Techniques Used	Best Techniques	Research Challenges
1	Automatic Semantic Annotation using Unsupervised Information Extraction and Integration	Information extraction, machine learning, wrapper extraction.	Information extraction.	In this paper propose a methodology to learn how to annotate semantically consistent portions of the web extraction and integration information from different sources.
2	Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web	Unsupervised learning, Wrapper methods, such as HLRT wrappers, Natural language-based techniques, such as Amilcare	Unsupervised learning	This assumption by exploiting reference sets to aid the extraction. These reference sets are chosen by the algorithm, removing the need for any human intervention.
3	Unsupervised Learning of Field Segmentation Models for Information Extraction	hidden Markov models (HMMs), general unsupervised HMM learning, HMMs using EM algorithm.	Unsupervised HMM learning.	This paper introduced in the MUC evaluations for the task of finding short pieces of relevant information with in a broader text that is mainly irrelevant, and returning it in a structured form.
4	A Probabilistic Model of Redundancy in Unsupervised Information Extraction	Supervised and Unsupervised method.	Unsupervised method.	This paper introduced a combinatorial URNS model to the problem of assessing the probability that an extraction is correct.
5	Unsupervised Discovery of Scenario-Level Patterns for Information Extraction	Supervised Information Extraction, Pattern Matcher.	Supervised and pattern matcher	An incremental discovery procedure to identify new patterns. We present experimental which show that the resulting patterns exhibit high precision and recall.

2.3. Semi-supervised Learning

The World Wide is a large reservoir of information that is still growing at a rapid rate. Unlike data found in a database, the vast majority of the web documents are only semi-structured, making information retrieval a challenging task [12]. In recent years, there has been a surge of interest in extracting structured information from Web documents and the extracted data into database objects [13,14,15,16].

It is beneficial to extract information from queries in a format that is consistent with the backend data structure. As one step more this object, this paper focuses on the problem of query tagging which is to allocate each query term to a pre-defined category. This problem could be approached by learning a conditional random field (CRF) model (or other statistical models) in a supervised fashion, but this would require substantial human-annotation effort [12].

Supervised methods for relation extraction perform well on the ACE Data, but they require a large amount of manually labeled instances [17]. Unsupervised methods do not need the definition of relation types and manually labeled data, but they cannot detect relations between entity pairs and its result cannot be directly used in many NLP tasks since there is no relation type label attached to each instance in clustering result [17]. Considering both the availability of a large number of untagged corpora and direct usage of extracted relations, semi-supervised learning methods has received great attention.

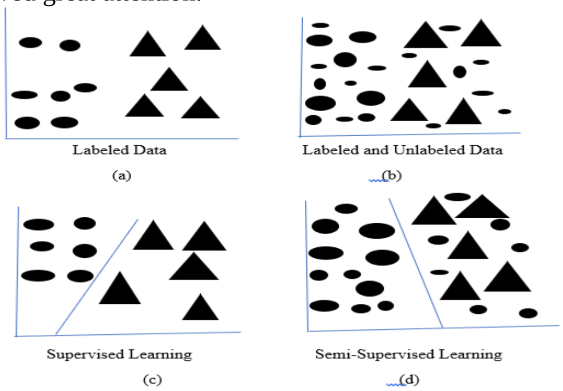

Fig:3 Semi-Supervised Learning

Semi-Supervised Learning (Fig:4), In semi-supervised leaning we have combination of label data and unlabeled data. So, this is label data belong to two different classes. So, one class is circle and other class is triangle. In semi-supervised learning apart from having data from two classes you also have unlabeled data which is indicated by the small circle.

Table 3 shows the comparative analysis of semisupervised information extraction and also discuss about the research challenges in those area.

Table 3: Comparative Analysis of Semisupervised Information Extraction

S. No	Research Paper Name	Techniques Used	Best Techniques	Research Challenges
1	Extracting Structured Information from User Queries with Semi-Supervised Conditional Random Fields	Condition random field (CRF) model in a supervised fashion, semi-supervised learning method for CRFs.	Semi-supervised learning method for CRFs.	presented semi-supervised CRFs that incorporate derived label information from additional resources.
2	Semi-supervised Relation Extraction with Large-scale World Clustering	statistical methods, kernel-based method.	Statistical and kernel method.	This paper described a semi-supervised relation extraction system with large-scale word clustering. We have systematically explored the effectiveness of different cluster-based features. we also demonstrated that the two proposed statistical methods are both effective and efficient in selecting clusters at an appropriate level of granularity through an extensive experimental study.

3	Aspect Extraction through Semi-Supervised Modeling	ME-SAS, ME-LDA, DE-LDA.	ME-SAS, ME-LDA, DE-LDA.	Our results showed that both models outperformed two state-of-the-art existing models ME-LDA and DE-LDA by large margins.
4	Improving Semi-Supervised Acquisition of Relation Extraction Pattern	kernel methods, iterative extraction pattern, semi-supervised learning,	Semi-supervised, iterative extraction.	Semi-supervised approaches to IE pattern acquisition benefit from the use of more expressive extraction pattern models since it has been shown that the performance of the linked chain model on the relation extraction task is superior to the simpler SVO model.
5	Coupled Semi-Supervised Learning for Information Extraction	Coupled Pattern Learner, Coupled SEAL, Bootstarp Extraction.	Bootstarp Extraction.	This empirical evidence leads us to advocate large-scale coupled training as a strategy to significantly improve accuracy in semi-supervised learning.

III.APPLICATION OF INFORMATION EXTRACTION

In this section, we introduce several extraction applications.

A. Information Extraction in Digital Libraries

In digital libraries (DL), “metadata” is structured data for serving users find and process documents and images. With the metadata information, search engines can retrieve required documents more exactly. Scientists and librarians need use greatly manual efforts and lots of time to create metadata for the documents. To reduce the hard labor, many efforts have been made toward the unmanned metadata generation based on information extraction. Here we take Cite seer, a popular scientific literature digital library [18].

B. Information Extraction from Emails

We also make use of information extraction methods to email data (Tang, 2005a). Email is one of the standards means for communication via text. It is approximate that an average computer user receives 40 to 50 emails per day (Ducheneaut, 2001). Many text mining applications need take emails as input, for example, email analysis, email routing, email filtering, information extraction from email and newsgroup analysis.

Unfortunately, information extraction from email has received little notice in the research community. Email data can contain different types of information specifically, it may contain headers, signatures, quotation, and text content. Furthermore, the text content may have program codes. Lists, and division; the header may have metadata information such as sender, receiver, subject, etc.; and the signature may have metadata information such as author name, author’s position, author’s address, etc. [18].

C. Information Extraction to improve Document Retrieval

As we have mentioned previously, one of the possible uses for FASTLETS is to improves the quality of document retrieval (DR) results. We discuss some of our past and current work, as well as future plans [19].

SRI’s MUC-6 system was used to recorder the retrieval results from the UMASS Inquery ad-hoc query system, based on the results of finite state patterns matching. This experiment produced a positive result, which, while far from being definitive, suggested that further investigation should be performed. Of course, the scenario that was being tested is not realistic, as such highly developed IE system will not generally exist for most information needs. A more reasonable scenario would be one in which a rapidly developed FASTLET is used to perform such a task [19]

IV. CONCLUSION

In this paper various techniques and methods are discussed for Information Extraction. In addition of that the more efficient processing of information algorithms is also cultured. Due to observation a promising approach is obtained given in. According to the analyzed methods an enhancement over this is submitted. In future the proposed techniques are implemented using JAVA technology and the comparative results are provided.

REFERENCES

[1] K. Thilagavathi, V. Shanmuga Priya, “A SURVEY ON TEXT MINING TECHNIQUES”, International Journal of Research in Computer Applications and Robotics, ISSN 2320-7345.

[2] Vishal Gupta, Gurpreet S. Lehal, “A SURVEY OF TEXT MINING TECHNIQUES AND APPLICATION”, Journal of Emerging technologies in web intelligence, Vol.1, No. 1 August 2009.

[3] R. Sagayam, S. Srinivasan, S. Roshni, “A Survey of Text Mining: Retrieval, Extraction and Techniques, “International Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5, ISSN 2250-3005(online), September|2012.

[4] R. Janani, Dr. S. Vijayarani, “TEXT MINING RESEARCH: A SURVEY”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016.

[5] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan, “Supervised Open Information Extraction”.

[6] Doug Downeva, Oren Etzionib, Stephen Soderlandb, “Analysis of a Probabilistic Model of Redudancy in Unsupervised Information Extraction”

[7] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. “Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence, 165(1):91-134, 2005.

[8] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. “Methods for domain-independent information extraction from the web: An experimental comparison.”, In Procs. Of the 19th National Conference on Artificial Intelligence (AAAI-04), pages 391- 398, San Jose, California, 2004.

[9] D. Freitag and A. McCallum, “Information Extraction with HMMs and shrinkage”, In proceedings of the AAAI-99 workshop on machine learning for information Extraction, Orlando, Florida, 1999.

[10] M. Banko, M. Cafarello, S. Soderland, M. Breadhead and O. Etzioni, “Open information extraction from the web, In procs, of IJCAI, 2007.

[11] M. Banko and O. Etzioni, “the tradeofits between traditional and open relation extraction”, In proceedings of ACL, 2008.

[12] Xiao Li, Ye-Yi Wang, Alex Acero, “Extracting Structure Information from User Queries with Semi-Supervised Conditional Random Fields”.

[13] C. Barr, R. Jones, and M. Regelson, “The linguistic structure of English web-search queries”, In proceedings of the 2008 conference on Empirical marhods in Natural Language Processing, page 1021-1030, 2008.

[14] P. Viola and M. Narasimhand, “Learning to extract information from semi-supervised text, using a discriminative context free grammer, In SIGIR’05: proceedings of the 28th annual International ACM SIGIR conference on Research and development in information retrieval, page 330-337, 2005.

[15] J. Zhu, B. Zhang, Z. Nie, J-R, wen, and H.W. Hon, ”Webpage understanding: an intergrated approach”, In proceeding of the 13th ACM SIGKDD international conference on knowledge Discovery and Data Mining, pages 903-912, 2007.

[16] T.-L. Wong, W. Lam, and T.-S. wong, “An unsupervised framework for extracting and normalizing product attributes from multiple websites”, In proceedings of the 31st annual International ACM SIGIR conference on Research and development in Information Retrieval, pages 35-42, 2008.

[17] Jinxiu Chen, Donghong Ji, Chew Lim Tan, Zhengyu Niu, “Relation Extraction Using Label Propagation Based Semi-Supervised Learning”.

[18] Jie Tang, Mingcal Hong, Duo Zhang, Bangyong Liang, and Juanzi Li, “Information Extraction: Methodologies and Applications”.

[19] Andrew Kehler, Jerry R. Hobbs, Douglas Applet, John Bear, Matthew Caywood, David Israel, Megumi Kameyama, David Martin, and Claire Monteleoni, “Information Extraction Research and applications: current progress and future directions”.