Computers in Industry 100 (2018) 137–147

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Industrial information extraction through multi-phase classification using ontology for unstructured documents

K. Rajbabu[a,*], Harshavardhan Srinivas[b], S. Sudha[b]

[a] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, TamilNadu, India
[b] Department of Electrical and Electronics Engineering National Institute of Technology, Tiruchirappalli, TamilNadu, India

ARTICLE INFO

Keywords:
Information extraction
Multi-phase classification
Ontology
Industrial unstructured documents
Feature transformation

ABSTRACT

The increased availability of unstructured text documents in industries such as e-mails, office documents, PDF files etc., has inspired many researchers towards Information Extraction. The objective of the proposal is to extract information from unstructured tender documents of power plant industries. The extraction efficiency of recent works depends on the linguistic structure and keyword taxonomy. Hence, these approaches are unsuitable for domain specific applications that demand semantic and contextual taxonomy together. In this paper, a two-phase classification approach for information extraction with feature weighing is proposed. The proposal performs sentence classification in first phase followed by word classification. As industries spans across multiple domains, a multi domain layered industrial ontology is used for knowledge representation. The unstructured documents are enhanced into DAG based semi-structured text with enriched features. A unique feature transformation approach based on the categorical data type of features is attempted to handle heterogeneous textual features. The proposal is evaluated with real time documents obtained from power plant tenders. The results showed minimal loss of precision which can be rectified by enriching the training data and customizing standard parser algorithms to suit the domain requirements.

1. Introduction

The availability of huge content across various text documents in the technical corpora has made the process of identifying crucial information a very tedious task. Recent researches have revealed that 80–98 percent of digital information such as e-mails, office documents, PDF-files and many other text-based documents are unstructured [1]. Nonetheless, the increased availability of such unstructured text documents in electronic format has led to the growth of many research problems in the field of Information Extraction.

Especially, the digitalization revolution in industry has stimulated enormous thrust on information extraction to proliferate business potential and overcome competition. Some of the ubiquitous challenges in extraction of key information from voluminous unstructured text are technical data summarization, complaint extraction and analysis, suggestions, feedback and failure analysis. Many information extraction methods for such unstructured documents are proposed. Some state-of-art approaches for web documents are developed [2–4]. However, these approaches are inapplicable to unstructured data as they rely heavily on the document structure.

Eiji et al. proposed a context-sensitive topical PageRank (cTPR) method to extract key phrases and summarize the data obtained from Twitter [5]. The key phrase ranking algorithm is based on a probabilistic ranking model built by considering the factors of relevance and interestingness of key phrases. However, due to the manual work involved in rating the keywords, this algorithm is non extendable to other domains. Fuchun Peng et al. employed Conditional Random Fields (CRFs) for extracting information from the headers and citation of research papers [6]. This algorithm out performed other implementations that used Hidden Markov Model (HMM) and Support Vector Machine (SVM) based models with significant reduction in error rates. But the effectiveness of CRF models depend heavily on header information and the results tend to be futile without semantic representation. Andrea Esuli et al. presented two methods based on linear chain CRF models combined with supervised machine learning approach to extract information from radiology reports [7]. The proposed work takes the positional attributes to predict the occurrence of a concept in specific places of a text by annotation of words in sentences. Donghui Feng et al. proposed a model, highlighting the importance of semantics in data extraction tasks [8]. This model based on semantic attributes, integrates them with a sequential labeling CRF model to extract data records. However, the above approaches cannot be applied to unstructured content.

* Corresponding author.
E-mail addresses: rajbabu@bhel.in (R. K.), harshsrinivas@gmail.com (H. Srinivas), sudha@nitt.edu (S. S.).

https://doi.org/10.1016/j.compind.2018.04.007
Received 13 October 2017; Received in revised form 30 March 2018; Accepted 10 April 2018

John Atkinson-Abutridy et al. proposed an information extraction approach involving genetic algorithm along with the usage of semantic and generic heuristic rules to optimize the extraction process [9]. However, the usage of semantic and rhetorical information is limited in the information extraction process.

Eric Tsui et al. proposed an extraction algorithm for intellectual capital-related information from unstructured financial documents by integrating rule-based reasoning and case-based reasoning techniques [10]. Siti Mariyah et al. employed multiple approaches to extract key information from financial documents [11]. In the first approach, a rule-based model was defined using orthographic, layout, and contextual features. The second approach used supervised classification algorithm to extract specific features from documents. However, rule-based reasoning is problematic when it comes to complex domains since thousands of rules are to be defined. Moreover, in situations where the rules depend heavily on the knowledge base, a huge problem arises when new knowledge is introduced. Jia-Lang Seng et al. proposed an algorithm based on intelligent word segmentation, Part Of Speech (POS) tagging and Named Entity Recognition (NER) to extract financial data from business valuation [12]. However, this algorithm is suitable only for structured documents.

Hui Han et al. proposed a SVM classification based method for extracting metadata from the header of research papers [13]. Chunguo Wu et al. defined a data preprocessing technique to extract keywords from documents using SVM [14]. The methodology of incorporating knowledge models and their transformation into analytical model for unstructured information extraction is discussed [15]. Though, the knowledge base developed is based on problem requirement and practical applicability, their results are undisclosed. Logistic regression and Naive Bayes are proved to provide competing results in the biomedical field with unstructured documents [16]. However, the approach is not applicable for domain oriented extraction as it ended with imprecise results. This is because of the usage of the orthodox term document inverse document frequency (td-idf).

WentaoCai et al. employ an ontology based conceptual information retrieval approach combined with approximate graph matching methods to extract location based information from the web automatically [17]. However, transforming plain texts into graphs is not an easy task, since concepts or their relationships are to be identified automatically. Moreover, graph matching is a tedious problem (NP-complete). These two drawbacks obstruct the application of graph structures in information retrieval.

Legal text analytics work on contracts have focused on classifying entire lines, sentences, or clauses, using smaller datasets or fewer classes [18–20]. Though the work has considered segmenting legal (mostly legislative) documents and recognizing named entities, they are not directly applicable in extracting contract elements. For example, they employ hand-crafted features, patterns, or lists of known entities that are to be tailored for contracts. Moreover, word (POS tag) embedding is left unconsidered.

Contract document analysis in industries have gained more importance due to its direct business impact in terms of finance and reputation. Reducing cycle time and improving the precision of summarizing the requirements from huge document set of tenders are the prime concerns in the competitive scenario. Extraction of contract elements such as contract title, start date, end date is done using hand crafted features, post processing rules and machine learning techniques [21]. However, this approach is applicable for similar contract documents. Further, when classification is to be carried out for large number of dataset, rule definition is cumbersome. The proposed approach uses multi domain ontology as knowledge base instead of hand crafting features, rules and applies multi class classifier for extracting attributes. Fine tuning of the result is done through feature weighing instead of the proposed sliding window approach.

Information retrieval from design document is carried out by associating domain ontology with the syntactic document representation generated from the document structure [22]. However, the lexical knowledge is not used and further it does not handle disambiguated context of the domain. Diana et al. proposed extraction of specific entities of interest for market analysis using domain ontology [23]. The approach uses the taxonomical structure and grammatical rules for extraction which is cumbersome for extensibility. Further the above approaches do not leverage the possibility of extending the domain knowledge by integrating with lexical and contextual elements of the domain.

Ontology based information extraction in construction tenders used document structure ontology which uses only the document structural pattern [24]. Mining service contracts using linguistic patterns are also explored [25]. Linguistic patterns help broader classifications such as phrases but, not deep classification of attributes with overlapping features.

Domain identification using multi domain ontology through keywords and synonyms is proposed [26]. However, it supports only web pages. An ontology based solution for word sense disambiguation (WSD) through a semi-supervised knowledge graph model within a domain was suggested [27]. However, it does not resolve contextual ambiguity of semantically similar text across multiple relevant domains.

From the literature, the existing supervised approaches of information extraction, particularly in industries use either linguistic, keyword, document or domain knowledge independently as the knowledge base. Moreover, the semantic and contextual taxonomy together are never considered in recent algorithms which is also an essential component of IE. Hence, it is evident that these approaches are quite unsuitable for industrial applications that demand multi domain knowledge in addition to linguistic, keyword and document structure. Further, the present supervised approaches are based on single level classification while the proposal is a multi-phase iterative classification approach aimed to support domain descriptions and contextual information along with document structure.

2. Problem statement

The goal of the proposal is to provide an end-to-end information extraction solution in the fields of design, manufacture and commissioning of power plants from unstructured tender documents. The nature of the various documents in the power plant industry, their functionalities and the information to be extracted from the respective documents are presented in Table 1. Almost 90% of the documents in this table are common to many industries. Extraction of potential allied businesses, financial implications due to policies and regulations, aggressiveness in competition from news articles are some of the most

Table 1
Information Extraction Challenges in power plant industry.

Functionality	Documents Extraction requirements
Engineering	Proposal, Tenders, Specification, Research articles Design parameters, Rules, exceptions, Failure Analysis
Production	Preventive, Correction, quality & Inventory plan, Operation manual Stock Analysis, Inspection prediction, Optimize utilization
Commercial	News articles, Guarantee conditions, schedule plan, deviation, bidding document Order prediction, guarantee analysis, deviation analysis, cost analysis
Finance	Taxation guidelines, Govt. policies, Budgeting Taxation analysis, budget and Expenditure analysis
Logistic	Load Analysis, Routing Procedure, safety Reports Route optimization, Load optimization
Business Analyst	New Articles, Technical presentations, Govt. Plans and Policies, Competitor Information Market Prediction, Diversification Analysis