IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.
www.ijiset.com
ISSN 2348 – 7968

Information extraction from Material Publications

Prof RushaliPatil,Amoolya S Kumar, AdityaCahaturvdi, Fiona Khatana, V Krishna
Department of Computer Science,Department of Computer Science,Department of Computer Science,Department of Computer Science,Department of Computer Science

Abstract—***Experiments in material engineering are not just costly to replicate but also time consuming. Therefore, material scientists prefer to analyze results that have already been derived and published in journals or papers pertaining to material science. Our project aims at extracting information from these papers and journals and providing them to the material scientist to save them the time and effort of going through all the papers related to their query. We provide a specific search box like interface to the user and mine the results from a database consisting solely of material engineering papers.***

.

***Index Terms*—Data mining,Data retrieval, Information extraction, Material publications.**

I. INTRODUCTION

Materials engineering deals with materials having different compositions, processing them through a sequence of manufacturing processes, evolution of their microstructure at various length scales and finally their properties. Researchers and designers in this community are often interested in questions of the following nature: how do material properties and structure change when we vary the material composition or subject it to a set of manufacturing processes?; and to achieve a certain set of properties or structure in a material, what composition of the material and which processes (along with their parameter set points) should be used? A large amount of information to answer these questions is present in materials publications in the form of models, research results, experimental results, and so on. Our objective is to mine knowledge about material compositions, their properties, processes and their parameters, process-structure-property relations, and so on from these publications. Some of this information may be present at sentence level, some at paragraph level, while some others may be in the form of tables, figures and images. Related pieces of information may be present in contiguous sentences/paragraphs/sections or they may be spread apart.

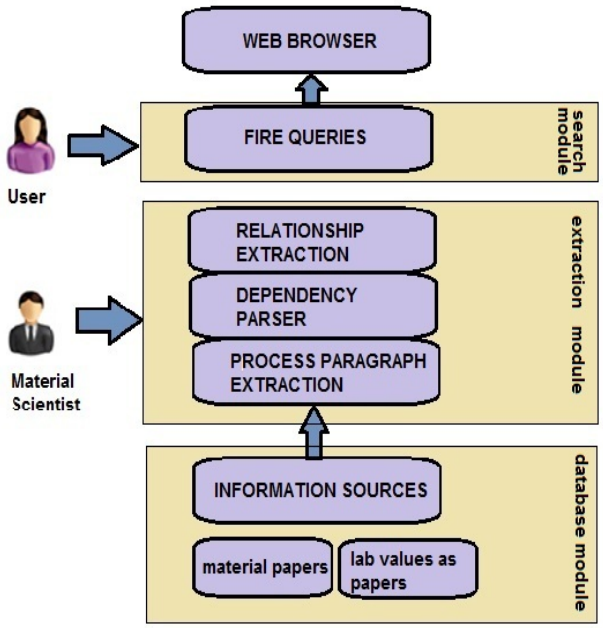
For this project, we limit our scope to the extraction of manufacturing processes and their parameters. The first step is to extract the entities of interest such as processes, parameters, values and units. This step mainly looks at the sentence level content. We use various entity extraction techniques such as patterns, rules, statistical techniques etc as appropriate for different entity types. Once the entities are identified, the next step is to find relations among the entities. The relations of interest are: process-parameter, parameter-value and value-unit.

II. LITERATURE SURVEY

The presently used technologies in text classification are involving Naive Bayes text classifier, regression trees, K means clustering algorithm etc. K means clustering algorithm was ruled out due to its poor efficiency in case of large databases. Thus classification algorithm is zeroed down to Naive Bayes. In comparison to s transformed weight normalized complement naive Bayes(TWNCNB), Standard multinomial Naive Bayes was found more suitable for the application that we are using[1]. Also for our application TF-IDF was found more suitable for our use[1]. In Naive Bayes there can be two approaches to calculate probalistic distribution of words. These methods are unigrams and bigrams. Unigrams approach will eliminate dependency of words occurring next to each other completely and will only depend on frequency of an individual word only. Some use a multi-variate Bernoulli model, that is, a
Bayesian Network with no dependencies between words and binary word features. Others use a multinomial
model, that is, a uni-gram language model with integer word counts [2]. The first supervised learning method we introduce is the multinomial Naive ayes or multinomial NB model, a probabilistic learning method [3]. The second step of the project proceeds towards working on extracted paragraphs which have been classified as process paragraphs by the Naive Bayes model. The next step is to apply dependency parsing on these paragraphs obtained. One of the most common choices would have been Stanford dependency parser. The Stanford dependencies provide a representation of grammatical relations between words in a sentence. They have been designed to be easily understood and effectively used by people who want to extract textual relations. Stanford dependencies (SD) are triplets: name of the relation, governor and dependent [4].
A stack-based nondeterministic algorithm with memory--based learning software was considered for application in this project [5]. Memory based algorithms for dependency parsing would be a suitable choice for applying dependency parsing to our project. It was chosen as it gives an attachment score of 87.1 % when experiments were conducted on the wall street journal[6]. Applying a standard algorithm to classify text into dependency trees can also be used for a simpler approach to dependency parser in our system. Unlike phrase-structure (constituency) parsers, this algorithm operates one word at a time, attaching each word as soon as it can be attached. There is good evidence that the parsing process used by the human mind has these properties [7].

ISSN 2348 – 7968

III. PROPOSED SYSTEM



SYSTEM ARCHITECTURE

Module 1:
At the first stage of design Material papers are kept in a database consisting of Material papers and lab values papers. This Information is kept together in any SQL database or read from a folder at the moment of calculation.

Module 2:

1. After fetching data from Material papers it is transferred to process extraction module of Information Extraction phase. This does the extraction of process paragraphs from the given non process and process paragraphs.
2. After fetching of these process paragraphs, they are subjected to dependency parsing stage, In this stage the relationship between a process and its corresponding parameter is extracted. This gives a clear cut relation as to which parameter belongs to which process.
3. The third stage is extraction of which value belongs to which parameter, this also requires some dependency parsing technique to judge the relationship.

After this judgment about which process and parameter are related and also which parameter and value are related, this information is stored in a database to be retrieved from later.

Module 3:
This stage is when the user uses a browser like interface to fire a query to the database. When such a query is ever fired, it searches the database that has the same match for a process, parameter and value. This can then be retrieved and displayed on the browser that is being viewed by the user currently.

IV. PROCEDURE

The first step to our task at hand is to classify the paragraphs of the journals in our database so that our search can be limited to paragraphs that describe experimental procedure and thereby, contain information relevant to our search. The paragraphs are classified into two categories: Process Description and Non Process Description. The classification was done using Naive Bayes Classification Algorithm.

Step 1: Paragraph Classification using Naive Bayes
In order to train the machine to classify paragraphs, we used a training set comprising of approximately 75 journals and a test set consisting of 40 journals. (..mention accuracy..). The paragraphs in the training set journals were manually tagged as “Process Description” and “Non Process Description”. Each word in the process description paragraph was considered as a possible feature for classifying the paragraph. The bag of words approach was used, in the sense that the words were considered to be independent of each other[1]. The prior probability of a process description paragraph was calculated as the total number of words in process description divided by the total number of words in the journals. Similarly the prior probability of non process description paragraphs was calculated. (i)<Insert formulae>. The probability of each word given a category was calculated as the total number of occurrences of the word in paragraphs of that category divided by the total number of occurrences of the word in all the journals. While testing, a paragraph is classified as process description or non process description by multiplying the prior probability of the category with the conditional probability of the words occurring in the class. The class which has the maximum probability out of the two is the category of the paragraph.

`

Feature Extraction: This step involves extracting features that play a role in differentiating the paragraphs. Feature selection reduces dimensionality by selecting a subset of original input variables, while feature extraction performs a transformation of the original variables to generate other features which are more significant.

Stopword Removal: There exist words that do not affect the classification of paragraphs such as prepositions, conjunctions, helping verbs etc. Such words must be removed before paragraph classification as they may affect the result.

Step 2: Relation extraction:
This is stage is of sentence level processing. We will be using Stanford NLP dependency viewer for finding out the relations between different words in a sentence. This viewer has been trained from a large database which gives the best accuracy in relations extraction till date. This dependency parser will find out the relationships between words in all sentences of paragraphs, which are tagged by naïve-Bayes classifier algorithm.

ISSN 2348 – 7968

For example, Steel is heated at 30 degree Celcius.
Here, heated is the root of a tree which is a process. Now as we move downwards of a tree we will get the parameters of process. And it's parameters are steel and 30C. There are two types of relations process-material and process-environment conditions like weight, length, temperature, duration, pressure, etc.

V. CONCLUSION

In conclusion we can say that the proposed system will be successfully able to search efficiently on a series of papers in a database and devise from them an extracted more efficient database that can be used for searching. This database will show a clear relation between every process and its parameter. Also it shows a relation between a parameter and its value. This storage of relationship will help in devising an optimized search algorithm and thus giving accurate and efficient results. Through this project the data scientist receiving numerous papers on a simple search will be completely eliminated and a more specific and targeted search will be achieved**.**

REFERENCES

[1] Geoffrey Holmes, Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, "Multinomial Naïve Bayes for Text Categorization Revisited "
[2] Andrew McCallum, Kamal Nigam, "A comparison of Event Models for Naïve Bayes Text Classification".
[3] The Stanford Language processing group",Naïve Bayes Text Classification".
[4] The Stanford Language processing group,"Dependency parsing"
[5] Benjamin Nash, "Implementation and Analysis of a Dependency Parser"
[6] JoakimNivre, Mario Scholz , "Deterministic Dependency Parsing of English Text"
[7] Michael A. Covington, "A Fundamental Algorithm for Dependency Parsing"