ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 10, December 2012

Review of issues in automatic labelling of formatted document

Pallavi Galgale [1], Priyanka Ahire [2], Snehal Ingavale[3], Dr. R.S. Prasad[4]

1, 2, 3 Graduate students, 4- Professor
Department of Computer Science and Engineering,
ZES's DCOER, Pune, Maharashtra, India

Abstract **- The labelling framework, which is proposed to label topic models, essentially consists of a multinomial word distribution, a set of candidate labels, and a context collection. Thus it could be applied to any text mining problems, in which a multinomial distribution of word is involved. To generate labels that are understandable, semantically relevant, discriminative across topics, and of high coverage of each topic, first extract a set of understandable candidate labels in a pre-processing step, then design a relevance scoring function to measure the semantic similarity between a label and a topic, and finally propose label selection methods. This paper presents all such issues involved in the problem of knowledge discovery using text mining. Our paper aims to review various issues described or presented by various researchers in this area.**

Keywords **– Knowledge discovery, Multinomial distribution, Semantic labelling, Frame Net.**

1. INTRODUCTION

Text mining, also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), refers generally to the process of extracting interesting and non-trivial Information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases [1]. Text mining is a new field that attempts to bring together meaningful information from natural language text. Automatic Text categorization and summarization is the process of assigning pre-defined class labels to incoming, unclassified documents. The class labels are defined based onset of examples of pre-classified documents used as a training corpus. This work comprises an automatic text categorization and labelling approach to analyze the structure of input text. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values.

2. A FRAMEWORK OF TEXT MINING

Text mining can be visualized as consisting of two phases**:** ***Text refining*** that transforms free-form text documents into a chosen ***intermediate form,*** and ***knowledge distillation*** that deduces patterns or knowledge from the intermediate form. The Intermediate form (IF) can be *semi-structured* such as the conceptual graph representation, or ***structured*** such as the relational data representation. Intermediate form can be ***document-based*** wherein each entity represents a document, or ***concept based*** where in each entity represents an object or concept of interests in a specific domain [1]. Mining a document-based IF deduces patterns and relationship across documents. A text mining analysis involves several challenging process steps mainly influenced by the fact that texts, from a computer perspective, are rather unstructured collections of words. A text mining analyst typically starts with a set of highly heterogeneous input texts. So the first step is to import these texts into one's favorite computing environment. Simultaneously it is important to organize and structure the texts to be able to access them in a uniform manner. Once the texts are

organized in a repository, the second step is tidying up the texts, including preprocessing the texts to obtain a convenient representation for later analysis. This is illustrated in **Figure 1**. This step might involve text reformatting (e.g., whitespace removal), stop word removal, or stemming procedures. Third, the analyst must be able to transform the preprocessed texts into structured formats to be actually computed [2].

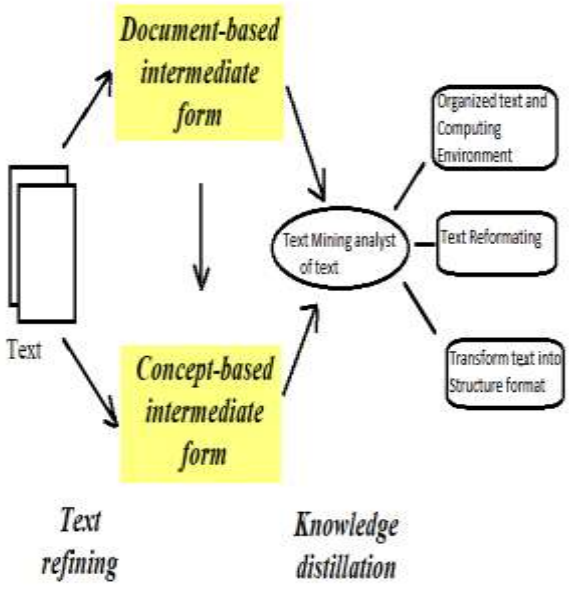

Figure 1. A text mining framework

3. PROBABILISTIC TOPIC LABELING

To generate labels that are understandable, semantically relevant , discriminative across topics, and of high coverage of each topic, we first extract a set of understandable candidate labels in a preprocessing step, then design a relevance scoring function to measure the semantic similarity between a label and a topic, and finally propose label selection methods to address the inter-topic discrimination and intra-topic coverage problems.[5] Statistical topic models are a class of probabilistic latent variable models for textual data that represent text documents as distributions over topics. These models have been shown to produce interpretable summarization of documents in the form of topics. Here we describe how the statistical topic modelling framework can be used for information retrieval tasks and for the integration of background knowledge in the form of semantic concepts [5]. We describe the special-words topic models in which a document is represented as a distribution of (i) a mixture of shared topics, (ii) a special-words distribution specific to the document, and (iii) a corpus-level background distribution.

4. AUTOMATIC LABELLING OF SEMANTIC RULE

The process of giving labels to the semantic tags is called senmantic labelingThe most widely used procedure for automatic labeling of semantic roles is the supervised machine learning technique. Approaches in supervised machine learning consist of two major steps which are shown in **Figure 2**. In the first step, the system is trained on the text where segments of texts are already correctly labeled (with semantic roles in this task). It reads the text (training input) and collects the knowledge about the occurrences of labels .In the second step, the system reads a new text for which the labels are to be automatically assigned (test input) and attempts to predict the correct label for every given segment of the text using the information available in the text and the knowledge acquired in training. The procedure usually also involves an evaluation of the performance of the system.There are various features, they are as follows [8].

A. Sentence type:

Several tasks approached by using text mining techniques, like text categorization, document clustering, or information retrieval, operate on the document level, making use of the so called bag-of-words model. Other tasks, like document summarization, information extraction, or question answering, have to operate on the sentence level, in order to fulfill their specific requirements [9].

B. Grammar:

Grammar defines the grammatical function of the constituent that realizes a particular semantic role. It is

based on the fact that some semantic roles are realized as the subject in a sentence.

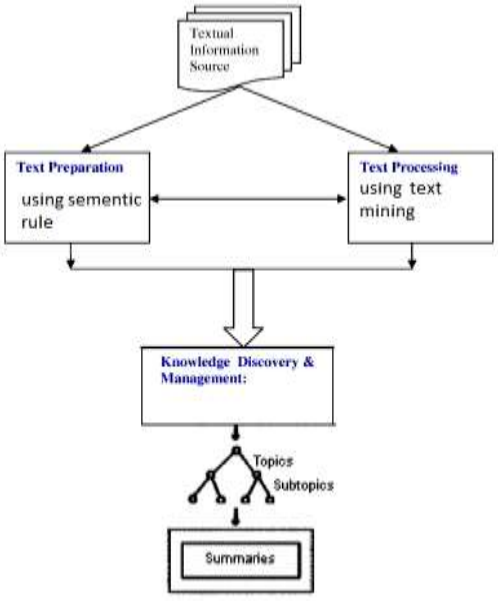

Figure2:Automatic labeling of system overview

C. Parse tree path:

Parse tree path defines the path in the syntactic tree which connects a given semantic role to its corresponding target word [9]. The value of this feature is the sequence of nodes that form the path, starting with the category of the target word and ending with the phrase that realizes the role. The direction of moving from one node to another is marked with arrows.

D. Location:

Location defines the position of the constituent bearing a semantic role relative to its corresponding target word, whether the constituent occurs before or after the target word. This is another way to describe the grammatical function of the constituent, since subjects tend to occur before and objects after the verb [9].

E. Sentence configuration:

Marking whether the verb is used as passive or active. This feature is needed to capture the systematic alternation of the relation between the grammatical function and semantic role of a constituent [9]. While agent is the subject and patient is the object in typical realizations, the reverse is true if the passive transformation takes place.

F. Root word:

Root word describes the relation between the lexical content of a constituent and the semantic role that it bears. The value of this feature is the lexical item that heads the constituent. For example, if a sentence contains two assigned semantic roles, speaker and topic, the constituent which is headed by Bill, brother, or he is more likely to be the speaker, while the constituent headed by proposal, story, or question is more likely to be the topic.[9].

5. KNOWLEDGE DISCOVERY PROCESS

The **knowledge discovery process** (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It consists of many steps (one of them is DM), illustrated in **Figure 3**, each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain. A common feature of all models is the definition of inputs and outputs. Typical inputs include data in various formats, such as numerical and nominal data stored in databases or flat files; images; video; semi-structured data, such as XML or HTML; etc. The output is the generated new knowledge — usually described in terms of rules, patterns, classification models, associations, trends, statistical analysis, etc.

KDP model consists of eight steps, which are outlined as follows:

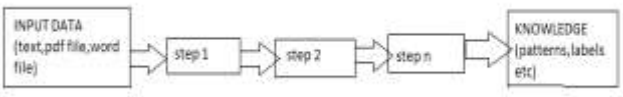

Figure: 3 Knowledge Discovery Process

1. Developing and understanding the application domain.

This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.

2. *Creating a target data set*.

This step usually includes querying the existing data to select the desired subset.

3. *Data cleaning and pre-processing*.

This step consists of removing outliers, missing values in the data, and known changes.

4. *Data reduction and projection*.

This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

5. *Choosing the data mining* task.

Here the data miner matches the goals defined with a particular DM method, such as classification, regression, clustering, etc.

6. *Choosing the data mining al*gorithm.

The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.

7. *Data mining*.

This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.

8. *Consolidating discovered knowledge*.

The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting.

CONCLUSIONS

In this paper, we formally study the problem of automatic labeling of multinomial topic models, and propose probabilistic approaches to label multinomial word distributions with meaningful phrases.

REFERENCES

[1] Heng Mui Keng Terrace, "Text mining: the state of art and challenges", Ah-Hwee Tan Kent Ridge Digital Labs, Singapore 119613

[2] Feldman, R. & Dagan, I," Knowledge discovery in textual databases (KDT)", proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, pp.112-117.

[3] Martin Scaiano, "Machine Learning for Automatic Labelling of Frames and Frame Elements in Text", School of Electrical Engineering and Computer Science University of Ottawa, Ottawa, ON, K1N 6N5, Canada

[4] Prajna Bodapati, Shashi Mogalla,"Document Clustering Technique based on Noun", Dept. Of CS&SE, College Of Engineering, Andhra University, Visakhapatnam, AP, India.

[5] Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai, "Automatic Labelling of Multinomial Topic Models", Department of Computer Science University of Illinois at Urbana Champaign Urbana, IL 61801

[6] Richard Fulton," CS229 Final Project Using Word Net and Clustering For Semantic Role Labelling" December 14, 2007

[7] Rion Snow, Daniel Jurafsky, Andrew Y. Ng, "Learning syntactic Patterns for automatic hypernym discovery" Computer Science Department Stanford University Stanford, CA

[8] Dr. Paola Merlo ,"Semantic roles in natural language processing and In Linguistic Theory Supervisor:" University of Geneva, Department de Linguistique October 9, 2009

[9] Gaurav Mishra, Siddharth Shimpi, Laxmi Bewoor,"Knowledge Discovery Using Semantic Labelling" Department of Computer Engineering, BRACT's Viswakarma Institute of Infromation Technology