ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Text Classification Using Symbolic Data Analysis

Sangeetha N[1]

Lecturer, Dept. of Computer Science and Applications, St Aloysius College (Autonomous), Mangalore, Karnataka, India.[1]

ABSTRACT: In the real world, an operational text classification system is usually placed in the environment where the amount of human-annotated training documents is small in spite of thousands of classes. In this environment text classifier are probably the most appropriate methods for the practical systems rather than other complex learning models. Text classifiers are basically used for free flowing texts that are basically unstructured text documents and classification is done with a statistical feature weighting method which involves a pre-processing- a method wherein texts are reduced by eliminating digits, punctuations, hyphens, stop words and high/low frequency words and by applying stemming. This strategy of text classification cannot be applied to the domain of unstructured texts describing the advertisements, since these texts give the description in terms of attribute values. Since none of the text classifiers are useful in classifying such texts in an unstructured text document, the concept of symbolic data analysis is introduced. Symbolic Data Analysis (SDA) is a new domain in the area of knowledge discovery and data management, related to multivariate analysis, pattern recognition, databases and artificial intelligence. In this method of Symbolic Data Analysis for classification of unstructured text documents, uses a symbolic database and querying processes are proposed. From the proposed technique it seems that it is one of the efficient techniques to classify texts in unstructured text documents and hence is introduced for the better result when dealing with unstructured text documents.

KEYWORDS text classification, symbolic data analysis, stemming, keyword extraction

I. INTRODUCTION

Text is an important and rich resource of data, information and knowledge. Text classification is the assignment of free text documents to one or more predefined categories based on their contents.

Here we focus on text categorization, which is the process of organizing a set of documents into different categories. The goal of classification is to build a set of models that can correctly predict the class of the different objects [1]. An important issue in text categorization is how documents are represented, and how features can be extracted from them which can be used for categorization. As the volume of information available on the internet and corporate intranets continues to increase, there is a growing need for tools helping people better find, filter and manage these resources. Text categorization, the assignment of free text documents to one or more predefined categories based on their content, is an important component in many information management tasks such as real time sorting of e-mail or files into folder hierarchies, topic identification to support topic-specific processing operations, structured search and/or browsing ,or finding documents that match long term standing interests or more dynamic task based interests .there are several similar tasks such as text filtering and routing. All of the above mentioned tasks require text classifiers that decide which class is more relevant to user interest. Thus, text classifier should be able to rank categories given a document & rank document given class. In many contexts trained professionals are employed to categorize new items which are very time consuming error prone and costly, thus limiting its applicability. Consequently there is an increasing interest in developing technologies for automatic text categorization.

The input to these methods is a set of documents (i.e., training data), the classes which these documents belong to, and a set of variables describing different characteristics of the documents. An important issue in text categorization is how documents are represented, and how features can be extracted from them which can be used

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

for categorization. A standard document representation is a vector of term occurrences, as used in the information retrieval field. Feature selection is used to extract a set of features which will aid in categorizing a document [2].

Applications:

Text classification techniques are used today in a variety of problems. Text classifiers can be used for classifying documents and web pages .Text classification plays an important role in certain business applications, such as content management, search and retrieval, user profiling and customer relationship management. Text classification can be used to organize information so as to make it navigable. Text categorization can help locate information, or route customer complaints. It can used to eliminate irrelevant search engine results, by ensuring that web pages are about the user's desired topic. In such an environment text classifiers are probably the most appropriate methods for the practical system rather than other complex learning models [3]. Text classifiers are basically used for free flowing text documents that are basically an unstructured text. Text classification for such unstructured text is done with a statistical feature weighting method. To get the better result, text documents are pre-processed. Under pre-processing, high dimensionality of text are reduced by eliminating digits, punctuations, hyphens, stop words and high/low frequency words and by applying stemming. It also covers reduction technique used for elimination of synonyms due to which we get the effective result. This strategy of text classification cannot be applied to the domain of unstructured texts describing the advertisements. This is because, the advertisement texts gives the description in terms of the attribute values [4]. For example, a classified matrimonial page as shown below gives the description of two kinds/classes of people interested in an alliance/marriage proposal. Consider an unstructured document, a classified matrimonial advertisement as follows:

"Alliance invited for tall, v_fair, beautiful girl- 172/ 28. MBA (London). Presently working in Mumbai. Belongs to well settled Punjabi business family, looking for smart, well educated/ well settled boy in business/ profession."

Here, each word acts as a value of an attribute of some pre-defined schema. These attribute values may belong to single-valued/crisp, multi-valued, qualitative, quantitative, category or an interval. For eg: tall, v_fair, beautiful features the looks of a bride and are termed as multi-valued attributes. Age is a crisp/single-valued attribute.
Hence the attributes in this document can be summarized as follows:

1	2
Looks	Tall, v_fair, beautiful
Height	172
Age	28
Qualification	MBA(London)
Occupation	Working in Mumbai
Caste	Punjabi

Seeks a groom with:-

1	2
Looks	Smart
Qualification	Well educated

Such type of unstructured documents can be represented efficiently as symbolic databases. Hence it is classified using Symbolic Data Analysis using a symbolic database, wherein each attribute has a symbolic meaning .Thus

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

the introduction of the concept of symbolic data analysis for an unstructured document is done here where the texts in the document gives the description in terms of attribute values.

Symbolic Data Analysis

Symbolic Data Analysis (SDA) is a new domain in the area of knowledge discovery and data management, related to multivariate analysis, pattern recognition, databases and artificial intelligence. SDA allows a more realistic description of the input units by taking into consideration their internal variation and their complex structure. SDA provides a better explanation of its results by an automatic interpretation which is closer to the user's natural language. SDA provides tools suitable for managing complex, aggregated, relational, and higher-level data described by multi-valued variables, where the entries of a data table are sets of categories, intervals, or probability distributions, which are usually related by logical rules and taxonomies. When observations in large data sets are aggregated into smaller more manageable data sizes, the resulting descriptions of the new units invariably involve ``symbolic data". By symbolic data, we mean that rather than a specific categorical or numerical value, an observed value can be a set of categories or numbers, an interval or a probability distribution or any kind or more complex information than the usual one. Hence, Symbolic Data Analysis generalizes classical methods of exploratory, statistical and graphical data analysis to more complex data issued from huge Conventional Data Bases [5].

Symbolic Data Tables

The input of SDA:

Symbolic Data Tables and Rules

Columns of the input data table correspond to symbolic variables which are used in order to describe a set of units called individuals. Rows are called symbolic descriptions of these individuals because they are not, as usually, only vectors of single quantitative or categorical values [5]. The cells of this symbolic data of different types, in particular:

(a) A single quantitative value:

For instance, if 'height' is a variable and w is an individual: height(w) =165.

(b) A single categorical value:

For instance, caste(w) = Brahmin.

(c) A set of values or categories (multi-valued variable):

For instance, height(w)={168,170,172} means that the height of w can be either 168 , 170 or 172 cms .

(d) An interval:

For instance age_range(w)=[23,26] means that the age varies in the interval [23,26].

II. RELATED WORK

Consider the problem of automatically classifying text documents. This problem is of great practical importance given the massive volume of online text available through the World Wide Web, Internet news feeds, electronic mail, corporate Databases, medical patient records and digital libraries. Existing statistical text learning algorithms can be trained to approximately classify documents, given a sufficient set of labelled training examples [1]. These text classification algorithms have been used to automatically catalog news articles (Lewis & Gale, 1994; Joachims,1998) and web pages (Craven, DiPasquo, Freitag, McCallum, Mitchell, Nigam, & Slattery, 1998; Shavlik & Eliassi-Rad, 1998) and automatically learn the reading interests of users (Pazzani, Muramatsu, & Billsus, 1996; Lang, 1995) automatically sort electronic mail (Lewis & Knowles, 1997; Sahami, Dumais, Heckerman, & Horvitz, 1998).One key difficulty with these current algorithms is that they require a large, often prohibitive, number of labeled training examples to learn accurately. Labelling must often be done by a person; this is a painfully time-consuming process.

Take, for example, the task of learning which UseNet newsgroup articles are of interest to a particular person reading UseNet news. Systems that alter or pre-sort articles and present only the ones the user ends interesting are highly desirable, and are of great commercial interest today. Work by Lang (1995) found that after a person read and labeled about 1000 articles, a learned classifier achieved a precision of about 50% when making predictions for only the top 10% of documents about which it was most condense. Most users of a practical system, however, would not have the patience to label a thousand articles especially to obtain only this level of precision. One would obviously prefer algorithms that can provide accurate classifications after hand-labeling only a few dozen articles, rather than thousands.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

III. MODEL CONSTRUCTION FOR THE MATRIMONIAL DATABASE

Pseudo code:

Step 1: Input the matrimonial advertisement unstructured text
Step 2: Scan the advertisement text word by word with the keywords from the keyword table and then perform splitting
Do{ (i) Keyword extraction : for eg: 'seeks alliance', 'looking for', 'alliance invited from'
(ii) Splitting: split the advertisement text before and after keywords
} Until (End of text);
Step 3: Trim texts
Do{
Extract word: Each word extracted after splitting
Do{
Trimming process: removes the unwanted symbols like commas, dots, brackets etc….
Except symbols like – (hyphen), " (double quotes), ' (single quote) etc..
} Until (End of word);
} Until (End Of Splitted Text);
Step 4: Scans the text word by word from two different strings and places it in the string of arrays
Do{
(i) Scan word by word with tables like caste, occupation, looks, height, age
(ii) Single valued quantities are directly transferred to the details of groom /bride
(iii) Multivalued attributes eg: for the text in advertisement as' requires groom aged between 30-35. '-' is scanned from table and values to the left and right are compared and stored in the requirement table as min_age= 30 max_age=35
} Until (End Of Splitted Text);
Step 5: Output the groom/bride details from the given advertisement by proposing querying processes
The matching groom/bride details are displayed

1. Construction of a symbolic database:

We start by entering the unstructured text document from a user interface and categorize it. The entered documents are stored in a symbolic database that allows users to store and retrieve data in a tabular form. It has a collection of tables, queries and reports in the database community. The matrimonial database consists of 2 table's brides and grooms and each table has their respective fields. Each table consists of a large amount of documents. Actually, one distinguishes the scheme of the database (which defines the structure of each table) from the contents of each table which may vary according to database updates. A relationship is set between the two tables

2. Classification based on giving a query to the database:

Now we perform operations on to the tables where each concept is associated with a class of units and which produces new tables. The operations are such that attribute values defined in one particular table are matched with the large sets of attributes defined in some other table, based on some condition. Then all these matched attributes are retrieved from the databases. These operations are defined in the form of queries. Hence Queries on the database are defined by a combination of the following operations; consequently the result of a query has also the structure of a table. As an example, we consider the matrimonial database where two particular kinds/classes of people (bride, grooms) are interested in an alliance/marriage proposal. Here the requirements of a particular bride are matched with the attributes of the large set of grooms and the respective matches are retrieved from the set of grooms. Here under the construction phase of the symbolic database, the requirements of the bride are defined only for the following set of attributes i.e., required_looks, required_group, required_min_age and required_max_age. The queries are defined as follows:
Query(1):

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Select all those records from the groom table that match the following criteria:
(1)Bride.required_looks = Groom.looks
(2)Bride.required_group= Groom.group
(3)Groom.age >= Bride.required_min_age AND Groom.age<=Bride.required_max_age.
Query(2):
Select all those records from the bride table that match the following criteria:
(1)Groom.required_looks = Bride.looks
(2)Groom.required_group= Bride.group
(3)Bride.age >= Groom.required_min_age AND Bride.age<=Groom.required_max_age.

IV. CONCLUSION

In this project of Symbolic Data Analysis for classification of unstructured text documents using a symbolic database and querying processes are proposed, since these unstructured texts gives the description in terms of the attribute values. Depending on the values of the attributes of the text document, the attributes are categorized into different groups. Then using a symbolic database the attributes are classified by giving significant queries to the database. From the proposed technique it seems that it is one of the efficient techniques to classify unstructured text documents.

REFERENCES

[1] Michael W. Berry and Malu Castellanos, Editors 'Survey of Text Mining: Clustering, Classification, and Retrieval' ,Second Edition, September 30, 2007
[2] Charu C. Aggarwal IBM T. J. Watson Research Center Yorktown Heights, NY and ChengXiang Zhai University of Illinois at Urbana-Champaign Urbana, IL "Mining Text Data, Chapter 4- A SURVEY OF TEXT CLUSTERING ALGORITHMS'
[3] M. Ikonomakis, S.Kotsiantis, V.Tampakas ,' Text Classification using Machine learning Techniques', WSEAS TRANSCATIONS on COMPUTERS, Issue 8, Volume 4, August 2005 , pp.996-974
[4] Grigorios Tsoumakas, Aristotle University of Thessaloniki, Greece and Ioannis Katakis, Aristotle University of Thessaloniki, Greece, 'Multi-Label Classification: An Overview', International Journal of Data Warehousing & Mining, 3(3), 1-13, July-September 2007
[5] Edwin Diday, 'An Introduction to Symbolic Data Analysis and the Sodas Software', Springer-2000, XVIII, 425pp, ISBN 978-3-642-57155-8