www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

"Analysis and Specifying Textcategorization Using a Bayesian Classification Approach"

Mrs.A.PAVITHRA

PG-Scholar

Department of Computer Science and Engineering

P.S.V College of Engineering and Technology

Karhsnigiri-635108, Tamilnadu, India

Prof.B.SAKTHIVEL, M.E,

Professor & Head

Department of Computer Science and Engineering

P.S.V College of Engineering and Technology

Karhsnigiri-635108, Tamilnadu, India

Mail id : sakthi_mnnit@yahoo.co.in

Mobile no : 9787447671

ABSTRACT

Traditional cluster ensemble approaches have three limitations: (1) they do not make use of prior knowledge of the datasets given by experts. (2) Most of the conventional cluster ensemble methods cannot obtain satisfactory results when handling high dimensional data. (3) All the ensemble members are considered, even the ones without positive contributions.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

In order to address the limitations of conventional cluster ensemble approaches, we first propose an incremental semi-supervised clustering ensemble framework (ISSCE) which makes use of the advantage of the random subspace technique, the constraint propagation the proposed incremental ensemble member selection process, and the normalized cut algorithm to perform high dimensional data clustering. The incremental ensemble member selection process is newly designed to judiciously remove redundant ensemble members based on a newly proposed local cost function and a global cost function, and the normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust and accurate results. Then, a measure is proposed to quantify the similarity between two sets of attributes, and is used for computing the local cost function in ISSCE. Next, we analyze the time complexity of ISSCE theoretically. Finally, a set of nonparametric tests are adopted to compare multiple semi-supervised clustering ensemble approaches over different datasets.

Keywords: approaches, dimensional data, conventional cluster, incremental

CHAPTER 1

1.1.INTRODUCTION

The wide availability of web documents in electronic forms requires an automatic technique to label the documents with a predefined set of topics, what is known as automatic Text Categorization (TC).Over the past decades, it has been witnessed a large number of advanced machine learning algorithms to address this challenging task. By formulating the TC task as a classification problem, many existing learning approaches can be applied [1] [2] [3]. The key challenge in TC is the learning in a very high dimensional data space. Documents are usually represented by the “bag-of-words”: namely, each word or phrase occurs in documents

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

once or more times is considered as a feature. For a given data set, a collection of all words or phrases forms a "dictionary" with hundreds of thousands features.

Learning from such high-dimensional features may lead to a high computational burden and may even hurt the classification performance of classifiers due to irrelevant and redundant features. To ameliorate the "curse of dimensionality" issue and to speed up the learning process of classifiers, it is necessary to perform feature reduction to reduce the size of features.

Feature selection is a common feature reduction approach for TC, in which only a subset of features are kept and the rest of them are discarded. In general, feature selection methods fall into the following three categories: the filter approach, the wrapper approach and the embedded approach [4]. The filter approach evaluates the importance of each individual feature with a score based on the characteristics of data, and only those features with the highest scores are selected. In contrast to the filter approach without involving the learning criteria, the wrapper approach greedily selects better features with the learning criteria. The greedily search in the wrapper approach, however, requires to train classifiers at each step and leads a high computational burden.

The embedded approach can be considered as the combination of both filter and wrapper an approach, which not only measures the importance of each individual feature but also employs a search procedure guided by a learning algorithm. In practice, because of the simplicity and the efficiency of the filter approach, it is predominantly used in TC. Most existing filter approaches first calculate class dependent feature scores, i.e., the feature importance for each class is measured. For example, the Mutual Information (MI) approach measures the mutual dependency between the binary feature and each predefined class label as the feature score. To measure the feature importance globally (for all classes), a combination

operations, such as summation, maximization and weighted average, is used. One major disadvantage is that using the combination operation may bias the feature importance for discrimination. Also, it lacks theoretical supports to choose the best combination operation, and thus, researchers and engineers usually need to explore the best one through extensive empirical studies for a specific TC task [3]. In this paper, instead of using the combination operation to select a global feature subset for all classes, we select a specific feature subset for each class, namely class-specific features. Previously existing feature importance evaluation criteria can still be applied in our proposed approach. Using Baggenstoss's PDF Project Theorem (PPT) [5] [6], we build the Bayes decision rule for classification with these selected class-specific features. The rest of the paper is organized as follows: In Section 2, we introduce background and previous work on document representation, naive Bayes, and feature selection techniques for automatic text categorization.

1.1. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

1.1.1 LITERATURE AND TECHNOLOGY SURVEY

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

In this chapter we will be taking a look at the different areas that need to be explored when creating a piece of software. We will first look at the software development methodologies that can be used and the decisions behind the final choice. The chapter then moves on to look at why there is a need to use models when developing a system and the models that have been chosen for use on this project. The last section is a review of the different technologies that will be used to build the system.

1.1.2 THE SDLC AND DEVELOPMENT METHODOLOGIES

The Systems Development Life Cycle (SDLC) is comprised of four fundamental phases that are followed in nearly all software development projects. These phases are planning, analysis, design, and implementation. The planning phase is concerned with making sure the software developers know how and why they are building a system. This is followed by the analysis phase which looks at the questions, who, what, where and when. To be more precise, who will be using the system, what it will be used for, and where and when it will be used. Next is the design phase, during this phase the developer is looking at how the system will actually work. Finally you get the implementation phases at which point the system is built.

Now that we know what is involved in the development of a piece of software we can look at the different types of methodologies available to us. When it comes to development methodologies it is found that they are split into two main groups. There are the Traditional Heavyweight ‘methodologies and the agile Lightweight ‘methodologies. The main difference between these types of methodologies is the extent to which each phase is considered and the order in which they consider them.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

An example of a traditional methodology is the waterfall development model. The approach taken in the waterfall development model is to tackle each phase of the process one at a time and in sequence. Before a phase is considered to be complete a large amount of documentation typically has to be produced and approved. The waterfall development model got its name from the flow it takes through the phase‘s. Due to the formality and rigid structure of such a method adapting to change is difficult to achieve,

Especially with the amount of time invested in the design. With the main constraint of this project being the lack of time available it would be unwise to follow a traditional methodology due to the large amount of documentation and the rigid structure that is difficult to adapt to change.

With the agile methodologies the development teams focus much less on the documentation side of the development process. With less documentation being written in an agile methodology you will find that project managers can direct their attention to what is considered to be the necessary documentation. Agile methodologies tend to try and work through all the phase of the SDLC in parallel and spend more time trying to make a working system that can be built on rather than planning the system in great detail from the start. The nature of agile methodologies are conducive to smaller development teams but often rely on working in a team environment which is not really a possibility for a solo project like this. This does not mean that there would be no advantages to using an agile methodology.

One of the main selling points is the iterative nature of the methodologies. The planning is conducted in an incremental manner that evolves as the project develops, thus it can accommodate change a lot easier. The use of short development cycles is also part of the iterative process, releasing a working piece of software at the end of each. This allows for the

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

client to view how the project is progressing at every stage and gives them the opportunity to offer feedback on whether the project is meeting the requirements or if new ones need to be added. Finally this project will make use of Agile Modeling (AM). AM is not considered a methodology in itself but works well in combination with agile processes such as DSDM. Simply put, Agile Modeling (AM) is a collection of values, principles, and practices for modeling software that can be applied on a software development project in an effective and light-weight manner‖. The main idea behind AM is that whatever approach to modeling you take during the development process its emphasis should be on making sure the model is there to help pass understanding between the team members and customers. The core principles that apply to this project are as follows:

Model With Others – This will be adapted slightly, the initial modeling will be carried out alone but it will then be refined in a group meeting with the project supervisor. Apply The Right Art effects – This simply means that the most suitable model for the job should be the one used. With so many models to choose from the chances are that there will be one suited to needs of the job. Iterate To Another Art effect – If while creating a model it comes to a point where it is difficult to go on, move to a different model. This can help because, by changing the viewpoint with which the system is looked at, it is possible to see what was causing the problem to start with whilst progressing with another model.

Use The Simplest Tools – When creating a model it is fine to use pen and paper as most models are thrown away once an understanding has been come to. Model in Small Increments

Create Several Models in Parallel – This is similar to the Iterate to Another Art effect principle. Create Simple Content – Try to keep models simple, only showing the content of things that will be in the system. Depict Models Simply – Similar to the principle above, keep

whatever is being modeled simple, it is not required to create models which cover every single detail. It should be noted that AM can be used with different forms of notation such as UML, the main modeling language used on this project.

1.2 EXPERIMENTAL RESULTS AND ANALYSIS

In our experiments, we tested our approach for text categorization on two real-world benchmarks: 20- NEWSGROUPS and REUTERS. These two benchmarks are widely used for text categorization in literature for performance evaluation. We used the data of these two benchmarks provided by Deng et al. [22] [23]. The benchmark 20-NEWSGROUPS collects 20, 000 documents that have been posted online with 20 different topics. In this benchmark, note that some topics are hierarchically categorized, and two different topics could be very closely related to each other, e.g., rec.sport.baseball and rec.sport.hockey, comp.sys.mac.hardware and comp.sys.ibm.hardware, etc. The original benchmark REUTERS consists of 21, 578 documents with 135 different topics. In our experiments, we used its Mod Apte version of REUTERS-21578 and removed those documents which belong to multiple topics.

The used benchmark contains 8, 293 documents with 65 different topics. However, the documents in this benchmark are highly unbalanced, and the topics are ranked with respect to the number of documents. Following the previous works in literature [7], we use its subset, named REUTERS-10, which consists of the documents from the first 10 topics. Before performing feature extraction and classification, we introduced a preprocessing stage in which we discarded those terms that occur in less than 2 documents, and ignored those terms in a stop list. For these two data sets, the officially split training and testing data are provided. We used the training data set for feature selection and parameter estimation, and applied the testing data set for performance evaluation.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

In our experiments, we compared our class-specific feature selection methods with other three non-class specific feature selection methods using the global combination functions of the sum, the weighted average and the maximum, and with another one class specific feature selection method using one-vs-all scheme. The one-vs-all classification scheme trains a binary-class classifier for each individual class and applies the maximum margin rule to make the final classification, which allows the use of either class specific or non-class-specific features to train binary class classifiers. The feature selection criteria of IG, Chi-square, RS, and MD [24] are used to measure class-specific feature scores for both conventional approaches and ours. Fig. 1 shows the overall testing accuracy on these two data sets with different number of features, when IG, Chi-square, RS, and MD are respectively used as feature selection criteria to compute feature scores for each class. It has been seen that the classification performance increases with more features are selected. The comparison results in Fig. 1 indicate significant improvements of our proposed approach for a small number of features, when IG, RS, and MD feature selection criteria are used.

1.2 EXISTING SYSTEM

The wide availability of web documents in electronic forms requires an automatic technique to label the documents with a predefined set of topics, what is known as automatic Text Categorization (TC).Over the past decades, it has been witnessed a large number of advanced machine learning algorithms to address this challenging task.

1.3 PROPOSED SYSTEM

In this paper, we present a Bayesian classification approach for automatic text categorization using class-specific features. Unlike the conventional approaches for text

categorization, our proposed method select a specific feature sub set of each class. To apply these class-dependent features for classification. we follow Baggenstos's PDF Projection Theorem to reconstruct PDFs in raw data space from the class-specific PDFs in low-dimensional feature space, and build a Bayes classification rule. One noticeable significance of our approach is that most feature selection criteria, such as Information Gain(IG) and Maximum Discrimination (MD), can be easily incorporated into our approach.

CHAPTER 2

WORK DONE IN PHASE TWO

2.1 SYSTEM ARCHITECTURE DESIGN

System Design is a solution, how to approach to creation of a new system. This important phase is composed of several steps. It provides the understanding and procedural details for implementing the system recommended infeasibility study. Stress in on translating performance requirement into design specification design goes through logical physical stages of development. Logical design reviews the present physical, prepare input and output specification. These steps are as follow:

1. Problem definition.
2. Input output specification.
3. Data based designed.
4. Modular program design.
5. Preparation of source code.
6. Testing and debug.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations. This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design. Input design is the process of converting the user created input into a computer-based format.

The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases. Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

Choosing the right output method for each user is another objective in designing output. Much output now appears on display screens, and users have the option of printing it out with their own printer. The analyst needs to recognize the trade-offs involved in choosing an output method. Costs differ; for the user, there are also differences in the accessibility, flexibility, durability, distribution, storage and retrieval possibilities, transportability, and overall impact of the data.

One of the most common complaints of users is that they do not receive information in time to make necessary decisions. Although timing isn't everything, it does play a large part in how useful output will be. Many reports are required on a daily basis, some only monthly, others annually, and others only by exception.

Designing Output to Serve the Intended Purpose

All output should have a purpose. During the information requirements determination phase of analysis, the systems analyst finds out what user and organizational purposes exist.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

Output is then designed based on those purposes. You will have numerous opportunities to supply output simply because the application permits you to do so. Remember the rule of purposiveness, however. If the output is not functional, it should not be created, because there are costs of time and materials associated with all output from the system.

Designing Output to Fit the User

With a large information system serving many users for many different purposes, it is often difficult to personalize output. On the basis of interviews, observations, cost considerations, and perhaps prototypes, it will be possible to design output that addresses what many, if not all, users need and prefer. Generally speaking, it is more practical to create user-specific or user-customizable output when designing for a decision support system or other highly interactive applications such as those using the Web as a platform. It is still possible, however, to design output to fit a user's tasks and function in the organization, which leads us to the next objective.

Delivering the Appropriate Quantity of Output

Part of the task of designing output is deciding what quantity of output is correct for users. A useful heuristic is that the system must provide what each person needs to complete his or her work. This answer is still far from a total solution, because it may be appropriate to display a subset of that information at first and then provide a way for the user to access additional information easily.

The problem of information overload is so prevalent that it is a cliché, but it remains a valid concern. No one is served if excess information is given only to flaunt the capabilities of the system. Always keep the decision makers in mind. Often they will not need great amounts

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

of output, especially if there is an easy way to access more via a hyperlink or drill-down capability.

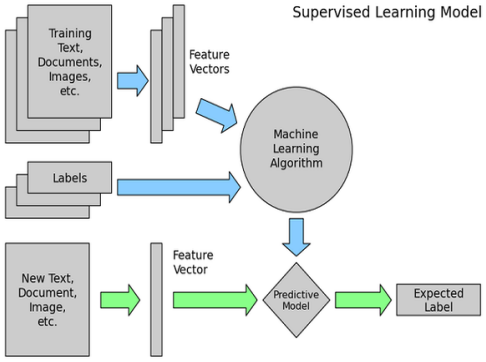

Figure 2.1: SYSTEM ARCHITECTURE

2.2 SYSTEM REQUIREMENTS

2.2.1 HARDWARE REQUIREMENTS

Processor	: Pentium dual core
RAM	: 1 GB
Hard Disk Drive	: 80 GB
Monitor	: 17” Color Monitor
Key Board	: 108 Keys
Mouse	: 3 Buttons

2.2.2 SOFTWARE REQUIREMENTS

Front End/GUI Tool	: Microsoft Visual studio 2012

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

Operating System	: Windows Family
Language	: C#, ASP.NET
Application	: Web Application
Back End	: SQL Server 2005

CHAPTER 3

SYSTEM ORGANIZATION

3.1 USE CASE DIAGRAM

The use case diagram that was finally designed for the Student Admissions system can be seen in Figure The most important step to fully understanding the requirements and scope of this system was the creation and subsequent refinement of this use case diagram. Many changes were made to the initial diagram before a consensus was met on the final design used.

3.2 COLLABORATION DIAGRAM

A collaboration diagram, also called a communication diagram or interaction diagram, is an illustration of the relationships and interactions among software objects in the Unified Modeling Language (UML). The concept is more than a decade old although it has been refined as modeling paradigms have evolved.A collaboration diagram resembles a flowchart that portrays the roles, functionality and behavior of individual objects as well as the overall operation of the system in real time. Objects are shown as rectangles with naming labels inside. These labels are preceded by colons and may be underlined. The relationships between the objects are shown as lines connecting the rectangles. The messages between objects are shown as arrows connecting the relevant rectangles along with labels that define the message sequencing.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

3.3 SEQUENCE DIAGRAM

Sequence diagrams are used to help quickly picture how the objects in a use case interact during the sequence of events in a use case . They do this by showing the behavior of the objects in the use case and the messages they pass. The main strength of sequence diagrams is the clarity with which they show what objects are making what calls and to whom, and which objects are doing what processing .

3.4 STATE CHART DIAGRAM

State chart diagram is one of the five UML diagrams used to model the dynamic nature of a system. They define different states of an object during its lifetime and these states are changed by events. State chart diagrams are useful to model the reactive systems. Reactive systems can be defined as a system that responds to external or internal events. State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. The most important purpose of State chart diagram is to model lifetime of an object from creation to termination. State chart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactive system.

3.5 CLASS DIAGRAM

Class diagrams are probably the most widely used of all the UML diagrams. Not only are they the most common but compared to any other UML diagram they have the largest range of concepts to deal with Although there are concepts that could be used to create deeper, more highly detailed class diagrams this project will only deal with the basics concepts that are most commonly encountered as this will instead create the simple, easy to understand models suggested by AM.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

There are three possible perspectives to take when creating class diagrams, these are conceptual, specification and implementation. When looking at a class diagram from a conceptual perspective the language used is that which anyone in the business domain of the system can follow. The class diagram is created with no regard to the programming language it will be implemented in. On the other hand there is the implementation perspective which is the opposite of the conceptual perspective and is created solely with how it will be implemented in a programming language in mind. The specification perspective is half way between the two, in this perspective the interfaces of each class are considered but not in relation to any specific programming language, this is the perspective that will be used to create the class diagram for this project.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. The following chapter looks at how the system implementation progressed over the course of the build. It starts off by detailing the plan that was used to order how the implementation would progress. The chapter then covers some of the important aspects of implementation that were carried out throughout the building of the system. The data-tier and then the web tier will be looked at in turn, and for the web-tier the

implementation of each aspect of the Model-View-Controller will be considered. Finally, the testing that was carried out on the system will be examined and the effect the results had on the system will be looked.

Designing Plan

As was stated in the Technology and Literature Survey chapter the methodology being followed for the development of this system is the Dynamic Systems Development Method (DSDM). Many of the core principles of the DSDM come about from the idea that the designing and building of a system should be an incremental process. To benefit most from the methodology being used it was decided that the implementation phase of the development would be carried out in an incremental fashion and to do this the system would have to be split into distinct sections. Due to the manner in which the requirements were gathered it was very easy to view the system in such a way. It was decided that the best way to split the system up would be by the main tasks that the system was required to achieve and it was hoped that by doing so each increment of the build would be kept to a manageable level.

4,2,MODULES DESCRIPTION

4.2.1 Class-Specific Feature Selection Method

In this section, we describe our Bayes classifier using class-specific features for TC. Considering a N-class classification problem, suppose that for each class ci, i = 1, 2, ⋯ , N, we select a class-specific feature subset zi = fi(x), where fi(x) could be a linear or nonlinear function such that the dimension of zi is much smaller than x.

4.2.2 Naive Bayes rule

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling..In this post you will discover the Naive Bayes algorithm for classification. After reading this post, you will know:. The representation used by naive Bayes that is actually stored when a model is written to a file. How a learned model can be used to make predictions. How you can learn a naive Bayes model from training data. How to best prepare your data for the naive Bayes algorithm. Where to go for more information on naive Bayes.

4.2.3.Bayesian classification Approach

It is based on applying Bayes' theorem with strong (naive) independence assumptions between in machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers the features.It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

CHAPTER 5

CONCLUSION & FUTURE WORK

In this paper, we have presented a Bayesian classification approach for automatic text categorization using class-specific features. In contrast to the conventional feature selection methods, it allows to choose the most important features for each class. To apply the class

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

specific features for classification, we have derived a new naive Bayes rule following Baggenstos's PDF Projection Theorem. One important advantage of our method is that many existing feature selection criteria can be easily incorporated. The experiments we have conducted on several data sets have shown promising performance improvement compared with the state of-the-art feature selection methods.

REFERENCES

[1] Lam, M. Ruiz, and P. Srinivasan, "Automatic text categorization and its application to text retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 6, pp. 865–879, 1999.

[2] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.

[3] G. Forman, "An extensive empirical study of feature selection metrics for text classification," The Journal of machine learning research, vol. 3, pp. 1289–1305, 2003.

[4] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005.

[5] P. M. Baggenstoss, "Class-specific feature sets in classification," IEEE Transactions on Signal Processing, vol. 47, no. 12, pp. 3428–3432, 1999.

[6] ——, "The pdf projection theorem and the class-specific method," IEEE Transactions on Signal Processing, vol. 51, no. 3, pp. 672–685, 2003.

[7] A. McCallum, K. Nigam et al., "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, vol. 752, 1998, pp. 41–48.

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

[8] V. Kecman, Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press, 2001.

[9] L. Wang and X. Fu, Data mining with computational intelligence. Springer Science & Business Media, 2006.

[10] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in Machine learning: ECML- 98, 1998, pp. 4–15.

[11] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in Proceedings of 14th International Conference on Machine Learning, 1997, pp. 170–178.

[12] Y. H. Li and A. K. Jain, "Classification of text documents," The Computer Journal, vol. 41, no. 8, pp. 537–546, 1998.

[13] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," Proceedings of the 10th European Conferenece on Machine Learning, pp. 137–142, 1998.

[14] B. Tang and H. He, "ENN: Extended nearest neighbor method for pattern recognition [research frontier]," IEEE Computational Intelligence Magazine, vol. 10, no. 3, pp. 52–60, 2015. [15] S. Eyheramendy, D. D. Lewis, and D. Madigan, "On the naive bayes model for text categorization," in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003, pp. 332–339.

[16] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in automated text categorization," in Research and Advanced Technology for Digital Libraries, 2000, pp. 59–68.

[17] I.-S. Oh, J.-S. Lee, and C. Y. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 1089–1094, 1999.

[18] X. Fu and L. Wang, "A GA-based RBF classifier with classdependent features," in Proceedings of the 2002 Congress on Evolutionary Computation, vol. 2, 2002, pp. 1890–1894.

[19] L. Wang, N. Zhou, and F. Chu, "A general wrapper approach to selection of class-dependent features," IEEE Transactions on Neural Networks, vol. 19, no. 7, pp. 1267–1278, 2008.

[20] S. Kay, "Asymptotically optimal approximation of multidimensional pdf's by lower dimensional pdf's," IEEE Transactions on Signal Processing, vol. 55, no. 2, pp. 725–729, 2007.

[21] B. Tang, H. He, Q. Ding, and S. Kay, "A parametric classification rule based on the exponentially embedded family," IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 2, pp. 367–377, 2015.

[22] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 12, pp. 1624–1637, 2005.

[23] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 911–920.

[24] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization," IEEE Transactions on Knowledge and Data Engineering, 2016.

[25] H. He, E. Garcia et al., "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.