Abstract:
The computerized arrangement of texts into decided categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital pattern and the subsequent need to establish them. In the analysis association the effective approach to this problem is based on machine learning techniques. A general inductive process naturally builds a classifier by research from a set of avert documents, the quality of the categories. The advantages of this approach over the knowledge engineering approach are a good capability, considerable savings in terms of expert lab or power, and straightforward portability to different realm. This analysis discusses the main access to text arrangement that decline within the machine learning archetype. We will argue in detail problem related to three different issues, namely, document representation, classifier construction, and classifier evaluation.

Keywords: Feature selection, text categorization, class-specific features, PDF projection and estimation, naive Bayes, dimension reduction.

I. INTRODUCTION
1.1 Project Idea

In the information systems field, due to the increased availability of documents in digital form and the ensuing need to access them in flexible ways, text categorization the activity of labelling natural language texts with thematic categories from a predefined set, is one such task. TC dates back to the early ’60s, but only in the early ’90s did it become a major subfield of the information systems discipline, thanks to increased applicative interest and to the availability of more powerful hardware[1-6]. TC is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching. Until the late ’80s the most popular approach to TC, at least in the “operational” community, was a knowledge engineering (KE) one, consisting in manually defining a set of rules encoding expert knowledge on how to classify documents under the given categories. In the ’90s this approach has increasingly lost popularity in favour of the machine learning (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of reclassified, the characteristics of the categories of interest.

The advantages of this approach are accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert labour power, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories. It is the ML approach to TC that this paper concentrates on. Current-day TC is thus a discipline at the crossroads of ML and IR, and as such it shares a number of characteristics with other tasks such as information/ knowledge extraction from texts and text mining.

There is still considerable debate on where the exact border between these disciplines lies, and the terminology is still evolving. “Text mining” is increasingly being used to denote all the tasks that, by analysing large quantities of text and detecting usage patterns, try to extract probably useful information. According to this view, TC is an instance of text mining. TC enjoys quite a rich literature now, but this is still fairly scattered. Although two international journals have devoted special issues to as a note, we should warn the reader that the term “automatic text classification” has sometimes been used in the literature to mean things quite different from the ones discussed here[7-13]. Aside from (i) the automatic assignment of documents to a predefined set of categories, the term has also been used to mean (ii) the automatic identification of such a set of categories or((iii) the automatic identification of

TEXT MINING APPROACH USING UNSUPERVISED LEARNING NETWORKS CLASSIFICATION

Dr V KHANAA, S.MALINI, S.RUBINI

Professor, BIST, Bharath Institute of Higher Education and Research. Bharath University,
drvkannan62@yahoo.com

P.G. Student, M.E CSE, RRASE College of Engg, Chennai, T.N, India.
Assistant Professor, Dept of CSE, RRASE College of Engg, Chennai, T.N, India.

Abstract: The computerized arrangement of texts into decided categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital pattern and the subsequent need to establish them. In the analysis association the effective approach to this problem is based on machine learning techniques. A general inductive process naturally builds a classifier by research from a set of avert documents, the quality of the categories. The advantages of this approach over the knowledge engineering approach are a very good capability, considerable savings in terms of expert lab or power, and straightforward portability to different realm. This analysis discusses the main access to text arrangement that decline within the machine learning archetype. We will argue in detail problem related to three different issues, namely, document representation, classifier construction, and classifier evaluation.

Keywords: Feature selection, text categorization, class-specific features, PDF projection and estimation, naive Bayes, dimension reduction.

I. INTRODUCTION
1.1 Project Idea

In the information systems field, due to the increased availability of documents in digital form and the ensuing need to access them in flexible ways, text categorization the activity of labelling natural language texts with thematic categories from a predefined set, is one such task. TC dates back to the early ’60s, but only in the early ’90s did it become a major subfield of the information systems discipline, thanks to increased applicative interest and to the availability of more powerful hardware[1-6]. TC is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching. Until the late ’80s the most popular approach to TC, at least in the “operational” community, was a knowledge engineering (KE) one, consisting in manually defining a set of rules encoding expert knowledge on how to classify documents under the given categories. In the ’90s this approach has increasingly lost popularity in favour of the machine learning (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of reclassified, the characteristics of the categories of interest.

The advantages of this approach are accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert labour power, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories. It is the ML approach to TC that this paper concentrates on. Current-day TC is thus a discipline at the crossroads of ML and IR, and as such it shares a number of characteristics with other tasks such as information/ knowledge extraction from texts and text mining.

There is still considerable debate on where the exact border between these disciplines lies, and the terminology is still evolving. “Text mining” is increasingly being used to denote all the tasks that, by analysing large quantities of text and detecting usage patterns, try to extract probably useful information. According to this view, TC is an instance of text mining. TC enjoys quite a rich literature now, but this is still fairly scattered. Although two international journals have devoted special issues to as a note, we should warn the reader that the term “automatic text classification” has sometimes been used in the literature to mean things quite different from the ones discussed here[7-13]. Aside from (i) the automatic assignment of documents to a predefined set of categories, the term has also been used to mean (ii) the automatic identification of such a set of categories or((iii) the automatic identification of
such a set of categories and the grouping of documents under them a task usually called text clustering, or (iv) any activity of placing text items into groups, a task that has thus both TC and text clustering.

1.2 Motivation of the Project

- The method divides the documents into sentences, and categorizes each sentence using keyword lists of each measure.
- A new automatic text categorization method based on unsupervised learning.
- This method can provide basic data for creating training documents from collected documents.
- That can be used in an application area to classify text documents in low cost.

Develop an automatic text categorization approach and investigate its application to text retrieval. The categorization approach is derived from a combination of a learning paradigm known as instance-based learning and an advanced document retrieval technique known as retrieval feedback. We demonstrate the effectiveness of our categorization approach using two real-world document collections from the MEDLINE database[14-18]. Next, we investigate the application of automatic categorization to text retrieval. Our experiments clearly indicate that automatic categorization improves the retrieval performance compared with no categorization. We also demonstrate that the retrieval performance using automatic categorization achieves the same retrieval quality as the performance using manual categorization. Furthermore, detailed analysis of the retrieval performance on each individual test query is provided.

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert lab or power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation.

The introduces concepts and algorithms of feature selection, surveys existing feature selection algorithms for classification and clustering, groups and compares different algorithms with a categorizing framework based on search strategies, evaluation criteria, and data mining tasks, reveals attempted combinations, and provides guidelines in selecting feature selection algorithms. With the categorizing framework, we continue our efforts toward building an integrated system for intelligent feature selection. A unifying platform is proposed as an intermediate step. An illustrative example is presented to show how existing feature selection algorithms can be integrated into a Meta algorithm that can take advantage of individual algorithms. An added advantage of doing so is to help a user employ a suitable algorithm without knowing details of each algorithm. Some real-world applications are included to demonstrate the use of feature selection in data mining. We conclude this work by identifying trends and challenges of feature selection research and development.

II. 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. By formulating the TC task as a classification problem, many existing learning approaches can be applied. 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 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[19-23]. 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. The embedded approach can be considered as the combination of both filter and wrapper approaches, which not only
measures the importance of each individual feature but also employ a search procedure guided by a learning algorithm. Most existing filter approaches first calculate class dependent feature scores, i.e., the feature importance for each class is measured.

III. PROPOSED SYSTEM

The categories are just symbolic labels, and no additional knowledge (of a procedural or declarative nature) of their meaning is available. No exogenous knowledge is available; therefore, classification must be accomplished on the basis of endogenous knowledge only. In particular, this means that metadata such as, for example, publication date, document type, publication source, etc., is not assumed to be available[24-28]. The TC methods we will discuss are thus completely general, and do not depend on the availability of special-purpose resources that might be unavailable or costly to develop. 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). Propose a new automatic text categorization method based on unsupervised learning. Without creating training documents by hand, it automatically creates training sentence sets using keyword lists of each category. And then, it uses them for training and classifies text documents. The proposed method can provide basic data for creating training documents from collected documents, and can be used in an application area to classify text documents in low cost.

IV. KNN ALGORITHM:

In cryptography, a keyd nearest neighbors algorithm: An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object[29-32]. This value is the average of the values of its k nearest neighbors. K-NN is a type of instance based learning or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Algorithm:

KNN Algorithm

STEP 1: BEGIN

STEP 2: Input: D = [(x1, c1) . . . (xN , cN )]

STEP 3: x = (x1 . . . xn) new instance to be classified

STEP 4: FOR each labelled instance (xi, ci) calculate d (xi, x)

STEP 5: Order d (xi, x) from lowest to highest, (i = 1 . . . N)

STEP 6: Select the K nearest instances to x: Dkx

STEP 7: Assign to x the most frequent class in Dkx

STEP 8: END

V. MODULES TITLE:

1. Preprocessing
2. Creating Training Sentence Sets
3. Feature Selection
4. Text Classifier
5. Query Analysis
6. Feature Extraction
7. Ranking
8. Recommendation System

MODULES DESCRIPTION:

Preprocessing:

First, the html tags and special characters in the collected documents are removed. The contents of the documents are segmented into sentences. We extract content words for each sentence using only nouns.

Creating Training Sentence Sets:

I define keywords for each category by hand, which contain special features of each category sufficiently. To choose these keywords, we first regard category names and their synonyms as keywords. We include several words that have a definite meaning of each category. The average number of keywords for each category is 3. (Total 141 keywords for 47 categories) . Next, the sentences which contain pre-defined keywords of each category in their content words are chosen as the initial representative sentences. The remaining sentences are called unclassified sentences. We scale up the representative sentence sets by assigning the unclassified sentences to their related category.

Feature Selection:
The size of the vocabulary used in our experiment is selected by ranking words. To measure the goodness of a word in a global feature selection, we combine the category-specific scores of a word.

**TextClassifier:**

The category predicted by the method for a given document is simply the category with the greatest score. This method performs exactly the same classifications as naive Bayes does, but produces classification scores that are less extreme.

**Query Analysis:**

A set of clusters that describe how the cases in a dataset are related. The decision tree that predicts an outcome, and describes how different criteria affect that outcome. This mathematical model that forecasts sales. A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together[29]

**FeatureExtraction:**

Determining optimal search keywords that result in relevant product related data being returned and search keywords are sufficient to minimize identification errors during the product feature extraction process[30-33]. These challenges exist because online data, which is primarily textual in nature, may violate several statistical assumptions relating to the independence and identical distribution of samples relating to a query.

**Ranking:**

Ranking the importance of web pages, evaluating the financial credit rating of a person, and ranking the risk of investments. The ranking problem is the task of learning a rank-prediction model that assigns an instance a discrete rank “as close as possible” to its actual rank. In this thesis we define the ranking problem as a multi-class classification problem with ordinal class labels and a distance-based loss function.

**RecommendationSystem:**

The term data mining refers to a broad spectrum of mathematical modelling techniques and software tools that are used to find patterns in data and user these to build models[34-36]. In this context of recommender applications, the term data mining is used to describe the collection of analysis techniques used to infer recommendation rules or build recommendation models from large data sets. Recommender systems that incorporate data mining techniques make their recommendations using knowledge learned from the actions and attributes of users.

VI. ARCHITECTUR

![Figure: System Architecture]

VII. CONCLUSION

I concluded from the results that there is good amount of perfection in accuracy of prediction and also good amount of fall in the percentage of classification error in both the proposed techniques. Importance of crop prediction is highly needed for agriculture and economy. Continuous research for improving new methods of prediction would be fruitful. The classification accuracy of the KNN approach is relatively consistent with the implementation of the different values of the parameter, as compared to the conventional KNN approach. Especially in the situation where the training samples are limited and insufficient for the preparation of the training set and the validation set. In the future, other alternative methods for calculating distance and similarity measurement, with lower computational cost, in order to propose a more effective and efficient classification approach.

VIII. REFERENCES


35. Kanniga E., Selvaramarathnam K., Sundararajan M., Kandigital bike
operating system, Middle - East Journal of Scientific Research, V-20, I-6, PP-685-688, 2014
