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. 3, Issue 3, March 2015

Survey on Text Classification Based on Similarity

Kavitha Sri.M, Hemalatha.P

Department of IT, IFET College of Engineering, Villupuram, India

Assistant Professor, Department of IT, IFET College of Engineering, Villupuram, India

ABSTRACT:The astonishing progress of computer technology in the few years has led to large supplies of powerful and reasonable computers**.**Text Mining is the detection by computer of new and previously unknown information, by automatically extracting information from different written assets**.**An efficient and effective text document classification is becoming a challenging and highly required area to capably categorize text documents into mutually exclusive categories. In this paper we discuss several approaches of text categorization, feature selection methods and applications of text categorization based on similarity.

KEYWORDS: Text mining, Extracting information, text classification, Feature selection

I. INTRODUCTION

Data mining is the process of extracting the implied previously unknown and potentially useful information from data.Text categorization [6] [7] is an upcoming and very important field in today's world which is mostimportantly required and demanded to efficiently categorize various text documents intodifferent categories.The capacity of storing data becomes enormous as the technology of computer hardware develops. So amount of data is increasing exponentially, the information required by the users become varies and actually users deal with textual data more than the numerical data. It is very difficult to apply techniques of data mining to textual data instead of numerical data.Therefore it becomes necessary to develop techniques applied to textual data that are different from the numerical data. Instead of numerical data the mining of the textual data is called text mining.A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects The similarity measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data It is important to measure a similarity/distance (Chim and Deng,2008). Choosing an appropriate similarity measure is also crucial for cluster analysis,especially for a particular type of clustering algorithms.Text Categorization (TC) is the classification of documents with respect to a set of one or more preexisting categories (Sebastiani, 2002). The classification phase consists of generating a weighted vector for all categories, then using a similarity measure to find the adjoining category.The similarity measure is used to determine the degree of likeness between two vectors. To achieve reasonable classification results, a similarity measure should generally respond with larger values to documents that belong to the same class and with smaller values otherwise. During the last decades, a large number of methods proposed for text categorization were typically based on the classical Bag-of-Words model where each term or term stem is an independent feature. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The similarity increases as the difference between the two values associated with a present feature decreases. To improve the efficiency, they have provided an approximation to reduce the complexity involved in the computation.

II. LITERATURE SURVEY

In this section, we are focus on the different methods for text classification based on similarityof [1] Feature selection methods have been successfully applied to text categorization but not often appliedto text clustering due to the unavailability of class label information. In this paper, we first giveempirical evidence that feature selection methods can improve the efficiency and performance of text clustering algorithm. Then we propose a new feature selection method called "Term Contribution (TC)" and perform a comparative study on a variety of feature selection methods for

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. 3, Issue 3, March 2015

text clustering, including Document Frequency (DF), Term Strength (TS), Entropy-based (En),Information Gain (IG) and _2 statistic (CHI).Finally, we propose an "Iterative Feature Selection (IF)" method that addresses the unavailability of label problem by utilizing effective supervised feature selection method to iteratively select features and perform clustering. Detailed experimental results on Web Directory data are provided in the paper.[2]Similarity is an important and widely used concept Previous definitions of similarity are tied to a particular application or a form of knowledge representation. We present an information theoretic definition of similarity that is applicable as long as there is a probabilistic model. We demonstrate how our definition can be used to measure the similarity in a number of different domains.[3]This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. This is important because in many text classification problems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available. We introduce an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation-Maximization (EM) and a naive Byes classifier. The algorithm first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. It then trains a new classifier using the labels for all the documents, and iterates to convergence. This basic EM procedure works well when the data conform to the generative assumptions of the model. However these assumptions are often violated in practice, and poor performance can result. We present two extensions to the algorithm that improve classification accuracy under these conditions: (1) a weighting factor to modulate the contribution of the unlabeled data, and (2) the use of multiple mixture components per class. Experimental results, obtained using text from three different real-world tasks, show that the use of unlabeled data reduces classification error by up to 30% [4] In the k-median problem we are given **a** set S of n points in **a** metric space and **a** positive integer k: The objective is to locate k medians among the points so that the sum of the distances from each point in S to its closest median is minimized[5]Clustering is one of the most important techniques in which the machine learning and data mining tasks. Similar data grouping is performed using clustering techniques. Hierarchical clustering model produces tree structured results. Partitioned clustering produces results in grid format. The documents are projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. The Euclidean distance is a dissimilarity measure describes the dissimilarities between the documents. Correlation indicates the strength and direction of a linear relationship between two random variables. A scale-invariant association measure is used to calculate the similarity between two vectors. Correlation preserving index (CPI) based clustering is used for document clustering process. The similarity-measure-based CPI method is used for detecting the intrinsic structure between nearby documents. In CPI method the documents are projected into a low-dimensional semantic space. Correlations between the documents in the local patches are maximized. Correlations between the documents outside these patches are minimized simultaneously. The spectral clustering is applied on the correlation similarity model with nearest neighbor learning process. The Ontology repository is used to manage the term concept relations. Local patch extraction is carried out with Ontology support. Term frequency based weight is replaced with concept weight based model. The document preprocess operations are carried out to extract term information. Stop word elimination and stemming process are applied on the term collection. Porter stemming algorithm is used for suffix analysis. Ontology is used to extract term relationships.

III. CONCLUSION

In this survey, the aim has been to explore and evaluate different techniques for similarity measures. Future research in the data mining similarity measure will try hard towards improving the accuracy, precision, and computational speed. There are three factors in text categorization: categorization model, similarity measure,and document representation. There are many alternatives for each one of these factorsAlthough the current scheme proved more accurate than traditional methods, there are still accommodation for improvement

REFERENCES

[1] An Evaluation on Feature Selection for Text Clustering TaoLiuLTMAILBOX@263.SINA.COM Department of Information Science, Nankai University, Tianjin 300071, P. R. ChinaShengping LiuLSP@IS.PKU.EDU.CN
Department of Information Science, Peking University
[2]An Information-Theoretic Definition of SimilarityDekang LinDepartment of Computer ScienceUniversity of ManitobaWinnipeg, Manitoba, Canada R3T 2N2
[3]Text Classification from Labeled and Unlabeled Documents using EM KAMAL NIGAM† knigam@cs.cmu.edu ANDREW KACHITES MCCALLUM‡† mccallum@justresearch.com SEBASTIAN THRUN† thrun@cs.cmu.edu TOM MITCHELL† tom.mitchell@cmu.edu †School of

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. 3, Issue 3, March 2015

Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 ‡Just Research, 4616 Henry Street, Pittsburgh, PA 15213 Received March 15, 1998; Revised February 20, 1999
[4]A nearly linear-time approximation scheme for the Euclidean k-median problemSG Kolliopoulos, S Rao - Algorithms-ESA'99, 1999 - Springer
[5]Document Clustering using Correlation Preserving Indexing with Concept Analysis 1M. Mohanasundari, 2P.Malathi 1,2Dept.of CSE, Velalar College of Engineering and Technology
[6] Jiawei Han, &Miche Line Kamber, (2006) Data Mining: Concepts and Techniques, 2nd ed.,Elsevier.[7] Margaret H. Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education
[8] Survey on Similarity Measure for ClusteringP. H. Govardhan, Prof. K . P. Wagh, Dr. P.N. Chatur Dept. of Computer Science and Engineering,

SGB Amravati University, India

[9]A SURVEY ON OPTIMIZATION APPROACHES TO TEXT DOCUMENT CLUSTERING R.Jensi1 and Dr.G.Wiselin Jiji2 1Research Scholar, ManomaniumSundaranar University, Tirunelveli, India 2Dr.Sivanthi Aditanar College of Engineering,Tiruchendur[10]A SURVEY OF TEXT CLUSTERING ALGORITHMSCharu C. Aggarwal*IBM T. J. Watson Research Center*Yorktown Heights, NYcharu@us.ibm.comChengXiangZhaiUniversity of Illinois at Urbana-ChampaignUrbana, ILczhai@cs.uiuc.edu
[11] A Survey of Text Mining Techniques andApplications
Vishal Gupta Lecturer Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India Email: vishal@pu.ac.in
Gurpreet S. LehalProfessor & Head, Department of Computer Science, Punjabi University Patiala, IndiaEmail: gslehal@yahoo.com
[12]A TECHNICAL STUDY AND ANALYSIS ON FUZZY SIMILARITY BASED MODELS FOR TEXT CLASSIFICATION Shalini Puri1 and SonaKaushik M. Tech. Student, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India eng.shalinipuri30@gmail.com 2M. Tech. Student, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India sonakaushik22@gmail.com

BIOGRAPHY

M.Kavitha sri Currently pursuing B.Tech Information technology at IFET College of Engineering,Villupuram,India.Her Area of Interests Includes Oops Concepts,Computer Networking.

Mrs.P.Hemalatha received her B.E in CSE P.S.N.A College of Engineering and Technology Pondicherry and M.E. CSE in IFET College of Engineering, Villupuram. Currently working as ASST.PROFFESOR in IFET College of Engineering in Villupuram. She has Published three papers in International Conference and also Journal paper. Her Area of interest includes Computer Networks ,Wireless Sensor Networks and Data Mining.