Volume 2, Issue 8, August 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Analysis of Query Based Text Classification Approach

Ankit Jain
Dept. Of Information Technology
MIT, Ujjain, India

Abhishek Raghuvanshi
Dept. Of Information Technology
MIT, Ujjain, India

Gourav Shrivastava
Dept. Of Information Technology
MIT, Ujjain, India

Abstract*— Text Classification (TC) is used to assign a target document into a well suited category or categories. Basically it is possible that a target document belong from various categories. To choose more appropriate category for target document is the main objective of TC. To search any topic using internet uses TC approach, and when user need to retrieve document and he enters his topic it may contain n words. This type of classification known as Query based TC (QBTC).This paper focus on QBTC while using K-nn approach and also shows shortcomings of it.

***Keywords*— Text Classification, K-nn Approach, Query based Text Classification.**

I. INTRODUCTION

Text Classification is used to sort data on the basis of predefined categories. Now a day's search engines are the part of individual life because it provides any information relevant to users need. When a user enters any keyword in search engine the information and document related to that keyword are retrieved. So, we can say that text Classification is the approach of information retrieval. Text classification is also known as categorization or topic spotting. There are three main components of text classification known as: Data Preprocessing, Classifier Construction and document categorization. Text Classification components are widely discussed in Section II.

From last ten years the popularity of TC increases. The main reason behind its popularity is that it decreases the overhead of manual working. Application developers are also fascinated from TC, because using this huge amount of data can be easily handled and data can be of any type and stored in well managed manner. On the whole in TC the data is divided on the basis of categories. For example books are divided in categories on the basis of authors, awards, by country, by publisher, by topic or alphanumeric form. The advantage of TC we can handle and organize any sort of data, even if any new document is come in light we can sort it in predefine suitable category after analysis of document.

This paper lime lights the basic concept of TC and QBTC. Section III presents Query Based Text Classification (QBTC), Section IV accent applications and Shortcoming of TC and QBTC. Finally we conclude in Section V.

II. TEXT CLASSIFICATION

TC is a task of assigning a document to one or more classes or categories. This task can be perform in two ways one via manual organization or intellectually and second way is algorithmically. Text classification problem arises in the field of library science, information science and computer science. So, generally manual classification is applied on information science and computer science. Document can be classified in categories of text, audio, video, images. Thus, for various types of TC technique is used.

If document is form of text we can classify any text on the basis of subject, author name, year of publication etc. Classification is based on two types of approach content based and query based. This paper emphasis on content based classification approach.

A. Text Classification Architecture

In Figure I the architecture of TC is shown, it contains mainly three stages: Data Preprocessing, Classifier Construction and document categorization. Each stage also contains sub stages. The main task of TC to assign a unclassified document into categories.

Input Data: Input for text classification is the combination of predefined document and unclassified document. Whereas document which are labeled and categorized known as predefine document. This type of document also called as training document. Second type of input data is unclassified document. This document is a target document, text classification process applied for this document.

Data Pre-processing: In this Stage initial document is transformed into a common format, because initial document may be collected from various resources. Thus, these documents have different formats. After converting all collected document into a uniform form all operations can be applied on these uniform format. Data pre-processing embraces six sub stages: *Document Converting:* Documents collected from various resources can be of different format such as Doc, PDF, XML; HTML converted into a plain text format.

Functional word Removal: In this step topic-neural such as articles (a, an, the), conjunction (and, or, nor), prepositions (in, of, at) are removed from the plain text format.

Word Stemming: It converts the words in standard form i.e. ranking→rank, Ranked→Rank. So, we can say that it convert words into proper verb.

Feature Selection: It reduces the dimensionality of the data space by removing irrelevant or less relevant features.

Dictionary Construction: In this stage uniform dictionary is constructed and it is used as a reference for converting the text document to a vector of features. Each feature in the vector corresponds to a word in the dictionary.

Feature Weighting: In feature weighting step assignment of weight is done using weighting function to different weights, which are present in dictionary.

Classifier Construction: This is the main phase of text classification. In this phase a classifier is build by learning via predefine documents, and classifier is constructed. This classifier is used to classify unknown document. There are many classification algorithms are present for TC. Some of them are SVM, Rocchio, K-NN. This work focused on K-NN algorithm. In this step the process of construction of classifier or learner is done when data which are comes from data preprocessing phase divided into three disjoint sets.

Training Set: The training set is the set of documents observing which the learner builds the classifier.

Validation Set: The validation set is the set of document on which the engineer fine-tunes the classifier.

Test Set: The test set is the set on which the effectiveness of the classifier is finally evaluated.

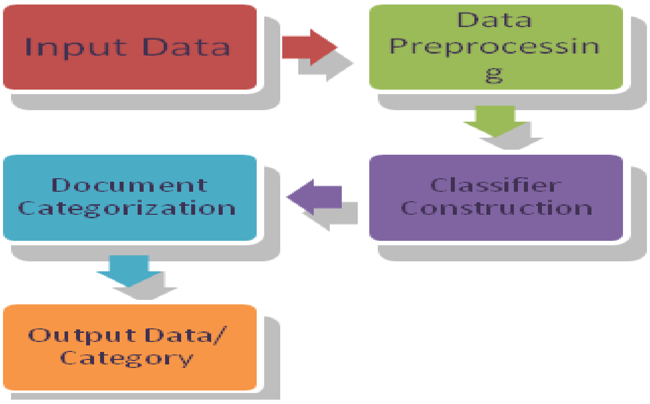

Figure I Text Classification Architecture

Document Categorization: In this phase Documents are classified and function of document classification is used. The result of this phase is the output of TC process. The output is analyze by the user and only user can say whether a given item of information is relevant to a query issued to a web search engine or to a private folder in which documents should be filled according to content [3].

B. Formal Definition of TC

TC may be formalized as the task of approximating the unknown target function

Φ: D x C →{T, F}……………..…. (1)

(That describes how documents ought to be classified, according to a supposedly authoritative expert) by means of a function Φ: D x C →{T, F} called the classifier, where C= {c1, c2, c3…cn} Is a predefined set of categories and D is a (possibly infinite) set of documents of

Ø(dj , cj) =T……………………..…..(2)

Then dj is called a positive example (or a member) of ci, while if

Φ(dj, ci) =F ………………………....(3)

it is called a negative example of ci[ ].In other language, Text classification is defined as for a given set of previously unseen documents D={ d1,d2,d3……dn} and a set of predefined classes or categories C= {c1,c2,c3…….ck}, a classifier(categorizer) is a function K that maps a document from set D to the set of all subsets of C[2].

C. Labels of TC

TC is applied on different sort of applications here labels of TC shows task i.e. a document belongs from how many labels or categories. Therefore as per users requirement TC can be of two types:

Single Label Task: In Single label task target document is assign to exactly one category.

Binary TC: Binary TC is the special case of single label task. So, here binary means a target document assign to category or complement of that category. For example: If ci is a category and dj is a target document then either dj is assign to ci or ci^{-}.

Multi-label TC: One document will be assign to many categories. For example: In India students of IIT more placed in MNC's rather than other colleges. This statement may be belonging from different categories such as: it may be assign to IIT label or MNC label, Placements in India. So, this sentence may be considered in different sort of categories.

In single label TC effectiveness is measured by accuracy A I.e. the percentage of correct classification decision. So, error is opposite to accuracy.

E=1-A…………………………….... (4)

In binary TC effectiveness of categories are measured by a combination of precision ci(πi) the percent of documents deemed to belong to ci that in fact belong to it. How many documents belong to category is a precision, where ci(pi), the percent of documents belonging to ci that are in fact deemed to belong to it[3].

D. Efficiency and Effectiveness of TC

Efficiency and effectiveness of TC is evaluated by the performance of TC and time taken by the process. Efficiency can be evaluated of training as well as classification.

Training Efficiency: Average time required to build a classifier from a given data.

Classification Efficiency: Average time required to classify a document shows efficiency of classification.

Effectiveness: Effectiveness of classification approach can be evaluated by average correctness of classification behaviour.

Two basic measures for a given document, in ranking based systems, are recall and precision. These are computed as follows:

$$Recall = \frac{Categories\ found\ correct}{Total\ categories\ correct}$$

$$Precision = \frac{Categories\ found\ correct}{Total\ categories\ found}$$

Very popular method for global evaluation of a classifier is an 11-point average precision measure.

It can be computed using following algorithm:

- For each document, compute the recall and precision at each position for the ranked list.
- For each interval between recall values (0%-10%, 10%-20%, 90%-100%- therefore 11-point), use the highest precision value as the representative precision.
- For each interval between recall values compute average precision over all test documents.
- Average eleven precision obtained from previous step to obtain a single number - 11-point average precision.

For evaluation of binary classifiers four are values important (for a given category):

a - number of documents correctly assigned to the category.

b -Number of documents incorrectly assigned to the category.

c -Number of documents incorrectly rejected from the category.

d - Number of documents correctly rejected from the category.

Following performance measures are defined and computed from these values:

Recall(r) = a/(a + c)…………………….(5)

Precision (p) = a/(a + b)………………...(6)

Fallout (f) = b/(b + d)…………………..(7)

Accuracy (Acc) = (a + d)/n……………. (8)

Where n = a + b + c + d

Error (Err) = (b + c)/n …………………. (9)

Where n = a + b + c + d

E. K-nn Approach

K-nn algorithm is very popular and widely used algorithm for problem domain for example: text classification. It is a similarity based learning algorithm. In any target document using K-nn algorithm k nearest neighbours from all training documents are retrieved, where all training document shows some weight on the basis of this weight of target or test document is evaluated. If several neighbour shares a category, then the pre neighbour weights of the category are added together, and the resulting weighted sum is used as the like hood score of candidate categories. After this ranking of weight is shown in ascending order for the test document, by threshold value on these scores, binary category assignments are obtained. In KNN (K-Nearest Neighbours') "Similar" item are searched and stored in a categories. So, we need a functional definition of "similarity" if we want to apply this automatically.

It is a Instance-Based Learning, also known as Lazy Learning, well known approach to pattern recognition, theoretical error bound analysis on it is done by Duda & Hart (1957), it provides strong baseline in benchmark evaluations, among top-performing methods in TC evaluations it is scalable to large TC applications. Figure II shows example of K-nn approach Example of k-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3(solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If k = 5(dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle). Various KNN Schemes are:

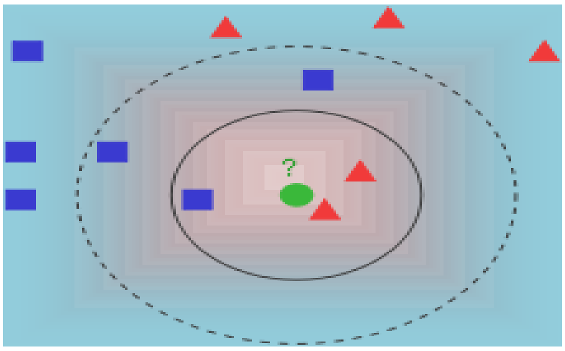

Figure II Example of k-nn approach

Decision Surface (Voronoi Diagram): In decision surface scheme of K-nn decision surface is used which is divided by several points. This scheme uses voronoi diagram, accordingly, it is also known as voronoi diagram technique. New points are classified in categories on the basis of its distance.

Majority Voting Scheme: In majority voting scheme the decision boundary is divided into circles and for target document the majority of document assign to it.

Weighted–Sum Voting Scheme: In weighted sum voting system each neighbour is contain some weight according to its nearness with respect to target document. For K=5we have two types of documents are of two colours white and black and for target document X we need to select a category. There are n documents of white and black colour, present in decision surface. So, in this technique for same categories i.e. black and white we calculate add of all weights of white documents and same technique used for black documents.

If total weight of white document is greater than black document we assign white to X or vice versa. The result of K-nn technique varies on the basis of k values. Figure II shows this concept [11].

III. QUERY BASED TEXT CLASSIFICATION

Automatic text classification is an effective tool to retrieve information very efficiently. Text documents have become the most common type of information repositories. Using WWW users can access documents as per their need. So, there are billions of documents stored in the form of text document. When a user wants to retrieve a document, and search document by giving a string of a word, known as query[1].

Query may contain n words so, for n length query many documents belongs from that query. For example: the query is “Analysis of Query Based Text Classification Approach”, there is possibility that multiple documents are present for: Analysis or Query or Query Based Text Classification or Text classification. In this case using K-nn approach and ranking the output is retrieved. Algorithm of query based classification [1] proposed the solution of this type of input as per this algorithm target document P are considered as a scalar point S.S1,S2,S3……Sn are vector points which represent the training document of the dataset in space S.

Input integer k which is the no. of vectors retrieved that are relevant to given point P. The training dataset are themselves are pre labelled into predefined categories. Result of this algorithm is the label name of maximum number of documents which are close to P from K points. To achieve this goal author suggest to divide query P into l words and apply K-nn approach on all l points, and also combination of their group so for every point resultant shows l*k values. After this values combination of l words are generated and apply k-nn on these when original query is generated k points which are actually near to P are generated.

IV. APPLICATIONS AND SHORTCOMING OF TC

Major Reason behind popularity of TC is the need to handle and organize huge quantity of documents like, which needs expenses and not feasible and time consuming because contains huge amount of data. So, if categories are pre define then it is useful to store and maintain these type of data in particular categories. Data can be of any type can be classified but need different type of classification strategy depend upon, nature of document, the structure of classification scheme and nature of the task means task may be single label or multi-label. The major applications of TC are [2]:

- Automatic Indexing
- Document Organization
- Text Filtering
- Hierarchical categorization of Web pages
- Word sense Disambiguation
- Automatic Survey Coding

- Spam Filtering

Shortcoming of TC:

TC has to face challenge of dealing with very large numbers of categories (e.g. in the tens of thousands) and challenge is the attempts at solving the labeling bottleneck, i.e. at coming to terms with the fact that labeling examples for training a text classifier when labeled examples do not previously exist, is expensive. As a result, there is increasing attention in text categorization by semi-supervised machine learning methods and the major challenge to choose the K value. When k = m, KNN becomes comparatively less, where m denotes the number of training queries. Datasets with different k values in terms of different groups as k increases, the performances first increase and then decrease. More specifically,

- When only a small number of neighbors are used, the performances of KNN are not so good due to the insufficiency of training data.
- When the numbers of neighbors increase, the performances gradually improve, because of the use of more information.
- However, when too many neighbors are used (approaching 1500), the performances begin to deteriorate[1].

V. CONCLUSION

In this paper analysis of QBTC using K-nn approach is shown. QBTC is very useful for finding relevant documents for a given query. This algorithm proposed to use ranking of documents, where ranking of each document is obtained using K-nn method. Research is require for to improve response time which increases due to P*K, accuracy and degree of relevancy. There is need to work on classification accuracy of marginal data that falls outside the regions of representations. The solution of this problem is if we create a index file of each document then the searching time will reduce and if weights are computed on the basis of occurrence of word in document then marginally data that falls outside the regions of representations also covers.

REFRENCES

[1] Suneetha Manne, Sita Kumari Kotha, Dr. S. Sameen Fatima" A *Query based Text Categorization using K-Nearest Neighbor Approach*", International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[2] Fabrizio Sebastiani *"Text Categorization"*, Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109-129.

[3] Gongde Guo, Hui Wang1, David Bell, Yaxin Bi, and Kieran Greer *"Using kNN Model-based Approach for Automatic Text Categorization"* European Commission project ICONS, project no. IST-2001-32429.

[4] Sebastiani, F.*,"Machine learning in automated text categorization".* ACM Computing Surveys, 34(1), pp. 1–47, 2002.

[5] XiuboGeng, Tie-YanLiu, TaoQin, AndrewArnold, HangLi and Heung-YeungShum, "Query Dependent Ranking Using K-Nearest Neighbor," ACM, SIGIR08, July20–24,2008,Singapore.

[6] Dik L. Lee, uei Chuang, H Ent Seamons*," Document Ranking and the Vector-Space Model",*a research theisis, March-April,1997.

[7] T.Y.Liu,Y.Yang,H.Wan,H.Zeng,Z.Chen,andW.Y.Ma*,"Support Vector machines classification with a very large scale taxonomy*. SIGKDD Explor. Newsl,7(1):36–43.

[8] Pascal Soucy, Guy W Minau, "*A Simple KNN algorithm for Text Categorization",* 0-7695-1119-8/01 IEEE 2001.

[9] Stavros Papadopoulos, Lixing Wang, Yin Yang, Dimitris Papadias, Panagiotis Karras, "Authenticated Multi-Step Nearest Neighbor Search"

[11] Yang, Y. & Pedersen, J.O., A comparative study on feature selection in text categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning, ed.D.H. Fisher,Morgan Kaufmann Publishers, San Francisco, US: Nashville, US, pp. 412–420, 1997.

[12] Guru, D. S., Harish B. S., and Manjunath, S. 2009. "Clustering of Textual Data: A Brief Survey", In the Proceedings of International Conference on Signal and Image Processing, pp. 409 – 413.

[13] Dr. Riyad Al-Shalabi , Dr. Ghassan Kanaan and Manaf H. Gharaibeh *"Arabic Text Categorization Using kNN Algorithm"* .

[14] Yu Wang, Zheng-Ou Wang*," A FAST KNN ALGORITHM FOR TEXT CATEGORIZATION",* Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.