# Survey on Text Similarity Measurements using Various Techniques

Indumathi A
PG Scholar,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Perumal P
Professor,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

**Abstract-Now a day's dealing with a tremendous measure of digital data in computerized form is essential in text mining applications. Text classification is a method of naturally arranging an arrangement of reports into classes from a predefined set. A noteworthy trademark or trouble of Text classification is high dimensionality of highlight space. The decrease of dimensionality by choosing new qualities which is subset of old properties is known as highlight determination. Highlight determination techniques are examined in this paper for lessening the dimensionality of the dataset by expelling highlights that are viewed as insignificant for the arrangement. In this paper we talk about a few methodologies of Text mining, highlight determination techniques and uses of Text order. The issue of characterization has been broadly examined in the data mining, machine learning, database, and data recovery groups with applications in various assorted spaces, for example, target promoting, restorative analysis, news aggregate separating, and report association. In this paper we have given an overview of a wide assortment of text classification algorithms.**

## I. INTRODUCTION

The capacity of storage 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. 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 is 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. Text mining [1] is procedure of synthesizing the information by analyzing relations, the patterns and rules from the textual data. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from some of the familiar things in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing of all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. The main functions [2] of the text mining include text summarization, text categorization and text clustering. The Text of this paper is restricted to text categorization.

The limit of putting away information winds up noticeably tremendous as the innovation of PC equipment creates. So measure of information is expanding exponentially, the data required by the clients move toward becoming fluctuates .really clients manage literary information more than the numerical information. It is extremely hard to apply strategies of information mining to literary information rather than numerical information. In this way it ends up plainly important to create methods connected to literary information that are not the same as the numerical information. Rather than numerical information the mining of the printed information is called Text mining.

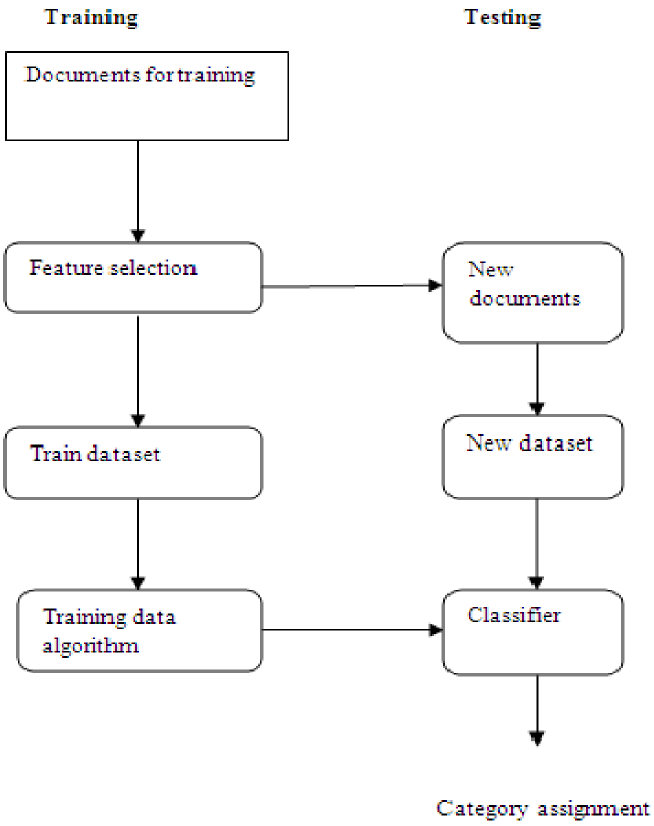

Fig.1 process of text categorization

Text mining [1] is system of integrating the data by examining relations, the examples and guidelines from the literary information. A key component is the connecting together of the extricated data together to frame new actualities or new theories to be investigated encourage by more traditional methods for experimentation. Text mining is not quite the same as what know about in web look. In seek, the client is regularly searching for something that is as of now known and has been composed by another person. The issue is pushing aside all the material that right now is not significant to your requirements keeping in mind the end goal to locate the pertinent data. In Text mining, the objective is to find obscure data, something that nobody yet knows thus couldn't have yet recorded. The capacities [2] of the Text mining are Text outline, Text classification and Text grouping. The substance of this paper is limited to Text classification.

## II. CATEGORIZATION METHODS

### *A. Decision Trees*

Decision tree techniques remake the manual classification of the preparation records by developing all around characterized genuine/false inquiries as a tree structure where the hubs speak to questions and the leaves speak to the comparing class of reports. In the wake of having made the tree, another record can without much of a stretch be ordered by placing it in the root hub of the tree and let it gone through the inquiry structure until the point that it achieves a specific leaf. The principle preferred standpoint of Decision trees is the way that the yield tree is anything but difficult to decipher notwithstanding for people who are not acquainted with the subtle elements of the model [6]. The tree structure created by the model furnishes the client with a solidified perspective of the classification rationale and is along these lines valuable data. A danger of the use of tree strategies is known as "finished fitting": A tree over fits the preparation information if there exists an option tree that orders the preparation information more awful yet would arrange the records to be sorted later better. This situation is the after effect of the calculation's expectation to develop a tree that sorts each preparation record effectively; notwithstanding, this tree may not be fundamentally appropriate for different archives. This issue is ordinarily directed by utilizing an approval informational collection for which the tree needs to perform also as on the arrangement of preparing information. Different strategies to keep the calculation from building colossal trees (that in any case just guide the preparation information accurately) are to set parameters like the most extreme profundity of the tree or the base number of perceptions in a leaf. On the off chance that this is done, Decision Trees demonstrate great execution notwithstanding for order issues with countless in the lexicon.

### *B. k-Nearest Neighbour*

The classification itself is normally performed by looking at the class frequencies of the k nearest documents (neighbours). The assessment of the closeness of reports is finished by measuring the point between the two component vectors or computing the Euclidean separation between the vectors. In the last case the component vectors must be standardized to length 1 to consider that the measure of the reports (and, in this manner, the length of the element vectors) may vary. A surely favourable position of the k-closest neighbour strategy is its effortlessness. It has sensible likeness measures and does not require any assets for preparing. K nearest neighbour performs well regardless of the possibility that the classification particular archives from more than one bunch on the grounds that the classification contains, e.g., more than one theme. This circumstance is seriously suited for most classification calculations. A disservice is the better than expected order time on the grounds that no preparatory venture (in the feeling of a learning stage) has been finished. Moreover, with various quantities of preparing records per classification the hazard expands that an excessive number of reports from a relatively vast classification show up under the k nearest neighbours and therefore prompts a lacking arrangement.

### *C. Bayesian Approaches*

There are two groups of Bayesian approaches in document categorization: Naïve [7] and non-naïve Bayesian approaches. The naïve part of the former is the assumption of word (i.e. feature) independence, that the word order is irrelevant and consequently that the presence of one word does not affect the presence or absence of another one. A disadvantage of Bayesian approaches [8] in general is that they can only process binary feature vectors and, thus, have to abandon possibly relevant information.

### *D. Regression-based Methods*

For this method the training data are represented as a pair of input/output matrices where the input matrix is identical to our feature matrix A and the output matrix B consists of flags indicating the category membership of the corresponding document in matrix A. Thus B has the same number of rows like A (namely m) and c columns where c represents the total number of categories defined. The goal of the method is to find a matrix F that transforms A into B' (by simply computing B'=A*F) so that B' matches B as well as possible. The matrix F is determined by applying multivariate regression techniques. An advantage of this method is that morphological pre-processing (e.g., word stemming) of the documents can be avoided without losing categorization quality. Thus, regression based approaches become truly language-independent. Another advantage in these methods can easily be used for both single category and multiple-category problems.

## III. APPLICATIONS OF TEXT CATEGORIZATION

The applications of text categorization are manifold. Common traits among all of them are

- The need to handle and organize documents in which the textual component is either unique, or dominant, or simplest to interpret component.
- The need to handle and organize large quantities of such documents, i.e large enough that their manual organization into classes is either too expensive or not

feasible within the time constraints imposed by the application.
- The fact that the set of categories is known in advance, and is variation over time is small.

An archive association is an accumulation of records made out of named bunches that contain comparable reports. Note that an accumulation of non-grouped reports is not a record association. On the off chance that the report association contains bunches with settled groups, it is known as a progressive record association. On the off chance that its groups don't have any settled bunches, it is known as a level archive association. It is important to construct an archive association, physically or consequently, for the productive administration of reports. There are two sorts of record associations, static archive association and dynamic report association. In the event that the bunches of the record association are settled forever, it is known as a static report association. In the event that it adjusts without anyone else's input, to the present circumstance, we allude to the report association as a dynamic archive association. Ordering with a controlled vocabulary is an example of the general issue of record base association. For example, at the workplaces of a daily paper approaching "characterized" advertisements must be, preceding distribution, sorted under classes, for example, Personals, Cars available to be purchased, Real Estate, and so forth. Daily papers managing a high volume of characterized promotions would profit by a programmed framework that picks the most reasonable class for a given advertisement. Other conceivable applications are the association of licenses into classes for making their hunt less demanding [18], the programmed documenting of daily paper articles under the suitable areas (e.g., Politics, Home News, Lifestyles, and so forth.), or the programmed gathering of meeting papers into sessions.

## IV. RESULT AND DISCUSSION

Regression-based Methods play a vital part in the diminished of the dimensional of the dataset by evacuating highlights that are viewed as unimportant for the characterization [10]. These element determination techniques have various preferences, for example, littler dataset measure, littler computational necessities for the text order calculations (particularly those that don't scale well with the list of capabilities estimate) and extensive contracting of the hunt space. The objective is the decrease of the scourge of dimensionality to yield enhanced arrangement precision. Another advantage of highlight determination is its propensity to lessen over fitting, i.e. the wonder by which a classifier is tuned likewise to the unforeseen qualities of the preparation information as opposed to the constitutive attributes of the classifications, and in this way, to build speculation.

| Algorithm / Method | Speed in seconds for 100 docs | Accuracy for 100 docs |
|---|---|---|
| Decision Trees | 192.3 secs | 81 % |
| k-Nearest Neighbour | 186.9 secs | 83 % |
| Bayesian Approaches | 163.7 secs | 85 % |
| Regression-based Methods | 141.5 Secs | 89 % |

Table 1:

## V. CONCLUSION

Text categorization assume an essential part in data recovery, machine learning , text mining and it have been fruitful in handling wide assortment of true applications. Key to this achievement have been the consistently expanding inclusion of the machine learning group in text arrangement, which has of late brought about the utilization of the exceptionally most recent machine learning innovation inside text order applications. Many methodologies for text order are talked about in this paper. Highlight choice strategies can effectively diminish the issue of dimensionality in text classification applications. Procedure of text characterization is all around explored, yet at the same time numerous changes can be made both to the element readiness and to the grouping motor itself to streamline the arrangement execution for a particular application. Research depicting what alterations ought to be made in particular circumstances is normal, yet a more non specific system is inadequate. Impacts of particular changes are likewise not all around inquired about outside the first territory of utilization. Because of these reasons, outline of text order frameworks is still a greater amount of a workmanship than correct science.

## REFERENCES

[1]. Berry Michael W., Automatic Discovery of Similar Words, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 2004, pp.24-43.

[2]. Vishal gupta and Gurpreet S. Lehal , “A survey of text mining techniques and applications”, journal of emerging technologies in web intelligence, 2009,pp.60-76.

[3]. Sebastiani F., “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, vol. 34 (1),2002, pp. 1-47.

[4]. Zu G., Ohyama W., Wakabayashi T., Kimura F., "Accuracy improvement of automatic text classification based on feature transformation": Proc: the 2003 ACM Symposium on Document Engineering, November 20-22, 2003,pp. 118- 120.

[5]. Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati , “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference , IEEE computer society,2005, 1-8.

[6]. D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz,“A decision-tree-based symbolic rule induction system for text categorization”, IBM Systems Journal, September 2002.

[7]. Kim S. B., Rim H. C., Yook D. S. and Lim H. S., “Effective Methods for Improving Naïve Bayes Text Classifiers”, LNAI 2417, 2002, pp.414-423.

[8]. Klopotek M. and Woch M., “Very Large Bayesian Networks in Text Classification”, ICCS 2003, LNCS 2657, 2003, pp. 397-406.

[9]. Joachims, T., Transductive inference for text classification using support vector machines. Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. I. Bratko & S. Dzeroski, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, ,1999 ,pp. 200–209.

[10]. Forman, G., an Experimental Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research, 3 2003, pp. 1289-1305.

[11]. Brank J., Grobelnik M., Milic-Frayling N., Mladenic D., “Interaction of Feature Selection Methods and Linear Classification Models”, Proc. of the 19th International Conference on Machine Learning, Australia, 2002.

[12]. Torkkola K., “Discriminative Features for Text Document Classification”, Proc.International Conference on Pattern Recognition, Canada, 2002

[13]. Forman, G., An Experimental Study of Feature Selection Metrics for Text Categorization.Journal of Machine Learning Research, 3 2003, pp. 1289-1305.

[14]. Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F., “Feature Selection Algorithms to Improve Documents Classification Performance”, LNAI 2663, 2003, pp. 288-296.

[15]. Soucy P. and Mineau G., “Feature Selection Strategies for Text Categorization”, AI 2003, LNAI 2671, 2003, pp. 505-509.

[16]. Kennt Ward Church and Patrick Hanks. Word association norms, mutual information andlexicography. In proceedings of ACL 27, pages 76-83, Vancouver, Canada, 1989.

[17]. J.W.Wilbur and k.sirotkin, “The automatic identification of stop words”, 1992, pp. 45-55.

[18]. Larkey, L.S., A patent search and classification system. Proceedings of DL- 99, 4th ACM Conference on Digital Libraries, eds. E.A. Fox & N. Rowe, ACM Press, New York, US: Berkeley, US, 1999,pp. 179–187.

[19]. Amati, G., D’Aloisi, D., Giannini, V. & Ubaldini, F., A framework for filtering news and managing distributed data. Journal of Universal Computer Science, 3(8), 1997 ,pp. 1007–1021.

[20]. Weiss, S.M., Apt´e, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T. & Hampp, T., Maximizing text-mining performance. IEEE Intelligent Systems, 14(4), pp. 63–69, 1999.