IBTC Model: A Model for Classification of Queries

Ankit Jain
Dept. of Information Technology
Mahakal Institute of Technology, Ujjain (M.P.)

Abhishek Raghuvanshi
Dept. of Information Technology
Mahakal Institute of Technology, Ujjain (M.P.)

ABSTRACT

Text classification is a way to categorize data in various classes, which shows some identical properties. Data can be easily utilized if once it is stored in well arranged manner. Classification task is used in text mining for mainly two reasons one for store data in well manner in categories and another is to retrieve data as per user requirement in few milliseconds i.e. searching. If data is store in proper place retrieving the data and its information can be retrieved in less time then unstructured data. In text classification if the target text which user wants to search is present in the form of sentence known as query, then it is difficult to search particular document [1]. This paper proposed IBTC model to provide efficiency to search when the target text in the form of query. The results and snapshots show validation and accuracy of proposed model.

Keywords

Text Classification; K-nn Method, VSM.

1. INTRODUCTION

Now a day's data can be collected from various sources and various systems are engaged to make it useful for various task. The collected data may contain useful data relevant to some criteria and rest data is un useful. But at last all collected data is useful for different sort of works. To utilize the usefulness of data the main aspect is the arrangement of this data. If data is arranged in well manner then it is useful otherwise the properties of data are hidden [2]. Classification is the technique to sort data on the basis of predefined categories[1].These categories and classes how some special property or information. For example In newspaper or magazine we find information regarding word, nation, state wise even topic wise like Hollywood, whether, lifestyle.etc [2]. So, it is a best example to understand the categorization process and its importance. When news is store in various categories user can read the news of same category in same place which is convenient and sensible [3].

Text classification is also used in various search engines to search relevant topics regarding given input. When the input is in the form of query which contains multiple words the searching process may take more time or sometimes even fails. So, there is requirement arises in this content to propose a new model which works efficiently for queries. There are some models already proposed which uses k-nn method for classification but the short comings of these models are present that is when the value of k increases the system performance degraded [1]. So, here arises a need of efficient system which performs effectively for any value of k. This paper proposed IBTC model for queries to overcome above mentioned drawbacks.

This paper lime light the basic concept and text classification model in section II. Section III contains the main part of this paper IBTC model and comparative study of IBTC with previous proposed models this section also highlights the advantages of IBTC. Section IV contains conclusion and future work.

2. BACKGROUND

2.1 Text Classification (TC):

From last ten years the popularity of TC increases .Text classification decreases the overhead of manual entry and categorization of data. Many industries which needs huge amount of data adopting this technique to store data in well mannered form and divided in multiple categories. For example in any book shop books are divided in categories on the basis of authors, awards, by country, by publisher, by topic or alphanumeric form. Data Preprocessing, Classifier Construction and document categorization are the main phases of text classification. Each stage contains sub stages. The main task of TC is to assign an unclassified document into categories [8].

Input Data: Input data may be a document or text. This document is a target document, text classification process applied on target document. The term target is used here because the predefined classes are assigned to target data after analysis [9] .

Data Pre-processing: To assign a predefine category to a target documents it is necessary that the target documents are in some specific format. So in this step all target documents are converted into a unified form so that it provides essential details easily for analysis [10]. Data pre-processing contains some sub stages: Document Converting, Functional word Removal Feature Selection, Feature Weighting.

Classifier Construction: In this stage a classifier is design by predefine documents, and thus classifier is constructed. This classifier is used to classify unknown document which are known as target documents. Some of text classifications algorithms are SVM, Rocchio, K-nn. IBTC model is based 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. 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 [11].

Document Categorization: In this phase Documents are classified. 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 or not [2,12].

2.2 K-nn Approach:

K-nn algorithm is a similarity based learning algorithm [4]. In any target document using K-nn algorithm k nearest neighbors 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 [5,13]. If several neighbor shares a category, then the pre neighbor 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 [6, 14].

In KNN (K-Nearest Neighbors') "Similar" item are searched and stored in a categories. So, we need a functional definition of "similarity" if we want to apply this automatically [7].

2.3 Vector Space Model:

Vector Space model is mainly used for text mining. IBTC model based on vector space model (VSM). VSM is consisting of four stages to perform classification task. In first stage extracts words tokenization method is applied. In tokenization process sentences are divided into words and each word is considered as a isolated element. The second step of VSM is stop words removal in this step prepositions are removed between elements which are generated from tokenization. Stemming is the third step of VSM; all steps are interrelated and process data further for classification.

In stemming phase the elements are converted into their root form. For example elements like going, gone are replace its root word go. The rest elements are the target elements are considered as features and used for searching. If any document contains these features is present then that document is retrieved as output [6].

3. IBTC MODEL

IBTC model for queries is a model for text classification when the input is in the form of query, where each query contains various words. The classification task is applied on two types of input, one when we want to search some document or text and another when we want to store a new text. In previous proposed models there is a contradictory condition for those documents which shows the different sort of properties then predefined classes. So, the assignment of this type of document to class is quite difficult. In some proposed models the marginal data is stored in a extra classes and it contains multiple documents which are not similar to each others.

In this case the searching of these kind of data is difficult and shows inefficiency to the system. IBTC model for queries is a model which covers both task of classification searching as well as classification of text. Figure 1 shows the architecture of IBTC model, which contains seven steps.

1. Input Data: The input data for IBTC is two types one is for searching. In searching process for a given query is based on k-nn method where the system find out the K documents based on the given query. Another task of TC is to classify the given text into already exist classes also known as training data. The given text is analyze with the existing text and if some terms match with the training set the input text is assign to that category. If the given text is different from all existing text then the text is assigned to new category which is set from the maximum index values of that text. Figure 2 shows masters table.

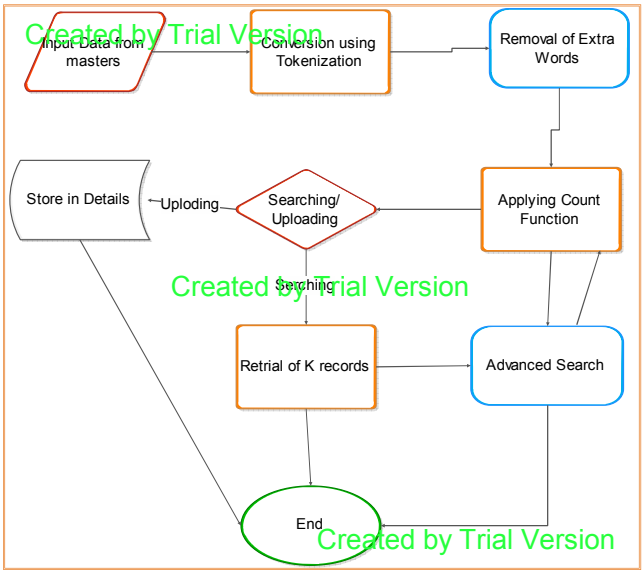

Figure 1. IBTC Architecture

2. Conversion using Tokenization process: Given input is in the form of query which contains multiple words. In this step the query is converted into words. Tokenization process is used in IBTC model to convert query into n words, for this the word is selected by detecting spaces between them and finally the results are stored.

MASTERS

SNO	AUTHOR	WEB_LINK	ABSTRACT
1	1 Jo	http://en.wikipedia.org/wiki/Computer_network	A computer network, or simply a network, is a collection of computers and other hardware interconnected by c
2	2 Mink	http://en.wikipedia.org/wiki/Test	BOB HOME FOR BOB by HOME the HOME at EAT
3	4 Jhon	http://en.wikipedia.org/wiki/DBMS	A Database Management System (DBMS) is a set of programs that enables you to store, modify, and extract
4	3 Bob	http://en.wikipedia.org/wiki/Operating_system	An operating system (OS) is a collection of software that manages computer hardware resources and provides

Figure 2. Masters Table

3. Removal of extra words: In this step the term extra word is used for special symbols and prepositions. The input query contains words as well as prepositions and special symbols. Using step 2 and 3 this model removes extra words to improve searching and indexing efficiency. Figure 3 and Figure 4 shows preposition table and special character Table.

PREPOSITION	
WORD	
1	A
2	BY
3	FOR
4	FROM
5	IN
6	OF
7	ON
8	THE
9	TO
10	WITH

Figure 3. Preposition Table

4. Applying count function: After removal of extra words the counting of all words is stores. It shows the count value for a single word present in text. Figure 5 shows results of count function.

SPECIAL_CHARACTER	
CHART	
1	,
2	.
3	:
4	;

Figure 4. Special Table

5. Retrieval of Result: When classification is used for searching a query all steps from 1 to 3 are performed. Step 3 applying count functions which is applied for all stored dataset and the retrieval of result is shown on the basis of highest count value for n words and k parameters which shows k-nearest neighbor approach. If K is set 10 then the result is retrieved and shows 10 texts which contains the words from input query's-nn method described in section II.

OUTPUT
Sno 1
TOTAL NO OF WORDS FOUND : 78 # TOTAL_NO OF WORDS INSETED : 60 # TOTAL_NO OF WORDS UPDATED : 18
Sno 2
TOTAL NO OF WORDS FOUND : 7 # TOTAL_NO OF WORDS INSETED : 4 # TOTAL_NO OF WORDS UPDATED : 3
Sno 3
TOTAL NO OF WORDS FOUND : 35 # TOTAL_NO OF WORDS INSETED : 23 # TOTAL_NO OF WORDS UPDATED : 12
Sno 4
TOTAL NO OF WORDS FOUND : 219 # TOTAL_NO OF WORDS INSETED : 123 # TOTAL_NO OF WORDS UPDATED : 96

Figure 5. Result of Count Function

6. Advanced Search: In IBTC model if the user is not satisfied from the retrieved results. There is a option of advanced search where user can select priority between various words and the system is again process for searching and retrieves results.

7. Categorization: If input is a text and used for classification after processing step 3, this step is performed. In this step categories are applied to the text on the basis of maximum count value of word. If words present in input text also present in another text which have a certain category system allot that category to input.

	SNO	WORD	CNT
2	1	COMPUTER	1
3	1	NETWORK	4
5	1	SIMPLY	1
7	1	COLLECTION	1
8	1	COMPUTERS	1
11	1	HARDWARE	1
12	1	INTERCONNECTED	1
13	1	COMMUNICATION	1
14	1	CHANNELS	1
16	1	ALLOW	1
17	1	SHARING	1
18	1	RESOURCES	1
19	1	INFORMATION	1
23	1	LEAST	2
24	1	ONE	3
25	1	PROCESS	2
26	1	DEVICE	2
27	1	ABLE	1
28	1	SEND/RECEIVE	1
29	1	DATA	2
30	1	TO/FROM	1
31	1	RESIDING	1
32	1	REMOTE	1

Figure 6. Retrieval of results

If any similar word not found then the system allot a new category on the basis of highest count value of word and that word will be new category. Figure 7 shows the algorithm for IBTC model.

Algorithm for IBTC Model:
//M master Table, P Preposition Table, S Special character table D details table x single record, i single word cnt count value which checks the occurrence of each i in D, p used in advanced search where user select some words from i which have more priority then otters. Cat value shows the presence of category allocation. If category is not allocated to any x then this algo. Will allocate category to that x.

Step 1: Text conversion using Tokenization method and removal of extra words
a. Select abstract from m
b. Compare i with P

For i if any word present in P then replace i value with space
c. Compare i with S
For i if any word present in S then replace i value with space
d. For x=1 to n.
Do
Replace each space with *.
e. If multiple * present in x then
Replace with single *
Else
Go to step 2.
f. Store i .

Step2: Apply count function
a. Insert i and cnt =1 in D.
b. for i=1 to n check cnt value
c. update current cnt for i in D.

Step 3: For upload a text go to step 6.

Step 4: Set K value

Step 5: Show results as per K value.

If data found is relevant to use need then go to Step 7.
else
go for advanced search
a. select p words from i.
b. go to step 4.

Step 6: Category allocation for x
a.If for x any three i values match with any category y of any document v then
set x[cat]=y[cat].
else
set x[cat]=max cat[i].

Step 7: Exit

Figure 7. IBTC Algorithm

4. CONCLUSION

Indexing based text classification model (IBTC) is used for classification of data which contains multiple words or sentence known as queries. Query may contain n words so, for n length query there is a possibility that many documents belongs from that query. For example: the query "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 so the decision of category of this is quite difficult [7]. IBTC model overcomes these types of drawbacks. IBTC is based on VSM model and uses k-nn method both to classify as well as searching of queries. The main advantage of this document that it not only checks the title of text it scans whole text and find out the documents based on incoming queries.

The performance of this system will not degrades with the increasing or decreasing value of K. the reason behind it is the system retrieves text details as per the count value of the elements of queries. And the count values are already calculated when the text is uploaded in the system. So, for searching purpose the system only retrieves the documents related to it. In section III. The results are shown which validates the implementation and accuracy of the system. In future the system is extended for scanning directly the whole document for generating indexing table.

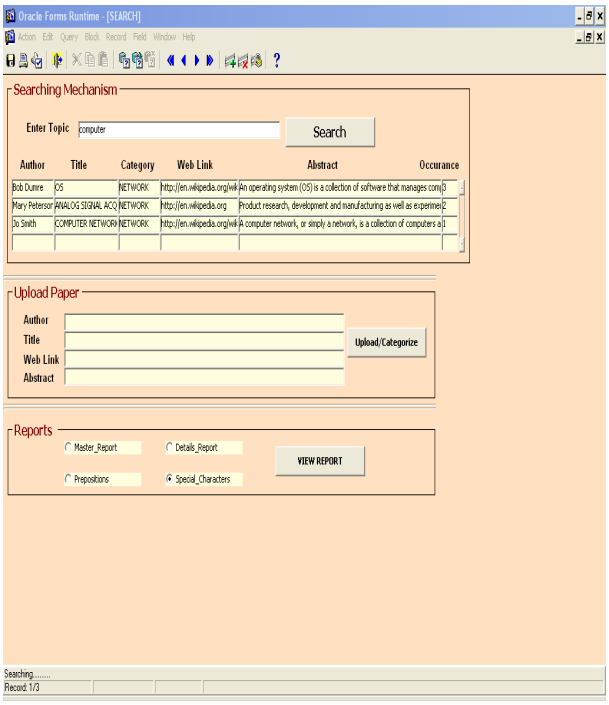

Figure 8. Snapshot of IBTC Model

Figure 9. Snapshot of IBTC Model to upload a new Paper

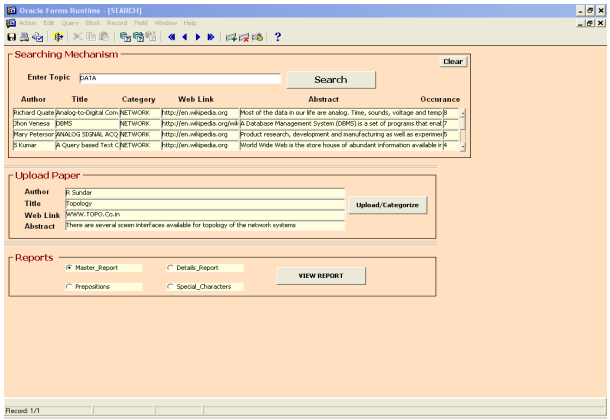

Figure 10. Snapshot of IBTC Model to show Tokenization Process

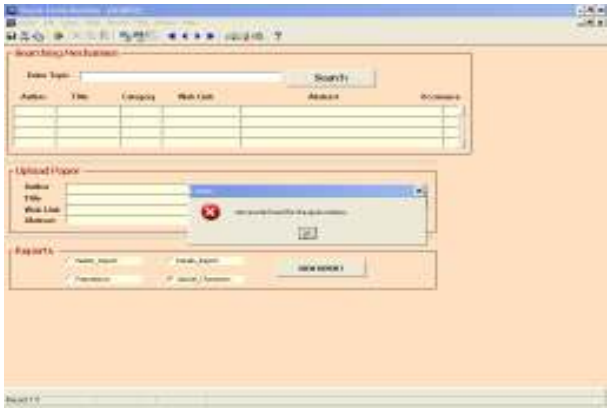

Figure 11. Snapshot of IBTC Model to show Result when no Record Found

5. REFERENCE

[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.