Semantic Classification, Keyword Mining and Search Space Optimization for digital ecosystems

Nikunj Yadav, Yanu Gupta, Manish Kumar, Ratna Sanyal
Indian Institute of Information Technology Allahabad, India
{nikunjyadav@rediffmail.com, yanu27@gmail.com, mani.iiita@gmail.com, rsanyal@iiita.ac.in}

ABSTRACT: *The volume of documents in the digital repositories numbers in thousands and is increasing constantly, in such a scenario it becomes a very important issue to organize and retrieve these documents in a way that relates to the human mind. In this paper, we present a novel approach to classify the documents in a digital repository and find the semantically significant keywords related to those documents to make the organization and the retrieval of the documents faster and more efficient. We approach this problem using Probabilistic Latent Semantic Analysis with incomplete training data to organize them and mark the relevant keywords. This approach makes the classification faster and instead of the unlabeled clustering gives classification with well defined topics relating to human logic.*

Keywords: Classification, Semantics, Query Retrieval, Keywords

© DLINE. All rights reserved

1.Introduction

Digital Repositories are constantly being worked on to preserve important documents and to give ease of access to these documents to a user. So generally there are important documents as well as books stored in the digital format. In this scenario when there are large repositories, the storing of these documents in an organized way and accessing is an aim that needs to be emphasized upon.

Although these documents and books contain images, diagrams and text but we focus on classifying the documents based on text only. Another issue involved in operating on these documents is that they are generally stored in formats like pdf or images etc, which cannot be directly used in our approach, we leave the details of converting these documents in the format which can be directly used in the classification process, although it is fairly easy to convert the pdf documents or scanned images into rich text format or techniques like Optical Character Recognition (OCR) [1] can be used with good efficiency to obtain text format file, interested readers can look into the details.

Searching documents and arranging them requires discerning at the semantic level using unsupervised learning, the best that can be done is to cluster the documents on probabilistic models but in that case the clusters are not tagged with their semantic meanings, hence we need some approach that will make the use of incomplete training and yet in the end we can relate to.

The crux of the approach lies in the fact that the training data does not need to be extensive, but based on this incomplete training some sort of manipulations can be done (as explained later). Based on the probabilistic model the documents will be assigned to a label which they belong to the most, using expectation maximization algorithm [2]. This approach will also help in finding the important words which are called keywords based on the fact that documents belong to some particular topics with some probabilities, based on this and the number of times the terms have occurred in the document, the terms can be arranged in their order of relevance. An observed major disadvantage of our approach is that since the approach maximizes the conditional probabilities based on the initial training documents do not belong to large number of topics, or have multiple levels of hierarchy.

Henceforth presented details are divided into 8 sections, section 2 explains about existing techniques and their advantages, section 3 explains model and training, 4 explains classifying documents and 5 explains keyword spotting, section 6 shows some experiments and results, section 7 explains conclusions and section 8 gives future research.

2. Related Work

There are many existing techniques for the document classification, but for keyword spotting in query retrieval there are no such fixed methods. Let us present some existing methods:

2.1 TF-IDF

It's a method of weighing terms in each topic such that their relevance is measured in terms of their frequency of appearance in the concerned topic/document as well as in the other topics/documents. This is measured in such a way that the weight of the term in the particular topic/document is decreased if the term appears in other topics/documents as well with considerable frequency [3].

2.2 PLSA

Probabilistic Latent Semantic Analysis (PLSA) [2][4] follows the statistical theory of learning the natural language, by means of which we can make the system distinguish between different topics and find out the relatively similar documents and thus clustering similar documents into a single topic. The topic has no definite meaning. They may or may not represent a real world entity/abstract subject etc.

This technique is based on the expectation maximization method [2][5] of finding maximum likelihood[2][6] in probabilistic models. We maximize the log of the functions of prob(terms | docs) as given in the research paper [2].
The input to PLSA[1] are random values of conditional probabilities P($wj|zk$) and P($zk|di$) [2], which are finally converged to a maxima and rarely a minima according to the Expectation Maximization algorithm [2][4]. Based on these final values, we can decide to which topic the document belongs and with what probability. Classification can further be carried out based on these values.

2.3 Support Vector Machines

Support vector machines are methods widely used for classification. It divides high-dimensional space by set of hyper planes, thus achieving classification [7]

2.4 kNN

K-nearest neighbor is technique used for classification. It achieves classification by comparing with closest training samples. It designates a class to a point by the votes from the nearest neighbors. The space in which the points (documents exist) can be high dimensional, the axes of which need to be chosen carefully. [5]

2.5 Latent Semantic Analysis

Latent Semantic Analysis is generally used in natural language processing, to analyze relationships between documents and terms with the help of concepts. Using the concept space thus defined the documents can be classified. This technique is also used to identifying relationships between documents and terms.

3.Training and Initialization

Although PLSA[2][4] is a probabilistic model which is primarily used for unsupervised classification, we observe that with a small amount of precise training the algorithm's probabilistic precision can be used to assign each document a topic.

Let us first explain what training step achieves, as explained before, EM algorithm makes use of the two steps to attain a maximum point in the probability function log curve. Training suggests that we will bias the initialization of probabilities prob(terms | topics) and prob (topics | docs) [2]. The biasing will not be arbitrary but it will be based on the training weights of each term. Biasing here means that instead of initializing the data randomly (choosing a random starting point) we are trying to use the incomplete information (training) to locate a point which is closer to the maxima (starting from an intelligent point). Thus we empirically state that training will locate the point on the curve which is closer to the maxima thus making sure that global maximum is obtained promptly, and topic is assigned a label like music, sports rather than abstract topics.

For training we suggest keeping in mind the following factors:

1. Documents have terms which precisely relate to each topic.

2. Results are better if the terms occurring for one topic occur in it exclusively.

3. Weighting techniques for each term in training data chosen meticulously.

Training is a crucial step as improper or clumsy training leads to ambiguity within topics and during the execution the PLSA [2] might conflate the documents within the categories, as the initial point takes a route to some local maxima. Another aspect to notice is that, since we are doing classification in huge digital repositories, we recommend that training need not be detailed in terms of the overlapping topics and multiple levels of hierarchies. Training data chosen such that the classification labels are not subsets of each other, or sets which are very similar suffices. Such an approach will be more than useful to classify the documents into general categories in the repositories. This can be extended to multiple hierarchies [7][10] and similar topics but in that case collecting data for training becomes a menial task. Fundamentally any amount of extensive training will work as we are suggesting that it does not matter as long as the training defines the topic, an optimum starting point will be located and PLSA will converge to maxima [2]. Extensive training in fact is conducive to differentiation between documents but an optimum level needs to be fixed.

As pointed before the weighting technique is an important issue. Weights will be the means to provide initial probability values to terms with respect to the topic. Now as explained in the research [3] there might be some terms which although frequent, if occur in many documents, become less important. For this purpose we make use of *tf-idf* [3] technique. Weights can also be assigned manually by the experts for an expert supervised classification.

Training can be done in two ways; either the expert can collect relevant terms for each topic and assign them weights which will give accurate results. Alternative is to use large number of documents already tagged with their classification topics and after stemming the documents and removing stop words, irrelevant terms are rejected (as explained before like using *tf-idf* [3]). The remaining terms are arranged in descending order according to relevance of weights and thus can be used for the initialization process. During initialization of matrix of prob (term | topics) first it is randomized, then normalized and then each term is looked up in all the topics and thus prob(term | topic) [2] accordingly is biased with help of the weights in the training file, i.e. if the term *i* is found in a training data of topic *j* with a weight *x* then prob(term *i* |topic *j*) will be biased according to *x*. Please note that for doing this a mathematical function would have to be made which will bias the prob(term | topic) automatically according to the scores.

Simultaneously prob(topic | doc) [2] is initialized with random values between a bound range and then normalized, please take a note that sometimes doing this leads to a local maxima instead of global maxima. Alternative approach can be adapted by first initializing this uniformly and then biasing for each prob(topic | doc) according to the number of terms it already has belonging to the topic and number of terms it does not, for each topic one by one.

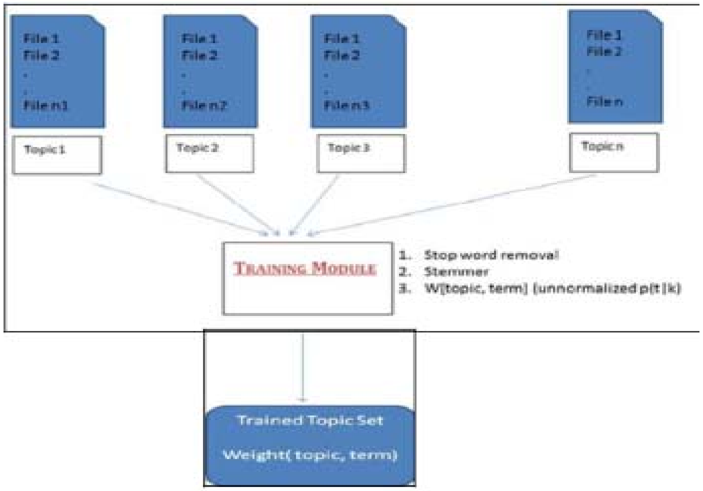
Algorithm for the Prob(term *i* | topic *j*)

a) Initially a set of documents are taken for the training, after filtering the words (stop words, and stemming of words), the frequency of different terms are calculated for the terms.

b) Now you have a set of vocabulary *Vj*. We assume that the vocabulary is fixed. The size of the *Vj* is *sj*.

c) Randomize the Prob(term *i* | topic *j*), (term is the *i*th term in *V*) and bias the value with the freq(*Vij*)/sum (freq(*Vj*)) in topic *j*.

d) Normalize the matrix Prob(term | topic)



Algorithm for the Prob(topic | doc)

a) Randomize the probabilities in a bounded range.

b) The set of the terms in doc *j* is set *Sj*.
Calculate sum (for all terms freq (term) in doc * score in training file (term))
Divide the sum /sum for all terms in training (score (term))
Bias with this value

c) Normalize the probabilities

4. Classification

After initialization, the matrices containing prob(term | topics) and prob(topic | docs) are obtained and tf-idf matrix for the documents is calculated. Now PLSA [2] or a similar probabilistic model iteratively executes on the maximum likelihood function in Expectation and Maximization steps to obtain the global maxima [2][5][8]. After the execution most optimum prob(topic | doc)[2] will be obtained, if matrix is stored with topic as row major then each row indicates for each topic how much is it significant in context to each document. Using mean and standard deviation we can select few documents for the topic in this row i.e. not taking the documents with lower probabilities as they belong to the topic less than others for example taking the documents within μ-s to μ+s where μ is mean and s is standard deviation , this will be done for each row similarly.

After doing this for each topic we will have certain number of documents belonging to it thus achieving classification. Readers should keep in mind that since the initialization was done based on the training which was tagged with the topics, the classification here is in well defined topics used in training. In repository thus the documents will be retrieved when a user enters the name of the category, a major advantage over unsupervised clustering.

Our main focus is yet to be explored though, i.e. how do we use this approach to classify the huge number of documents and the books. To achieve this we empirically state that in digital repositories, instead of the whole book or documents, considering specific carefully picked parts yield accurate results. Our results propose that for a book we should consider introduction, preface, author's name, index and the introductory paragraphs of chapters.

Similarly for research papers abstract, introduction and conclusion should be considered. Doing this would reduce the number of computations hence leading to expeditious classification.

5. Keyword Spotting and Query Retrieval

5.1Keyword Spotting

Keyword spotting means finding out words that are most relevant for a particular document, they can either be the most prevalent terms or the latent terms which are associated to the topic that document belongs to. This is mining semantic information from the document, we propose to include terms which although might have very less frequency in that particular document. These terms are included with respect to their relevance to the topic which the document belongs (through weights in the training data). This ensures that the less frequent terms are inducted amidst the frequent terms since they should be significant to the document if they are very significant to the topic which the document belongs to. This gives the system an abstract way to see the keywords not on the basis of their occurrence but on the basis of their relevance to the document. This is done in following way:-

1. For each document let *K* be set of keywords. K={ Φ };

2. Let *T* be the vector of topics that the document belongs to, i.e. for every document, *i* document belongs to topics *Ti* with some probability.

3. For each topic set *S*, every term *Si* of set *S* is inducted in *K* with score:-

$$\sum_{j} Pj$$

Pj= (tf-idf score of each term) *x* (weight of *Si* w.r.t topic *Tj*) *X* prob(*Tj*|*doc*)

4. For each word in document not belonging to any of the trained topics in *T*, induct it into the *K* with the score : -

(*tf-idf* of term *Si*) *x* (average weight of trained topic)

Average weight of trained topic=

$$\frac{1}{Y}\ \frac{\Sigma\ j(\Sigma\ \forall i \in_{T} \text{Score of } S_i}{length\ of\ S\ of\ Tj} * Prob\ (Tj\,|\ doc)$$

$$Y = \sum_{i}\quad prob\ (Tj/doc)$$

5. Arrange the *K* in decreasing order of scores, choose some number of terms as keywords or take the terms having score between μ-σ to μ+ σ , this set is *K*`

Readers should take a note that if a document does not belong to the topic prob(topic | doc) for that pair of topic and document is zero. Also mentioned above when we are storing spotted keywords, we have a whole set of words *K* and then we shortlist it to obtain a smaller set which are finally called keywords *K*`. Both are stored for query retrieval.

5.2 Query Retrieval

The main focus of the paper is to present approach for keyword spotting and document classification, but although keyword spotting might seem like a tedious task it very efficiently supports query retrieval. Once keyword spotting is done, it is fairly easy to return the results for a query on semantic grounds, which will be very helpful in retrieving documents from the huge repository. This is done on the basis of the scores calculated in above manner. Again since the calculations above maintain a semantic approach, the results will be again on the basis of the their relevance to the document taking into the account the kind of topic document falls into and not on the basis of occurrence of the terms in query. Query Retrieval mentioned here is very primitive, readers should please refer to the future work. Query Retrieval is done in the following way.

1. The Query is input, the stop words are removed

2. Query is stemmed.

3. Each word is taken from the query and its score is looked up in the set K where all the keywords with their scores which we originally stored before taking a few of them as keywords (not K‘). Score of the query is Σ (score of each terms),

4. The documents are arranged in the decreasing order of their scores

5. The results are viewed

Note that here we assume that the query has no semantic meaning, the score of a document for a query, depends on the sum of score of the document for each query word, not on the semantic meaning of the query which means it doesn’t matter if the terms of the query are shuffled or the query as a sentence makes no sense at all. Though it has been taken care of during the training itself that the document would assign the score to each term based on its context in the training document. So, at the document level semantics of each term or context is available, but not on the query level, which allows the user to enter Boolean queries too. For eg, he can enter a query like :

Which means give higher score to the document that contain both A and B but not C. Now if the term C has very low score in some document's context, it won't make much difference on the overall result but if its very important in some other document's context, its score would decrease significantly, thus lowering its priority in the result.

As mentioned before, the motive of the work presented is keyword spotting and classification. What the process suggested here can do is, rather than applying the usual query retrieval algorithms after indexing the documents, the search space can be reduced manifold. With this process since you know the context of the query with respect to the documents, the importance of calculating the context at the query level for that particular document can be decided upon based on the score the method generates here. If a document is not relevant to the query as it doesnt contain those keywords, or it is not semantically related to the query with respect to the terms then the documents need not be calculated for, at all.

We also suggest one more capability here, if a user after using this kind of classification for a purpose wants to input a document and gather information about which category it might belong to in the digital library, we need not do the PLSA iterations[2] again, instead we can assume the document as a query and put it straightforward in the topic that the document retrieved with the highest score (in response to the query) belongs to for the time being, it will be true since the retrieval is based on the words which are relevant to the document instead of frequent as mentioned above. Although proper classification should be done by redoing the whole process again with the new set of documents, although it makes sense to repeat the process again only if there is considerable number of documents being added.

5. Experiments and Results

Repeated testing after the last submission [9] has fetched out very successful results, of course the testing of the approach is yet at a smaller level and has to be scaled manifold to be applied in an application. We tested documents ranging from number 70-100, training data for 20 tagged topics with training documents with 30 short documents from each. The topics are listed below

Topic
1. Health
2. Politics
3. Business
4. Sports
5. Religion
6. Bio-Informatics
7. Terrorism
8. Technology
9. Entertainment
10. Academics
11. Fashion
12. Family
13. Books and Magazines
14. House Related
15. Environment
16. Women
17. Children
18. Holidays
19. Video Games
20. Love and Relationships

Table 1. Table containing experimental topics

As it is fairly visible that the topics that we used are very different, so when we used PLSA [2][4], it used to converge the documents completely to the topics, if the topics are overlapping then the prob(topic | docs) for a document is not totally biased for one topic. Therefore because of this reason sometimes the documents might cross the boundary between the overlapping topics for example here bio-informatics and health.

Since we are claiming that biasing in the matrices i.e. starting from a point on the curve which might be closer to the solution we made sure in the experiments to vary the number of topics from 8-20, and with each number of topics the results achieved were efficient with around 7-12% errors. But this might be too ambitious since the topics in the digital libraries will be larger in numbers and a lot more closely related.

We also experimented on the PLSA without any prior biasing, in that case the efficiency reduced considerably, also in the unsupervised version the topics were abstract hence they are of no use in the digital library whatsoever because if a user searches the documents based on the category then the topics must be concrete.

During our experiments we changed the testing data set 200 times by totally changing the documents, and we obtained errors ranging from 7-12%, with the exception of 40 test cases where keywords spotted got a bit mixed up, we suggest that when conducting experiment this sometimes happens if the training data set is ambiguous or insufficient. The efficiency factors can vary from person to person for example we took some documents which were the market strategies of Dell and a document which explained some exchange offer strategy of HP, hoping to get it classified as technology but it was classified in Business, many may claim otherwise. We sometimes also faced difficulties between the documents where there were documents relating to the recent cricket league and rock band documents because there were keywords like stage, dancers, music, player, public, crowd etc (because of the recently the cricket leagues are conducted). Hence during the query retrieval when inputting query like "guitar live on stage" the results were ambiguous and contained documents from both the categories, although the ranking of entertainment documents was higher and if there were a larger set of documents the result would spread over larger number of documents and thus entertainment documents would have higher ranking.

Conducting the experiments we suggest that documents with widely different topics like religion and terrorism never create difficulties. This is because the documents within these categories are so uniquely identified by their words especially when other topics are so vastly disjoint from them. Any document which are very negative, mythological, and religious, etc are easily classified as compared to the documents talking about morality, specific person, geographical, etc.

Some excerpts from the result are shown below:

Student
College
CBSE
Board
Subject
Syllabus
Secondary
Higher

Table 2. Table showing keywords spotted for document talk about cut-off marks for universities falling under academics

Art Brut releases album
Broken Family Band releases 7th album SlumDog Goes toOscars
About Robert Plant, the singer from Led Zeppelin
Jimi Hendrix's influence on guitars
Kasm's Debut album
'Ongiara' by Canda Great Lake Swimmers Ben Harper

Table 3. Showing the documents falling under Entertainment, the names are the headings documents were about

market
point
Rupees
Stock
Company
share
close
cash
plan
equity
bse (bombay stock exchange)

Table 4. Showing the keywords for typical Business document, this one talking about bombay sensex status in October

Health	Bio- Informatics	0.1
Health	Environment	0.0085
Entertainment	Video Games	0.0057
Family	Women	0.086
Family	Children	0.07
Business	Politics	0.08
Religion	Politics	0.05
Academics	Books and Magazines	0.04
Technology	Video Games	0.065
Entertainment	Sports	0.08
Environment	Holidays	0.064

Table 5. Conflations among the topics

The table above shows the mixing up of documents between several categories. The third column is the probability by which the documents can travel from one topic to other and vice-versa. This is assuming that the documents taken into consideration belong to strictly one topic as they were taken from several news articles and information pages and description articles about the different aspects of a topic. Some instances of the problems in the mixing up of the documents is for example a detailed version of holiday spots and document describing a similar beautiful location or history about a location. Technology talking about computing power, graphics computing, relating to ergonomics, mix up with the articles specifically talking about video games. Of course apart from the mixing up of the documents above we have so many conflicts between other categories but they are minimal to be described. Apart from really close topics and some initialization problems, another difficulty which causes mixing up of the documents is problem in identifying entities, the context of entities and clubbing together entities to form one entity. For example stock in terms of books store is very different from the stock in stock exchange where both the words have to be clubbed, of course the word will have different probabilities for both the topics because of the training but still issues like this caused deviations in the training.

6. Conclusion

Based on the experiments conducted by us we conclude that incorporating incomplete training leads to faster execution of PLSA towards maxima, such that it always converges to maxima where the document classification is obtained with documents being classified in the topics labeled so that the user can retrieve them back by specifying the category or topic which is much needed in digital repositories. Deploying this approach in a tool will help the prompt retrieval of queries as well as the management of the digital library using semantics.

This approach or an approach similar to this kind is strongly recommended for digital libraries instead of unsupervised clustering.

7. Future Research

The research focuses mainly on document classification and keyword spotting, subsequent work is desired to focus on query retrieval, using keyword spotting to make retrieval more efficient, making the semantics part of metadata for the repository. Currently we are treating each query as a new document and finding out its score with respect to the already available document. This score is sum of the individual scores of each query term with respect to a particular document which doesn't reflect any semantic meaning in context to the relativity between the terms of the query, in case query terms may mean differently in different context. Future researches are expected to find out a way for each query to be interpreted semantically and present their score as a group rather than sum of individual scores. Currently the work only focuses on reducing the search space for query retrieval by document classification and keyword spotting.

We can also move to hierarchical classification, which means first classifying the documents based on some categories, and then further classify documents belonging to a particular category into sub-categories, for eg. Cateogory sport may have sub-categories like cricket, soccer, hockey, tennis etc.

We can refine our research by training according to the sub- categories and then clustering those sub-categories together. Numerous ideas can be invited on this thought.

8.Acknowledgement

Authors acknowledge Indian Institute of Information Technology Allahabad for support in carrying out the research.

References

[1] Handel P W (2007). U.S. Patent 1,915,993, *Optical Character Recognition;* Sanyal S, Dhingra K D, Sharma P K: Optical Character Recognition for Degraded Text Documents. IMECS'07 p. 1988-1993.

[2] Hofmanm, T (2001). *Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning*, Vol. 42. 177–196.

[3] Nikolov, S. *A, Novel Approach to Automatic Document Similarity Measurement and Categorization web.mit.edu/snikolov/www/topicweb_verbose.pdf*

[4] Hofman, T (1999). Probabilistic Latent Semantic Indexing, in the proceedings of twenty second annual international SIGIR conference on research and development in information retrieval.

[5] Dasarathy, B. V. (2006). Nearest Neighbor (NN) Norm : NN Pattern Classification Techniques, *IEEE Computer Society* Press, Los Alamitos, CA.

[6] Hung Chi-chun, A novel gray-based reduced NN classification method, *Journal Pattern Recognition Archive*. 39 (11).

[7] Alexei V and Mark G, A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections, in Information Processing and Management, journal of Intelligent Information Systems archive Vol. 18 Issue 2-3, 2002

[8] Cortes, C., Vapnik, V. (1995) Support-Vector Networks , *Machine Learning*, 20 (3) 273-297.

[9] Dempster, P., Laird, N.M., Rubin, D.B (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm in *Journal of the Royal Statistical Society. Series B (Methodological),* 39 (1) 1-38.

[10] Nikunj, Y,. Yanu, G,. Ratna, S,. Manish (2009). K Semantic Document Classification and Keyword Spotting, In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems ACM New York, NY, USA, MEDES '09, p. 157-161.

[11]Fu, Huaiguo (2008). Scalable conceptual hierarchy based algorithm for knowledge sharing in digital ecosystem, *In*: 2nd IEEE International Conference on Digital Ecosystems and Technologies, Phitsanulok, Thailand, February.

Author Biographies

Nikunj Yadav previously a student of Indian Institute of Technology, Allahabd (IIITA) has been working in the areas of Natural Language Processing and Information Retrieval. He has been associated with the Technology Development Lab for Indian Languages for around 2.5 years, a lab dedicated to research in Natural Languages. During his stay at IIITA, he worked on the projects related to knowledge organization, document classification, and information retrieval based on probabilistic algorithms to compute semantically. Nikunj apart from being a software development engineer is working on developing new techniques in the field of Asian Language Processing.

Yanu Gupta, an alumni of Indian Institute of Information Technology Allahabad. She at the IITA was working on Information Retrieval and Natural Language Processing. She has been associated with Technology Development Lab for Indian Languages for around 2.5 years, a lab funded by HRD ministry of India dedicated to Natural Languages Processing. Yanu apart from being a software engineer is also currently working on developing new techniques in the field of Asian Languages Processing.

Manish Kumar, an alumni of Indian Institute of Information Technology Allahabad. He at IITA was working on Information Retrieval. He was associated with the project Semantic Document Classification and Keywords spotting for a long time. He is currently a student at Indian Institute of Management Ahemdabad.