Scientific Journal of Impact Factor (SJIF): 3.134

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 12, December -2015

Ontology based Information Search Platform for Comparative Analysis of Educational Data using NLP

,Prof. Mrs. R.A Rane[1], Analyst Pratik S. Deshmukh[2]

[1,2]*Computer Engg Department, MIT College, Pune*

Abstract *— The exponential growth of educational information on the World Wide Web makes it increasingly difficult to discover relevant data about a specific topic.*
There is no platform available to extract information of Indian educational universities and compare multiple universities on the basis of mostly searched factors. In this paper I have proposed a system to build such platform where user will be able to extract accurate information about multiple educational universities and also will be able to get comparisons between them.
To implement this system in practice we suggested a NLP Driven search engine on ontology based training data set of three different universities. We implemented Navie Bias classifier and updated it with a relevancy method to get ranking of mostly searched universities.

Keywords*-NLP: Natural Language Processing.*

I. INTRODUCTION

The challenge now is to develop systems capable of simulating human reasoning and enable machines to interpret the information, the ontology by its capacities of representation of the knowledge and the mechanism of reasoning and inference which he offers represents a solution to this need, thus it constitutes the main core of knowledge management systems.
Ontology's are commonly used in the process of information retrieval where the goal is twofold: "understand" the contents of the documents and "understand" the need for the user to be able to put them in relation. Indeed, Thanks to the semantics they provide, ontology's may be involved in the reformulation or the extension of user requests or in terms of indexing and annotation of documents and web resources.
This paper main objective is the proposal of a knowledge management system based on ontologies. This system has for ambition the capitalization and dissemination of knowledge in a university system. The approach presented here aims to create an organizational memory based on ontology's. The ontology produced will lead to index the documents and thus enable an extension of the classical-based metadata to a search based on semantic criteria. The main content of this ontology result from a manual extraction of the knowledge from a number of documents resulting from the daily work of the university system actors. These documents are validated and then indexed and classified using an ontology. The ontology created will be used to facilitate search and navigation within the field of knowledge capitalized [27].

1.2 Motivation for making use of NLP and ontology model for IR

Imagine we could interact with a university intranet search engine just like with a human person in a natural dialogue. The search engine would automatically extract knowledge from the Web site so that a searcher can be assisted in finding the information required. A student who asks for a particular subject can be directed to the most recent lecture notes or the contact details of the lecturer. An external searcher typing in "PhD NLE" could be assisted by allowing him to explore the space of experts and projects available in the area of natural language engineering. Obviously, this information can change any day and the idea is to have always the most up-to-date facts and relations available to assist a searcher. Currently, we do not have systems which support this type of interaction. However, our aim is to automatically acquire knowledge (a domain model) from the document collection and employ that in an interactive search system [2].
One motivation for a system that guides a user through the search space is the problem of "too many results". Even queries in document collections of limited size often return a large number of documents, many of them not relevant to the query. Part of the problem is the fact that both on the Web and in intranet search queries tend to be short and short queries always pose ambiguity and uncertainty issues for information retrieval systems. Some form of dialogue based on feedback from the system could be very useful in helping the user find the right results. This combination of NLP and IR we assume is particularly promising and scalable in smaller domains like university intranets or local Web sites[3].

Goals and Objectives

- To design and implement a system for comparative analysis of Distributed Educational knowledge.
- To develop a efficient user based query retrieval process.

- My objective is to make human query highly expressive and highly representative of reality.
- To provide mechanism that gives freedom of expression.
- To make user free from unwanted search results on his input query.
- To provide a system for comparative analysis of Indian educational universities.

II. SYSTEM DESIGN AND ARCHITECTURE

In propose system we are using the concept of ontology and natural language processing for information retrieval using Naive Bayesian algorithm from the data of different universities with accurate and correctly with ranking and paging concepts. User enter the query with respect to course, stream and branch for some important information regarding universities for education point of view for that user query your system provide correct and accurate result to the user. And one most of the advantage of our system is i am implementing ranking and paging concept means most visited and according to user requirements universities information show based on ranking and finally going to display it on user interface in the form of tables.[6]

2.1 System Design:

Below is the High level system design of proposed system. Query will be converted into pure query using NLP techniques. Naïve bias classifier probability formula will identifying that query belongs to which university from the datasets. Information will be retrieved from ontology schema developed for 3 major universities with limited datasets. A randomly generated dataset will be used to calculate rank of the university in case of multiple universities. Else data from all the universities will be displayed.

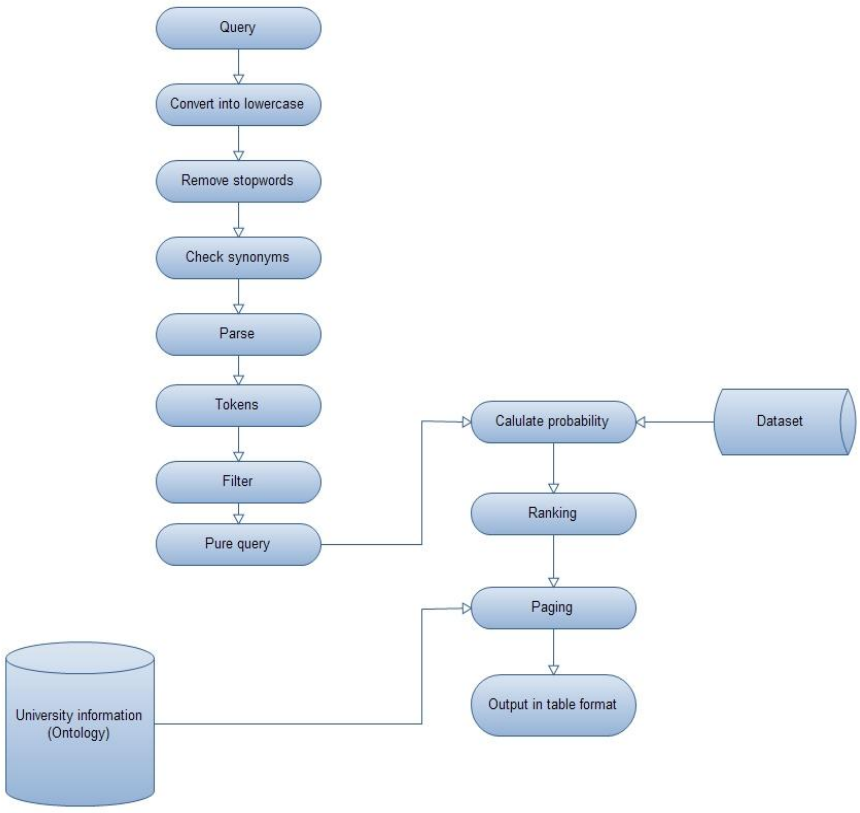


Fig 1.1 System design

2.2 System Architected:

Below diagram represents the architecture of the information retrieval platform.

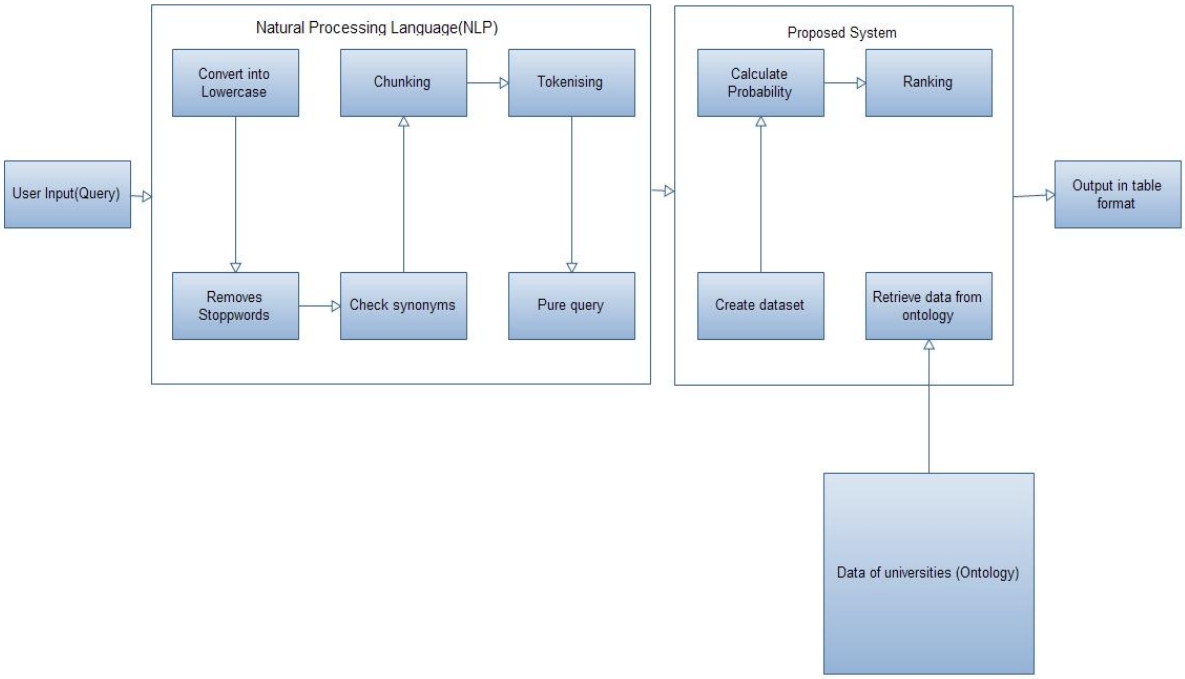


Fig 1.2 System Architecture

Fig 1.2 shows the process of parsing the user query using NLP techniques. Query content is first converted into lowercase, then Removing stop words, Check synonyms, chunking, and tokenizing, finally the pure query is then passed to next module. Next First block is consist of NLP Module, and second block represents the proposed work. Module consists of proposed algorithm, where first the probability will be calculated to find out the input items belongs to which class. Then the relevance algorithm will find out the ranking of the output results on the basis of maintained dataset which gets randomly generated.

2.3 Component Diagram

Component diagram is to show the main components/elements in the system. There are three interfaces and among that, Application User Interfaces are two. In the first Application User Interface two components are there. Response Bundle Creation is connected to Request Processing. In the next interface, i.e. Data Set Processing, it contains two elements, Ontology and Raw Data Set (it is in form of File System). Another Application User Interface holds Java Server Page (JSP) and Static Resource. Whereas, JSP is connected to Request Processing of the first Application User Interface.[9]

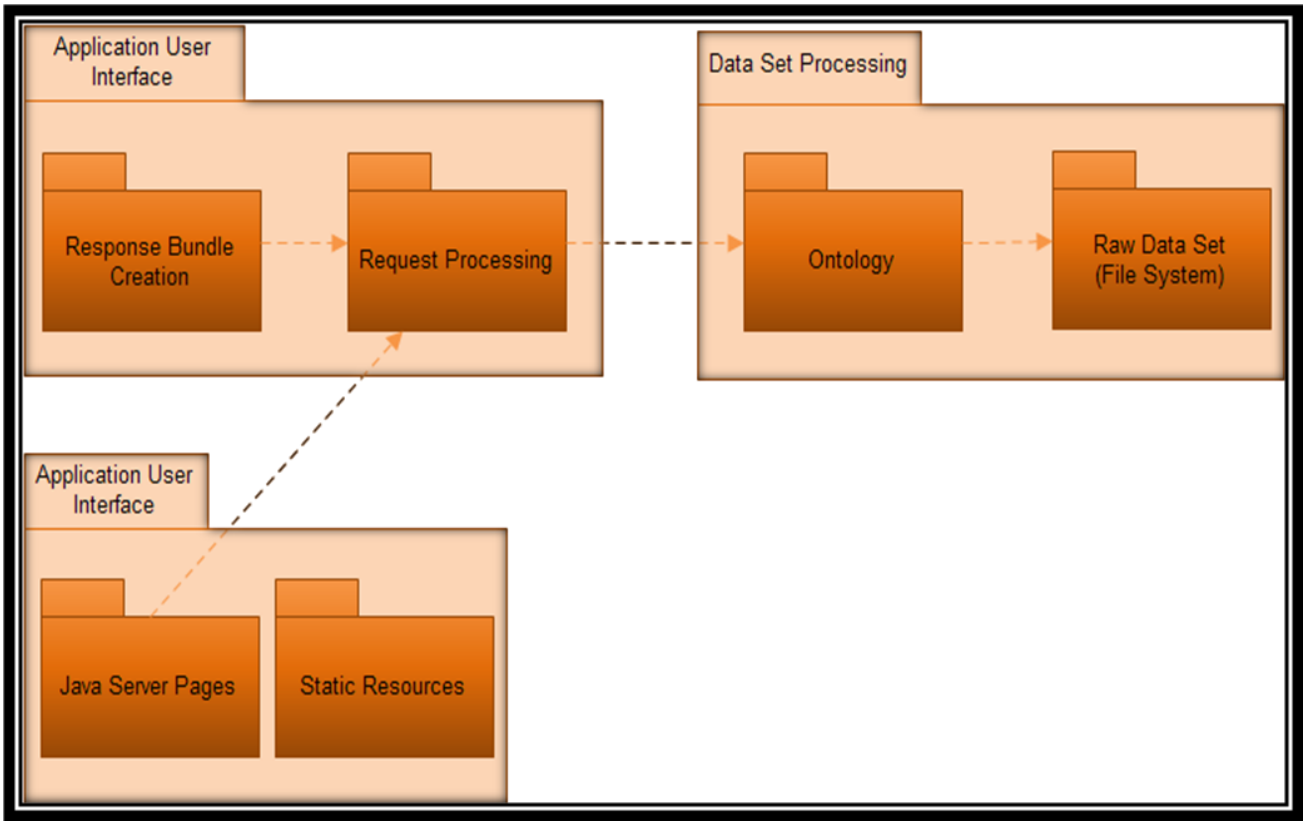


Fig 1.3 Component Diagram

III. RELEVANCE ALGORITHM WITH NAIVE BAYESIAN CLASSIFIER

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. [11]
Bayes theorem provides a way of calculating the posterior probability, P (c|x), from P(c), P(x), and P (x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.[17]

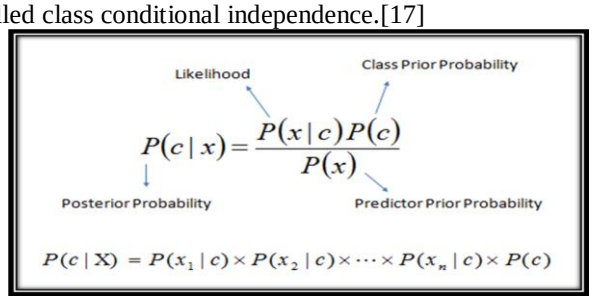


Figure1. 1: Naive Bayesian Algorithm

- P (c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P (x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Algorithm Steps:

1) Start
2) Declare variables course, stream, branch, query, pure query, fix_path_display, user_input, segments, contains, information, university ranking, Aspect_through_NB_order.
3) Access course, stream, branch and query from search page.
4) Query = query.lower_case();
5) Pure_query = removeStopwords(query);
6) Check synonyms for universities and assign to pure_query.
7) Pure_query = synonyms.checkSynonyms(pure_query);
8) Removes extra words except stop words from query and assign to pure_query variable.
9) Assign select course ,stream and branch to user_input variable
10) Identify contents from pure query
11) Initialize variable i=0
12) Repeat the steps until i<Information.size();
    - 12.1) If pure_query.contains (Information of ith position) then
    - 12.2) add information of ith position to contains
    - 12.3) increment i by 1
    - 12.4) Go to step 12.
13) Initialize variable i=0
14) Repeat the steps until i<university. size();
    - 14.1) If pure_query.contains university of ith position) then
    - 14.2) add university of ith position to , Aspect_through_NB_order
    - 14.3) increment i by 1
    - 14.4) Go to step 14.
15) Create object of proposed algorithm for ranking of university
16) Create dataset for university
17) Maintain dataset
18) Retrieve dataset for calculation
19) Call ranking method using proposed algorithm object and pass parameter as user_input and contains of query
    - 19.1) relevant.rank_university(user_input,contains)
20) Calculate probability for each university
    - 20.1 initialize variable i=0, uni_probability[10];
    - 20.2 Repeat the steps until i<university.size();
    - 20.3 Calculate probability using given formula

      $p(\underline{cj}| d) = p(d|\underline{cj}) * p(\underline{cj})/p(d)$

      d = attribute/(instance)

      $\underline{cj}$ = Class attribute

      $p(d|\underline{cj})$ = Probability of attribute is assign class in dataset

      $p(\underline{cj})$ = count of class in dataset

      p(d) = probability count of attribute(instance in dataset)

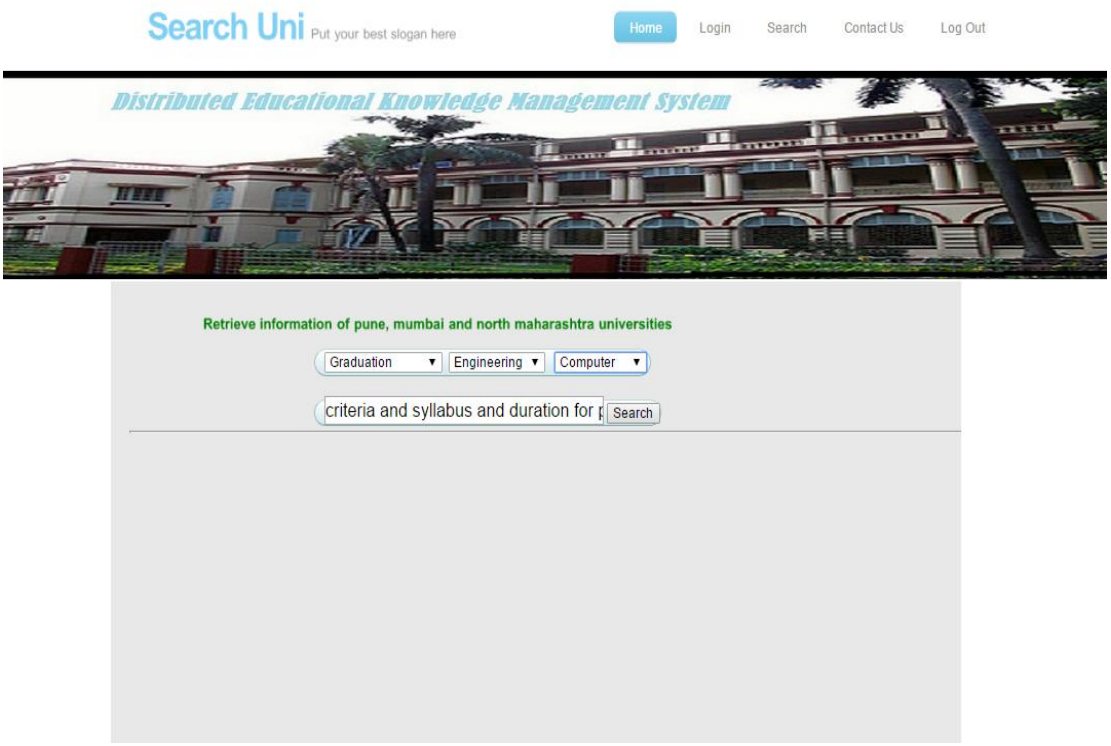
    - uni_probability[i]=p;
    - 20.4 Increase i by 1
    - 20.5 Go to step 20.2
21) Ranking of universities
22) Initialize n=size pf uni_probability and k = i,temp,university_rank[];
    - 21.1 Repeat the steps until i<n;
      - 21.1.1 Initialize j=i+1;
      - 21.1.2 Repeat the steps until j<n
      - 21.1.3 If uni_probability[j]< uni_probability[k] then
      - 21.1.4 K=j;
      - 21.1.5 Invariant: a[k] smallest of a[i..n]
      - 21.1.6 university_rank =swap a[i,k]
      - 21.1.7 invariant: a[1..i] in final position
      - 21.1.8 Go to step 21.1
23) Paging
    - 23.1 Initialize i=0; fix_path_display[]=null;
    - 23.2 Repeat the steps until i< university_rank .size();
    - 23.3 fix_path_display [i]=Retrieve path of data from ontology

23.4 increment i by 1
23.5 Go to step 23.2

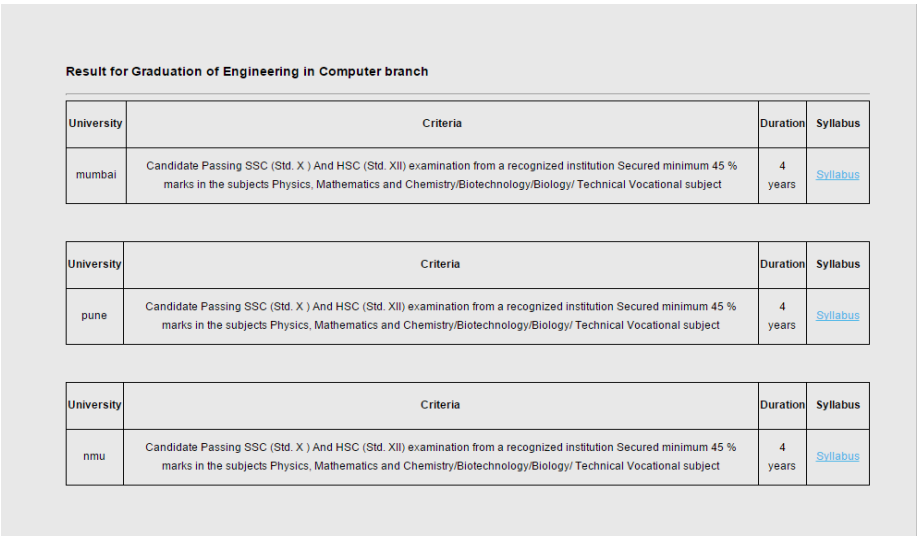
24)Display that data into jsp page as an output in th table format.
25)Stop

IV. USER INTERFACE AND RESULTS

User will select the required fields and enter the Query:



Output – Results of Input Query



Result for Graduation of Engineering in Computer branch

University	Criteria	Duration	Syllabus
mumbai	Candidate Passing SSC (Std. X) And HSC (Std. XII) examination from a recognized institution Secured minimum 45 % marks in the subjects Physics, Mathematics and Chemistry/Biotechnology/Biology/ Technical Vocational subject	4 years	Syllabus

University	Criteria	Duration	Syllabus
pune	Candidate Passing SSC (Std. X) And HSC (Std. XII) examination from a recognized institution Secured minimum 45 % marks in the subjects Physics, Mathematics and Chemistry/Biotechnology/Biology/ Technical Vocational subject	4 years	Syllabus

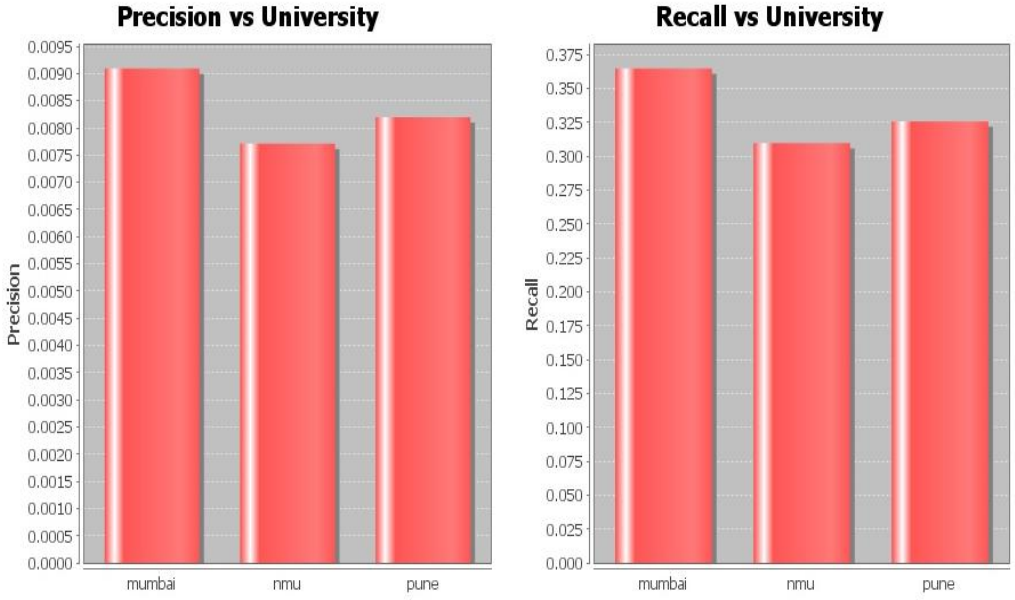
University	Criteria	Duration	Syllabus
nmu	Candidate Passing SSC (Std. X) And HSC (Std. XII) examination from a recognized institution Secured minimum 45 % marks in the subjects Physics, Mathematics and Chemistry/Biotechnology/Biology/ Technical Vocational subject	4 years	Syllabus

Result Analysis using Precision and Recall Metrics:

Precision: The actual retrieval set may not perfectly match the set of relevant records.
Recall: The ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually ex pressed as a percentage.

Precision, recall, and the F measure are set-based measures. They are computed using unordered sets of documents. We need to extend these measures (or to define new measures) if we are to evaluate the ranked retrieval results that are now standard with search engines. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each such set, precision and recall values can be plotted to give a precision-recall bars, such as the one shown in Figure 1.4. Precision-recall bars have a distinctive saw-tooth shape: if the $(k+1)^{th}$ document retrieved is non relevant then recall is the same as for the top k documents, but precision has dropped. If it is relevant, then both precision and recall increase, and the curve jags up and to the right. It is often useful to remove these jiggles and the standard way to do this is with an interpolated precision: the interpolated precision p_{interp} at a certain recall level r is defined as the highest precision found for any recall level $r' \ge r$:
\begin{displaymath}
p_{interp}(r) = \max_{r' \ge r} p(r')
\end{displaymath}

The justification is that almost anyone would be prepared to look at a few more documents if it would increase the percentage of the viewed set that were relevant (that is, if the precision of the larger set is higher).[10]



V. CONCLUSION

By constructing a framework of knowledge management system based on ontology, this paper expounds the function of each layer, and analyses the implementation of this system from the knowledge organization and expression and knowledge retrieval. Finally, it provides a case which implements the management system and realizes some parts of retrieval modules. This management system establishes a sharable ontology that can be understood both by human and computer, which people can found more relations of different concepts through a better circumstance of knowledge retrieval interface. In addition, the system is also open to some extent, so it can accumulate tacit knowledge constantly and polymerize explicit knowledge efficiently, which can lead to a better management and application of knowledge, to support the innovation for the designers.

- By constructing a NLP driven platform for efficient information retrieval on ontology based training data set I have proposed a system to achieve accurate and efficient data retrieval to provide comparative analysis of Educational Universities.
- Ontology schema is scalable to large extent which provides a great scope of improvement and usability as a standalone as well as a web hosted application.

- We have been working and we will continue our work towards the design of entirely ontology based structure and the development of our own reasoning methods to operate with it.

VI. REFERENCES

[1] A. Maalel, L. Mejri, H. H. Mabrouk and H. B. Ghezela, ―Toward a Knowledge Management Approach Based on an Ontology and Case-based Reasoning (CBR)‖, In Proc. Sixth International Conference on Research Challenges in Information Science (RCIS), IEEE, pp. 1-6, 2012.
[2] A. Uszok, L. Bunch, J. M. Bradshaw, T. Reichherzer, J. Hanna and A. Frantz, ― Knowledge-Based Approaches to Information Management in Coalition Environments‖, Intelligent Systems, IEEE, Vol. 28, Issue 1, pp. 34-41, 2013.
[3] Akbik and J. Bross, "Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns," in Workshop on Semantic Search in Conjunction with the 18th Int. World Wide Web Conference (WWW2009), Madrid, Spain, 2009.
[4] Caputo, P. Basile, and G. Semeraro, "Boosting a semantic search engine by named entities," in Proceedings of the 18th International Symposium on Foundations of Intelligent Systems. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 241–250.
[5] "An Overview of a Role of Natural Language Processing in An Intelligent Information Retrieval System"' Asanee Kawtrakul.
[6] Bollegala, Y. Matsuo, and M. Ishizuka, "Unsupervised extraction of semantic relations between entities on the web," in 19th Int'l World Wide Web Conf. (WWW 2010), Raleigh,North Carolina, USA, 2010, pp. 151–160.
[7] D. Fensel, 'Ontology-based knowledge management', Computer, vol. 35, no. 11, pp. 56-59, 2002.
[8] GUO Rong and WUJun, ―Design and Implementation of Domain Ontology-based Oilfield Non-metallic Pipe Information Retrieval System‖ In Proc. 2012 International Conference on Computer Science and Information Processing (CSIP), IEEE, pp. 813-816, 2012.
[9] H. Aust, M. Oerder, F. Seide, and V. Steinbiss, "The Philips automatic train timetable information system," Speech Commun., vol. 17, no. 3-4, pp. 249–262, November 1995.
[10] H. Li, Y. Cao, J. Xu, Y. Hu, S. Li, and D. Meyerzon, "A new approach to intranet search based on information extraction," in Proceedings of CIKM'05, Bremen, Germany, 2005, pp. 460–468.
[11] J.Weizenbaum, Eliza "a computer program for the study of natural language communication between man and machine", Commun. ACM, vol. 9, no. 1, pp. 3645, January 1966.
[12] J. Zhang, W. Zhao, G. Xie and H. Chen, 'Ontology- Based Knowledge Management System and Application', Procedia Engineering, vol. 15, pp. 1021-1029, 2011.
[13] Jing FAN, Xiuying LIU, Ying SHEN and Tianyang DONG,―Ontology-based Knowledge Management for Forest Channel‖, In Proc. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), IEEE, pp. 1523-1527, 2012.
[14] M. Gaeta, F. Orciuoli and P. Ritrovato, 'Advanced ontology management system for personalised e-Learning', Knowledge-Based Systems, vol. 22, no. 4, pp. 292-301, 2009.
[15] M. Zhou and J. Tao, ―A Framework for Ontology-Based Knowledge Management‖, In Proc. 2011 International Conference on Business Management and Electronic Information (BMEI), IEEE, pp. 428-431, 2011.
[16] P. Quaresma and I. P. Rodrigues, "Using dialogues to access semantic knowl- edge in a web ir system", in Proceedings of the 6th international conference on Computational processing of the Portuguese language, 2003.
[17] Perez and V. Benjamins, 'Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods', Estocolmo, 1999.
[18] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev, "Beyond basic faceted search," in Proceedings of the international conference on Web search and web data mining. New York, NY, USA: ACM, 2008, pp. 33–44.
[19] R. Mizoguchi, M. Ikeda and K. Seta, 'Ontology for modeling the world from problem solvingperspectives', Montreal, 1995.
[20] R. Neches, R. Fikes and T. Gruber, 'Enabling Technology for Knowledge Sharing', AI Magazine, vol. 12, no. 3, pp. 36-56, 1991.
[21] R. Sutcliffe, K. White, and U. Kruschwitz, "Named entity recognition in an intranet query log," in Proceedings of the LREC Workshop on Semitic Languages, Valletta, Malta, 2010, pp. 43–49.
[22] S. Seneff, L. Hirschman, and V. W. Zue," Interactive problem solving and dialogue in the atis domain", in Proceedings of the workshop on "Speech and Natural Lan- guage". Stroudsburg, PA, USA: Association for Computational Linguistics, 1991.
[23] Suma Adindla and Udo Kruschwitz "Combining the Best of Two Worlds: NLP and IR for Intranet Search", International Conferences on Web Intelligence and Intelli- gent Agent, University of Essex Wivenhoe Park, Colchester, C04 3SQ, UK,2011.

[24] Sun, P. Liu, and Y. Zheng, "Short query refinement with query derivation," in Proceedings of the 4th Asia information retrieval conference on Information retrieval technology. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 620–625.
[25] R. Studer, V. Benjamins and D. Fensel, 'Knowledge engineering: Principles and methods', Data & Knowledge Engineering, vol. 25, no. 1-2, pp. 161-197, 1998.
[26] T. Gruber, 'A translation approach to portable ontology specifications', Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.
[27] U. Kruschwitz and H. Al-Bakour, "Users want more sophisticated search assistants: Results of a task-based evaluation," Journal of the American Society for Information Science andTechnology, vol. 56, no. 13, pp. 1377–1393, November 2005.
[28] U. Kruschwitz, "Intelligent Document Retrieval: Exploiting Markup Structure, ser. The Information Retrieval SeriesSpringer", 2005, vol. 17.
[29] Xiaohui Tao, Yuefeng Li, and Ning Zhong,"A Personalized Ontology Model for Web Information Gathering" ,IEEE Transactions on knowldge and data Engineer- ing, Vol.23,No.4,April2011.

