An Implementation of Novel approach for Information Extraction using Ontology

Varsha Manwade[1], Mr. Sachin Yele[2]

[1]*M.Tech student, Dept. of CSE,* [2]*Asst.professor, Dept. of CSE*

Sanghavi Institute of Management Science, Indore, R.G.P.V. University, Bhopal (M.P.)

Abstract — ***Need of information from the web is a rich domain of study and research. The information flood on the web is most of the time is available in unstructured format. Therefore search processes are becomes more complicated for finding the appropriate documents from the web. A number of research efforts are found when exploring the document search techniques but most of the technique time consuming, less accurate and high resource consuming. Therefore the information processing and information retrieval from the unstructured data source becomes more complicated. On the other hand the unstructured query processing and finding the user query relevance data from such data sources are also a complex issue. In this presented work the unstructured data handling and retrieving the user query relevance data is the key domain of research work. Therefore a new technique for retrieving data in structured query format is presented in this work. In addition of that for demonstrating the issues of data extraction and managing them is also an area of the proposed study. After addressing the primary issues a new solution using the ontology based information parsing, fuzzy keyword based search processes is proposed. That technique is implemented on the real world application of resume search. The given case study is limited on two different users where primary user uploads their resume for processing and information processing. Additionally the secondary user utilizes the extracted features and perform constrain based search on the extracted knowledge. The implementation of the proposed technique is performed using JSP application and their comparative performance study is performed with the TF-IDF based information retrieval technique. For comparative study the precision, recall and f-measures are taken as primary performance factors. According to the experimental observations the proposed technique found more efficient and able to produce more accurate results as compared to TF-IDF based technique.***

Keywords— ***Information Extraction, Information Retrieval, Ontology, Resume.***

I. INTRODUCTION

The invention of the computer promotes the use of digital data to store and retrieve the information from the computer storages. Additionally the uses of computers are also rapidly increases in different aspects. Therefore a significant amount of data is generated every day. Among them most of the data generated in the computers are found either structured manner or in unstructured manner. The structured data is always found in the form of relational manner or predefined manner where the storage and retrieval made easy. On the other hand the unstructured format of data is most of the time found in text files where for storing the similar kind of information in different manners can be used.

As the manners of information storage are growing the techniques of retrieval are also tries to improve self. In study a number of information retrieval [1] and extraction [2] systems are studied. In these systems most of the work and efforts are found for finding the structured data but too fewer techniques are developed for extracting the knowledge from unstructured format of data. Beside the data extraction is different from the information retrieval, because the information extraction only interested to find the targeted kind of information from the raw set of data. Therefore the hurdle is to first scan all the data and then refine them to obtain the target data.

In this presented work the key focuses is to study about the different information extraction methods additionally need to develop a method by which the unstructured manner of data becomes transformable to the structured format of data to be make effective and efficient applications. Therefore a resume search technique is demonstrated for finding the target information from unstructured files; extract the hidden attributes from them; and store them into a structured manner to make efficient information retrieval application.

The need of data and the accurate required information from the unstructured data is a complicated task because the data may be placed in any format and any place in entire documents. Therefore a technique required to scan the entire document, target the specified contents, extract them, transforms the data and consumes with the search techniques. Thus the accurate data retrieval algorithms for the unstructured data need long running processes to find the actual contents and retrieve them whenever required. Therefore accuracy and efficiency is the primary goal of any information retrieval technique. The data bases designed for any organization can include the information in structured and also in unstructured manner. During the experiments and review that is found that the structured information can easily retrieved as compared to unstructured formats of the data. In this presented work the key focus is placed on the information retrieval and

extraction techniques. Therefore a resume information retrieval system is developed. The key aim of the application development is to demonstrate the information retrieval technique in real world environment. This application includes the implementation of the client side user interface for upload the resumes and the administrator panel by which the resumes are searched on the pre-specified criteria. Additionally the algorithms and other techniques by which the information is transformed from unstructured data during the data upload or training of the system and the search process is implemented during the administration query processes.

In order to complete the implementation and design of the required information extraction technique and the information retrieval technique, the fuzzy based keyword search processes are included with the system. The given section reports the basic overview of the proposed study and the next section involves the entire system development and the data processing techniques description.

II. PROPOSED SYSTEM

The proposed system can be defined using the given figure 1. This diagram contains all the components and their sub-components that are organized to develop the complete required system.

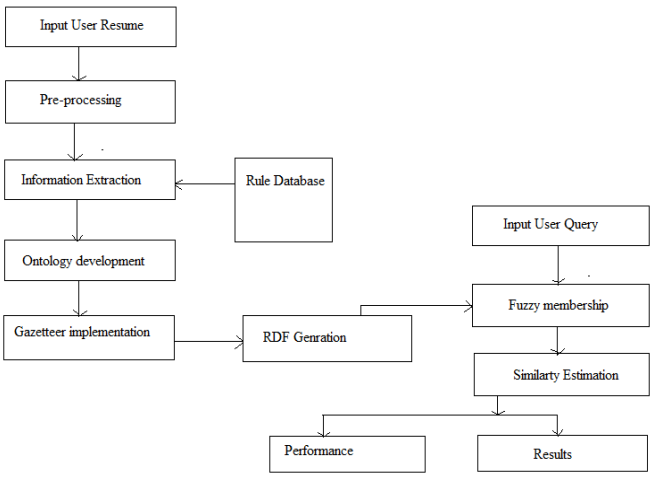

Figure 1 proposed system

Input user resume: in this given phase the end user need to upload resume into the server in web application. The resumes are always found in unstructured format and user put the information in resumes according to their own. Therefore the uploaded information by the end user needs to transform and filter according to the system needs. The resume uploading process needs to develop a simple user interface by which user can select the resumes from the local disk and upload to the server for further processing.

Pre-processing: In this phase when the data is uploaded on the server the background process is initiated. During this the unwanted stop words are removed from the entire document and the different special unused characters are also removed from the data.

Rule database: That is the phase of the information extraction is performed therefore some pre-defined rules are prepared and stored on the data base. These rules are used to targeting the specified information which are hidden in input user data and need to extract them for further processing of data.

Information extraction: In this phase the stored rules or the predefined rules are applied on the data. Additionally the required information is extracted from the input documents. These rules are applied on the data by different text processing techniques exist on the JAVA development technology.

Ontology development: After targeting the required information that is validated the input data is resume or not. If the input data is a resume and contains the information that is required in the resume search methodology the ontology is prepared using the input data that provides the hierarchical attributes that are used in information extraction.

Gazetteer implementation: In order to obtain the named identity form the given domain ontology the Gazetteer technique is implemented to find the accurate contents from the raw data.

RDF generation: The previous phase of data analysis produces the RDF (Resource Description Framework). That format of data is used to convert the unstructured information into the structured format to make query on the data. The generated RDF file is preserved into the data repository for the further use with the information retrieval system.

Input user query: That is the second phase of information retrieval at the administrator end, in this phase for finding the information user can place the query and their keywords as constrains.

Fuzzy membership: In this phase the user query inputs and the previously stored information in the format of the RDF is used. Basically the RDF document attributes are processed for the finding the named elements of user query. Therefore the RDF documents are processed using the fuzzy probability functions and the membership values of the named instances are estimated. That is further used to identify the required data from the RDF.

Similarity estimation: In this phase the user query similarity is measured from the evaluated named elements of ontology therefore Sparql query language is used to find the best matched data from the available set of data.

Results: In best matched results from the above given process is obtained as results of the information retrieval system.

Performance: After fetching out the results from the information retrieval system the performance of the entire system is computed and listed for further results analysis. .

III. PROPOSED ALGORITHM

The proposed algorithm is functioning in two major phases first the data processing and ontology development. And in the next phase the information retrieval form the available data bases. This section provides the understanding of both modules of the retrieval processes.

Information extraction
Input: user input document D_i
Output: RDF data RD_i
Processes:
1. D = read_input_file(D_i)
2. $PD_i = preprocess\_Data(D)$
3. $For(i = 1; i \leq numRule;\ i++)$
4. $RI_i = ExtractData(rule_i, PD)$
5. end for
6. RDF = Gazetteer_List(RI_i)
7. Return RDF

Table 1 information extraction

The above given table 1 shows the process involved in data extraction and the ontology development from the given input document and the given table 2 shows how the user keywords are used to retrieve the required information form the unstructured data.

Information retrieval
Input: RDF Database RDF, User Query Q
Output: information info
Process:
1. $F_{rule} = FuzzyRules(RDF)$
2. $for(i = 1; i \leq F_{rule}.lenght; i++)$
3. Sim= computeSimilar(F_{rule}, Q)
4. $if(Sim \geq .75)$
5. info= F^i_{rule}
6. end if
7. return info

Table 2 information retrieval

The above given process results the required data from the unstructured to the structured format during these processes the system performance is also investigated.

IV. RESULTS ANALYSIS

This chapter provides the detailed discussion about the proposed system evaluation for the accurate data retrieval. Therefore the different performance parameters are evaluated and compared with the traditional approach of data retrieval.

A. Precision

Precision of the information retrieval technique can be defined by the part of data retrieved that are relevant to the search query. That can be evaluated by the following formula:

$$precision = \frac{relevant\ document\ \cap retrieved\ document}{retrieved\ document}$$

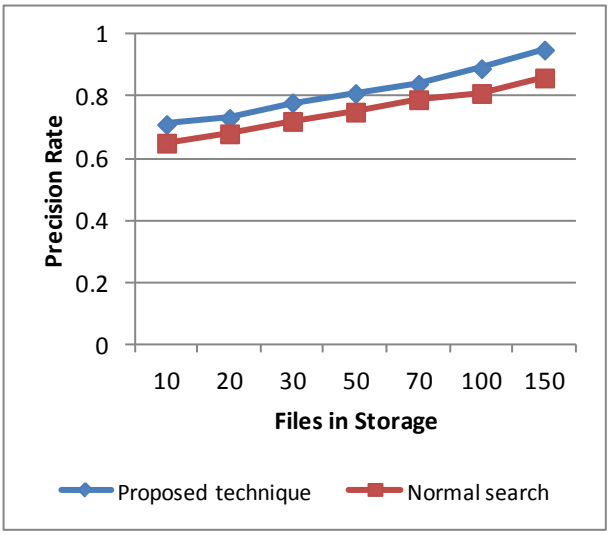

Figure 2 precision rate

The precision rate of the proposed and the traditional search system is given using the figure 2 and the table 3. In this diagram the X axis contains the amount of documents stored in the storage space and the Y axis demonstrate the corresponding precision value. The performance of the proposed technique is given here using the blue line and the performance of the traditional technique is given by the red line. According to the obtained performance the precision of the proposed technique is much enhancing as compared to the traditional method. Additionally the proposed algorithm demonstrates the high precision rate as compared to the traditional simple search methods, therefore the proposed technique efficient and accurate than normal search method.

Dataset size	Proposed technique	Normal search
10	0.71	0.65
20	0.73	0.68
30	0.78	0.72
50	0.81	0.75
70	0.84	0.79
100	0.89	0.81
150	0.95	0.86

Table 3 precision

B. Recall

Recall is the amount of data that are extracted during the search is relevant to the user query. That can be estimated using the following formula:

$$recall = \frac{Relevant\ Documents \cap Retrieved\ Document}{Relevant\ Documents}$$

The given figure 3 graphically represents the estimated recall rate of implemented information retrieval systems. In this diagram the blue line represents the performance of the proposed technique and the red line shows the performance of traditional algorithm. By using the different experiments the table as given 4 is developed and their points are aggregated with the figure. The X axis of the given diagram contains the file stored in the server for evaluation and the Y axis shows the amount of results correctly obtained. According to the comparative performance the proposed system provides the accurate outcomes as compared to the traditional technique.

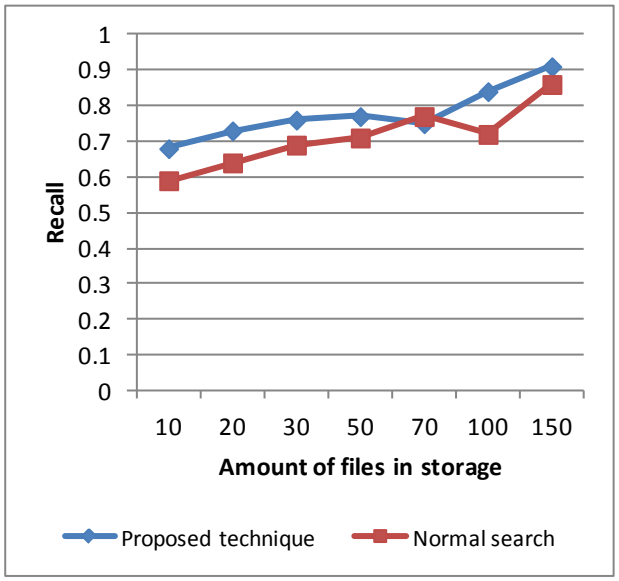

Figure 3 recall rate

Dataset size	Proposed technique	Normal search
10	0.68	0.59
20	0.73	0.64
30	0.76	0.69
50	0.77	0.71
70	0.75	0.77
100	0.84	0.72
150	0.91	0.86

Table 4 recall rate

C. F-measure

That measure and combines precision and recall in terms of harmonic mean of precision and recall rate of the obtained results, that can also be termed F-measure or balanced F-score:

$$F - measure = 2.\frac{precision * recall}{precision + recall}$$

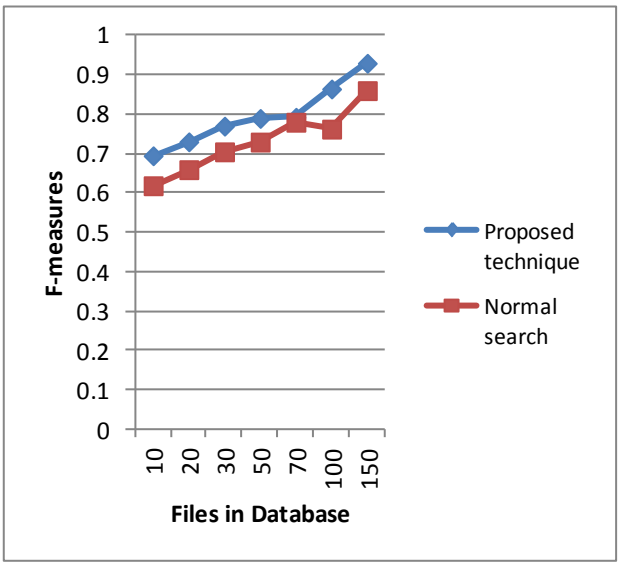

Figure 4 f-measures

The obtained f-measures of the proposed and traditional approach of information extraction are given using table 5 and the figure 4. In the given diagram the X axis contains the size of data that are stored on server for processing the user request in the similar manner the Y axis contains the estimated f-measures of both the systems. The performance of the proposed system (given by blue line) is more stable than the traditional approach (denoted by red line). According to the obtained performance of the system the proposed fuzzy based information extraction and retrieval system is much accurate and efficient.

Dataset size	Proposed technique	Normal search
10	0.6946	0.6185
20	0.73	0.6593
30	0.7698	0.7046
50	0.7894	0.7294
70	0.7924	0.7798
100	0.8642	0.7623
150	0.9295	0.86

Table 5 f-measures

D. Memory consumption

The amount of main memory required to execute the algorithm is termed as the space complexity of the system. That is sometimes also called the memory consumption of the algorithms. The comparative memory consumption of both the algorithms namely traditional and fuzzy rule based information extraction technique is given using figure 5. In this diagram the X axis includes the amount of files that stored on server for information extraction, and the Y axis shows the amount of memory consumed during processing of the data. The given memory consumption is provided here in terms of kilobytes. According to the

evaluated results the performance of the traditional technique is less efficient as compared to the traditional technique.

Dataset size	Proposed technique	Normal search
10	26881	27719
20	27193	28917
30	27918	29887
50	29027	30857
70	30184	32853
100	31573	33857
150	32848	34718

Table 6 memory consumption

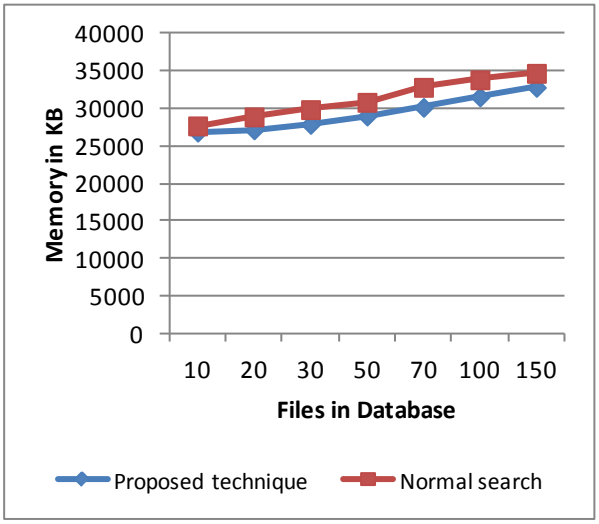

Figure 5 memory consumption

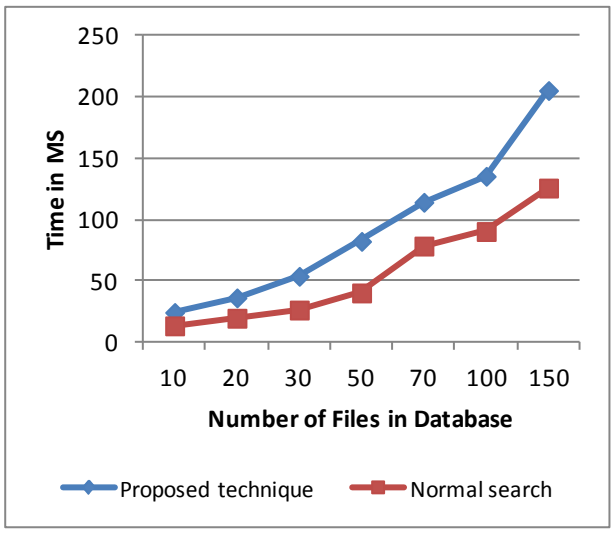

Figure 6 Time consumption

E. Time consumption

The amount of time required to retrieve the required data from the storage is known as the time consumption of the proposed information retrieval system. The figure 6 and table 7 contains the time consumption of normal information retrieval system and the proposed fuzzy based information retrieval system. According to the given results the proposed technique needs more time as compared to the traditional system. Therefore due to additional work load the proposed system is time consuming.

Dataset size	Proposed technique	Normal search
10	24.34	13.42
20	36.25	19.63
30	53.83	26.31
50	82.33	40.26
70	114.21	78.51
100	135.42	90.37
150	205.37	125.82

Table 7 Time consumption

V. CONCLUSIONS

The main aim of the proposed work is to study and demonstrate the real world information processing and information extraction from the unstructured data sources. Therefore a named ontology based scheme is implemented and their application in resume retrieval application is presented. The given chapter provides the summary of the entire performed work as the conclusion of work and some future extension of the work is also suggested to improve the given method further.

A. Conclusion

The information is need of current age human; therefore a number of information extraction and retrieval applications are developed recently that supports the current age information needs. During the implementation of different kinds of data search and retrieval techniques a number of methods for structured and similarly unstructured information processing is developed recently. Among most of the work is devoted for the structured data processing but their limited efforts are made for retrieving information from the unstructured data sources. The traditional unstructured data processing techniques either not much efficient, or not accurate for adopting and using in real world application therefore a new technique is required to investigate and develop by which the user query relevancy and performance both are improved.

In this presented work an ontology based information processing technique is developed and utilized in the resume search application. The given application demonstrates the real world data processing issues and need. Additionally that also addresses the issues of high complexity of data processing. The proposed technique involves two step procedures for information retrieval from unstructured data sources. In first phase the data is pre-processed, transformed for converting the information

into unstructured to structured format additionally the information extraction is taken place to identify the hidden named entity that are subject to information retrieval. In the next phase the user query processing is taken place and the data is searched over the pre-processed data using fuzzy concept. The second modules are responsible for generating the search outcomes and performance of the system.

The implementation of the proposed technique is performed using the JAVA technology and with the help of JSP environment. After the implementation of the system performance of the proposed information processing technique is measured and compared with the traditional content based text information search (tf-idf) based technique. The comparative performance is summarized using the given table 8.

S. No.	Parameters	Proposed technique	Normal search
1	Precision	High	Low
2	Recall	High	Low
3	f-measures	High	Low
4	Memory consumption	Low	High
5	Time consumption	High	Low

Table 8 performance summary

According to the obtained performance the proposed technique is adoptable and efficient additionally able to produce the high user query relevant outcomes. Therefore as compared to Tf-idf based information search the proposed technique able to extract and retrieve more accurate outcomes.

B. Future work

The proposed work is dedicated to find an efficient and accurate approach for extracting the knowledge from the raw database. Therefore a named instance based schema matching technique using fuzzy theory is developed and designed. The proposed technique is found accurate and efficient as compared to the traditional methodology. But the current approach has some limitations therefore future improvements on proposed technique are required.

1. The presented system need to define a set of rules by which the required information is targeted for data extraction need to find some way by which this limitation is removable.
2. The proposed system works on the unstructured data but when the formats are changed and information names are changed then the technique provides less effective results.
3. The proposed technique consumes more time for pre-process the data thus in near future the pre-processing time is required to reduce.

VI. ACKNOWLEDGMENTS

I would like to thank to my guide Prof. Sachin Yele , H.O.D., Computer Science & Engineering Dept. and dissertation Coordinator Mr. Kapil Sahu for providing me a support and proper guidance for the completion of this work. I am also thankful to all faculty members for help and support, whenever I faced problem during my work. I render a deep sense of gratitude towards my friends and well-wishers who gave me the support and help as and when I needed during the making of this project.

VII. REFERENCES

[1] Dr. Bode Prasad, D. Raveendra, P. PrudhviKiran, “Efficient Information Retrieval System using Incremental Approach”, IJCSMC, Vol. 4, Issue. 11, November 2015, pg.295 – 300

[2] L.Tari, PhanHuyTu, JorgHakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, and ChittaBaral, “Incremental Information Extraction Using Relational Databases”, IEEE Transactions on Knowledge and DataEngg, Vol. 24, No. 1, Jan 2012

[3] Daya C. Wimalasuriya, Dejing Dou, “Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches”, Journal of Information Science, XX (X) 2009, pp. 1–20 DOI: 10.1177/0165551506

[4] Sunil Kumar Kopparapu, “Automatic Extraction of Usable Information from Unstructured Resumes to Aid Search”, 978-1-4244-6789-1110/$26.00 ©2010 IEEE

[5] D Çelik, A Karakaş, G Bal, CGültunca, A Elçi, B Buluz, M Can Alevli, “Towards an Information Extraction System based on Ontology to Match Résumés and Jobs”, 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops

[6] Hongsheng Wang, Lu Yuan, Hong Shao, “Text Information Extraction Based on OWL Ontologies”, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 978-0-7695-3305-6/08 $25.00 © 2008 IEEE

[7] Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, FraunhoferAiS, “A Brief Survey of Text Mining”, Knowledge Discovery Group Sankt Augustin, May 13, 2005s

[8] Hien Nguyen, Eugene Santos, and Jacob Russell, “Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization”, IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6, November 2011

[9] MilošRadovanović, MirjanaIvanović, “Text Mining: Approaches And Applications”, Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234

[10] Vishal Gupta, Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009

[11] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, “Knowledge Extraction Using Rule Based Decision Tree Approach”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[12] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, “Exploiting Social Relations for Sentiment Analysis in Microblogging”, WSDM ’13, February 4–8, 2013, Rome, Italy, Copyright 2013 ACM 978-1-4503-1869-3/13/02

[13] RoshaniChoudhary, JagdishRaikwal, “An Ensemble Approach to Enhance Performance of Webpage Classification”, International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5614-5619

[14] Yudong Zhang, Shuihua Wang, Preetha Phillips, GenlinJi, “Binary PSO with mutation operator for feature selection using decision tree applied to spam detection”, Knowledge-Based Systems, 2014 Elsevier B.V. All rights reserved.

[15] GrigoriosTzortzis, AristidisLikas, “The MinMax k-Means clustering algorithm”, Pattern Recognition, 2014 Elsevier Ltd. All rights reserved.

[16] Bhushan M. Kulkarni, Prof. S. A. Kinariwala, “Review on Fuzzy Approach to Sentence Level Text Clustering”, International Journal Of Scientific Research And Education, Volume 3, Issue 6, Pages-3845-3850, June-2015, ISSN (e): 2321-7545

[17] R.S. Zhou, Z.J. Wang, “A Review of a Text Classification Technique: KNearest Neighbor”, International Conference on Computer Information Systems and Industrial Applications (CISIA 2015), 2015. The authors - Published by Atlantis Press

[18] R. Sagayam, S. Srinivasan, S. Roshni, “A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques”,International Journal of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5

[19] F. S. Gharehchopogh, Z. A. Khalifelu, “Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing”, 978-1-61284-832-7/11/$26.00 ©2011 IEEE

[20] Z Wang, A D. Shah, A. R Tate, S Denaxas, J S-Taylor, H Hemingway, “Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning”, PLoS ONE, 1 January 2012, Volume 7, Issue 1

[21] Jiawei Han, Chi Wang, “Mining Latent Entity Structures from Massive Unstructured and Interconnected Data”, Request permissions from permissions@acm.orgSIGMOD’14, June 22–27, 2014, Snowbird, UT, USA Copyright 2014 ACM

[22] Gubanov Michael, Michael Stonebraker, and Daniel Bruckner, “Text and Structured Data Fusion in Data Tamer at Scale” 2014 IEEE 30th International Conference on Data Engineering, ICDE (March 31-April 4, 2014), Chicago, IL. IEEE p 1258-1261.

[23] Xiaojing Yao, Xiang Li, Ling Peng, Tianhe Chi, “A Novel Fuzzy Chinese Address Matching Engine Based on Full-text Search Technology”, Copyright owned by the author(s) under the terms of the Creative Commons Attribution-Non-Commercial-No-Derivatives 4.0 International License (CC BY-NC-ND 4.0).

[24] Farman Ali, EunKyoung Kim, Yong-Gi Kim, “Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system”, Published online: 14 November 2014 © Springer Science+Business Media New York 2014

[25] Balasubramaniam K, “Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), Procedia Computer Science 50 (2015) 135 – 142, Elsevier

[26] Jun Hu, Xinzhou Lu and Chun Guan, “A Semantic Information Retrieval Approach Based on Rough Ontology”, The Open Cybernetics &Systemics Journal, 2014, 8, 399-404

[27] G. Nagarajana, R. I. Minub, “Fuzzy Ontology Based Multi-Modal Semantic Information Retrieval”, International Conference on Intelligent Computing, Communication & Convergence, 2015 The Authors Published by Elsevier B.V.