Extraction of Entities from Web with Knowledge Mining

Utkarsha Daradmare [1], Sagar Bhakre [2]

[1]P.G. Student, [2]Associate Professor
[1] Department of Computer Science and Engineering,
[1] Ballarpur Institute of Technology, Ballarpur, India

***Abstract*— a lot of information is rapidly growing on the web, extracting this valuable information of real-world entity is most tedious task. Search engine plays a vital role in collecting and understanding this valuable information. In order to reach more accurate resulted information, there is need to develop and utilized unique characteristics of a web. This paper introduces the concept of search engine for entity. It also describes about architecture, iknoweb framework and entity linking task adopted for it. It presents the summary of information about real-world entity.**

***Index Terms*— Entity, Entity disambiguation, Entity extraction, Crawler, Entity linking, Knowledge mining.**

I. INTRODUCTION

The main purpose of developing SEE (Search Engine for Entity) is to present summary of relevant information about the searched entity i.e. person, location, organization etc. instead of navigating through number of web pages. By applying question-answering system in iknoweb framework solves the problem of name disambiguation. Entity linking task is carried out in knowledge base to link entity with its corresponding entity. Here entities with similar names will linked with resulted entities and displayed to user. And the Generation of Entity-Relationship Graph which will shows the entities related to (person, organization, location etc.) are interconnected according to relationships with each other. This proposed methods and techniques tends to view the search results more accurately. Search Engine for Entity targets to extract all the related web information about the same entity together and integrate it as an information unit.

II. ARCHITECTURE

An architecture of entity search engine is shown in figure 1.

A Crawler

Crawler is the program which will visits the web pages and fetches the data according to targeted entities. Most of the search engines have such type of program called spider or boat.

Here is the process that a web crawler follows [3]:

- Using the available training data, machine learning model will automatically extracts the information about the entity.
- Extract all the links on that page.
- Follow each of those links to find new pages.
- Extract all the links from all of the new pages found.
- Follow each of those links to find new pages.
- Extract all the links from all of the new pages found.

Classification

The crawled data is classified into different entity types, such as papers, authors, products, and locations and for each type, a specific entity extractor is built to extract structured entity information from the web data ([1], [2]).

Aggregation

The data about the same entity will be aggregated.

Entity Linking and Disambiguation

Once the entity information is extracted and integrated, it is put into the web entity store, and search engines for entity can be constructed based on the structured information in the entity store ([1], [2], [4]).

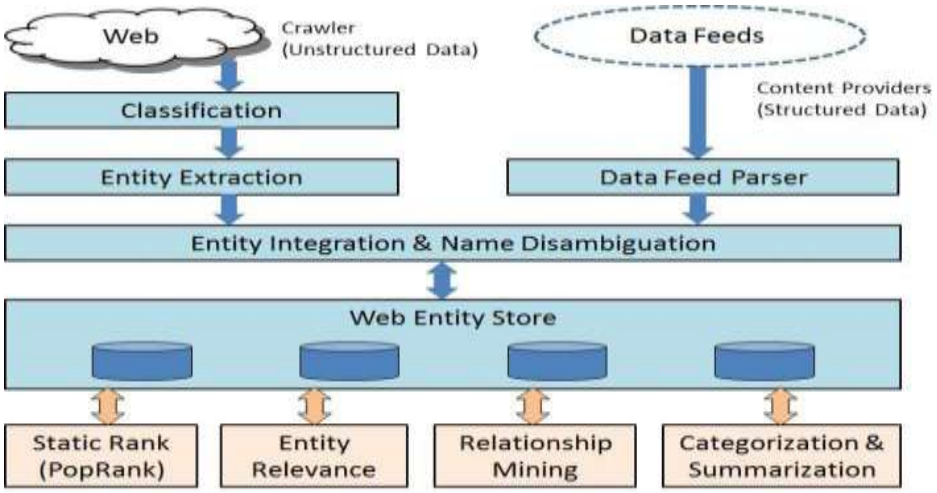


Figure 1: Architecture of Entity Search Engines [1]

The characteristics of entity search engine are as follows ([1], [2]):

Here is the process that a web crawler follows [3]:

- Entity search engines can return a ranked list of entities most relevant for a user query ([1], [9]).
- Entity search engines enable users to explore highly relevant information during searches to discover interesting relationships/facts about the entities associated with their queries.
- Entity search engines detect the popularity of an entity and enable users to browse entities in different categories ranked by their prominence during a given time period.

Entity Linking

Once the entity information is extracted and integrated, it is put into the web entity store, and search engines for entity can be constructed based on the structured information in the entity store ([1], [2], [4]).

An entity may have many different names or different entities may have similar names. Our focus is on to link targeted entity with its corresponding entity in knowledge base. This task is challenging due to variations in names, absence and entity ambiguity [3].

Key issues:

There are 3 challenges to entity linking ([3], [8], [9], [12]):

- **Name variations:** An entity often has multiple mention forms, including abbreviations (International Standard Organisation vs. ISO). Entity linking must find entity despite changes in mention string.
- **Absence:** Processing large text collections virtually guarantees that many entities will not appear in the KB (NIL), even for large KBs.
- **Entity ambiguity:** A single mention, like Springfield, can match multiple KB entries, as many entity names, like people and organizations, tend to be polysemous.

Entity linking performs several different task such as information retrieval, extraction, knowledge base population, question-answering [3] which are explained as follows:

Information Retrieval

Named entities are ambiguous and appear in search queries. This can be explained with the example of "Smith" entity. In the search query "smith" could mean many different entities, such as the name of person, organisation and many novels whose names are "Smith". Linking these ambiguous entity with a knowledge base will improve quality of search results [9].

Information Extraction

Named entities and relations extracted by information extraction systems are usually ambiguous. Linking them with a knowledge base is a good way to disambiguate and fine-grained typing them, which is essential for their further exploitation.

Question Answering

Question answering systems used to take maximum advantage of their supported knowledge bases to give answer to questions of user. Answering the question such as "Mr. Smith belongs to California?", then first task of system is to disambiguate the mentioned entity "Smith". Then the linking technique will compare the search query "Smith" to the California, and then it retrieves information from the knowledge base directly and presents the answer to the user's question.

Knowledge Base Population

Automatically extracting and populating this knowledge from knowledge base are the issue of techniques adopted for knowledge management. Entity linking is inherently considered as an important subtask for knowledge base population [6]. If the named entity entered for search have any relationship with any other entity in knowledge base, then entity linking task is carried out. The

mentioned entity is linked with its corresponding entity. Therefore, the knowledge base population task could potentially benefit from the entity linking problem [3].

Advanced entity search techniques and frameworks (such as iknoweb) are applied to make search more accurate.

An information about the single entity may distributed over more number of web pages. Entity extraction is the most tedious task because of the name disambiguation problem. Name disambiguation is also the major problem in entity integration. This is the challenging task to improve the search quality of search engine.

III. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

To overcome the problem of name disambiguation, we propose a novel entity disambiguation framework (called iKnoweb) to add people into the knowledge mining loop and to interactively solve the name disambiguation problem with users ([1], [2], [4]).

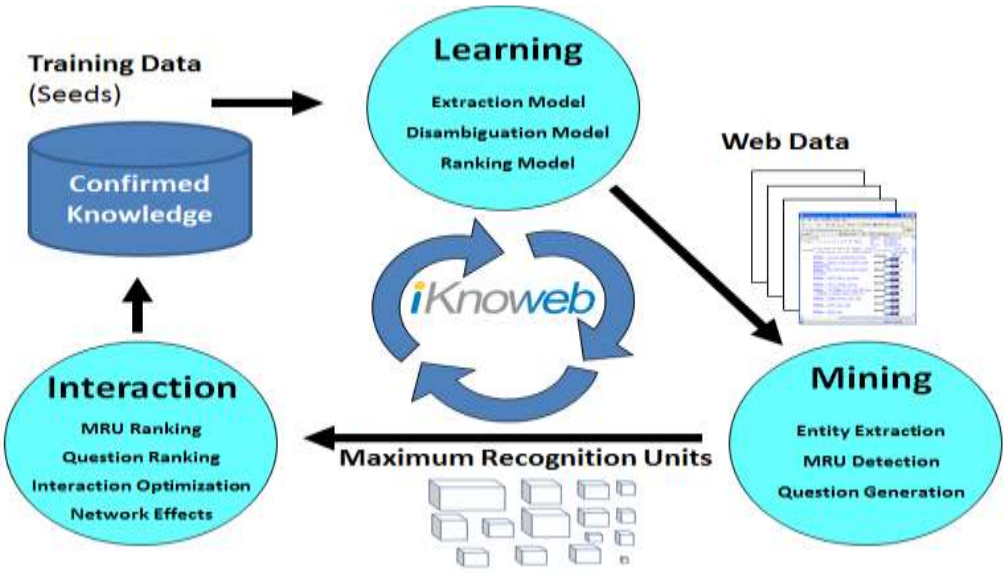


Figure 2: The iknoweb Framework [1]

Figure 2 shows the iKnoweb framework and it is explained as follows:

- Using the available training data, machine learning model will automatically extracts the information about the entity.
- The information obtained from extraction process is then merged into MRU's.
- When user enters the query for searching entity, he/she will go through the selecting some MRU or question/answering system to get result more accurately.
- The confirmed knowledge gained through question/answering system is stored into entity store.
- This confirmed knowledge can be used as a seeds for further improvement of entity extraction process.

.***Main Modules***

- Entity Extraction
- Detecting Maximum Recognition Units
- Question Generation
- Network Effects
- Interaction Optimization

Module description [1]:

- **Entity extraction:** Entity extraction is the task of extracting knowledge pieces of an entity and integrating all the pieces of the entity together [1].

Following figure 3 shows the Entity extraction:

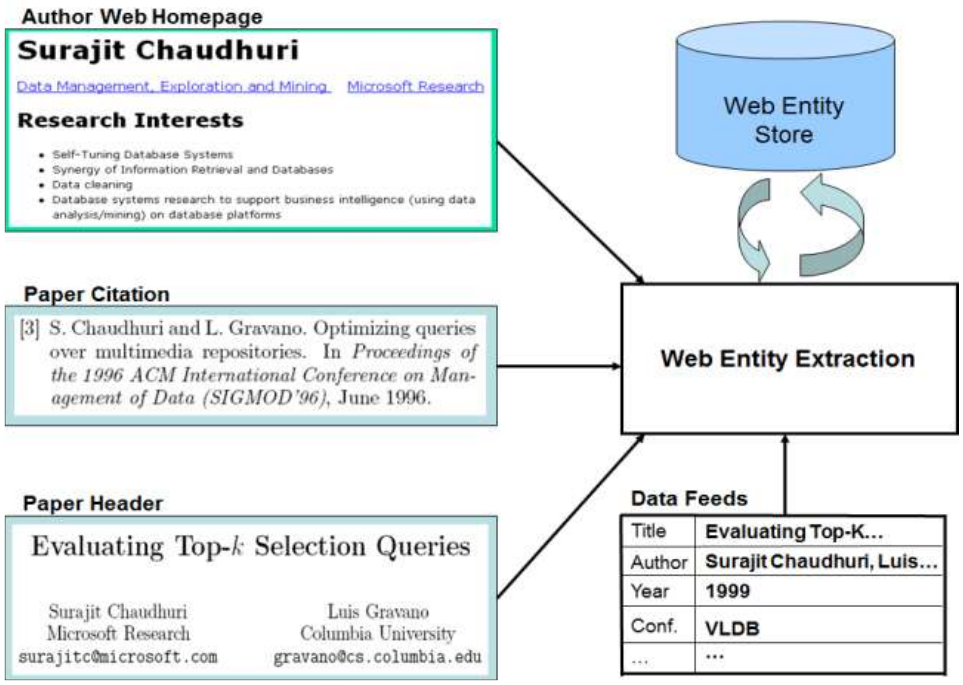


Figure 3: Entity Extraction examples [1]

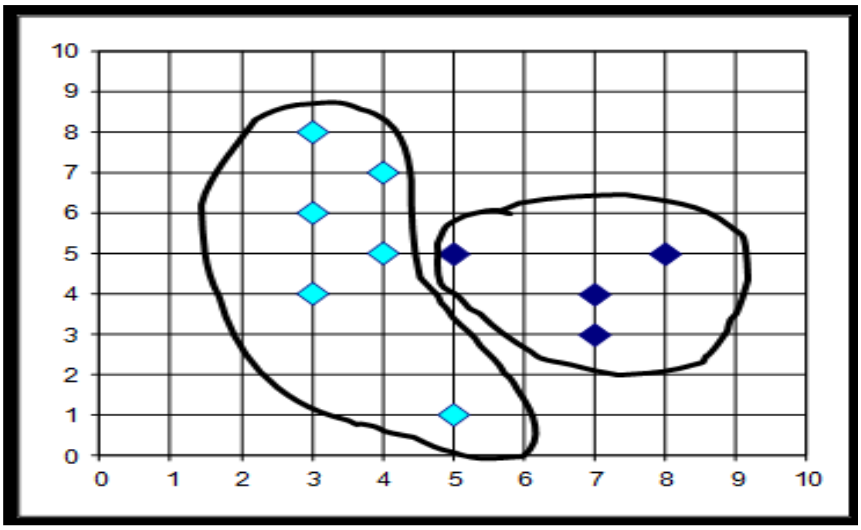
- **Maximum Recognition Unit:** We need to automatically detect highly accurate knowledge units, and the key here is to ensure that the precision is higher than or equal to that of human performance.
- **Question Generation:** By asking easy questions, iKnoweb can gain broad knowledge about the targeted entity. An example question could be: "Is the person a researcher? (Yes or No)", the answer can help the system find the topic of the web appearances of the entity.
- **MRU and Question Re-Ranking:** iKnoweb learns from user interactions, and the users will see more and more relevant MRUs and questions after several user interactions.
- **Network Effects:** A new User will directly benefit from the knowledge contributed by others, and our learning algorithm will be improved through users' participation.
- **Interaction Optimization:** This component is used to determine when to ask questions, and when to invite users to initiate the interaction and to provide more signals.

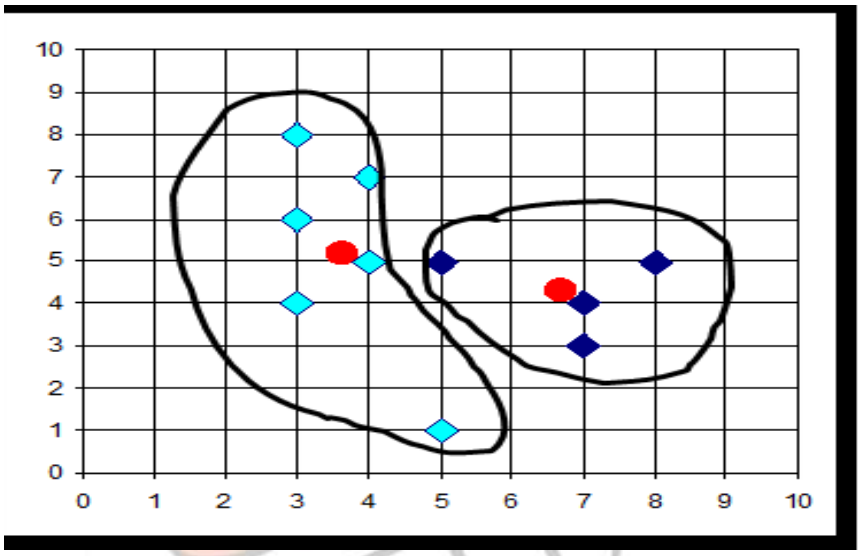
K-means clustering [7]

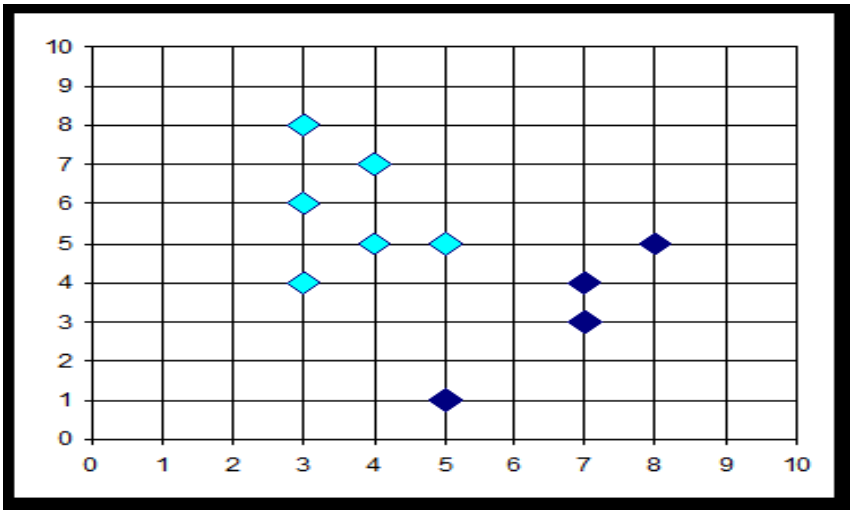
Here we have used K-Means clustering algorithm. The input to the algorithm is a data from multiple entries in table of database to be clustered with same property. The output from the clustering algorithm provides the average distance from cluster members to the center of each cluster. This computation is used to obtain the similarity between the entries in table. The similarity is obtained from the distance from the center of the cluster.

In our approach we have used K-Means Cluster based algorithm. K-Means is a widely use clustering algorithm. In that we use random seeds data and according to that arrange the clusters. Given k, the k-means algorithm is implemented in 4 steps:

- Partition objects into k nonempty subsets
- Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
- Assign each object to the cluster with the nearest seed point.
- Go back to Step 2, stop when no more new assignment.


10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10


10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10


10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10

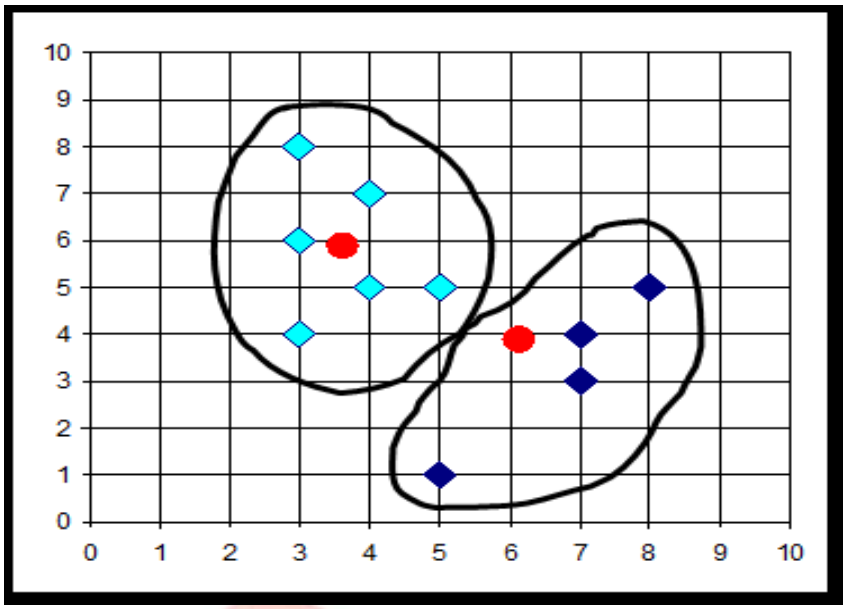


Figure 4: Clustering Process using K-Means Algorithm

IV. PROJECT SCREENSHOT

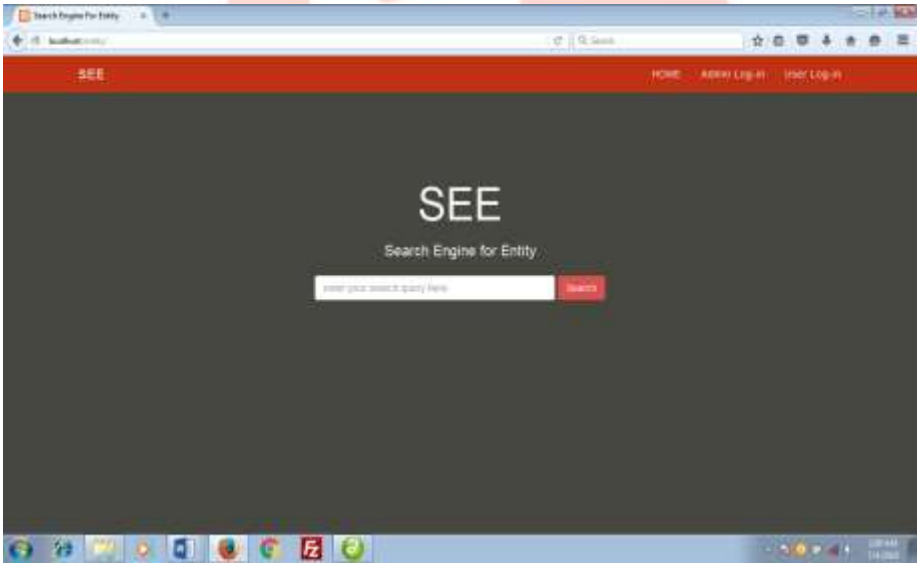


Figure 5: Screenshot of Home Page

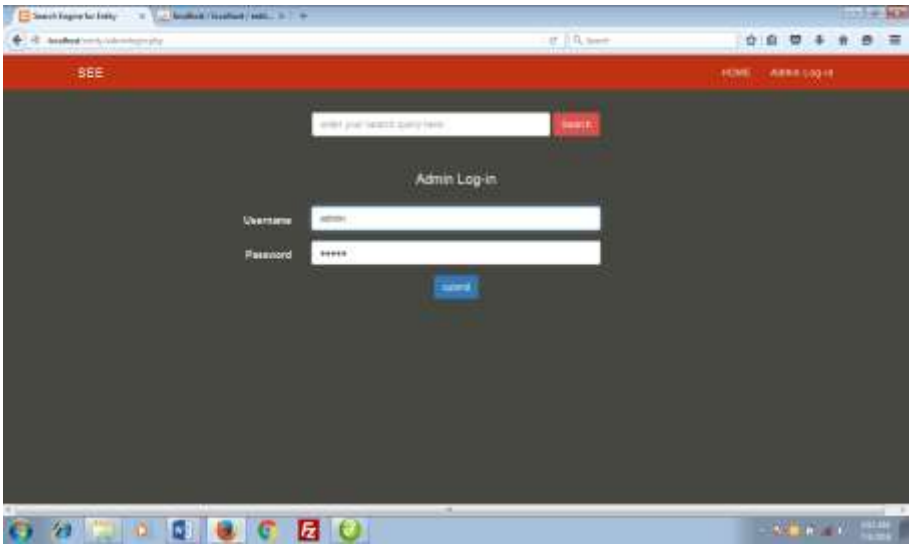


Figure 6: Screenshot of Admin login page

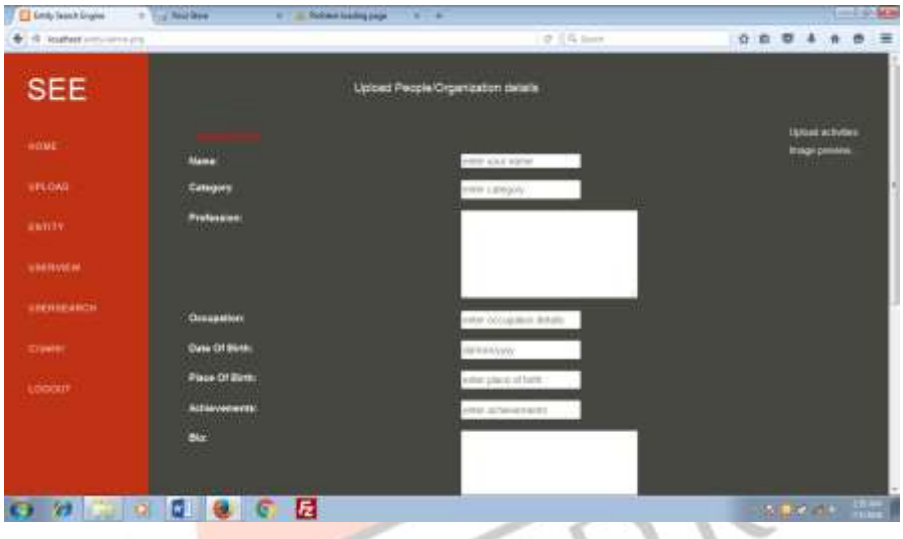


Figure 7: Screenshot of admin page

Figure 8: Screenshot of Result Page

V. RESULT

The implementation process used for the algorithm implementation and clustering the results. K-mean algorithm and iknoweb framework were implemented for clustering the given query and the results were compared. Based on the results the K-means Algorithm and iknoweb framework is efficiently forms the thematic groups of cluster in minimum response time compare to exiting method.

Various types of user queries are processed into the academic domain and categorized the final results in Table 1. The query processing time and the number of clustering groups are the two main factors to evaluate the efficiency of the two methods. Based on this result, the k-means method is more effective and efficient for data clustering techniques.

Table1. Comparison results in processing time and cluster count

User Query	Entries In cluster	Processing Time for Entries
Obama	25	1.025 seconds
smith	15	1.014 seconds
peter	20	1.022 seconds

VI. CONCLUSION

In search engine for entity which aims and targets to extract valuable and important information as an information unit and solves the problem of name disambiguation with iknoweb framework with more accuracy than traditional searches. By presenting summary of the information about the entity, it removes the necessity to navigate through all the web pages for getting complete view of entity.

Entity linking task is carried out to link the targeted entity in knowledge base which improves quality of search result.

VII. FUTURE SCOPE

The work can be extended to solve the ambiguity problem which can be solved by training the system with a large training corpus of various kinds of news, so that it contains a variety of combinations of names. The work can also be extended to solve unknown words problem which can be solved by using some lists that contain names, especially foreign names.

The work can also be extended to make NER more general. The work can also be extended in making more efficient transliteration rules. Since the Hindi POS tagger is not currently available to us, we can't do so much work in Hindi NER .The work can also be further extended in case of Hindi NER.

It is expected that more research or even a better understanding of the entity linking problem may lead to the emergence of more effective and efficient entity linking systems, as well as improvements in the areas of information extraction and Semantic Web.

VIII. ACKNOWLEDGEMENTS

The authors duly acknowledge the support provided by the Management and Principal of BIT Engineering College-Maharashtra by means of providing all the study related facilities.

REFERENCES

[1] Zaiqing Nie, Ji-Rong Wen and Wei-Ying Ma, "Statistical entity extraction from web," *IEEE*, vol. 100, No.9, Year 2012.
[2] Pinky Paul and Mr. Thomas George, "Entity Search Engine," IJCSMC, vol. 3, Issue. 2, Feb. 2014, pp. 877–880.
[3] Wei Shen, Jianyong Wang, and Jiawei Han, Fellow, "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions," IEEE transactions on knowledge and data engineering vol. 27, No. 2, Year 2015.
[4] Nilesh jain, Priyanka Mangal, "An approach to build a web crawler using clustering based k-means algorithm," journal of global research in computer science, vol. 4, No. 12, Dec. 2013.
[5] Shaik Muneeb Ahamed, Sd.Afzal Ahmad, and P.Babu, **"**Entity Extraction Using Statistical Methods Using Interactive Knowledge Mining Framework," IJCSN, ISSN: 2231 – 1882, Vol. 2, Issue. 1, Year 2013.
[6] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, "Open Information Extraction from theWeb," IJCAI-07,pp. 2670-2676.
[7] Daljit Kaur and Ashish Verma," Survey on Name Entity Recognition Used Machine Learning Algorithm," IJCSIT, Vol. 5 (4), Year 2014, pp. 5875-5879.

[8] Michal Laclavik, Stefan Dlugolinsky and Marek Ciglan.” Discovering Relations by Entity Search in Lightweight Semantic Text Graphs,” Computing and Informatics, Vol. 32, Year 2013, pp. 1001-1028, V 2014-Jul-24.
[9] Renaud Delbru,” Searching Web Data: an Entity Retrieval Model,” Digital Enterprise Research Institute, National University of Ireland, Galway, Sept. 2010.
[10] Manika Nanda,” The Named Entity Recognizer Framework,” IJIRAE, Vol. 1, Issue. 4, May 2014.
[11] Alexandros Komninos and Avi Arampatzis, **“**Entity Ranking as a Search Engine Front-End,” IJAIT, Vol. 6, No.1&2,Year2013.Available:http://www.iariajournals.org/internet_technology/
[12] Ashwini Zadgaonkar,” An Overview of Entity Relation Extraction Techniques,” IJARCSSE, Vol. 5, Issue. 11, Nov. 2015.