International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 05, May 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

ONTOLOGY LEARNING FROM UNSTRUCTURED INFORMATION SOURCES FOR USER QUERY

Mulla Nilofar[1], Pathan Ayesha[2]

Shahapurkar Namrata[3], Tayde Dipali[4], K.P.Moholkar[5]

[1,2,3,4,5]Students, Computer Engineering , Savitri Bai Phule Pune University (India)

ABSTRACT

With the fast growth of information volume through the World Wide Web causes an increasing requirement to develop new automatic system for retrieval of documents and ranking them according to their relevance to the user query. There are many search engines available. Most of the search engine results are usually presented in a list and are commonly called hits. There are ontology based search engines like Swoogle, Ontokhojand Semoogle. And these available search engines can't retrieve the proper results as per the user request. The proposed system will overcome this drawback as well as it retrievesdocuments in two languages, English and Marathi. The System retrieves the documents that falls into tourism. Results will be processed by web adviser which will implement ontology to remove unnecessary results and create a précised result list by expanding the query. Resulting webpageswill be ranked based the semantic similarity between ontological concepts extracted from web pages and ontological concepts represented by the user query.

Keywords: Ontology, Semantic Similarity, Web Page Ranking, Multilingual Information Retrieval

I.INTRODUCTION

One of the first Multi-Language Information Retrieval (MLIR) systems was implemented in 1969 by Gerard Salton who enhanced his SMART system to retrieve multilingual documents in two languages, English and German. Majority of information retrieval systems are monolingual and more precisely English-based. Our proposed system presents a Multi-Language Information Retrieval (MLIR) approach that falls into the area of Domain Specific Information.

There are many search engines available. The drawback of current conventional web search engines is the knowledge gap between users and computers. The knowledge and work of computer is much more limited than the knowledge of user. Our proposed System uses the ontology learning which extracts documents from Wikipedia. This methodology is used to semantically model the significant concepts of a query along with its weighted semantic relations to other related concepts. The resulting ontology can be viewed as a benchmark of a topic that can be used to classify or re-rank documents based on the degree of similarity to the original query.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

II. BACKGROUND AND RELATED WORKS

Current Web search engines' ranking approaches are classified into two categories: ontology-based and non-ontology based approaches. Non-ontological ranking approaches like Page Rank and HITS are the most common approaches. Both approaches depend on the analysis of link structure between nodes in order to rank the returned documents.The limitation of such approaches is that user still required to browse through long list of Web pages to select those that are actually considered to be of interest. However, users usually check only one or two pages of the results returned by the search engine. Another shortcoming is that the resultant list of documents from a search engine does not make distinctions between the different concepts that may be present in the query, as the list inevitably has to be ranked sequentially.Therefore, an approach is needed to minimize the number of hits that a user needs to revise in order to give the user faster access to information he needs.

Ontology-based search engines like Swoogle, and OntoKhoj indexes ontologies that capture concepts, their properties, and their relationships for specific domain which can be used by computers to process within the data of those domains semantically.Ontology learning means the acquisition of domain model from data it aims to accelerate the time and reduce the effort of knowledge markup or ontology population to construct the ontology; this can take place by gaining concepts and relations semi-automatically or automatically from different information sources such as databases, documents, and/or Web pages.In the proposed work, ontology learning concerned more specifically with knowledge acquisition from unstructured information sources (e.g., text files, HTML files). Obviously, much of the work in this area therefore related to the large body of work in this direction within Natural Language Processing (NLP), Artificial Intelligence (AI), and machine learning.The proposed work presents a methodology for resolving ambiguity of words.

III.PROPOSED SYSTEM

The proposed System (Figure. 1), presents a method for extracting ontology concepts from Web pages. It builds a semantic graph for any topic indicating the semantic relatedness between extracted concepts. Finally, it generates a list of ranked concepts based on their relatedness to initial search topic.In this system, a document is modeled as a graph of semantic relations between its concepts. A remarkable notice is that, concepts that are related to each other and to the main topic; tend to cluster up into densely interconnected communities or clusters, while non-related concepts become isolated.The proposed system is divided into four main stages namely Key phrase Extraction, Search and matching function, Semantic similarity computation and graph generation, Cluster ranking and ontology

The Google translate API will convert the language which will provides accurate results that is query conversion of Marathi language to English language and if input is English there is no problem.Google translate is a free multilingual statistical machine translation service provided by Google to translate text, speech, images, sites or real time video from one language into another. It offers a web interface, mobile apps for Android and iOS and an API that developers can use to build browser extensions, applications and other software. Google translate supports over 100 languages at various levels and servesover 200 million people daily.

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 05, May 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

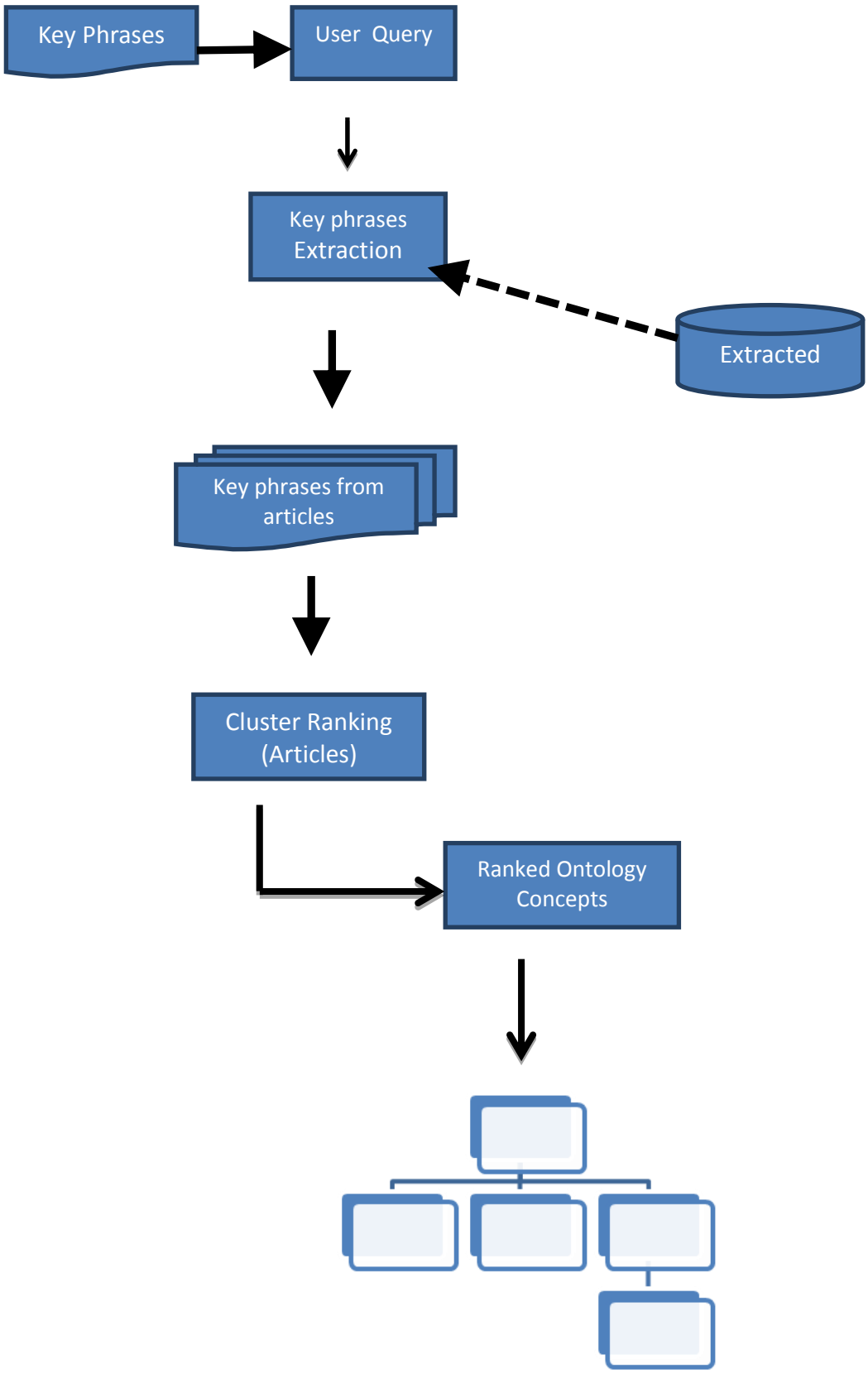


Fig : 1

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Ontology based search engine will provide accurate results depending on the literal meaning of the query from that user will relate keyword with Wikipedia documents and accurate results will be shown.The proposed system will search query results for the tourism domain. Ranking of query results is one of the fundamental problems in information retrieval (IR), the scientific/engineering discipline behind search engines. Given a query q and a collection D of documents that match the query, the problem is to rank, that is sort the documents in D according to some criterion so that the "best" results appear early in the result list displayed to the user. Classically, ranking criteria are phrased in terms of relevance of documents with respect to an information need expressed in the query. The proposed system will rank the results per the given query.

IV.EXPERIMENTAL ENVIRONMENT AND METHODOLOGY:

4.1 Keyphrase Extraction

The criterion of user understanding of an article depending on his awareness of article's important keywords or keyphrases, and the transition or relations between keywords inside it.

Following the same idea, the proposed model start by extracting important keyphrases from highly authoritative information sources that represent the query.

Following are the results for the example of semantic keyphrase extraction for topic "Pune".

Shaniwar Wada Palace

Shaniwar Peth, Pune, Maharashtra 411030, India point_of_interest

Mahadji Shinde Chhatri

Wanwadi Nagar Fatima NagarWanwadi Nagar, Fatima Nagar, Vikas Nagar, Wanwadi, Pune, Maharashtra 411001, India point_of_interest

Famous Spots, Pune

Sr. No. 3, Opp. I.E.S. School,, Yogeshwar Soc, Vadgaon Sheri,, Pune, Maharashtra 411014, India point_of_interest

Pu La Deshpande / Okayama Friendship Garden

Sinhagad Road, Dattawadi, Pune, Maharashtra 411030, India park

Rajiv Gandhi Zoological Park and Wildlife Research Centre

Pune - Satara Road, Near Bharati Vidyapeeth, Pune, Katraj, Maharashtra 411046, India zoo

ISKCON NVCC Temple

New Vedic Cultural Center, Katraj-Kondwa Bypass, Pune, Maharashtra 411048, India hindu_temple

Rajgad Fort

Balekilla Road, Pune, Maharashtra 412213, India point_of_interest

Katraj Lake

Katraj Vasahat, Katraj, Pune, Maharashtra 411046, India point_of_interest

Sinhagad Fort

Sinhagad Ghat Road, Thoptewadi, Pune, Maharashtra 411025, India point_of_interest

Gandhi National Memorial Society

Agakhan Palace, Pune Nagar Rd, Kalyani Nagar, Pune, Maharashtra 411006, India museum

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Dagdusheth Ganpati Temple

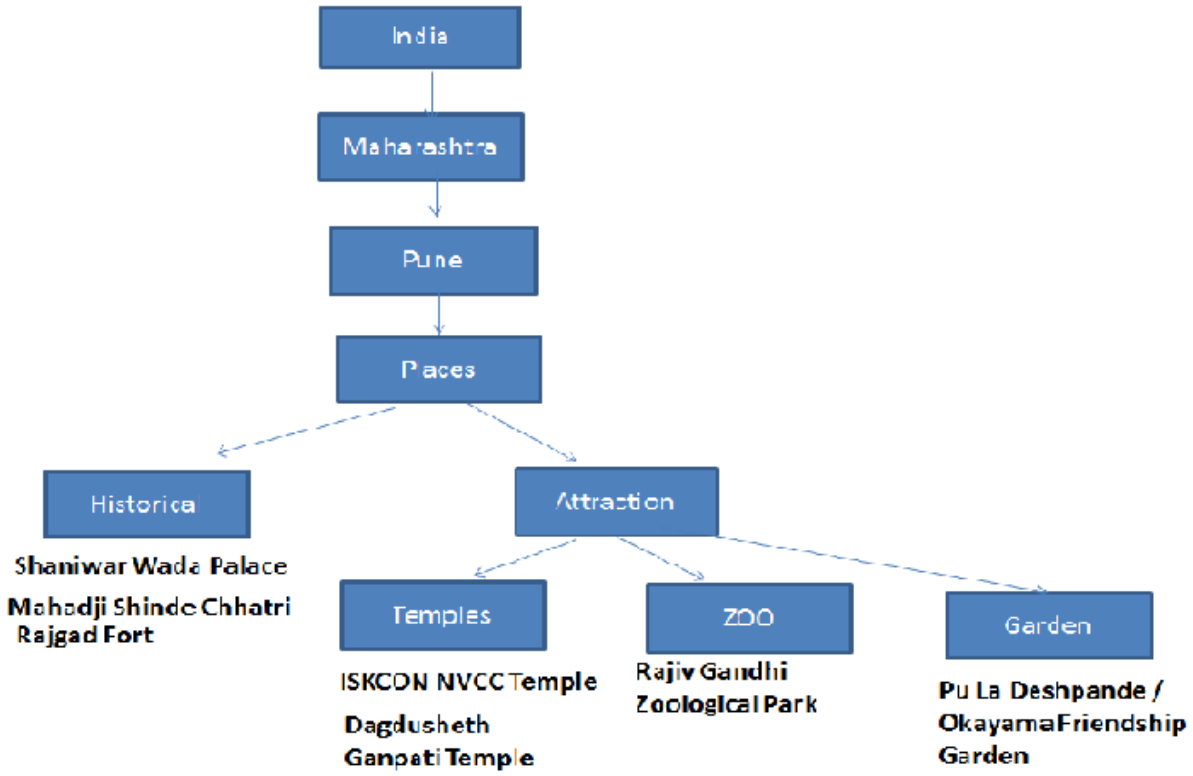
Ganpati Bhavan, 250, Budhvar Peth, Shivaji Road, Pune, Maharashtra 411002, India

hindu_temple

Butterfly Park

11, Aranyeshwar Marg, Aranyeshwar Nagar, Parvati Paytha, Pune, Maharashtra 411009, India park

4.2 Ontology Creation



V.EXPERIMENTAL RESULT

- The system provide a solution to process data semantically & It uses ontology learning methodology.It semantically model the query along with its weighted semantic relations to other related concepts.Also the system searches documents in Multiple languages.The system will able to search any query specific to Tourism Domain. Interoperability plays the major role in multilingual ontology. Here, the matching methods are important because it requires automatic searching and pattern matching of words of similar pattern or dissimilar pattern.

VI.CONCLUSION

The proposed system automates the generation of ontology by extracting semantic relationships between concepts from unstructured information sources. The system is fully unsupervised as it requires neither training, nor user annotation.The system searches documents only in English & Marathi.The main objective is to

ISSN (O) 2454-9665
ISSN (P) 2454-0663

allowing highly relevant pages to a query to be placed on the top positions of search results returned by a search engine.

ACKNOWLEDGEMENT

We would like to acknowledge the guidance of Mrs. K. .P. .Moholkar for her insightful support and inspiration throughout the various stages of this paper. We sincerely appreciate the help and advice given by her which went a long way in helping us understanding the key concept of this paper.

REFERENCES

[1] Elizabeth Liddy. "Enhanced text retrieval using natural language processing". Bulletin of the American Society for Information Science, 24, pp. 14-16, 1998.

[2] Philipp Cimiano. Ontology Learning and Population from Text Algorithms, Evaluation and Applications. Springer. 2006.

[3] L. Ding, T. Finin, A. Joshi, R. Pan, R. Cost, Y. Peng, et al., "Swoogle: a search and metadata engine for the semantic Web", in Proceedings of the 13th ACM Conference Information and Knowledge Management, ACM Press, New York, USA, pp. 652–659, 2004.

[4] Aliaa A.A. Youssif, Atef Z. Ghalwash, and Eslam Amer. "KPE: An Automatic Keyphrase Extraction Algorithm", IEEE proceeding of International Conference on Information Systems and Computational Intelligence (ICISCI 2011), pp. 103 -107, 2011.

[5] Chintan Patel, KaustubhSupekar, Yugyung Lee, E. K. Park," OntoKhoj: a semantic Web portal for ontology searching, ranking and classification", Proceedings of the 5th ACM international workshop on Web information and data management, pp. 58-61,2003.

[6] Fortuna, B., Grobelnik, M., Mladenic, D. "Background Knowledge for Ontology Construction". In: Proceedings of the 15th International World Wide Web Conference WWW 2006, Edinburgh, Scotland, pp.949-950, 2006

[7] Maryam Hazman, Samhaa R. El-Beltagy, and Ahmed Rafea. "Ontology Learning from Textual Web Documents". Proceeding of INFOS2008, pp. 113-120, 2008.

[8] OlfaNasraoui, Elizabeth Romero, Robert Wyatt, LeylaZhuhadar, "Multi-language Ontology-Based Search Engine", *International Conference on Advances in Computer-Human Interaction*, vol. 00, no. pp. 13-18, 2010, doi:10.1109/ACHI.2010.43