INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Semantic text analysis using machine learning

Bhargavi Joga
jogabhargavi21136@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Sarath Sattiraju
sarath.sattiraju@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Venkatesh Kandula
kandulavenkatesh222@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Narayana Murthy Kallempudi
murthykallempudi023@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Mandava Kranthi Kiran
mkranthikiran.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

ABSTRACT

As the amount of information on the World Wide Web grows, it becomes increasingly burdensome to and just what we want. While general-purpose search engines such as Ask.com and Bing high coverage, they often provide only low precision compared to others, even for detailed and relative queries. When we know that we want information about a certain type, or on a certain topic, a domain-specific search engine can be a powerful tool. Like www.campsearch.com allows complex queries over summer camps by age-group, size, location, and cost. Domain-specific search engines are becoming increasingly popular because they increase accuracy not possible with general, Web-wide search engines. Unfortunately, they are also burdensome and time-consuming to maintain. In this paper, we use machine learning techniques to greatly automate the creation and maintenance of domain-specific search. It describes new research in semi-supervised learning, text classification, and information extraction. We have built a demonstration system using these technics like Web Scrapping, Fuzzy C-Means and Hierarchy Clustering for a search engine which gives accurate results which is a more advantage when compared to other Search engines. Searching with a traditional, general purpose search engine would be extremely tedious or impossible to perform search operations. For this basis, domain-specific search engines are becoming popular. This article mainly concentrated on Project an effort to automate many aspects of creating and maintaining domain-specific search engines by using machine learning techniques. These techniques permit search engines to be created quickly with less effort and are suited for re-use across many domains.

***Keywords**— Machine learning, Web scrapping, Reference management software*

1. INTRODUCTION

The amount of text content available over the web is so abundant that anybody can collect the information related to any or particular topic. But, the performance of the retrieval system is still far below the level of our expectation. The automatic text categorization has received a very high demand by many applications to well organize a huge collection of text content in hand. For better understandable class structure is known as semi-supervised clustering. Generally, there are two different approaches to semi-supervised learning; one is a similarity-based approach and the other one is a search based approach. In similarity-based approach, an existing clustering algorithm that uses a similarity matric is applied, while in search-based approach, the clustering algorithm itself is modified so that the user entered labels are used to bias the search for an appropriate partition.

User-specific suggestions can be provided in any form. In our system, the user searches for the research papers and the suggestions are provided based on the domain he prefers obtained from his previous searches. The above mentioned is achieved by semantic analysis of the query and filtering out the required information from the analysis. The obtained information with previous searches can be used to provide better search results for the user. Today, almost every system is using machine learning techniques to provide a better and personalised experience for its users. Examples of user-personalised suggestions are YouTube suggested videos, Song recommendations in various music applications, Search results in Google. Google is close in relation to our system which uses semantic analysis for the queries by the user to provide search results. We intend to implement such system in our software, to provide a local search system which understands the user's queries even in general, daily used English and also refines the suggestions based on the previous search history and the papers uploaded by the user to the software. The system uses semantic analysis by extracting the relevant keywords from the query and obtaining the semantics of it by using the “Thesaurus” dictionary. Using these meanings, the domain is extracted and the search results are reduced to the relevant domain and sometimes to the specific paper that the user searches for.

1.1. Motivation

There are many file sharing systems in the current scenario and a few for sharing research articles. All these systems use employ a central server for storing the documents and send the files to the requested user. This client-server architecture requires high-cost maintenance on the server side as the server must have sufficient amount of computational power and memory for processing the requests and storing the huge collection of files. Scalability of the application also becomes an issue as the server has to be upgraded to serve more requests and store number of files. This results in high maintenance in cost and also the infrastructure for the server. The application is also vulnerable as the crash of the server causes the entire application to crash.

To resolve this problem, we intend to implement a peer-to-peer based file sharing system. Peer-to-peer based networks do not require a central server to connect between two systems, rather facilitate the transfer of data over the network directly from one system to another. This removes the problems faced by maintaining a high configuration server and also can be scaled easily.

For such a system, the search suggestions provided by the application for a particular user search query becomes an important aspect of the application, as the files might have similar keywords or titles associated with them. The application is mainly intended for the research articles, hence there arises a necessity that the intended domain has to be extracted from the query of the user. This methodology opens doorways to more user-customized search results and also makes the searching for a required from a huge collection of documents easier.

1.2. Existing System

According to “Mendeley” software, it is a reference manager that uses keyword-based search to retrieve the required research paper. The major drawbacks of Mendeley are that it maintains a central server for storing the data and uses a keyword-based search for retrieving the research papers requested by the user. Another application that uses another form of the searching algorithm is “Springer”, the retrieval of research papers in Springer is complicated and cannot be done by using simple queries.
Google uses a semantic-based approach and hence gives out appropriate results for the user’s queries.

2. PROPOSED SYSTEM

We propose a system that uses a semantic-based approach for retrieving the articles based on the user’s queries. The semantic part of the query is extracted and is mapped to the domain which gives the appropriate result. An ontology is built using the dump of the Wikipedia page and the keywords for each domain are extracted. The keywords are used as cluster points and the queries are then segregated into the clusters using Fuzzy C-Means Algorithm. Recursive Fuzzy C-Means is applied and the model is trained with the data. This model can be applied to real-time data and can give appropriate results for the user queries.

2.1 Architecture of the proposed approach

Our main motive is to build a reference management software which includes corpus-based and knowledge-based semantic measures for effective organization and retrieval of relevant research articles as desired by the user. In this view we considered two aspects to be important in our reference management software, one is an advanced semantic search for effective information retrieval and the other is effective storage and organizing of metadata retrieved from the incoming research articles. In our earlier work in building the metadata extraction includes Title, Author, Author Details and Reference for establishing a reference linking mechanism in the reference management software that works on a standalone personal computer. But in this work, we have mainly given importance to the Title and Concept of a research paper.

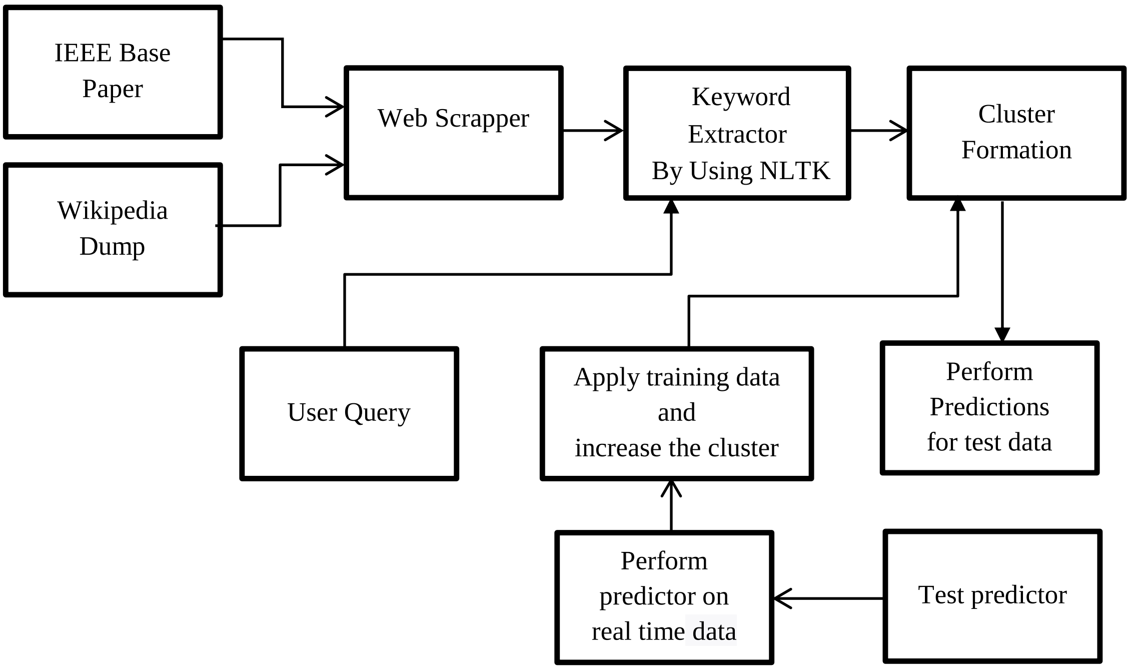
IEEE Base Paper
Wikipedia Dump
Web Scrapper
Keyword Extractor By Using NLTK
Cluster Formation
User Query
Apply training data and increase the cluster
Perform Predictions for test data
Perform predictor on real time data
Test predictor

Fig. 1: System architecture

As shown in figure 1, comprise four main modules IEEE Base Papers and Wikipedia Dump, which includes Title Extractor and Feature term extractor, which are very helpful for Web Scrapping, In Web scrapping module. Initially, an incoming research journal article in Wikipedia and IEEE Base Papers form is encountered to PDF parsing where the title extraction and feature term extraction is done. The job of the Term Keyword Extractor module is to expand each term and then stop words were removed from the extracted title with its synonyms if found in the WordNet and store them in the synonym store. If a word is not found in the WordNet, which it means it isn't a word that is generally used and is specifically technical, is also stored in the synonym store as a metonym to the concept of the research paper to which it belongs. Now each term along with its family of synonyms is given an index number and is stored in the term-number store. The job of a Cluster Formation is explained by manually forming a cluster initially by taking three domains and making them as groups with keywords for further prediction analysis given by a user through a query. Feature terms that are extracted from each incoming research paper and decide, to which concept the research paper be-longs by interacting with the computer science domain built with the help of Wikipedia dump and IEEE Base papers. The job of Keyword Extractor, which is explained in detail and is to analyse whether the query given by the user is a title or a set of important keywords belonging to a concept. Based on this analysis, the keyword Extractor module will either search for a match using a query to title comparison or search for a concept to which the mentioned keywords in the query match to and display the results with all the research articles that belong to that particular concept.

3. BUILDING COMPUTER SCIENCE DOMAIN

In the process of conceptualizing computer science domain, we have chosen to build Cluster, as it is defined as the specification of a conceptualization. As said in Cluster Formation provide a common vocabulary for all the researchers who are in need of sharing the information in the domain, therefore constructing an Cluster in a machine-readable formation can help the system we develop, to understand the basic concepts or the terms and the relation between them, used in the particular domain of target. There is also a progress in information retrieval, which has been improved from a key-word-based retrieval to knowledge-based search. As said in the literature survey we are not following a long process mentioned in where the authors have followed a seven-step method explained by. Instead, we have considered using Wikipedia and IEEE which is the free encyclopaedia and consists of best knowledge in the human perspective, as I said in earlier Wikipedia and IEEE Consists of vast amounts of highly organised human knowledge. As the resources from Wikipedia is growing day by day, it can also be considered as a source for new concepts that are being included throughout its expansion. Wikipedia and IEEE are mostly considered for a general sense, which might not deal with in-deep scientific knowledge, but it has the capacity to hold many important concepts and the hierarchical links and relations between them for a certain targeted domain, in our example, it is about computer science domain.

To help us in our task of building computer science domain Cluster Formation, which should consist of important concepts and the relations among the concepts, we have taken the help of Wikipedia dump and IEEE Base Papers. Our Cluster Formation module initially parses the Wikipedia dump which is in the form of IEEE papers and Wikipedia dump which consists the string between < title> </title> as a concept. From each concept that was found in the above process, their links were gone through to the other concepts, which are linked to them. Initially, Cluster Formation was built using Wikipedia and IEEE Base Papers with the help of Web scrapping.

Since semantic web provides a representational infrastructure for the metadata representation and RDF (Resource Description Framework) is a semantic web data model that is adapted to represent the information of the resources available on web and according to Cluster are populating the emerging semantic web and are the core data structure of the big web data and in other words helps in managing huge amounts of data. As the research publications have metadata and an increase in a number of research publications that are being stored in the reference manager on a personal computer or a desktop, will increase the metadata that becomes huge day by day. In this particular situation, it is desirable to use semantic web technologies for managing desktop data where each research publication can be identified by a URI (Uni- form Resource Identifier) and the metadata can be represented using RDF graphs. So, not only for building computer science domain, it has been decided to represent other data that has to be stored, as Cluster.

4. PARSING IEEE AND WIKIPEDIA DUMP RESEARCH ARTICLE

Parsing the Wikipedia and IEEE Base Papers is done for two main categories, title extractor and keyword extractor. These two categories have their own importance in building up an innovative Storage and retrieval system for reference management software.

4.1 Key Word Extractor

In research papers, keywords form an important component since they provide a short representation of the content of the paper. Keywords also play an important role in locating the paper from information retrieval systems, bibliographic databases and for search engine optimization. Keywords also help to categorize the paper into the relevant subject or discipline.

Conventional approaches to extracting keywords involve the manual assignment of keywords based on the article content & the author's judgment. This involves a lot of time and effort & also may not be accurate in terms of selecting the actual keywords. With the emergence of the Natural Language Tool Kit (NLTK), keyword extraction has evolved into being effective as well as accurate.

When a user enters a query in search of base papers, then the title was extracted from the given query than by using a National Language Tool Kit (NLTK) stop words were removed and then keywords were extracted. Then the keywords are compared with the predefined domain keywords to find the probabilities, the maximum probability value along with the relative domains serves as the accurate results.

5. EVOLUTION AND RESULT

The important aspects of our system are to evaluate title and extract the keywords either exact or partial if the user wants to search using the title from the raw query posed by the user, keywords were extracted from the title. The second one is the concept identification which is subjected to parsing and then compared with the clusters which were formed previously and displays the result with relative domain along with probability values.

For extracting the keywords from the query given by the user, we have initially taken 3 domains like unsupervised learning, Cybersecurity and y belonging to the computer science domain. For each selected concept, we have compared them with the given title from the user after extracting the keywords. Five different titles of those 3 domains were compared and results were displayed with values. After analysing the results a few domains were added from Wikipedia dump and IEEE base papers and the same above process was repeated with the different titles and the results were compared. By using this we made a precision testing with one of the search engines for finding the difference between the results and the following are the observations observed.

Sodhana: **Text Analyser:**

Queries:

1. **Machine Learning Model for Hemoglobin Estimation.**
2. **Intelligent Monitoring System.**
3. **Diabetes Diagnosis using Machine Learning.**

1.'Machine Learning'-0.7	**1. 'Machine Learning'-0.7**
2. 'Control Theory'-0.58	**2. 'Control Theory'-0.66**
'Document management system'-0.0	**'Document management system'-0.0**
'Un-supervised learning'-0.0	**'unsupervised learning'-0.0**
3. 'Automata theory'-0.0	**3. 'Automata theory'-0.3**
'MACHINE Learning'-0.5	**Machine Learning'-0.66**
'Supervised learning'-0.4	**'supervised learning'-0.6**
'Graphics processing unit'-0.2	**'Graphics processing unit'-0.2**

6. FUTURE SCOPE AND CONCLUSION

6.1 Future Scope

We intend to scale the algorithm even further by employing the algorithm to all kinds of files and scaling the application for all kinds of files. It causes a significant impact on the way search engines work and gives rise to different kinds of applications. We intend to generalize the algorithm to work with all kinds of data such as music files, video files etc.

6.2 Conclusion

The World Wide Web is a large resource of textual data and finding something on the web is a very tedious task. When it comes to searching for articles on the web, the intended or a useful article is never found in the first instance. It requires a certain amount of time to dig through various results and refine the search query to finally get the appropriate result. Many applications employ keyword-based searches which never give accurate results when the article that is to be searched gives the problem of ambiguity with another having similar keywords. We intend to provide a semantic-based search engine that provides a domain extraction of the query and gives the appropriate results related to the domain. It also considers the previous search history if there still occurs an ambiguity for the query.

7. REFERENCES

[1] "Natural Language Processing for Information Retrieval: the time is ripe (again)", Matthew Lease, Brown Laboratory for Linguistic Information Processing (BLLIP) Brown University, Providence, RI USA

[2] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore, "A Machine Learning Approach to Building Domain-Specific Search Engines"

[3] "Search Personalization using Machine Learning", Hema Yoganarasimhan, University of Washington

[4] "Semi-Supervised Text Categorization using Recursive K-means clustering", Harsha S Gowda, Mahamad Suhil, D S Guru and Lavanya Narayana Raj

[5] SodhanaRef: A reference management system built using hybrid semantic measure, Mandava Kranthi Kiran, Dr K Thammi Reddy

[6] https://medium.com/udacity/natural-language-processing-and-sentiment-analysis- 43111c33c27e