Multilingual Search Engine in Commercial websites

G. Jaikishore, Vignesh.T, Dasari Vijaya Rani, Priyanga.A.

Abstract:

In this modern internet world has developed with many cloud the cloud with more advanced services. The internet world is used by well-educated persons and people aware of internet. The main object of this project is to make the illiterate make use of resources available in the internet and make them to involve in online shopping site by providing the speech to text service for performing search operations and further more operations and item identification is also there for the so far the Google, Bing…and many other search engines are powered with speech based searches. But in case for Commercial websites textual based searching medium is only supported for searching products. In this project the online websites are made search through voice based search operations and Multilingual based search using virtual keyboard for the Indian languages. For the language identification performed through the encoding processing and then the search is made in the selected language. Thus the users can search the product through the voice based input. Thus in country like India with more diversity there will be more number of language used among various regions then among the various languages there will be different dialect.

So for different dialect machine learning is implemented for keyword identification in different dialects through displaying the pictures of random products pictures in the database. Thus the main objective is to bring the illiterates to the internet world for the initiation the multilingual and the speech recognition support is to be given in commercial website with NLP support and Machine learning to fetching the appropriate product from the queries given in multiple languages and displaying the product.

Introduction:

Online shopping is the process whereby consumers directly buy goods, services etc. from a seller interactively in real-time without an intermediary service over the internet. Online shopping is the process of buying goods and services from merchants who sell on the Internet. Since the emergence of the World Wide Web, merchants have sought to sell their products to people in online. Customers can visit web site from the comfort of their homes and shop as they sit in front of the computer. Consumers buy a variety of products from online stores. In fact, people can purchase just about anything from companies that provide their products online.

The given input in any language is used. To quickly get an overview of the contents of such datasets, tools for exploratory analysis are essential. We propose a method for extracting from a set of texts the relevant words that distinguish these documents from others in the dataset. We can then summarize the texts belonging to each product by visualizing the extracted relevant word thereby enabling one to grasp the contents of the documents at a glance. This paper shows how to identify topics in the dataset and then selecting and visualizing relevant words, which distinguish a group of documents from the rest of the texts, to summarize the contents of the documents belonging to each topic. The main idea is about to bring the illiterates to the modern computer world through machine learning algorithms and various supports provided. The term search engine is general would search for metadata and use page rank algorithm for searching through the search engine.

For Key word extraction NLP (Natural Language Processing) is used. Natural language processing is a branch of computer science and artificial 0intelligence which is concerned with interaction between computers and human languages. Natural language processing is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. These includes the spoken language systems that integrate speech and natural language. Natural language processing has a role in computer science because many aspects of the field deal with linguistic features of computation. Natural language processing is an area of research and application that explores how computers can be used to understand and manipulates natural language text or speech to do useful things. The applications of Natural language processing includes fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence(AI) and expert systems.

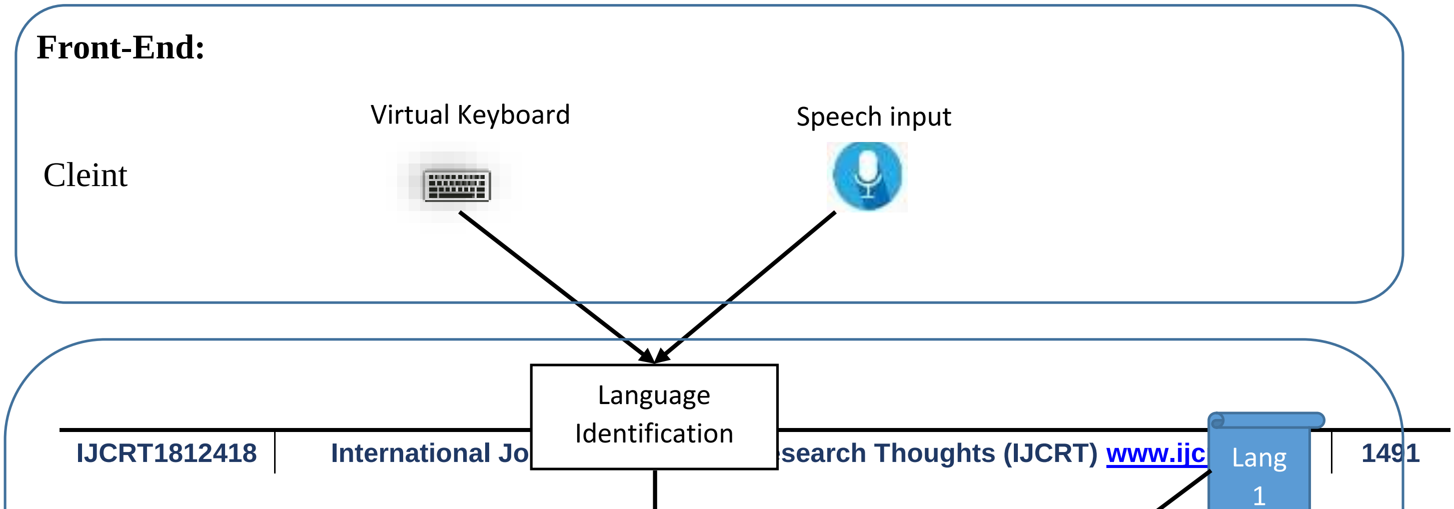
Literature Survey:

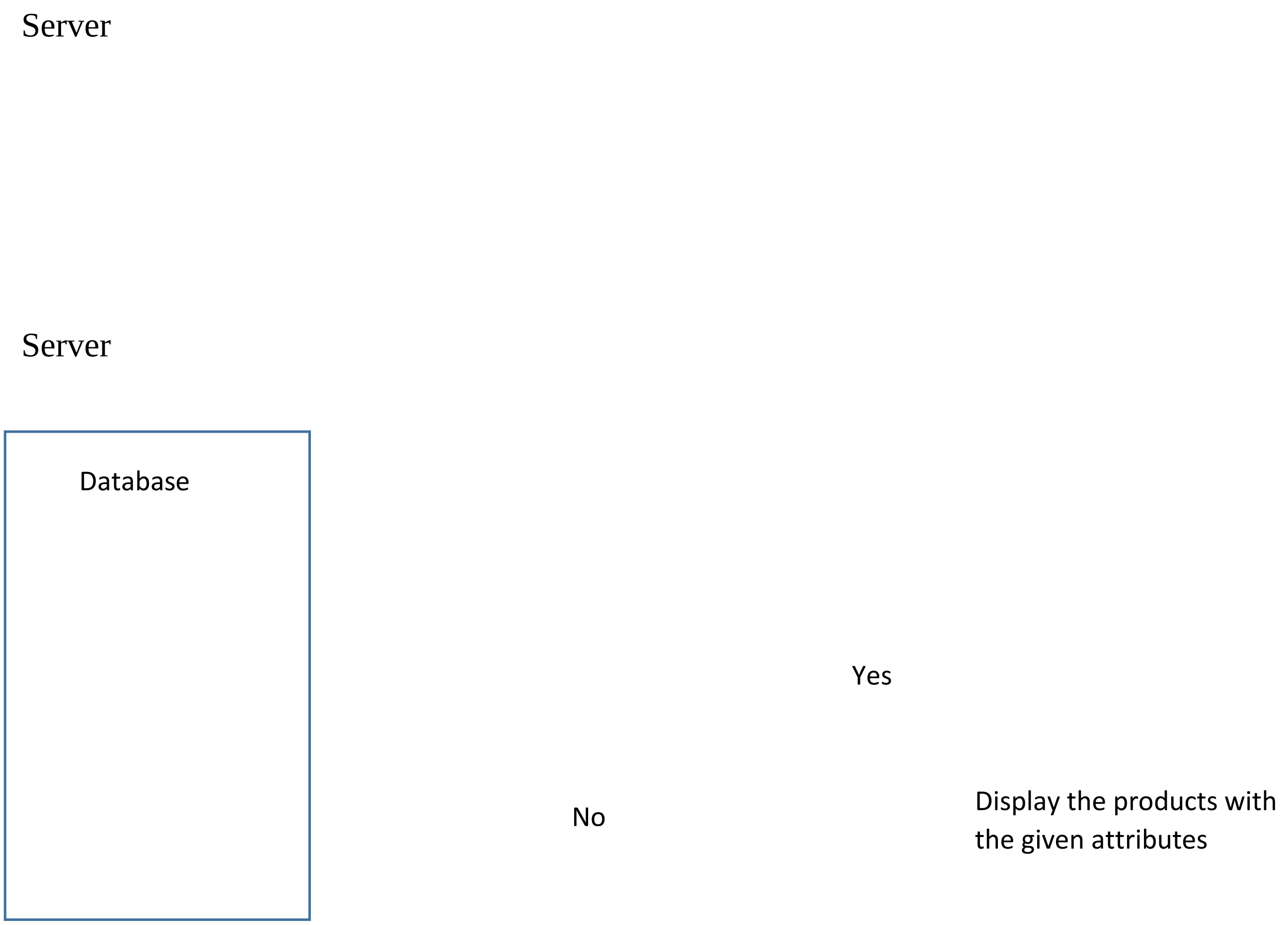
Si .no	Author Name	Paper Topic	Techniques	Problem	Solution	Ref.
1.	David A. Hanauer, Danny T.Y. Wu, Lei Yang, Qiaozhu Mei, Katherine B. Murkowski-Steffy, V.G. Vinod Vydiswaran, Kai Zheng	Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine	Information retrival using the metamap and indexing through lemur search engine	The patients get awareness about the disease with the symptoms	Sematic based information retrival from given queries	[1]
2.	Ilya Segalovich	A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine	Morphology analysis for unknown keyword using dictionary based analysis	In multiple languages the unknown key word may occurs due to some lexical grammatical mistakes.	Trie based dictionary algorithm	[2]
3.	Vidyanand Choudhary , Imran Currim, Sanjeev Dewan, Ivan Jeliazkov Ofer Mintz, John Turner	Evaluation Set Size and Purchase: Evidence from a Product Search Engine	Machine Learning	The results reveal that evaluation set size and purchase are negatively correlated and that factors typically presumed to be associated with purchase	Conceptual framework linking four categories of antecedents to ESS and purchase by integrating elements of search cost, information, two-stage, context effects, and consumer behavior theories.	[3]
4.	C.M.M. Mansoor and H.M. Nasir	Tamil Search Engine for	Multiple languages	They use standard	Answering the challenge question	[4]

		Unicode Standard		and improved reinforcement learning based memory-network architectures to solve QRAQ problems in the difficult setting where the reward signal only tells the Agent if its final answer to the challenge question is correct or not.	requires multi-turn interactions in which a good agent should ask only non-deducible and relevant questions at any turn.	
5.	Hideo Joho, Adam Jatowt, Rio Blsnco	Temporal information searching behavior and strategies	Access the Metadata	Temporal information retrieval, the user gives the query to the search engine which seek for time related information whether it is past, present, future for more effective results.	Accessing the data Cookies and metadata from Client-side which used for analyzing the temporal information	[5]
6.	Xiaoxiao Guo, Tim Klinger, Clemens Rosenbaum	Learning to Query, Reason, and Answer Questions On Ambiguous Texts	Machine learning to understand query from given memory and environment	Interactive agent for understanding the query and answering to the query	Reinforced learning is used to learn from given queries and respond to the given queries. base reinforced learning is used to learn from memory network and imp Reinforced learning which improve the better result by software-	[6]

					attention mechanism.	
7	MarcoBasaldella ,Muhammad Helmy, ElisaAntolli MihaiHoriaPopescu ,GiuseppeSerra and CarloTasso	Exploiting and Evaluating a Supervised, Multilanguage Keyphrase Extraction Pipeline for Under-Resourced Languages	Low-level NLP, Candidate generation, Features extraction, Candidate scoring	Key word extraction for the given query in under resourced language	Automatic keyword extraction agent is built with various steps like stemming, lemmatization and learning word phrase patterns from various documents.	[7]
8	Eva Katta, Anuja Arora	An Improved approach to English-Hindi based Cross Language Information Retrieval System	Naïve Bayes and particle swarm optimization algorithm	The user give the query in their user native language which the key word is extracted for query processing	Naïve Bayes and particle swarm algorithm which suggest the result in which the query given in same format as in document	[8]
9	Pascale Fung and Lo Yuen Yee	An Approach for translating new words from nonparallel ,comparable texts	NLP	Our algorithm is the first to have generated a collocation bilingual lexicon	Algorithm has good precision, lint the recall is low due to the difficulty in extracting unambiguous Chinese and English words.	[9]

Architecture Diagram:





Proposed System:

The Proposed system consists of following models:

1. Input interface.
2. Language Identification.
3. Extract keyword.
4. Product Search.
5. Learning Process.

1. Input interface:

The input to the system is given through Voice command or either virtual keyboard. The Virtual keyboard is made to avail in all Indian languages and then user have to choose the language to give the text based input in their native language.

In addition to it speech recognition google api is integrated in for voice command on their user own language they need to provide in which language they are going to speak and the user can provide voice based command this speech api is integrated in the point of allowing the illitrated people to perform online shopping and in voice based command about their product.

1.1. Google Speech Recognition api [11]:

Google Speech to recognition api[11] convert the speech dictated to the text using powerful neural network model and deep learning. This recognize module supports over 110 languages. This API can filter inappropriate content and recognize the speech in noisy environment. Automatic speech recognition is powered with deep learning. They have made this service allocated through API for integrating with various services.

1.2. Virtual Keyboard:

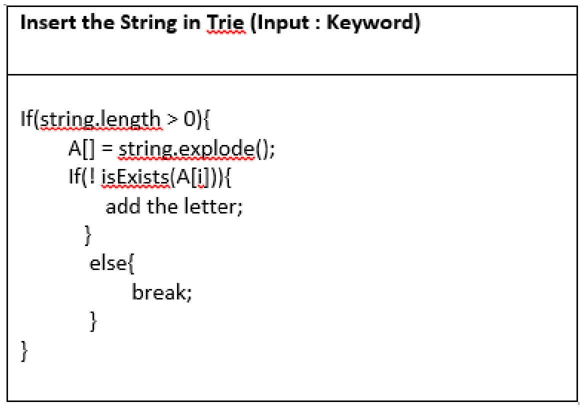
The virtual keyboard act as the interface for input given in multiple languages**.** Various languages is given in the keyboard like tamil, telugu, hindi kandam for the input given in multiple languages. Thus the virtual keyboard is build with jquery.

2. Language Identification:

Language identification is perform to identify in which language the user have given the input either the virtual keyboard or google speech recognition api both return the text based input for analysis .Thus from the given text the language is identified using the Unicode in the text then the key word is used to search the product in appropriate language.

3. Extract Keyword:

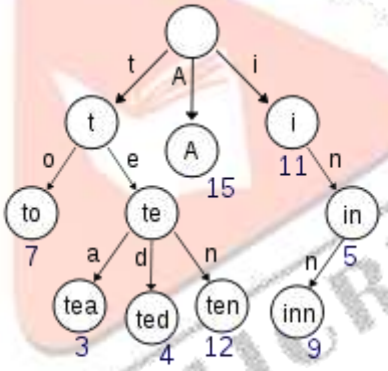
The key word is extracted using trie based dictionary algorithm demonstrated in [2]. Thus in this they have performed key word extraction using dictionary based trie algorithm. Trie is effective information retrieval algorithm.



Insert the String in Trie (Input : Keyword)

```
If(string.length > 0){
      A[] = string.explode();
      If(! isExists(A[i])){
            add the letter;
         }
         else{
               break;
         }
}
```

Thus the algorithm used search the product on tree based on key extracted in multiple languages.



The key word extracted is in tree based structure

4. Product Search:

Thus the keyword extracted using trie based algorithm then the extracted keyword is search through the database with the given attributes like color, type, materials, price…etc.,

5. Learning Process:

Learning process is done using supervised learning algorithm and which train the system using the various datasets and then in the case of search for unidentified product by asking the user to identify the products showing some random images then the system is trained accordingly using the prediction in [10].

Reference:

1. Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine, March 2017.

2. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, January 2003.

3. Perceived irritation in online shopping: The impact of website design characteristics, March 2015.

4. Tamil Search Engine for Unicode Standard, Nov 2004.

5. Temporal information searching behavior and strategies, March 2015.

6. Learning to Query, Reason, and Answer Questions On Ambiguous Texts, March 2017.

7. Exploiting and Evaluating a Supervised, Multilanguage Keyphrase Extraction Pipeline for Under-Resourced Languages, Sep 2017.

8. An Improved approach to English-Hindi based Cross Language Information Retrieval System, 2017.

9. An Approach for translating new words from nonparallel, comparable texts

10. Predicting good probabilities with supervised learning,2017.

11.Google Cloud, from https://cloud.google.com/speech/.