Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 3– No.8, December 2015 – www.caeaccess.org

Document Retrieval using Multilingual Keywords

Deepika Singh T.
Computer Engineering
VESIT
Mumbai, India

Meghna Peswani
Information Technology
VESIT
Mumbai, India

Piyush Mantri
Computer Engineering
VESIT
Mumbai, India

Sagar Bhojwani
Computer Engineering
VESIT
Mumbai, India

ABSTRACT

Maintenance of information is a serious issue. Easy retrieval of information is often deprived when it is needed, and it is not always feasible to remember the exact file name when searching for a specific document. To address this issue keywords can be used in a system that help in retrieving the desired files by ensuring proper management of data. The designed system will help the user to retrieve data instantly and efficiently. A list of multilingual keywords is defined for every document on the database. The user input keywords are compared with the keywords from the database and the relevant searches are formulated in the form of query which is then output to the user as a list of relevant documents which have been retrieved. This paper will give an in depth knowledge on how keywords in more than one language can be used to access data instantly and effectively.

Keywords

Keyword mapping, multilingual, indexing, information retrieval

1. INTRODUCTION

Information retrieval and relevant extraction has become a very important part in today’s world of data abundance. Keywords form a pivotal role in information extraction, having found application in document extraction, online searches and summarization of texts. Using keywords or indexed terms makes hunting for specific results easier and the entire document to be searched is retrieved efficiently. Thus, selecting effective keywords is a necessity. It is required to ensure that the keywords are selected correctly and precisely, So as to say that the keywords chosen are independent and appear frequently in the document to be retrieved. Such techniques are used to identify documents accurately, but the pre-requisite is that the entire collection be made available beforehand.

The challenge arises when keywords are to be extracted from a single document. This is because when a document collection is considered, the keywords are expected to be independent for information retrieval. However, in case of single document, there is no assurance that the keywords of the given document will be independent from keywords of other documents. In the past, single document extraction has been achieved by using methods such as co-occurrence and machine learning.

Previously, algorithms were assessed by using human judges and evaluators. The authors would get their solutions examined most often by the precision and recall methods, as a measure of accuracy between the algorithm’s keywords and those determined by the human judge. This would be the most genuine evaluation if the aim were to make keywords that would mimic the human brain’s thought process, or in other words, create keywords that the human would use. But in a case where the keywords will be used by other information retrieval algorithms and techniques, then human judged keywords are not considered the best form of evaluation. Moreover, these methods were designed and implemented to work only on one language. This means that the authenticity of the algorithm on other languages is unknown. In contrast, the evaluation method inscribed in this paper highlights an information retrieval technique, keyword searching, which determines effective keywords by distinctively describing the documents they are extracted from. The testing is done on documents containing the languages English and Marathi to estimate the applicability on different languages. This will be further examined in the later sections of the paper.

A keyword extraction tool for multi-lingual intercommunication is needed. This is the early stage of developing a cross-lingual information extraction and management system, which reduces the problems faced due to language barriers. The usage of linguistic information is helpful in various IR related jobs. As a result, the algorithm presented in this paper requires the use of language dependent components, such as a morphological analyzer and simple noun phrase (NP) grammar. While these components are readily available for most languages, there is a provision of software that can be used for languages for which the components are not designed. The TnT Tagger is a helpful tool to handle any language with segmented white spaces. In this paper, the algorithms have been used in multiple ways, from topic analysis to finding similar documents. The goal is to analyze how well the keywords accurately describe the document they were extracted from rather than the way humans perceive the keywords.

The paper will proceed as follows; Section 2 will give the background information about the system that is being developed. Next, in section 3 the multilingual tool will be briefly discussed. In section 4 the algorithm will be explained. Experimental results will be shown in section 5. In section 6 related work and inferences will be discussed. Finally, in section 7 and 8 concluding remarks and future works will be discussed respectively.

2. BACKGROUND

Presently, the proposal is to build a multilingual system for extraction, management and presentation of information. The aim is to allow the user to search any keyword from, say their native language, and retrieve all the relevant information from documents in any language. This would be portrayed to the user in a transparent environment, meaning that the user would not know what the original language was for which the answer has been obtained. The main focus will be in the news and educational fields. English and Marathi are the targeted languages for time being. It will consist of the following parts:

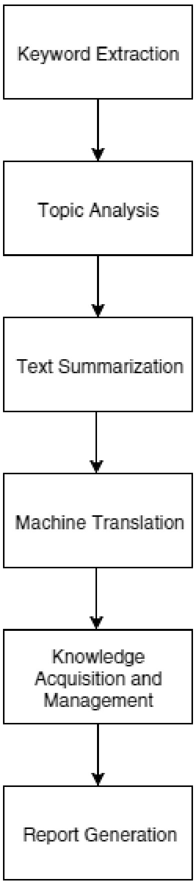


Figure 1: Parts of the system

Document summaries will be created by Keyword Extraction and Topic Analysis. The translation of keywords, answers, and all other kinds of texts will be majorly done by machine language and Google's multilingual tool. There would be a scope of learning from reports and questions that are asked, which will be done through the knowledge acquisition and management modules. This information will be extracted and used later on. The input and output components of the system are essentially the question/answers and reporting modules. The user will be given a brief one sentence answer to the question asked and the report generation module will present a summary of the same to the user. Typical keyword extraction algorithms are not suitable for this proposed system as the underlying base knowledge is not made available beforehand. Moreover, previous extraction methods were more focused on how the humans would agree with the keywords rather than extracting those keywords which would actually be useful in IR tasks. These reasons, along with the fact that keywords play a pivotal role in the system as they will be used by us very extensively in various applications; a change to the pre-existing algorithm was needed.

Machine translation will be required for translating keywords, answers, documents, etc. The knowledge acquisition and management modules will be an attempt to learn from the questions that are asked or reports that are asked for and extract information to be used later on. The question and answering and report generation modules will be the input and output of the system. In question and answering the user will be presented with a sentence or two giving the answer to the question they asked. The report generation module will present a summary on a topic to the user. The first step in this system is to build an effective keyword extraction algorithm. Since the underlying knowledge base is not able to be collected beforehand typical keyword extraction algorithms may not be suitable. Moreover, the previous methods for keyword extraction from single documents were more focused on extracting keywords that humans would agree with than they were with extracting keywords that are useful in IR tasks. Because of these reasons and the fact that the extracted keywords will be play a pivotal role in the system.

3. MULTILINGUAL TOOL

In order to differentiate the proposed system from the older ones, it was necessary to incorporate the flexibility of language to the user. The user would be able to type in one language and retrieve the information in any other language. For this, the system made use of the Google Multilingual Tool. A database of translated words was maintained at the server end. Depending on the user request, a call for multilingual keywords was made and relevant search results were extracted and presented to the user. Presently, this search tool has been implemented in two languages, English and Marathi.

4. KEYWORD EXTRACTION AND WORKING

The keyword extraction algorithm was programmed on simple querying languages which are easy to understand by an entry level technician. Basic web programming tools aided the development of an efficient searching tool for documents which contained multilingual keywords. Programming languages like HTML and CSS were used to design the front end and interface of the local host based extraction tool. For intensive working and robust performance, the back end development made use of PHP codes with MySQL databases to store the documents which would be searched by the end user.

The general working is divided into two broad categories. Before searching for any document which is to be retrieved or extracted, it needs to be uploaded to the database. This is made possible by enabling two modes of working – the upload mode and the search mode. Both of these modes are accessible by the user, depending on whether or not he is authorized and registered personnel with the organization, that is, has a personalized login or is just a guest user. The functioning of the two modes of operation is as follows:

4.1 Upload Mode

Before extracting or searching for a document, one needs to be able to upload a collection of documents into the database which can ultimately be used as an inference for the searching tool. The upload mode is visible to only those users who are authorized with the permission to manage the functionalities of the extraction tool. This is done to avoid the chance of security issues and resolve conflicting interests. If any guest user were to upload documents, the authenticity of documents would be compromised and would need higher degrees of scrutiny. It would be difficult to maintain a check on the quality of documents that are being uploaded. To eradicate this error, the upload privileges are given to only the admin users or other authorized personnel. Few of the code snippets from the upload form are shown below:

```
// Check if file already exists
```

```
if (file_exists($target_file))

{

echo "<script> alert('Sorry, file already exists.')</script>";

        $uploadOk = 0;

}
```

The above code checks if the uploaded file previously exists in database or not. This is to avoid duplication of the same or similar data, so as to provide better results to the end user. If the file is found in database, the system generates an error message and displays it to the screen, stating that the file is pre-existing in the database and it is not possible to recreate the intended file. This sets the uploadOk pointer to 0, which means that error has been generated.

```
// Check if $uploadOk is set to 0 by an error

if ($uploadOk == 0)

{
echo "<script> alert('Sorry, your file was not
uploaded.')</script>";
// if everything is ok, try to upload file
}
else
{
echo "<script> alert('The file ". basename(
$_FILES["fileToUpload"]["name"]). " has been
uploaded.')</script>";

}
```

Further on, a check is generated to verify whether an error has been generated by the system or not. In case of an error, user gets an intimation that the upload has failed. However, if there hadn't been any errors generated by the system, it goes ahead and uploads the file into the database by calling and writing the contents of the file as a new entry in the database. Once that is done, keywords are generated and searching for this uploaded file can now be implemented.

4.2 Search Mode

Once a collection of documents have been uploaded into the database, it is possible to search the documents and retrieve them depending on the searched keyword. This facility is not restricted to registrations or authorizations and can be availed by any user. Considering a guest appears on the search tool page, he is not asked for a login. The search can directly be performed by using a single keyword or multiple keywords in more than one language. Presently, the search is made available in two languages; English and Marathi. Code snippets from the search form are shown below:

```
if($srchkey != "")
{
                $start = microtime(true);
                $sql = "SELECT name, date FROM `files`
WHERE  `name` LIKE '%$srchkey%' OR `keyword` LIKE
'%$srchkey%'";
                $retval = mysql_query( $sql, $conn );
                $end = microtime(true);
                $num_rows = mysql_num_rows($retval);
                echo "<center>". $num_rows ." results found in
". round(($end - $start),3) ." seconds.</center>";
                if($num_rows > 0)
                {
echo "<table border=1px> <tr><th>File
Name<th>Date<th>View File</tr>";
                while($row = mysql_fetch_assoc($retval))
                {
                                echo "<tr><td><a
href='Files/".$row["name"]."'>". $row["name"] ."<td>".
$row["date"] ."<td><a
href='Files/".$row["name"]."'>View</a></tr>";
                }

                echo "</table>";

                }else{

                echo "0 results";

                }

}
```

The above code represents the logic behind the search tool. Every time the user enters a keyword to be searched, the algorithm goes into the loop and starts executing. If all the data entries match with the required information of the user, the system generates a message that the document has been retrieved in a certain amount of time and further continues to display all the relevant files in a tabular manner. However, if the search key does not return any relevant results, it gives rise to a search fail and a message of no results found is intimidated on the screen. The same logic can be implemented on a search based on date.

5. EXPERIMENTAL RESULTS

After working out an extensive algorithm to meet the demands of an efficient searching system based on multilingual keywords, the results were analyzed. The basic guidelines as identified during implementation stated that only the authorized personnel of an organization will be allowed to upload the documents along with their respective keywords or identifiers. The admin will have to login before uploading any document. Users need not have a unique username to use the system. Search can be done anonymously. The system has the capability of caching the most recent keywords. This will enable the user to search the desired data quickly. The retrieval of the information can be done by the input of any one of the many keywords and not necessarily all of them. Information can be managed easily due to a secure and expandable database. The system exists independently or can be hosted on a website. The user capacity of the system will be dependent on the server capacity of the host. The results will be displayed in a tabular format comprising of columns of name, date and the option to view the file. Given below are a few snapshots from the experimental results:

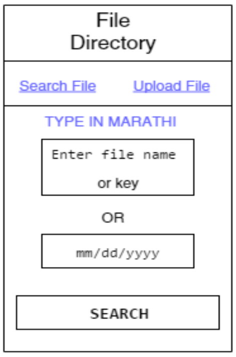
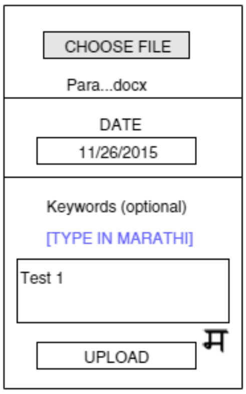


Figure 2: Admin screen with options

This image is viewed by the admin only. The registered admin has the option to either search for the file in the directory or upload the file and update the database. Based on the selection of the option by the admin the respective screen appears.



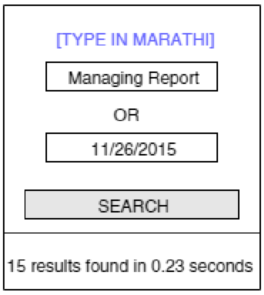
Para...docx

11/26/2015

Test 1

Figure 3: Admin screen while uploading

This image is seen only by the admin while uploading the file into the system directory. The keywords that the admin enters can be in Marathi or English. The language of input can be toggled very easily by the admin using the option given. The Devnagiri script is currently active then it can be seen just below the keywords text box. The date of the file is also a keyword. All the keywords along with the date are stored in the database which is then further used to retrieve the data by the user.



Managing Report

11/26/2015

15 results found in 0.23 seconds

Figure 4: User screen while searching

This image pops up when the user wants to search for any file present in the directory of the system. The files can be inserted into the database of the system only by the admin. The file can be retrieved by the keyword reference or by the reference of date on which the file was created. The input of data in the search box can be multi-lingual. Currently the system supports English and Marathi languages only. The Google tool can convert the English data to Marathi i.e. Devnagiri script in case the keyboard of the user does not support Devnagiri script. This makes the system flexible.

6. IMPROVEMENTS FROM PREVIOUS MODELS

On comparing this system with pre-existing ones, there are a few new features that have been established, namely:

6.1 Measuring the performance metrics

After every successful search, the retrieval time is displayed thereby giving an insight into the internal working of the system. If it takes longer to retrieve, then either the arrangement of the files in the repository need to be structured well or the crawling code needs to be improved accordingly.

6.2 Flexibility to attach any file system

Given that the archives would grow in volume with greater velocity and comprise of complex variety of data, there will arise a need to using distributed file system frameworks namely Hadoop, MongoDB and Spark. This would aid in decreasing the retrieval time and ensuring accurate results.

6.3 Device independent

The software package can be embedded onto any browsing application or in a mobile application source code as well. This feature was kept in mind while designing the code, in order to avoid any hassles arising due to platform dependency.

7. INFERENCE

The system has been tested on sample domain of computer science containing books of chapters and it is now possible to reduce the number of words to be searched in the file, thereby minimizing the search space. This effectively reduces the searching time as well. Reduction causes search spaces to be reduced more than 90% as the formula implemented for selecting high frequency words finally used for index creation selects only such amount of words. There are two type of comparisons are also performed which affects the search results. These are – search using ontology or without it – phrase based vs. term based search. In the first comparison using ontology, recall increases more than 70% than without ontology, if user enters words related to the domain in the query. In the second comparison term based approach’s results turned out be less relevant to the query in comparison to phrase based approach.

For an example query operating system would fetch file system file as most relevant document while other would fetch operating system concept file as most relevant which is logically correct. This is because former approach uses both word operating and system as distinct term while later treats them as single. But recall in term based approach would be more as there are more terms to find out in the repository, also in case of single word query only term based approach would fetch a result. Also second comparison may enjoy the benefits of using ontology in both case and hence can improve recall. Although this searching method takes some time to index and then search the query but it reduces time of overall search in comparison to time required to search a document as a whole. By varying threshold of index creation it is possible to vary the number of words in document descriptive i.e. index table. Moreover, the threshold value above a particular limit can eliminate some important words which are not desirable for the search. This limit depends upon the size of the documents that are being used in the system.

8. CONCLUSION

This paper described a technique which uses the concept of stemming so that a word can be searched using its root form and hence no need to be worried about query word’s lexical forms. It also reduces search space by removing stopwords which are not helpful in search. It is possible to vary the number of words in document descriptive i.e. index table by varying threshold of index creation. The matrix multiplication approach finds out comparative results as which file are more relevant to the query and thus useful in ranked retrieval of documents. Use of ontology made a 70% recall for this system. The use of phrase based approach with traditional term based approach increased relevancy between query and the result opted by user. Thus it shows an easy and fast approach to information retrieval.

9. FUTURE WORK

The next step would be to attach semantic meaning to this IR system both in query as well as documents. The aim is to develop an information retrieval system which will be able to take the traditional word based indexing and add word semantics to it—an intelligent system that automatically concatenates to the base repository of indices and updates it regularly. The prime tasks being, the indexing and retrieval components, which will use a combined word based and sense-based approach. Main focus of the system would be a methodology for building semantic representations of open text, at word and collocation level. Also by using ontology, the attempt will be to incorporate different relationships like is-a, has-a, part-of etc and make use of domain knowledge to make efficient search. Thus incorporating semantic indexing approach with improved keyword based search approach overall efficiency of IR system can be improved.

10. REFERENCES

[1] Using Structured Queries for Keyword Information Retrieval; Rajasekar Krishnamurthy Sriram Raghavan Shiva Kumar Vaithyanathan Huaiyu Zhu IBM Almaden Research Center, San Jose, CA 95120

[2] Toward industrial-strength keyword search systems over relational data; Baid, A.; Rae, I.; Anhai Doan; Naughton, J.F.; Comput. Sci. Dept., Univ. of Wisconsin, Madison, WI, USA.

[3] Callan, J. P. and W. Bruce Croft. An Evaluation of Query Processing Strategies using TIPSTER collections. In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, 347-356

[4] A.Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge," In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), 2003.

[5] P. Turney, "Learning algorithms for keyphrase extraction," Information Retrieval, vol. 2, no. 4, 303-336, 2000.

[6] J. Fry, "Parallel Japanese-English corpus," [Online], Available at http://johnfry.org/blog/.

[7] P.Resnik and N.A.Smith,"TheWebasaparallel corpus,"Computational Linguistics, vol. 29, no. 3, 349 - 380, 2003.

[8] D. D. Lewis, Y. Yang, T. Rose, and F. Li, "RCVI: A New Benchmark Collection for Text Categorization Research," Journal of Machine Learning Research, vol. 5, 361-397, 2004.

[9] Kumar Sourabh Vibhakar Mansotra, “Factors Affecting the Performance of Hindi Language searching on web: An Experimental Study”, International Journal Of Scientific & Engineering Research, Volume 3, Issue 4, April-2012 1 ISSN 2229-5518, pp-1-15.

[10] Kumar Sourabh, Vibhakar Mansotra, “An Experimental Analysis on the Influence of English on Hindi Language Information Retrieval”, International Journal of Computer Applications (0975 – 8887) Volume 41– No.11, March 2012

[11] Roger Bradford, John Pozniak, “Combining Modern Machine Translation Software with LSI for Cross-lingual Information Processing”, 2014 11th International Conference on Information Technology: New Generations, pp-65-72.

[12] Maria Pia di Buono, Mario Monteleone, Federica, Johanna Monti, “Knowledge Management and Cultural Heritage Repositories Cross-Lingual Information Retrieval Strategies”, 2013 IEEE, pp-295-302.

[13] Sandeep Chaware, 2Srikantha Rao,”Ontology Supported Inference System for Hindi and Marathi”, 2012 IEEE

[14] Fuminori Kimura, Akira Maeda, Kenji Hatano, Jun Miyazaki, “Cross-Language Information Retrieval by Domain Restriction using Web Directory Structure”, Proceedings of the 41st Hawaii International Conference on System Sciences – 2008, pp-1-8.

[15] Hassan Alam, Aman Kumar, “Multi-Lingual Author Identification and Linguistic Feature Extraction – a Machine Learning Approach”, 2013 IEEE, pp-386