iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** I **Month of publication:** January 2019

DOI: http://doi.org/10.22214/ijraset.2019.1117

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue I, Jan 2019- Available at www.ijraset.com

Content Based Page Ranking by using some Natural Language Processing Techniques

Gumduboina Seshu Babu[1], Sachin Saj T K[2]
[1]*Department of Computational Engineering and Networking, Amrita Vishwa Vidyapeetham*
[2]*Department of Computational Engineering and Networking, Amrita Vishwa Vidyapeetham*

Abstract: ***World Wide Web is a collection of huge data which is increasing exponentially day by day and it consists of different formats of data's. When the user tries to search anything in search engine, it gives many results, which can be relevant as well as irrelevant information related to user query. Web mining is a process of extracting information from this web resources and show only relevant data to the user from all the unstructured /irrelevant data's. One of the traditional method used by search engine is page ranking algorithms to analyse and rank the web pages. In this paper, we are ranking the web pages on bases of content in that webpages related to the user query and this is called as content based page ranking algorithm, in which user will give a keyword and our architecture will extract 'N' links related to that keyword and by using some natural language processing techniques, similarity and frequency estimations techniques, ranking of the web pages is been done.***

I. INTRODUCTION

The web is the largest source of data. The data can be structured as well as unstructured data plus it can be noisy also. Since the data present in today's world is so huge, the search engine is facing a big challenge of providing the user with the proper information that he/ she meant to get. Web mining is a technique which can resolve this problem to a larger extent.
Web content mining is a process of extracting important information's related to certain kind of user query from the web pages. There are two kind of approaches used in web content mining

A. Agent approach
B. Database approach

There three types of agents
1) Intelligent search agents
2) Information filtering agent
3) Personalised web agents
In intelligent search agent it automatically searches for the information's related to the user query using domain characteristics
In web contents mining there are different approaches to mine the data

a) *Unstructured Text Data Mining:* most of the web content that is available for us is unstructured text data. It can be mined by either data mining or text mining techniques. Some of the text mining techniques is
i) Information extraction topic tracking
ii) Summarization
iii) Clustering
b) *Structured Data Mining:* Structured data are much easier to extract compared to unstructured text. The techniques available for mining unstructured data are
i) Web Crawler
ii) Wrapper Generation
iii) Page Content mining
c) *Semi- Structured Data Mining:* HTML is one of the big example for the special case for such intra-document structure. The techniques which is available for mining semi-structured data are
i) Object Exchange Model
ii) Top Down Extraction
iii) Web Data Extraction language

d) Multimedia Data Mining: The technique available for mining multimedia data is

i) SKICAT

ii) Colour Histogram Matching

a. Traditionally page ranking is used for finding the relevance of the web pages. In this paper, we are proposing content based page ranking, which show much better results. The entire page ranking algorithm can be written in 3 steps

b. Query Interface: This is graphical interface, where the user can enter the input and the input can be searched and the data's can be extracted

c. Search Technique: There can many techniques which can be used for searching

d. Page Ranking: After extracting the web pages through the search techniques. Then web pages are being ranked on the bases of page ranking algorithm.

In this paper, we are doing content based page ranking technique, where the user will be asked to type a keyword, on the basis of the keyword 'N' web pages will be extracted. By using some Natural language processing techniques and similarity and frequency estimating techniques, Score will be generated giving content based page ranking.

II. METHODOLOGY

A. Proposed System

Based on the user's query, search engine will provide us with important/relevant information. Each results are then individually analysed, based on the keyword (user query) and the contents related to that. When the user gives certain keyword, our architecture will extract 'N' number of websites related to that from python library 'gsearch.googlesearch' and save them in CSV file for the future reference and the following process is been carried out

B. Pre-Processing

The aim of data cleaning process is to remove all the unwanted contents from the extracted websites. The extracted data (Web pages) will be completely unstructured and inconsistent data. So, after retrieving the results (web pages) for the user query, the pre-processing is done by using information retrieval techniques. The aim of doing this is to generate relevant results for the search query, which can be achieved by stop word removal

Unwanted Content	Action
Punctuation(!?,.:;)	Removed
Emoticons	Removed
Uppercase characters	Lowercase all content

Table 1: Pre-processing steps

Table 1, are the pre-processing steps, such as if unwanted contents appears –action required, if Punctuations is seen- it is being removed, if Emoticons is seen- it is being removed, all the uppercase characters is being converted to lowercase character because to avoid the change in their meaning.

C. Stop Word Elimination

it is very much necessary to remove the stop words from the data. Stop words are filtered either before or after processing of the data, here the data comes from the website, that is been extracted by using the input keyword search. The reason for removing of stop words, it does not carry any meaning to the content, so therefore, eliminating the stop words will give a better result. For example, the words such as on, and, the, and, in, among, others do not provide any kind information's to the contents and this are called as stop words. After pre-processing phase of the extracted data are considered for further processing. Thus in this phase the stop words, punctuations and symbols are removed.

1) Tokenization: Tokenization is the process of chopping away certain sentences into pieces, called as tokens and at the same time throwing away certain characters, such as punctuations. Here is the example of how tokenization works.

2) Consider The Input: Friends Romans, Countrymen, lend me yours ears;

3) Output: Friends Romans Countrymen lend me yours ears

So, in this output the sentences are broken into different tokens and all punctuations are removed through the process. Thus, from this process it will remove all the unwanted data and provide better accurate results.

III. DATA FLOW DIAGRAM

So, this is the complete data flow diagram of our proposed architecture:

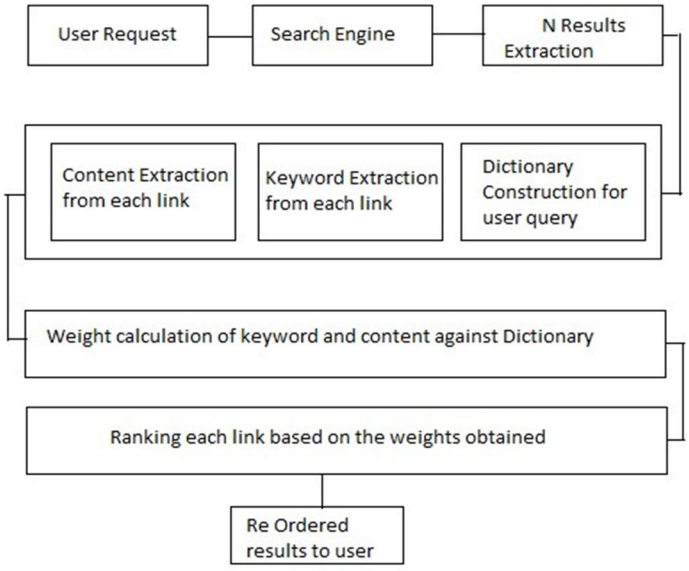


Table 2: Design and Algorithm

From table 2, Design and architecture. It is clear that, first the user will be asked to type a user query. Then after processing in search engine 'N' results (web pages) is been obtained. After extracting 'N' results, the algorithm will extract contents from each link/page, then it will extract keywords from each link/page. Then our algorithm will create a dictionary for storing all keywords related to the user query. Consider, the user has type a user query, 'N' websites will be extracted related to it, then pre-processing is done such as tokenization, removal of stop words and punctuations. After that is been stored (tokens) into a dictionary. Each corresponding website will have their own dictionary and finally a master dictionary will be created and all the words related to all the websites that we have taken will be copied to the major dictionary. And weight calculation of the keyword and the content against the dictionary is done through this formula

Score = (Summation of word- weight)/ (Total number of unique words)

Then ranking of each link is been done on the bases of the score obtained by the links that is been extracted. Then the link highest score will be ranked first and rest of the link will be ordered in descending order.

IV. RESULTS

Experiments were carried out by using different keywords for the user query. the first keyword which is used is {Laptop} against a specific search engine. Then top 5 web pages after typing the keyword is taken as input data set and is listed in table 3.

S No	ID	URL
1	SR1	https://en.wikipedia.org/wiki/Laptop
2	SR2	https://www.amazon.in/Laptops/ b?ie=UTF8node=1375424031
3	SR3	https://www.flipkart.com/laptops-store
4	SR4	https://www.gadgetsnow.com/laptops
5	SR5	https://www.snapdeal.com/Computers-Peripherals

Table 4: URLS(First)

After doing all the pre-processing steps like tokenization, removal of stop words, punctuations. Word-Weights is being calculated, then by using the formula for score. The score is being calculated and web pages is listed in the descending order of page ranking.

ID	SCORE	RANK
1	0.0222	SR2
2	0.0166	SR3
3	0.0081	SR1
4	0.0077	SR4
5	0.0079	SR5

Table 5: Score and Ranks

From the above results, it is been clear that SR2 have the highest score which is "https://www.amazon.in/Laptops/b?ie=UTF8node=1375424031 " and SR5 has the least score "https://www.snapdeal.com/Computers-Peripherals" .The amazon site consists highest amount of contents that is related to the keyword {Laptop} and the least amount of contents related to the keyword{Laptop} is in snap deal site. Thus we were able to get content based page ranking using keyword frequency

Now the experiment is carried with keyword {Air} against specific search engine. top 5 web pages related to that is taken as the input data set and is been listed in the table 6

S No	ID	URL
1	SR1	https://allindiaradio.gov.in/
2	SR2	https://www.newsonair.com/
3	SR3	https://www.airindia.in/
4	SR4	https://www.airindia.in/online-booking-faqs.html

Table 6: URLS(Second)

After doing all the pre-processing steps like tokenization, removal of stop words, punctuations. Word-Weights is being calculated, then by using the formula for score. The score is being calculated and web pages is listed in the descending order of page ranking.

RANK	SCORE	ID
1	0.090	SR3
2	0.045	SR2
3	0.027	SR1
4	0.018	SR4

Table 7 : Scores and Ranks

From the table it is clear that SR3 as the highest rank "https://www.airindia.in/" and SR4 has the lowest rank "https://www.airindia.in/online-booking-faqs.html" Thus air India website consists of highest amount of contents related to the keyword that user query and air India online booking consists of least amount of contents related to the user input{air} thus content based page ranking is being done.

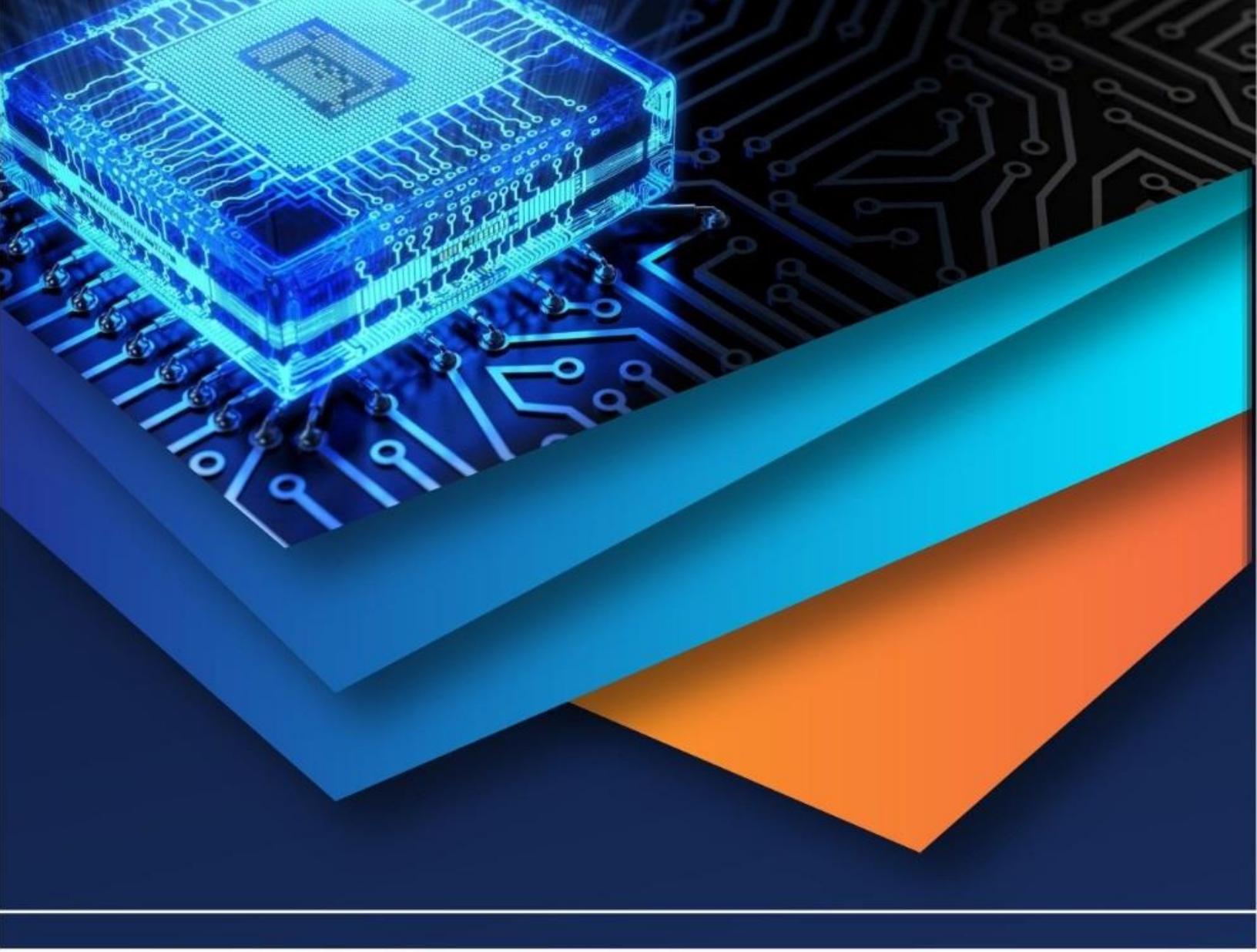
V. CONCLUSION

So, Content based page ranking algorithm is a type of algorithm in which the ranking of the web pages is done based on content weight, With the help of content based page ranking algorithm, we will get most relevant web pages related to the user query, thus allowing the user not to get misguided to some other web pages, where they lack in information. In this paper, we have used some basic NLP techniques such as tokenization, stop words removal, clean data and finding weights for each keywords based on frequency of occurrence of the keyword for achieving content based page ranking.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue I, Jan 2019- Available at www.ijraset.com

REFERENCE

[1] Jayendra Singh Chouhan, Anand Gadwal," Improving Web search Query Relevance using Content Based Page Rank", IEEE International Conference on computer,communication and Control, 2015
[2] Mladen Stanojevic, Sanja Vranes, " A Natural Language Processing for Semantic Web Services", EUROCON 2005.
[3] Anurag Kumar, Ravi kumar Singh, "A Study on Web Content Mining", International Journal of Engineering and Computer Science, volume-6, Issue-1, 2017, pp-20003-20006



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY