ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Quality based Web information extraction approach using NLP and Text Mining

Ashish Kumar Ray[1]**, Mr. Ajay Kushwaha**[2]

M.Tech. scholar, Department of CSE, RCET,Bhilai, India [1]

Associate Professor, Department of CSE, RCET,Bhilai, India[2]

Abstract: Today for discovering information regarding any matter people give preference to web resources rather than other resource. But World Wide Web contains billions of web pages and various other documents containing information about various topics. So we use a search tool for retrieving information about any particular topic. And performance of these search tools is mostly described by the quality of search result. For improving quality of search result search strategies uses various techniques. Despite of these techniques the quality of search results is degraded due to uncertainty in intending context of fired query. Another problem of search result is redundant pages that contain contents that are already contained by previous pages; degrade the quality of search result. In this paper we proposed an approach for improving quality of search results with the help of Natural language Processing and Text Mining. Here quality of search results is described by relevancy precision of search results return by search strategy. The relevancy precision can be defined by ratio of relevant articles to total retrieved articles.

Keywords: WWW, NLP, WSD, SYNSETs, SERPs

I. INTRODUCTION

Now a day's web has become the most popular source in terms of availability of content related to almost every field of life. Now web represents every field of society so its size increases very rapidly to huge amount. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. So for finding information regarding any topic we need a tool or strategies that take input of queries and returns list of urls as result. While searching for context search tool returns all the results related to the query. Spotting the relevant result is most monotonous task for a user. Various search tools uses different methods to provide high quality information; quality in terms of relevancy of search result to query.

II. RELATED WORK

Search strategies are continuously improving their techniques and approach to provide more relevant, accurate information for a given query. They are working on different segments of search engine to improve the search results. Enhancing quality by improving crawling: Early search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day. Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day. There is a series of modification in crawling mechanism to improve the quality of web information extraction:

Keywords oriented strategy: It start crawling pages based on keywords and then proceed with static list of URLs of web pages that contain keywords of query. Crawling with candidate URL structure: It uses contents of web pages pointing to current page, candidate URL structure, and other behaviours of siblings web pages in order to estimate the probability that a candidate is beneficial for a given crawl [4].

Distributed crawling with static assignment: There is a set of fixed rules that defines how new URLs will be assigned to the crawlers. For static assignment, a hashing function can be used to transform URLs into a number that corresponds to the index of the corresponding crawling process. Distributed crawling with dynamic assignment: In this approach a server works as a scheduler that assign URLs to crawlers at run time. Improvement on indexing segment: Different search engine uses different indexing mechanism to provide more relevant information in less amount of time. Different index mechanism is evolved in order to support search engine architectures to provide more relevant information. Types of indices include:

Suffix tree: It supports linear time lookup. It built by storing the suffixes of words. The suffix tree is a type of trie. Tries support extendable hashing, which is important for search engine indexing.

Inverted index: It stores a list of occurrences of each atomic search criterion as hash table or binary tree.

Citation index: It stores citations or hyperlinks between documents to support citation analysis, a subject of Bibliometrics.

Ngram index: Stores sequences of length of data to support other types of retrieval or text mining.

Document-term matrix: Used in latent semantic analysis, stores the occurrences of words in documents in a two-dimensional sparse matrix.

Using temporal dimension to improve the quality of search: The Web is not a static environment. It changes constantly. Quality pages in the past may not be quality pages now or in the future. Previous techniques favour older pages because these pages have many in-links accumulated over time. New pages, which may be of high quality, have few or no in-links and are left behind. Bringing new and quality pages to users is important because most users want the latest information [13].

III. PROBLEMS AFFECTING QUALITY OF SEARCH

Although search engines are continuously modifying and adding new techniques to improve the quality of search but still search results are suffering with some problem. Some of them are given below:

Immaterial Results: Some times when we search for a particular topic but we get results which are completely different from searched query. This problem exists because most words in natural languages are polysemy that is they have multiple possible meanings or senses. Redundant Content: Another problem associated with searching web is that SERPs sometimes contains links to redundant web pages which provides duplicate content. The presence of near duplicate web pages plays an important role in this performance degradation while integrating data from diverse sources. Web mining faces huge problems due to the existence of such documents.

IV. APPROACHES THAT CAN RESOLVE PROBLEMS STATED IN SECTION III

Natural language Processing: Natural language processing (NLP) is a field of computer science that handles the interactions between computers and human (natural) languages. It allows computers to derive meaning from human or natural language input.

Word Sense Disambiguation: Word Sense Disambiguation (WSD) is a branch of NLP which is the task of selecting the most appropriate meaning for a polysemy word, based on the context in which it occurs. Word sense disambiguation (WSD) can be used to overcome the problem of immaterial results. All natural languages contain words that can mean different things in different contexts. In English, for example, the word bark can refer to the sound made by a dog or the covering of trees. Such words with multiple meanings are potentially ambiguous, and the process of deciding which of their several meanings is intended in a given context is known as Word Sense Disambiguation (WSD) [2].

Word Sense Disambiguation (WSD) is the process of choosing the most suitable meaning for a polysemy word, based on the context in which it occurs. For example, in the phrase the bank down the street was robbed, the word bank indicate a economic organization, while in The city is on the Western bank of Jordan, this word indicates to the shoreline of a river . WSD is an interior process in the natural language processing (NLP) chain. It is used in many applications such as text similarity check machine translation and information retrieval [3]. An approach that makes use of word sense disambiguation can achieve more accuracy in retrieving information for a given query.

Text Mining: Text mining also called as text data mining which is the process of deriving high-quality information from text. It analyses the text at character level and use mathematical expression to find the similarity of two pages.

Text mining is a combination of the process of structuring the input text (usually parsing, and addition of some derived linguistic features and the elimination of others, and succeeding insertion into a database), deriving patterns within the structured data, and finally assessment and understanding of the output [9]. Text mining can be used to overcome the problem of redundant content and by giving importance to web content mining the searching process can be improved [5] .And it provides relevant information by eliminating the redundant and irrelevant contents [6]. Removal of duplicate pages from search result will enhance the performance of strategy.

V. PROPOSED METHODOLOGY

The proposed work describes an approach for web information retrieval that gives high quality result in terms of relevancy. This approach integrates multiple mechanisms to improve the quality of search engine result pages. First, it resolve the ambiguity in given search query that may occur due to polysemus words .This mechanism restrict the unwanted web pages from listing in search engine result pages. Second, it removes pages that have redundant contents from search engine result pages. Third, after performing these two levels of optimization it applies ranking technique of web pages. By this approach of search we will get search results of higher quality in terms of relevancy.

This approach integrates following mechanism to improve the quality of information:

1. Removing context ambiguity of query before start search process.

2. Eliminated pages that contains duplicate content while crawling web pages

3. With above two mechanisms we use Pagerank implementation for ranking of web pages.

Steps followed by proposed system while processing a query:

1. First of when a user enters a query, system read the query as it is without any modification.
2. After reading query system applies process of tokenization in which it extracts all the words present in the query.
3. After the process of tokenization system extracts keywords of query that constitute the meaning of query and rest of unnecessary words are left here.
4. After the process of keywords extraction system tries to identify the intending meaning of query. For this purpose we are using one branch of Natural Language Processing (NLP), called Word Sense Disambiguation (WSD).For removing uncertainty in the sense of query system uses a variant of Lesk approach that make use of online SYNSETs, which are sets of cognitive synonyms ,

Architecture of Proposed method:

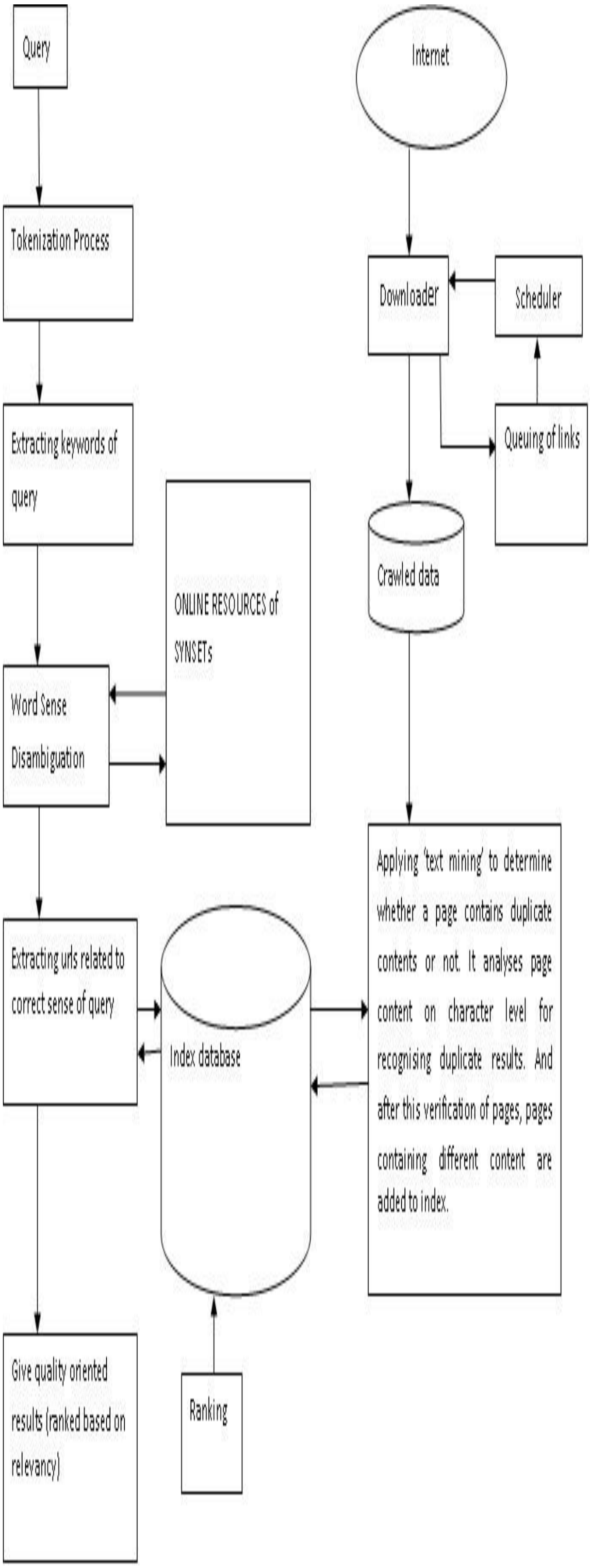

Query
Internet
Tokenization Process
Downloader
Scheduler
Extracting keywords of query
Queuing of links
ONLINE RESOURCES of SYNSETs
Crawled data
Word Sense Disambiguation
Applying 'text mining' to determine whether a page contains duplicate contents or not. It analyses page content on character level for recognising duplicate results. And after this verification of pages, pages containing different content are added to index.
Extracting urls related to correct sense of query
Index database
Give quality oriented results (ranked based on relevancy)
Ranking

Fig 5.1: Architecture of proposed system

5. each conveying a different conception. Synsets are connected by means of semantic relationships. The basic idea of Lesk algorithm is words in a given neighbourhood suppose to share a common subject. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighbourhood [9]. Here Lesk approach makes use of online SYNSETs, which are sets of cognitive synonyms, each conveying a different conception.

6. After obtaining intending meaning of query system communicate with index database to give results for fired query where pages are ranked base on Pagerank algorithm.

7. After getting search results system return the search result to the user.

Steps followed by proposed system while crawling down the pages from World Wide Web (WWW):

1. First the crawling system take the seed url to start the process.
2. The seed url feeds to downloader which crawls the content of the document existing at that url. After crawling contents of the document it recognises all the urls presented on that document. All the urls are transferred to scheduler and remaining contents are transferred to crawled database [4][8].
3. The scheduler after getting list of urls, checks whether these urls are crawled previously or not. If a url is not crawled previously then this url is scheduled for crawling otherwise this url is eliminated.
4. And above two steps are repeated till the stoping criteria.
5. From the crawled databases pages are retrieved to verify whether a particular page is containing different content or redundant content. This is done by application of Text mining that analyses the text at character level and use mathematical expression to find the similarity of two pages. And after this verification of pages, pages containing different content are added to index.
6. For ranking of web pages we use the Pagerank algorithm which can be described as

$$PR(A)= (1-d)+d(PR(T1)/C(T1)+.....+PR(Tn/C(Tn))$$

Here

PR (Ti): is the PageRank of the Pages Ti which links to page A,

C (Ti): is number of outlinks on page Ti and

d: is damping factor. It is used to stop other pages having too much influence. The total vote is “damped down” by multiplying it to 0.85 [1] [7].

VI. RESULTS AND DISCUSSION

In this paper we present an approach for web data retrieval that enhances the relevancy measure of search results especially for queries that contains words of multiple meaning. For such query the presented approach gives better results that have more relevant results than other approach. Here we present a comparative analysis of the presented approach and Google results.

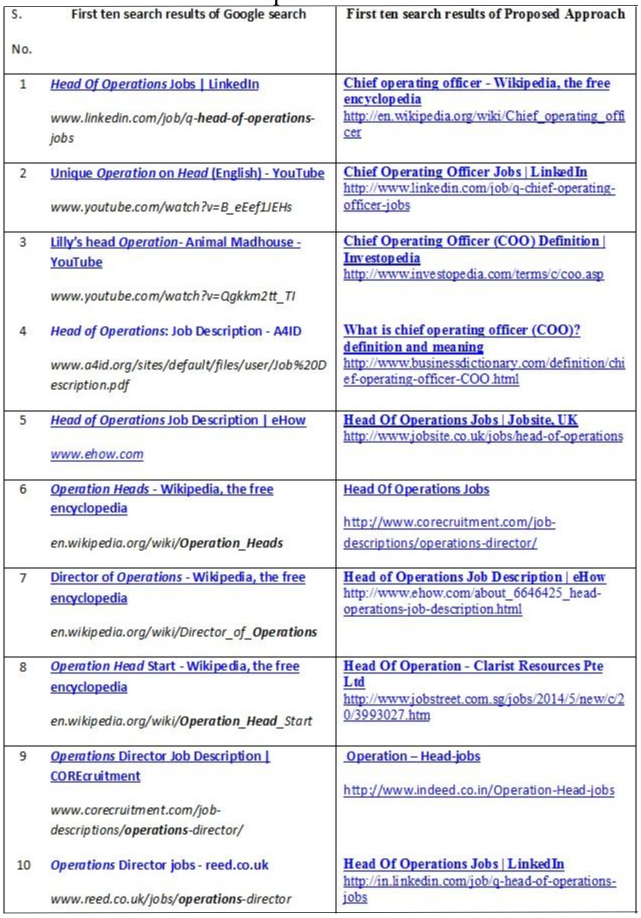
For this comparative analysis we consider a collection of queries that contains the words of multiple meaning .We will make an analysis of results of both strategies (Google and proposed method) to evaluate performance measures. The collection of such queries is given below:

Table 6.1: Collection of queries

Query Id	Query Text
A	"head of operation"
B	"huge bank of earth"
C	"mole structure into water"
D	"cause of march"
E	"interest in doing business"

For this comparative analysis we will take only first ten ranked web pages from both strategies. On the basis of these observations we will calculate performance measures. For one of the queries, query- A (head of operation) search results of both strategies are listed in following table:

TABLE 6.2: first ten ranked results for QUERY "head of operation"



S. No.	First ten search results of Google search	First ten search results of Proposed Approach
1	*Head Of Operations* Jobs | LinkedIn www.linkedin.com/job/q-**head-of-operations**-jobs	Chief operating officer - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Chief_operating_officer
2	Unique *Operation* on *Head* (English) - YouTube www.youtube.com/watch?v=B_eEef1JEHs	Chief Operating Officer Jobs | LinkedIn http://www.linkedin.com/job/q-chief-operating-officer-jobs
3	Lilly's head *Operation*- Animal Madhouse - YouTube www.youtube.com/watch?v=Qgkkm2tt_TI	Chief Operating Officer (COO) Definition | Investopedia http://www.investopedia.com/terms/c/coo.asp
4	*Head of Operations*: Job Description - A4ID www.a4id.org/sites/default/files/user/Job%20Description.pdf	What is chief operating officer (COO)? definition and meaning http://www.businessdictionary.com/definition/chief-operating-officer-COO.html
5	*Head of Operations* Job Description | eHow www.ehow.com	Head Of Operations Jobs | Jobsite, UK http://www.jobsite.co.uk/jobs/head-of-operations
6	*Operation Heads* - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/**Operation_Heads**	Head Of Operations Jobs http://www.corecruitment.com/job-descriptions/operations-director/
7	Director of *Operations* - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Director_of_**Operations**	Head of Operations Job Description | eHow http://www.ehow.com/about_6646425_head-operations-job-description.html
8	*Operation Head* Start - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/**Operation_Head**_Start	Head Of Operation - Clarist Resources Pte Ltd http://www.jobstreet.com.sg/jobs/2014/5/new/c/20/3993027.htm
9	*Operations* Director Job Description | COREcruitment www.corecruitment.com/job-descriptions/**operations**-director/	Operation – Head-jobs http://www.indeed.co.in/Operation-Head-jobs
10	*Operations* Director jobs - reed.co.uk www.reed.co.uk/jobs/**operations**-director	Head Of Operations Jobs | LinkedIn http://in.linkedin.com/job/q-head-of-operations-jobs

Similarly after performing comparative analysis of both method (Google and proposed approach) for all the queries of given collection of queries that contain words of multiple meaning ,we obtain following table that show how many results from first ten results are relevant to query for both the approaches:

TABLE 6.3: Relevant results from first ten ranked pages

Query Id	Number of relevant search result among first ten search results for a given query in Google search	Number of relevant search result among first ten search results for a given query in proposed approach
A	8	10
B	2	9
C	0	8
D	7	9
E	9	10

Based on data provided by the above illustrated table we can make a comparative analysis and evaluate the performance of both search strategies. For this comparative analysis we can draw a bar graph of existing scenario. This graphical representation facilitates us to evaluate the performance measure of both approach (Google and presented approach). The graphical representation is given below:

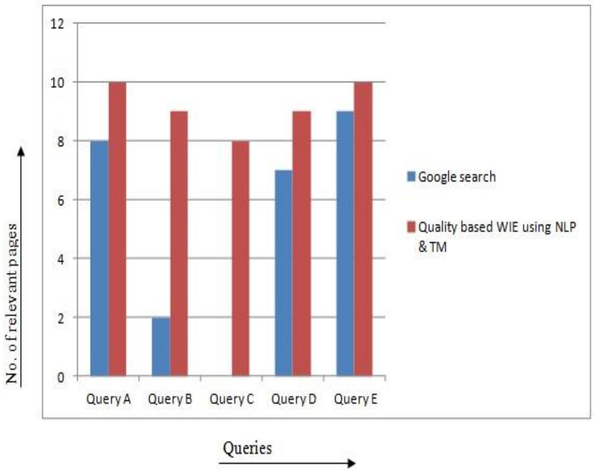
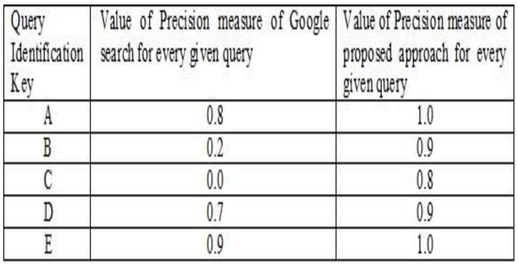


Fig 6.1: Relevant results from first ten ranked pages

Now on the basis of data provided by the above bar chart and table we can calculate precision measure of both search strategies and compare performance of both strategies. The table containing values of precision measures is given below :

TABLE 6.4: Precision Measures for first ten results



Query Identification Key	Value of Precision measure of Google search for every given query	Value of Precision measure of proposed approach for every given query
A	0.8	1.0
B	0.2	0.9
C	0.0	0.8
D	0.7	0.9
E	0.9	1.0

For this comparative analysis we can draw a bar graph of precision values of both strategies for given queries. This graphical representation facilitates us to evaluate the performance measure of both approach (Google and presented approach). Bar graph of precision measure is given below:

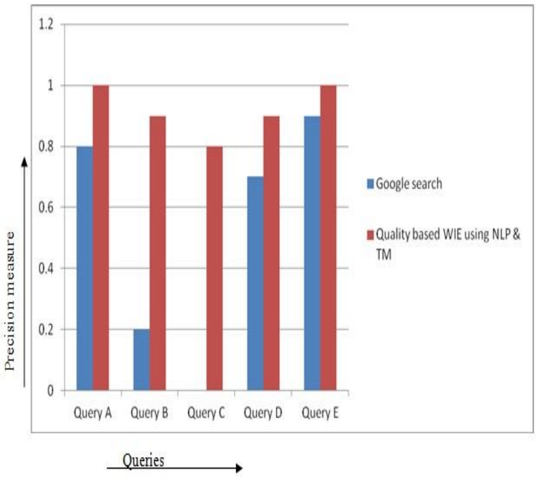


Fig 6.2: Graph of Precision Measures for first ten results

The precision measures of search strategies help us in analysing the quality of search results. In the above section tabular and graphical representation provides a clear view of the difference between precision measures of proposed approach and other strategy which indicate that the quality of search results greatly affected by existence of polysemy words in query and redundant contents.

VII. CONCLUSION

By observing the comparative analysis of proposed approach with other search strategy we can say that the existence of polysemy words in query which cause uncertainty in intending context of fired query and existence of pages with redundant content degrade the quality of search engine result pages (SERPs) to a great extent. Strategy that resolves these problems improve the quality of search results.

ACKNOWLEDGMENT

We are very thankful to respected Mr. Santosh Rungta, Chairman, RCET, Bhilai, Dr. S. M. Prasanna Kumar, Director, RCET, Bhilai, Mr. Toran Verma, M.Tech.(CSE) coordinator RCET, Bhilai, India, for providing excellent computation facilities in the University campus and for giving their moral support and help to carry out this research work.

REFERENCES

[1] Sandhya,Mala Chaturvedi, A survey on Web mining Algorithms, IJES vol.-2 2013

[2] Miguel Angel Rios Gaona, Alexander Gelbukh, Sivaji Bandyopadhyay ,Web-based Variant of the Lesk Approach to Word Sense Disambiguation, IEEE 2009.

[3] Kenneth W. Church and Lisa F. Rau,Commercial Application of Natural Language Processing, ACM 0002- 0782/95/1100

[4] Charu C. Aggarwal ,Fatima Al-Garawi, Philip S. Yu], Intelligent Crawling on the World Wide Web with Arbitrary Predicates, WWW10 May 1-5, 2001, Hong Kong. ACM 1-58113-348-0/01/0005.

[5] P. Ravi Kumar and Ashutosh Kumar Singh, Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of Applied Sciences 7 (6): 840-845, 2010 ISSN 1546-9239 ©2010Science Publications.

[6] Raymond Kosala,Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Explorations. 2000 ACM SIGKDD, July 2000.

[7] Preeti Chopra, Md. Ataullah, A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms, IJEAT ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

[8] Christopher Olston and Marc Najork, Web Crawling, Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010).

[9] http://www.wikipedia.org/.

[10] Philip S. Yu, Xin Li, Bing Liu, On the Temporal Dimension of Search, WWW 2004, May 17-22, 2004, New York, NY USA.