Global Journal of Computer Science and Technology: B
Cloud and Distributed
Volume 18 Issue 1 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 | Print ISSN: 0975-4350 | DOI: 10.17406

Implementing a Search Engine for Bangladeshi E-Commerce Product

By Md. Mijanur Rahman, Razia Sultana Rupa & Hasan Mahmud Tuhin
Jatiya Kabi Kazi Nazrul Islam University

Abstract- This project is concern with the practical implementation of Information Retrieval, where the main focus is on the algorithmic challenges in efficiently representing large data sets while supporting fast searches in the Web. The application was developed by using a system approach where analysis, design and development were carried out by the incremental model. The main aim of this work to introduce an efficient Information Retrieval System for Bangladeshi Ecommerce Product Search Engine. This search engine base on the technology of public search engine and is built specific for the structure of Bangladesh E-commerce Search Engine. The system provides the relevant searching results from Bangladeshi web domains. The proposed system has been designed and developed using Python programing language tools and methods.

Keywords: *e-search engine, indexing process, information retrieval, query process, web crawler.*

GJCST-B Classification: *K.4.4 , J.1*

Strictly as per the compliance and regulations of:

Implementing a Search Engine for Bangladeshi E-Commerce Product

Md. Mijanur Rahman [α], Razia Sultana Rupa [σ] & Hasan Mahmud Tuhin [ρ]

Abstract- This project is concern with the practical implementation of Information Retrieval, where the main focus is on the algorithmic challenges in efficiently representing large data sets while supporting fast searches in the Web. The application was developed by using a system approach where analysis, design and development were carried out by the incremental model. The main aim of this work to introduce an efficient Information Retrieval System for Bangladeshi E-commerce Product Search Engine. This search engine base on the technology of public search engine and is built specific for the structure of Bangladesh E-commerce Search Engine. The system provides the relevant searching results from Bangladeshi web domains. The proposed system has been designed and developed using Python programing language tools and methods.

Keywords: e-search engine, indexing process, information retrieval, query process, web crawler.

I. Introduction

Generally, people use search engine for one of three things, such as, research, shopping and entertainment [1]. For example, Amazon, Yahoo! Shopping, Google product search etc. handles all of the purchasing, shipping and ordering info. People also look up things like videos, movies, games, social networking sites etc. There are many E-commerce site in Bangladesh, they are increasing by size and number day by day [2]. There is no special search engine to search for product all over these sites. Though the size of information rapidly increasing in this field, so we need an efficient Information Retrieval System to make better use. So, a new Information Retrieval System named as "Bangladesh E-commerce Product Search Engine" is introduced in this project. The proposed system can help to find best relevant data from those sites within a seconds.

II. E-search Engine

The search engine is a software system [3] that compares queries to documents and produces ranked result lists of the documents. Search engine must be able to capture, or crawl, many terabytes of data, and then provide sub second response times to millions of queries submitted every day. This software is a program that search documents for specified keywords given through the E-commerce sites [4] and returns a list of the documents where the keywords were found. Typically, Web search engines work by sending out a *spider* to fetch as many documents as possible [5]. Another program, called an *indexer [6]* then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each *query.* The search engine activities and the core information retrieval issues are shown in Figure-1.

The architecture of e-Search Engine is designed to ensure that a system will satisfy the application requirements of search engine are effectiveness (quality), means it able to (i) effectiveness (quality), that is we want to able to retrieve the most relevant set of documents possible for a query; and (ii) efficiency (speed), that is, we want to process queries from users as quickly as possible [7].

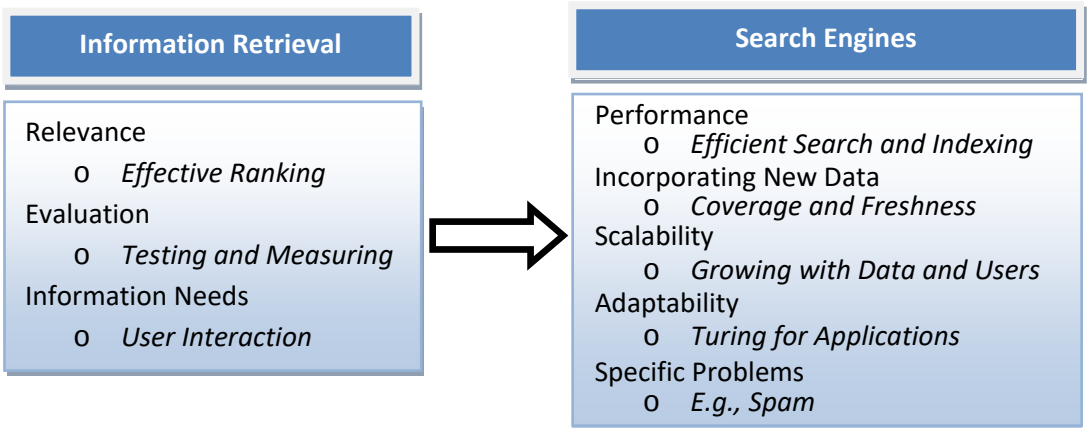

Figure1: Search engine design and the core information retrieval issues [8].

Author α: PhD. Dept. of Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Bangladesh. e-mail: mijanjkkniu@gmail.com
Author σ ρ: Dept. of Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Bangladesh.

III. Functions of E-search Engine

Search engine components support two major functions, such as, the indexing process and the query process [8].The indexing process builds the structures that enable searching. The query process uses those structures and a person's query to produce a ranked list of documents.

a) Indexing Process

Indexing process involves on acquiring data from various sources. Index processing includes several major components, such as, text acquisition, text transformation and index creation, as shown in Figure-2 [8].

The *text acquisition* component is used to identify and make available the documents that will be searched. Text acquisition will more often require building a collection by crawling or scanning the sources of information. Then the documents to the next component *data store*, which contains the text and metadata for all the documents. The *text transformation* components transforms documents into index terms or features and then the documents are stored in *index* and used for searching. Also, the *index creation* component takes the output of the text transformation component and creates the indexes or data structures that enable fast searching.

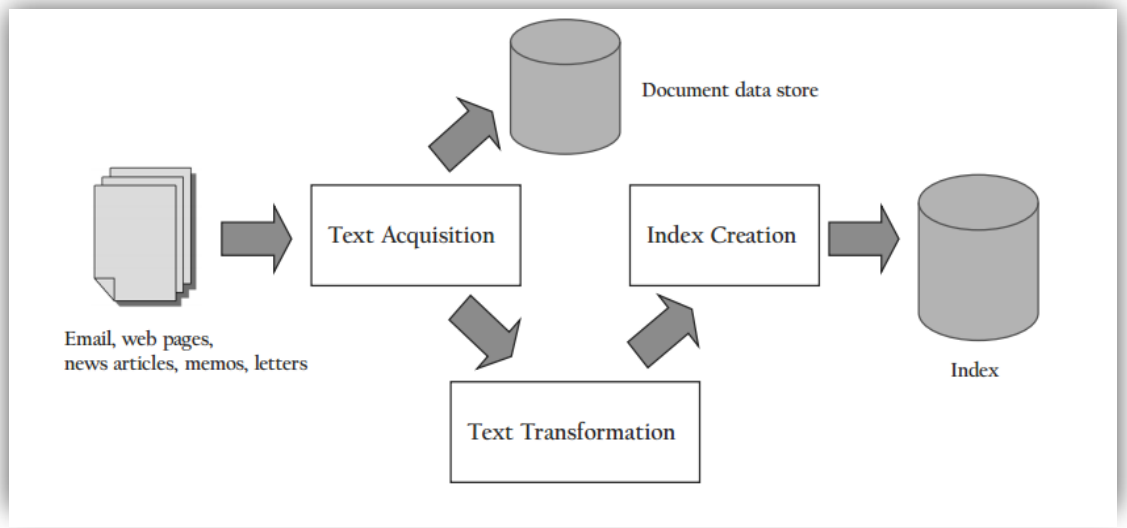

Figure 2: The indexing process

b) Query Process

This part process user query and find document using index and rank the document for the efficiency of search engine. The proposed search engine involves BM25 ranking algorithm [9] for calculating the rank of a document, shown in Figure-3.The major components of the query processing are user interaction, ranking and evaluation, as shown in Figure-4 [8].

The *user interaction* component provides the interface between the person doing the searching and the search engine. The main task is to accept user's query and transform it into index terms and then to take the ranked list of documents from the search engine and organize it into the results shown to the user. The *ranking component* takes the transformed query from the user interaction component and generates a ranked list of documents using scores based or a retrieval model [9].The task of *evaluation component* is to measure and monitor effectiveness and efficiency. An important part of that is to record and analyze user behavior using log data. Evaluation is primarily an offline activity, but it is a critical part of any search application.

Rank Calculation

Given a query Q containing keywords $(q1, q2, q3, \ldots, qn)$.

The BM25 score of a document is given by

$$score(D,Q) = \sum_{i=1}^{n} IDF(qi) \times \frac{f(qi,D).(k1+1)}{f(qi,D) + k1\left(1 - b + \frac{b|D|}{avdl}\right)}$$

Where,

$f(qi,D) = is\ qi\ \ is\ term\ frequency$

$|D|\ is\ the\ length\ of\ the\ document$

$K_1\ and\ b\ are\ free\ parameters\ chosen\ in\ absence\ of\ an\ advanced\ optimization$

$IDF(qi)\ is\ given\ by$

$$IDF(qi) = \log \frac{N - n(qi) + 0.5}{n(qi + 0.5)}$$

Where,

$N = total\ number\ of\ documents\ in\ the\ collections$

$n(qi) = number\ of\ documents\ containing\ qi$

Figure 3: Calculation of rank of a document using BM25 ranking algorithm

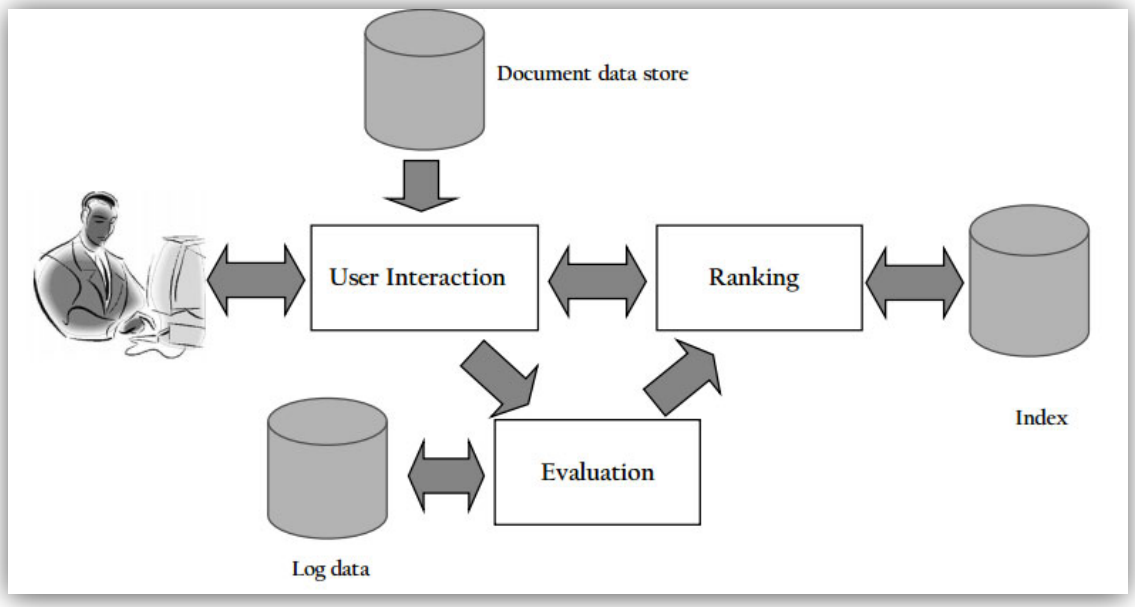

Figure 4: The query process

IV. The Web and Web Crawler

Web crawler, database and the search interface are the major component [10] of a search engine that actually makes search engine to work in the Web. The structure of the Web is shown in Figure-5 which include the crawler, the parser and the downloader [11].Search engines are programs that search documents for specified keywords and returns a list of the document. The web crawler program is to search for the information in the database. Once web crawler finds the pages, the search engine then shows the relevant web pages as a result. A Web crawler is an Internet bot which systematically browses the WWW [12]. A web crawler is designed to follow the links on web pages to discover and download new pages. Web crawlers are also known as ants, automatic indexers, bots, spiders, web robots, and worms. The crawler is adapted to discover and update all documents related to a company's operation.

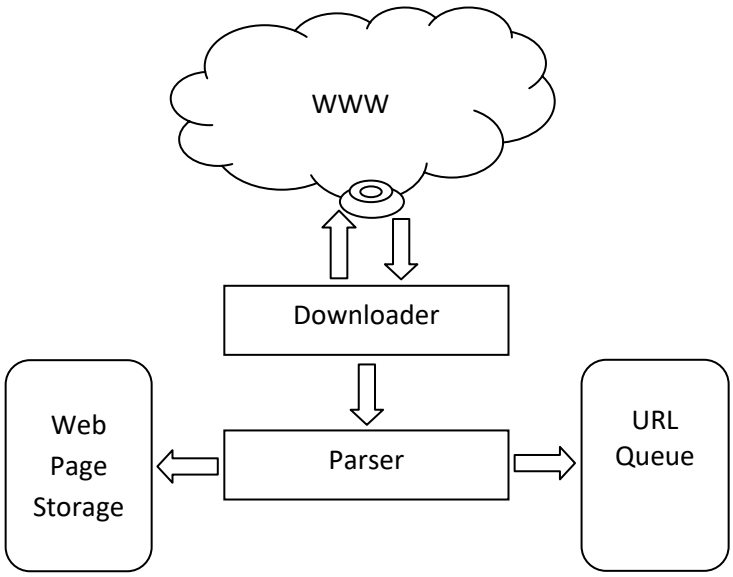

Figure 5: Structure of the web for search engine

V. Proposed System Design and Implementation

The design methodology of the proposed system is described by the data flow diagram, as shown in Figure-6.To develop the proposed system, several methods have been implemented in Python programs, such as, Downloader, Parser, Crawl, Tokenize, Build_Index, Query_Transformation, Scoring_Algorithm, etc.

The Downloader(url,retry) method is used to download resource from given URL. The Parser (HTML, ALLOWED_DOMAIN<DEPTH) method parses resource returned by download. The Crawl(allowed_domain, depth) method uses other two methods Parser and Downloader for crawling. The Tokenize (data) method performs text transformation and makes them ready to build index. The Build_index(data)method builds inverted index from transformed data. The Query_transformation(query) method is used to transform user query into tokens so that we can use them with inverted index. The scoring_algorithm(query) transforms user query so that to make them compatible to indexed data. Also a user friendly interface (see Figure-7) has been developed to interact with the proposed that the users can easily access the system.

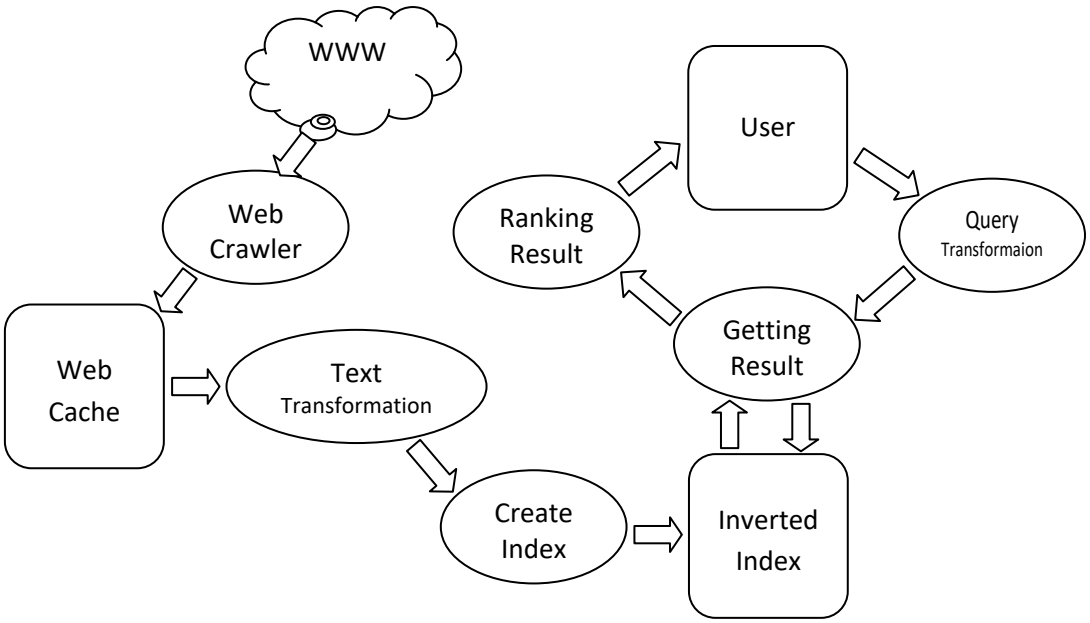

Figure 6: Data flow diagram

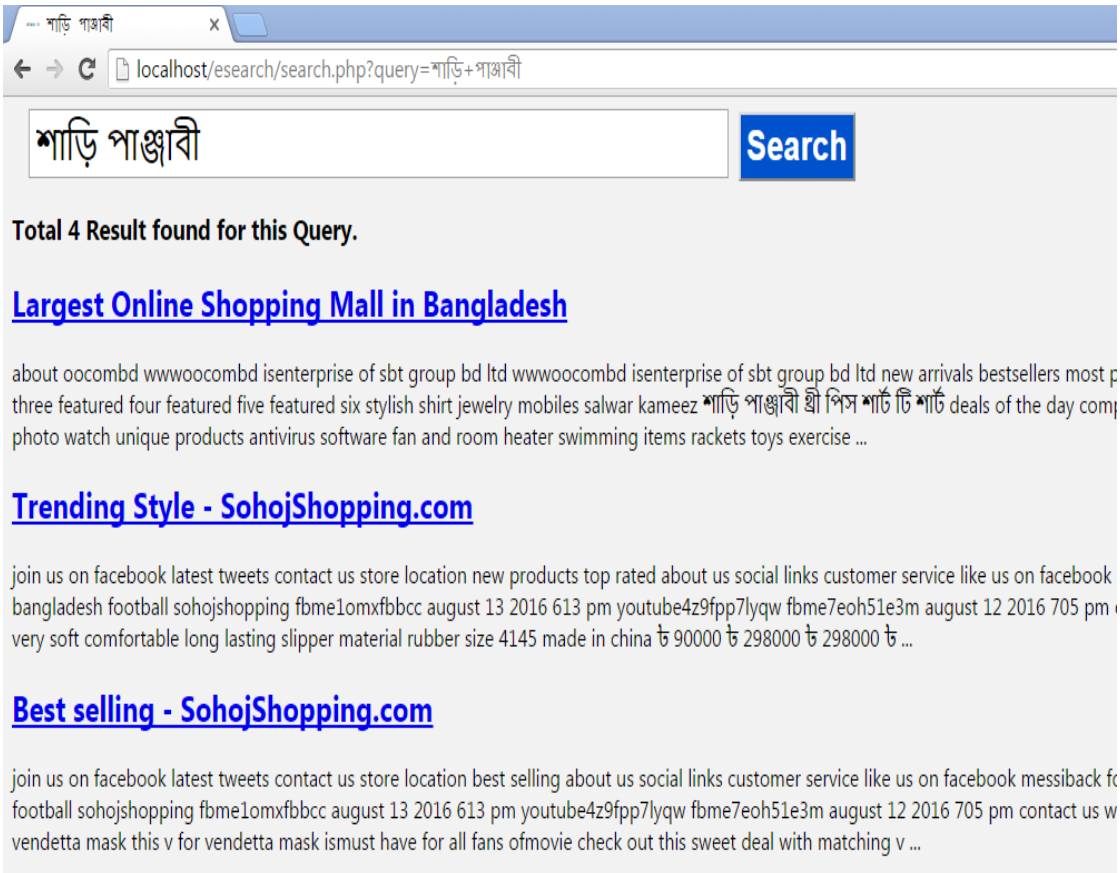

Figure 7: User Interface for e-Search Engine.

VI. Results and Conclusion

Total 2228 documents are indexed and there has 12858 Bangla keywords used in the proposed system. The indexed documents are retrieved from following Bangladeshi domains: rang-bd.com (2), rokomari.com (790), aarong.com (278), oo.com.bd (2), sohojshopping.com(215),priyoshop.com(2),daraz.com.bd (441), bdhaat.com (398), kaymu.com.bd (2), bagdoom.com(2),hutbazar.com(61),fortunabangladesh.com (46), ajkerdeal.com(2). The outcomes of the proposed system is shown in Figure-7.

The main aim of this work to introduce an efficient Information Retrieval System for Bangladesh E-commerce Product Search Engine, so that the proposed system can help to find best relevant data from those sites within a seconds. Though small amount of document and poor text transformation technique search result are not efficient as we expected. The web crawler could not fetch many information correctly, because E-commerce sites use a lots amount of JavaScript, and to get that data the system need to interact with some JavaScript event. This may be improved by using an intelligent web crawler and better text transformation technique. Also spell checking and making query suggestions may be employed further.

References Références Referencias

1. Search Engine Optimization Article (2018). "Why People Use Search Engines: Research, Shopping and Entertainment". Dummies: A Wiley Brand Publication. http://www.dummies.com/.
2. Sharif Ahmed (August 20, 2017). "How online shopping is changing Bangladesh". The independent Article. Published by Independent Publication Ltd, Bangladesh.
3. JoãoMagalhães. "Information Retrieval: Hands-on guides". Dep. Computer Science, NOVA FCT, Universidade NOVA, Lisboa.
4. SwapnaKodali (2007). "The Design and Implementation of an E-Commerce Site for Online Book Sales". MSc Thesis, Dept. of Computer Science and Information Technology, Indiana University South Bend, May 2007.
5. Amir Manzoor (2017). "E-Commerce 2018". ISBN: 978-969-9443-06-0. Printed in United States of America.
6. Andrew Trotman (2003). "Compressing Inverted Files". Information Retrieval, Vol. 6 (1), January, 2003, pp. 5-105. Kluwer Academic Publishers. Manufactured in The Netherlands.
7. Vipul Narayan1, R.D.S.Yadav, R.K.Mehta, MahendraRai, Musheer Ahmed, Rahul Maurya, AbhishekKanujiya and Dharmpal (2017). "A Novel Approach for Information Retrieval using Web Based Search Engine". International Journal of Current Engineering and Technology, Vol.7, No.3, June 2017.
8. Croft, D. Metzler, T. Strohman (2015). "Search Engines: Information Retrieval in Practice". Published by Pearson Education, Inc.
9. JoydipDatta (April 2010). "Ranking in Information Retrieval". MTech Seminar Report, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay Powai, Mumbai – 400076.
10. Web Article (© Copyright 2018). "Search Engines". Published in Tutorials Points; Simply Easy Learning. Retrieved from https://www.tutorialspoint.com/internet_technologies/search_engines.htm.
11. Khawshik Pal (2016). "*How to create your own search engine with PHP and MySQL*". Retrieved from: http://mrbool.com/how-to-create-your-own-search-engine-with-php-and-mysql/32733.
12. Adi Omari, Sharon Shoham and EranYahav (2016). "Cross-Supervised Synthesis of Web-Crawlers". ICSE'16 Proceedings of the 38th International Conference on Software Engineering, Pages 368-379, Austin, Texas - May 14 - 22, 2016.