Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 35-38

Structuring of Web Pages using XML Framework for Information Filtering

Mu. Annalakshmi[1] and A. Padmapriya[2]
[1]Research Scholar, [2]Associate Professor
[1&2]Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu, India
E-Mail: annalakshmi_mu@yahoo.co.in, mailtopadhu@yahoo.co.in

***Abstract* - Finding the required information in the vast area of web has been increasingly difficult in recent days since the web is overloaded with enormous content in the form of text, images, audio and video. Search engines help in this context to some extent but there are difficulties with them also. This paper proposes a framework in XML for the web pages in results of the search engines which helps in information filtering and search engine optimization.**
Keywords: **Search Engines, Information Filtering, Search Engine Optimization, XML Framework**

I. INTRODUCTION

Though the web contains several billions of pages, the required information has to be found at appropriate time. Search engines have become an integral part of web as they assist the users in finding the relevant information in the web. They are actually software systems that search the entire web for the user's need that are represented in the form of queries and return a list of matching documents. This resulting list is usually called as SERPs (Search Engine Result Pages. Some of the popular search engines in use today are Google, Yahoo, Bing, AOL, Baidu, Excite, DuckDuckGo.

The major processes involved in search engines [5] are
Web crawling is done by specialized software components called spiders or bots. The spider begins with a set of seed URLs. The links from seed set are traversed recursively until no more URLs are reached or a threshold is reached.

After a webpage is crawled, it is stored in the data center of the search engine which is the repository of such pages collected. Then indexing is done on such pages so as to enable easy retrieval of web pages for a search query at minimum cost and time. The web pages that match the given search query are retrieved for the data repository with the help of indexing and then they are ranked according to their relevance to the query. This is done with the help of specialized ranking algorithm used by the search engine.

The Extensible Markup Language (XML) [6]-[7] is a simple text-based format for representing and sharing structured information. It is currently the most used and sophisticated format for data that is distributed in different remote servers. The basic building block of an XML document is an element, defined by tags. Each element in XML has a starting and an ending tag. The outermost element known as the root element contains all the other elements in the XML document. XML also facilitates nested elements, or elements within elements which allows XML to support hierarchical structures. Element names represent the content of the element, and the structure represents the relationship between the elements. XML's platform independent nature simplifies data sharing between various systems because XML data doesn't require any conversion when transferred between different systems. Semantic web languages like RDF and OWL use XML to represent their syntax.

Even though the search engines display the SERPs of the user's query, the user has to browse the web pages in the SERPs to determine whether the page is relevant to their query. This information retrieval in web pages is difficult mainly because of their unstructured or semi-structure nature. So bringing structure to the web pages can be useful in the area of web mining. Databases are structured but they occupy more space when compared with XML. So this paper proposes an XML based framework for structuring the web pages.

II. LITERATURE SURVEY

Information extraction is the task of extracting structured information from unstructured text in an automatic fashion. Information extraction is typically considered as a preliminary step dealing with the preprocessing activity in other text mining applications such as question answering [9], hypothesis generation [10] and summarization.

The ranking algorithm uses several different factors for ranking the web pages which differs from one search engine to another. Same web pages are ranked differently by different search engines because of these factors and the ranking algorithm used. [1] analyzes the ranking factors used by some major search engines. If a page has to be ranked higher by a search engine then the web pages developer has to take into account these factors.

The data from the web pages can be gathered and converted into a tabular format using boundary extractor and pattern generator algorithms [2]. The authors of [3] have introduced an XML tree named as Steiner tree to represent the user preferences in a hierarchical manner. The tree contains

information about the domain, topics in the domain and the keywords in each topic in the Search engine optimization [4] is a technique that includes right strategies and tactics used to increase the amount of visitors to a website by obtaining a high ranking placement in the search results page of a search engine (SERP) of top search engines like Google, Bing. Some of the on page SEO factors are title tag, meta tag, Alt attribute, header tags, URL, hyperlinks, keyword density. Off page SEO factors include the number of back links.

The research work [8] proposes a web page change detection system with a node signature algorithm. The algorithm compares the XML structure formed from HTML source of the web pages to detect the changes in the web page.

III. STRUCTURING OF WEBPAGES USING XML FRAMEWORK

This paper proposes an XML framework for the web pages by examining their HTML source code. For information filtering the user query will be given as input. The list of URLs for the given search query are retrieved from the web sources. These URLs are given as input to this framework. The framework will generate the XML documents by analyzing the URL and its meta data.

The workflow of the framework is depicted in the figure given below.

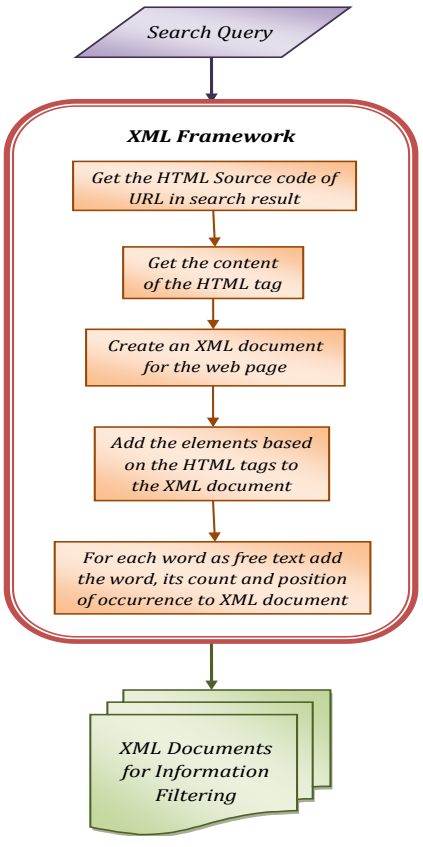


Fig.1 XML Framework for webpage structuring

The analysis about web pages will be done based on the tags in the source code of the URL in the search result. In the proposed framework fields for URL, contents of Title, Meta tag, headings, bold, italic and underline (BIU), image tags are considered for the analysis. The analysis is done on the basis of the occurrence (position) and number of occurrences of the keywords in the search query.

The analysis results are included in the XML document by the number and position of occurrence of each word in HTML source code found as free form text. These values can be used to find the keyword density which is required by Term Frequency - Inverse Document Frequency (TF-IDF). The sample XML document format is given below.

```
<Web pages>
        <url>  …        </url>
        <Title> …       </Title>
        <Meta Description> …      </Meta Description>
        <Meta Keywords> …         </Meta Keywords>
        <Headings> …    <Headings>
        <BIU>…    </BIU>
        <Image>…        </Image>
        <Words>
                <Word1>
                        <count>  …   </count>
                        <position> …        </position>
                </Word1>
                <Word2>
                        <count>  …   </count>
                        <position> …        </position>
                </Word2>
                ..
                ..
        </Words>
</Web pages>
```

Fig.2 XML webpage structuring format

The framework proposed in this paper is a generic one. The parameters considered for XML file creation are Meta tag, Title tag, Bold, Italic, Underline etc. Depending upon the requirements the tags used for analysis may be included or excluded.

The XML documents generated will be used by the further phases of information filtering. The structure of the web page generated in XML format will be very helpful for improving the search in either the “Keyword-based” search or “Ontology-based” search. The proposed system is implemented using .NET framework.

The sample queries used for XML documents generation are
Query 1: Transposition Ciphers
Query 2: Pointers in C

The sample outputs of the XML document generated using this proposed framework for query 1 - “Transposition Ciphers” are shown in fig.3 and fig.4.

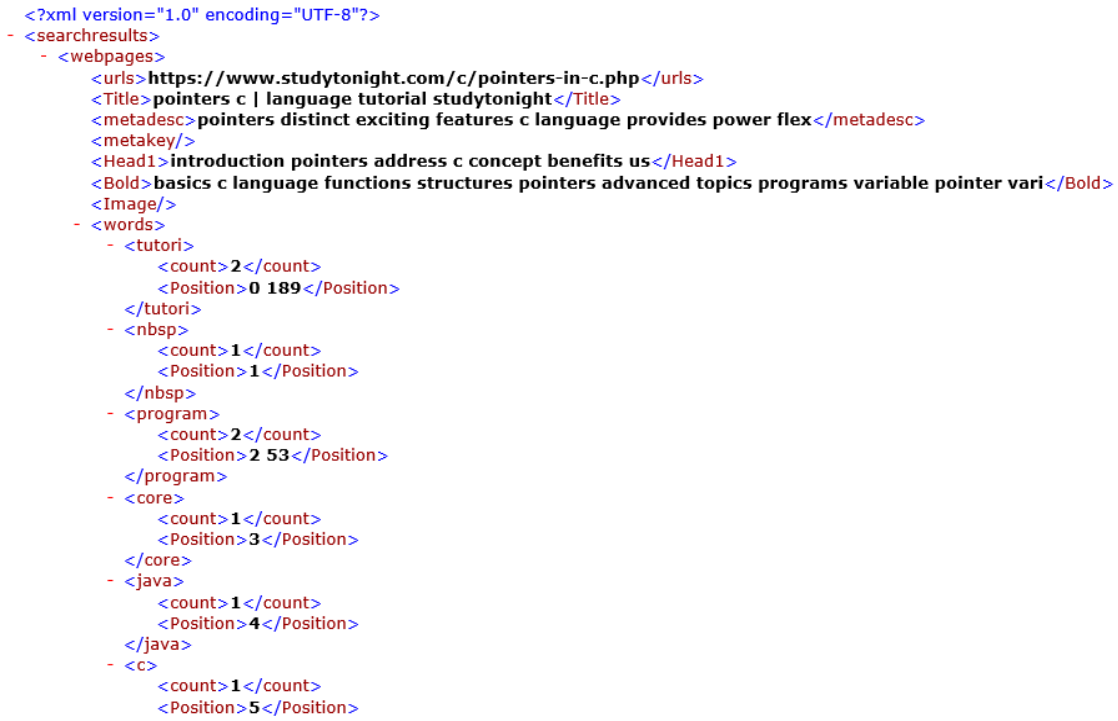
The sample outputs of the XML document generated using this proposed framework for query 2 - “Pointers in C” is shown in fig.5.

```
<?xml version="1.0" encoding="UTF-8"?>
- <searchresults>
  - <webpages>
      <urls>https://www.dcode.fr/transposition-cipher</urls>
      <Title>transposition cipher - decoder encoder solver transl</Title>
      <metadesc>tool decrypt/encrypt transposition cipher called columns permutation technique change order letters text placing grid</metadesc>
      <metakey/>
      <Head1>transposition cipher decoder encoder answers questions encrypt using decrypt recognize ciphertext decipher key variants source code com</Head1>
      <Bold/>
      <Image/>
    - <words>
      - <search>
          <count>1</count>
          <Position>0</Position>
        </search>
      - <tool>
          <count>2</count>
          <Position>1 23</Position>
        </tool>
      - <dcode>
          <count>1</count>
          <Position>2</Position>
        </dcode>
      - <keyword>
          <count>1</count>
          <Position>3</Position>
        </keyword>
      - <form>
          <count>1</count>
          <Position>4</Position>
        </form>
      - <transposit>
          <count>1</count>
          <Position>5</Position>
        </transposit>
```

Fig.3 XML document generated for the query - “Transposistion Ciphers” and URL - https://www.dcode.fr/transposition-cipher

```
<?xml version="1.0" encoding="UTF-8"?>
- <searchresults>
  - <webpages>
      <urls>https://en.wikibooks.org/wiki/Cryptography/Transposition_ciphers</urls>
      <Title>cryptography/transposition ciphers - wikibooks open books world</Title>
      <metadesc/>
      <metakey/>
      <Head1>cryptography transposition ciphers columnar edit double grille navigation menu personal tools namespaces variants views search community languages
        sister projects print export</Head1>
      <Bold>transposition cipher z e b r </Bold>
      <Image/>
    - <words>
      - <wikibook>
          <count>1</count>
          <Position>0</Position>
        </wikibook>
      - <open>
          <count>1</count>
          <Position>1</Position>
        </open>
      - <book>
          <count>2</count>
          <Position>2 231</Position>
        </book>
      - <world>
          <count>1</count>
          <Position>3</Position>
        </world>
      - <lt>
          <count>1</count>
          <Position>4</Position>
        </lt>
      - <cryptographi>
          <count>1</count>
```

Fig.4 XML document generated for the query - “Transposistion Ciphers” and URL– https://en.wikibooks.org/ wiki/ Cryptography/Transposision_ciphers



```
<?xml version="1.0" encoding="UTF-8"?>
<searchresults>
  <webpages>
    <urls>https://www.studytonight.com/c/pointers-in-c.php</urls>
    <Title>pointers c | language tutorial studytonight</Title>
    <metadesc>pointers distinct exciting features c language provides power flex</metadesc>
    <metakey/>
    <Head1>introduction pointers address c concept benefits us</Head1>
    <Bold>basics c language functions structures pointers advanced topics programs variable pointer vari</Bold>
    <Image/>
    <words>
      <tutori>
        <count>2</count>
        <Position>0 189</Position>
      </tutori>
      <nbsp>
        <count>1</count>
        <Position>1</Position>
      </nbsp>
      <program>
        <count>2</count>
        <Position>2 53</Position>
      </program>
      <core>
        <count>1</count>
        <Position>3</Position>
      </core>
      <java>
        <count>1</count>
        <Position>4</Position>
      </java>
      <c>
        <count>1</count>
        <Position>5</Position>
```


Fig.5 XML Document generated for the query - “Pointers in C” and URL - “https://www.studytonight.com/c/pointers-in-c.php

IV. CONCLUSION

This framework is a generic one and any further improvement can be incorporated easily without affecting the existing system. This framework shows the distribution of words in the web pages and also the inner text in various parts of the HTML code. It introduces a structure to the web pages which are semi-structured or unstructured in nature. It is very useful in information filtering in web pages since it provides a clear picture of how words are distributed in the different part the web pages as well as its HTML source code. This analysis of web pages would definitely aid in search engine optimization which helps in improving the ranking of web pages by search engines. Optimized web pages are found in the top of the search engine results which helps the users to find their required information much easier.

ACKNOWLEDGMENT

This article has been written with the financial support of RUSA – Phase 2.0 grant sanctioned vide Letter No.F.24-51/ 2014-U, Policy(TNMulti-Gen), Dept. of Edn. Govt. of India, Dt.09.10.2018.

REFERENCES

[1] Mu. Annalakshmi, and Dr.A. Padmapriya, “Search Factors used by Search Engines”, *International Journal of Advanced Research Trends in Engineering and Technology*, Vol. 3, No. 20, pp. 652-655 , April 2016.

[2] Prashant M. Ahire, Anil P. Gagare, Yogesh B. Pawar, and Savan S. Vidhate, “Extract and Analysis of Semi Structured Data from Websites and Documents”, *International Journal of Computer Science and Mobile Computing*, Vol. 4, No. 2, pp. 314–318, February 2015

[3] Sheetal Patil, and Prof.M.B.Ansari, “User Profile Based Personalized Research Paper Recommendation System Using Top-K Query”, *International Journal of Emerging Technology and Advanced Engineering*, Vol.5, No. 9, Sept. 2015.

[4] Ayush Jain, “The Role and Importance of Search Engine and Search Engine Optimization”, *International Journal of Emerging Trends & Technology in Computer Science*, Vol. 2, No. 3, May –June 2013

[5] (2019) Computer how stuff works website. [Online] Available: https://computer.howstuffworks.com/internet/basics/search-engine1.htm

[6] (2019) The search micro services website. [Online] Available: https://searchmicroservices.techtarget.com/definition/XML-Extensible -Markup-Language .

[7] (2019) The W3 resource website. [Online] Available: https://www.w3resource.com/xml/uses-of-xml.php

[8] H.P. Khandagale and P.P. Halkarnikar, “A Novel Approach for Web Page Change Detection System”, *International Journal of Computer Theory and Engineering*, Vol. 2, No. 3, June, 2010

[9] Sofia J Athenikos and Hyoil Han, “Biomedical question answering: A survey”, *Computer methods and programs in biomedicine*, Vol.99, No.1, pp. 1–24, 2010.

[10] Aaron M Cohen and William R Hersh, “A survey of current work in biomedical text mining”, *Briefings in bioinformatics*, Vol. 6, No.1, pp. 57–71, 2005.