ROOTINDEXING
JOURNAL ABSTRACTING AND INDEXING SERVICE
INTERNATIONAL ADVANCE JOURNAL OF ENGINEERING, SCIENCE AND MANAGEMENT (IAJESM)
January-June 2023, Submitted in January 2023, iajesm2014@gmail.com, ISSN -2393-8048
Multidisciplinary Indexed/Peer Reviewed Journal. SJIF Impact Factor 2023 =6.753

A Review Paper on Web Indexing

Gaurav Rastogi, Department of Applied science, RD Engineering College Ghaziabad, India. gauravrastogi@gmail.com

Abstract

Web indexing plays a crucial role in information retrieval and search engine operations. It involves the process of collecting, analysing, and organizing web pages to make them easily searchable and accessible to users. This research paper presents a comprehensive study of web indexing techniques and strategies used by search engines. The paper discusses various indexing techniques such as crawling, content analysis, link analysis, and metadata analysis. It also examines the challenges faced by search engines in indexing the web, such as duplicate content, spamming, and content freshness. Finally, the paper analyses the future of web indexing, considering the emerging technologies and trends in the field.

Introduction

The internet has become an essential part of our lives, and the amount of information available on the web is growing exponentially. Search engines are the primary tools used to access this information, and their effectiveness depends on their ability to index web pages accurately. Web indexing involves collecting, analysing, and organizing web pages to make them easily searchable and accessible to users. The purpose of this research paper is to provide a comprehensive study of web indexing techniques and strategies used by search engines. The paper discusses the challenges faced by search engines in indexing the web and analyses the future of web indexing.

The World Wide Web has grown exponentially in recent years, with millions of web pages being added every day. This has made it increasingly difficult to find relevant information on the web. Web indexing is the process of organizing and categorizing web pages to facilitate their retrieval by search engines. The goal of web indexing is to make it easier for users to find the information they need on the web.

The World Wide Web is a vast and ever-growing collection of information, consisting of billions of web pages, documents, images, videos, and other types of content. As the number of web pages continues to increase exponentially, it becomes increasingly difficult for users to find the information they need quickly and easily. Web indexing plays a critical role in enabling users to access relevant information from the web efficiently. Search engines such as Google, Bing, and Yahoo rely on sophisticated web indexing techniques to crawl, index, and rank web pages based on their relevance to user queries.

Web indexing means creating indexes for individual Web sites, intranets, collections of HTML documents, or even collections of Web sites.

Indexes are systematically arranged items, such as topics or names, that serve as entry points to go directly to desired information within a larger document or set of documents. Indexes are traditionally alphabetically arranged. But they may also make use of hierarchical arrangements, as provided by thesauri, or they may be entirely hierarchical, as in the case of taxonomies. An index might not even be displayed if it incorporated into a searchable database.

Indexing is an analytic process of determining which concepts are worth indexing, what entry labels to use, and how to arrange the entries. As such, Web indexing is best done by individuals skilled in the craft of indexing, either through formal training or through self-taught reading and study.[1]

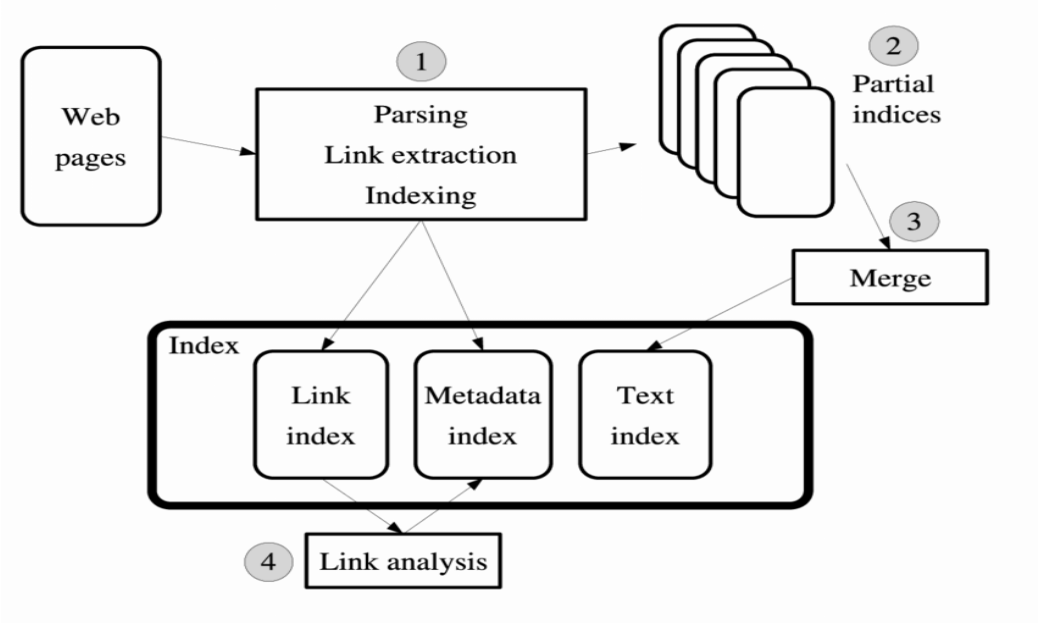

Figure 1

Web Indexing Techniques

Web indexing techniques can be broadly categorized into four types: crawling, content analysis, link analysis, and metadata analysis.

Crawling: Crawling is the process of collecting web pages by following hyperlinks from one page to another. Search engine crawlers or bots start with a set of seed URLs and recursively follow the links on the pages to discover new pages. Crawling is a time-consuming process, and search engines need to prioritize which pages to crawl first based on factors such as page rank and relevance.

Content Analysis: Content analysis involves analysing the text and other content on a web page to determine its relevance and importance. Search engines use various algorithms to analyse the content, such as keyword frequency, semantic analysis, and machine learning.

Link Analysis: Link analysis involves analysing the links between web pages to determine their importance and relevance. Search engines use algorithms such as PageRank to determine the importance of a page based on the number and quality of links pointing to it.

Metadata Analysis: Metadata analysis involves analysing the metadata of a web page, such as its title, description, and keywords, to determine its relevance and importance.

Web Crawling:

Web crawling is the process of systematically visiting web pages and extracting information from them. Search engines use web crawlers, also known as spiders or bots, to automatically traverse the web and collect information about web pages. Web crawlers typically start by visiting a set of seed URLs and then follow links on those pages to discover new pages to crawl. Web crawlers need to be efficient, as there are billions of web pages on the internet, and crawling all of them can be time-consuming and resource-intensive.Web crawling is the process of indexing data on web pages by using a program or automated script. These automated scripts or programs are known by multiple names, including web crawler, spider, spider bot, and often shortened to crawler.[2]

Indexing:

Once a web crawler has collected information about a web page, it needs to be processed and stored in a way that enables efficient retrieval. Indexing involves parsing the contents of web pages and creating an index, which is a data structure that stores information about the web pages in a way that enables efficient searching. Search engines typically use one of two types of indexing: keyword-based indexing and semantic indexing.

Keyword-based indexing:

Keyword-based indexing is the most common type of indexing used by search engines. In this technique, web pages are indexed based on the keywords that appear in their content. When a user enters a search query, the search engine looks for web pages that contain the keywords in the query and ranks them based on their relevance to the query.

Semantic indexing:

Semantic indexing is a more advanced form of indexing that uses natural language processing and machine learning algorithms to analye the meaning of web page content. Semantic indexing attempts to understand the context and meaning of the words and phrases on a web page and create a semantic representation of the page. This allows search engines to return more accurate search results by understanding the user's intent behind their query.

Ranking:

Once web pages have been indexed, search engines need to rank them based on their relevance to user queries. Ranking algorithms use a variety of factors to determine the relevance of a web page, including the number of links pointing to the page, the quality of those links, and the relevance of the page's content to the user's query. Page ranking algorithms, such as Google's PageRank algorithm, are used to determine the authority and relevance of a web page.

THE RDF FRAMEWORK FOR SEMANTIC INDEXING

The semantic indexing process is substantially characterized by the following steps:

(i) Web Search: this activity has the task of retrieving a set of HTML documents that satisfy some search criteria using a Search Engine external component;

(ii) Text Extraction: the sentences are extracted from the several web sources by parsing the related HTML pages using the JSOUP API;

(iii) NLP and Triplets Extraction: NLP processing techniques are performed on the input sentence to detect a set of triplets using Stanford Libraries 1;

(iv) Semantic Distance Matrix Builder: a matrix containing the semantic distance values for each couple of triplets is computed;

(v) Rk Mapping and Semantic Index Building: this activity performs the mapping of the triples in a metric space and the building of the final indexing structure based on K-d Tree.

HTML Priority System

HTML Priority System is content based ranking mechanism and this kind of system is vulnerable to content spamming. Content spamming can be classified into 5 subtypes [11]:

1. Title spamming involves including too many keywords to achieve an overall higher ranking.

2. Body spamming involves either including a lot of keywords that appear frequently in queries/titles to achieve higher ranking or including diverse set of keywords in order to get indexed under as many result tables as possible. Quite often the spam text is kept hidden from the user using various strategies.

3. Meta Tag spamming, as discussed earlier meta-tag content is given low priority for ranking purpose. Since it is not visible to the end users, it is vulnerable to heavy spamming.

4. Anchor Text spamming is done by creating links with desired anchor text that contain certain keywords as text to trick the system into assuming content credibility.

5. URL spamming occurs when spammers design URLs with certain keywords to boost ranking for those keywords. The system proposed however, is resistant to this kind of spamming[7].

Related Work

Owing to the dynamic nature of the web, it is difficult for search engine to find the relevant documents to serve a user query. For this purpose, search engine maintains the index of

downloaded documents stored in the local repository. Whenever a query comes search engine searches the index in order to find the relevant matched results to be presented to the user. The quality of the matched result depends on the information stored in the index. The more efficient is the structure of index, more efficient the performance of search engine. Generally, inverted index are based solely on the frequency of keywords present in number of documents. In order to improve the efficiency of the search engine, an improved indexing mechanism to index the web documents is being proposed that keeps the context related information integrated with the frequency of the keyword. The structure is implemented using Trie. The implementation results on various documents show that proposed index efficiently stores the documents and search is fast[3].

Nowadays, advent of web service is considered as a technology bringing a revolution operations of online B2B (Business to Business) and B2C (Business to Customer) applications. However, the user requirement for web services is diverse and complex. In most cases, a simple web service cannot meet the user requirement. Therefore, the composition of web services (or Web Service Composition - WSC) is set out as an indispensable. When the number of web services in repository increase, the computational cost of composition problem also increases significantly. The performance of composition process depends on the way we organize web services in the repository. Indexing is a technique of data mining field which can help us organize the web services and retrieve them optimally. In this paper, H. T. Khai, B. H. Thang and Q. T. Tho et.al.proposed a bitwise-based indexing mechanism for fast location of suitable web services. This approach is implemented and proved its effectiveness over the real datasets with thousands of web services. As a result, significant improvement of performance has been made, especially as compared to other existing works in the same field[4].

Designing and developing an effective web crawler is a challenging role in a large search engine. This paper proposes component based web crawler along with the indexer. The WebCrawler consist of crawler services and indexer services and realized as web services. The communication between the services is sent and received using XML, SOAP and WSDL. In the crawler service, the web pages are fetched and parsed for retrieving all the hyperlinks. The process is carried out recursively using Breadth-First strategy. The extracted URLs are downloaded and those web pages are sent to the indexer services by passing the message. In the indexer service, HTML pages are parsed, stop words are removed, stemming of keywords are carried out as pre-processing steps and the result is stored in the form of inverted index.A. Vadivel, S. G. Shaila, R. Devi Mahalakshmi and J. Karthika et.al.evaluated the performance of the proposed design specification of the crawler with indexer and found that the number of pages retrieved is notably on the higher side[5].

Managing efficiently and effectively very large amount of digital documents requires the definition of indexes able to capture and express documents' semantics. In this work, F. Amato, V. Moscato, F. Persia, A. Picariello and F. Gargiulo et.al. propose an RDF based framework for semantic indexing of web pages considering the related textual information. In particular, we propose to capture the semantic nature of a given document, commonly expressed in natural language, by retrieving a number of RDF triples and to semantically index the documents on the base of meaning of the triples' elements (i.e. subject, verb, object). Preliminary experiments are reported to evaluate the proposed index strategy[6].

The unstructured nature and the sheer size of the World Wide Web make it a challenging task to index. This paper will discuss about how web can be incrementally indexed using Inverted Indices and Distributed Hash Table for efficiently organizing the data while incrementally build the index using the search mechanism itself, and HTML Priority System for ranking the pages to improve precision and recall. It also discusses certain challenges that a content-based ranking system must face to counter spam[7].

Challenges in Web Indexing

Search engines face various challenges in indexing the web, such as:

Duplicate Content: The web contains a vast amount of duplicate content, which makes it difficult for search engines to determine which version of the content to index.Duplicate content is content which is available on multiple URLs on the web. Because more than one URL shows the same content, search engines don't know which URL to list higher in the search results. Therefore they might rank both URLs lower and give preference to other webpages[8].

Spamming: Spammers use various techniques to manipulate search engine rankings, such as keyword stuffing, cloaking, and link spamming.Web Spam, in other words, are techniques that are used by some websites to try and cheat their way to the top of the search engine results page[9].

Freshness: The web is constantly changing as content is added, deleted, and modified. In order for a crawler to reflect the web updates it will encounter it, needs to recrawl content soon after it changes. This need for freshness is key to providing a good search engine experience. For instance, when breaking news develops, users will rely on your search engine to stay updated. It's also important to refresh less time-sensitive documents so the results list doesn't contain spurious links to deleted or modified data[10].

The Future of Web Indexing

The future of web indexing is closely tied to the emerging technologies and trends in the field. Some of the emerging trends in web indexing include:

Semantic Search: Semantic search involves understanding the meaning of search queries and web pages to provide more relevant search results.Semantic search describes a search engine's attempt to generate the most accurate SERP results possible by understanding based on searcher intent, query context, and the relationship between words[11].

Personalization: Personalization involves tailoring search results to the user's preferences and search history.Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behaviour (usage data) in correlation with other information collected in the Web context, namely, structure, content, and user profile data[12].

Voice Search: Voice search is becoming increasingly popular, and search engines need to adapt to this new way of searching.Voice search or voice-enabled search is the means of searching by using the most natural input channel, human speech. Voice search can refer to Google searches performed by using voice, but in this article, we'll investigate how voice search could be used in touch screen applications and websites[13].

Conclusion

Web indexing is a critical process that enables users to efficiently access relevant information on the web. In this research paper, we have provided a comprehensive study of web indexing techniques, including web crawling, indexing, and ranking. We have discussed the advantages and disadvantages of different indexing techniques, such as keyword-based indexing and semantic indexing, and provided a critical analysis of their effectiveness in delivering relevant search results to users.

References

[1]https://www.web-indexing.org/about-web-indexing/ retrieved on 29-04-2022
[2]https://research.aimultiple.com/web-crawler/ retrieved on 29-04-2022
[3]P. Mudgil, A. K. Sharma and P. Gupta, "An Improved Indexing Mechanism to Index Web Documents," 2013 5th International Conference and Computational Intelligence and Communication Networks, Mathura, India, 2013, pp. 460-464, doi: 10.1109/CICN.2013.101.

[4]H. T. Khai, B. H. Thang and Q. T. Tho, "An Application of Bitwise-Based Indexing to Web Service Composition and Verification," 2016 International Conference on Advanced Computing and Applications (ACOMP), Can Tho, Vietnam, 2016, pp. 51-58, doi: 10.1109/ACOMP.2016.017.

[5]A. Vadivel, S. G. Shaila, R. Devi Mahalakshmi and J. Karthika, "Component based effective web crawler and indexer using web services," IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), Nagapattinam, India, 2012, pp. 792-797.

[6] F. Amato, V. Moscato, F. Persia, A. Picariello and F. Gargiulo, "An RDF-Based Framework for Semantic Indexing of Web Pages," 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, 2013, pp. 395-396, doi: 10.1109/ICSC.2013.76.

[7]Y. Sagar, "Web indexing using HTML priority system," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015, pp. 581-584, doi: 10.1109/ABLAZE.2015.7154929.

[8]https://yoast.com/duplicate-content/ retrieve on 30-04-2022.

[9]https://acs-web.com/digital-marketing-lexicon/search-engine-optimization-terms/web-spam/ retrieve on 30-04-2022.

[10] https://course.ccs.neu.edu/cs6200sp15/slides/m05.s06%20-%20freshness.pdf retrieve on 30-04-2022.

[11] https://www.searchenginejournal.com/semantic-search-seo/264037/ retrieve on 30-04-2022.

[12]https://dl.acm.org/doi/10.1145/643477.643478 retrieve on 30-04-2022.

[13]https://www.speechly.com/blog/voice-search/ retrieve on 30-04-2022.

[14] Dharamveer, Samsher, Singh D.B., Singh A.K., Kumar N. (2019) “Solar Distiller Unit Loaded with Nanofluid—A Short Review”. In: Kumar M., Pandey R., Kumar V. (eds) Advances in Interdisciplinary Engineering. Lecture Notes in Mechanical Engineering. Springer, Singapore. pp 241-247, Paper Published. Scopus Index, Springer Publication. https://doi.org/10.1007/978-981-13-6577-5_24

[15]Dharamveer, Samsher “Comparative Analysis of Energy Matrices and Enviro-economics for Active and Passive Solar Still”. Journal Materials Today proceedings, Elsevier publication. https://doi.org/10.1016/j.matpr.2020.10.001

[16] Dharamveer, Samsher, Anil Kumar “Performance analysis of N-identical PVT-CPC collectors an active single slope solar distiller with a helically coiled heat exchanger using CuO nanoparticles”, Water supply, Vol. 22 (201) 02 1306-1326, October 2021 https://doi.org/10.2166/ws.2021.348

[17] Dharamveer, Samsher, Anil Kumar “Analytical study of photovoltaic thermal (PVT) compound parabolic concentrator (CPC) active double slope solar distiller with helical coiled heat exchanger using CuO Nanoparticles” Desalination and water treatment, vol. 233 (2021) 30–51 https://doi.org/10.5004/dwt.2021.27526