IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 7, Issue 1, January 2020

Page Relevance Computation in Web Crawlers: Techniques and Challenges

Ms. Kompal

Govt. College, Panchkula

Abstract: As the internet continues to grow exponentially, efficient web crawling and effective page relevance computation have become crucial for information retrieval systems. The objective is to determine how relevant a web page or document is to a given query or user context. This paper explores various techniques employed in assessing page relevance, including link-based, content-based, and hybrid methods. Additionally, it discusses the challenges associated with relevance computation, such as the dynamic nature of web content, scalability issues, and the presence of low-quality pages.

I. INTRODUCTION

Web crawlers are automated agents that navigate the internet to collect and index information for search engines. The ability of these crawlers to prioritize pages based on relevance directly influences the quality of search results. Web crawlers, often referred to as spiders or bots, systematically browse the web to index content for search engines.

They play a vital role in organizing information, ensuring users can efficiently retrieve relevant data. Relevance is a measure of how closely the content of a web page aligns with user queries. High relevance leads to better user satisfaction and engagement, making effective relevance computation essential for search engines.

II. TECHNIQUES FOR PAGE RELEVANCE COMPUTATION

2.1 Link-Based Approaches

2.1.1 PageRank

PageRank is a widely recognized algorithm that assesses the importance of a page based on the number and quality of links pointing to it. Developed by Google founders Larry Page and Sergey Brin, the algorithm operates under the assumption that important pages are likely to receive more links from other pages.

2.1.2 HITS Algorithm

HITS (Hyperlink-Induced Topic Search) evaluates pages based on two scores: authority and hub. An authority page is one that is linked to by many hubs, while a hub page links to many authoritative pages.

2.2 Keyword Analysis

- TF-IDF (Term Frequency–Inverse Document Frequency): Measures the importance of terms by balancing their frequency in a document against their occurrence across the corpus.

- BM25 (Best Matching 25): A probabilistic extension of TF-IDF that incorporates term saturation and document length normalization.

2.3 Semantic Analysis

Advanced semantic analysis techniques, including Latent Semantic Analysis (LSA) and word embeddings, allow crawlers to assess page relevance based on the contextual meaning of words rather than mere keyword matching. This approach enhances understanding of synonyms and related concepts.

2.4 Hybrid Approaches

Hybrid methods combine link-based and content-based techniques to provide a more comprehensive view of page relevance. Machine learning models can integrate multiple features, such as link structure, content quality, and user engagement metrics, to yield more accurate relevance scores.

DOI 10.17148/IARJSET.2020.7121

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 7, Issue 1, January 2020

III. CHALLENGES IN PAGE RELEVANCE COMPUTATION

3.1 Dynamic Nature of the Web

The internet is continuously changing, with new content being added and old content being updated or removed. Maintaining up-to-date relevance scores is a significant challenge for crawlers, requiring constant re-evaluation of indexed pages.

3.2 Scalability

As the volume of web content increases, computing relevance efficiently becomes more complex. Techniques that function well on smaller datasets may struggle to scale effectively, necessitating innovations in data processing and storage.

3.3 Spam and Low-Quality Pages

The proliferation of spam and low-quality content complicates relevance computation. Effective filtering mechanisms are essential to ensure crawlers prioritize high-quality pages, thereby enhancing overall search result quality.

IV. FUTURE DIRECTIONS

To address the challenges of page relevance computation, future research may explore the following areas:

- AI and Deep Learning**:** Leveraging advanced AI techniques could significantly improve relevance assessment accuracy.
- User Personalization: Tailoring search results based on individual user behaviour and preferences can lead to enhanced relevance.
- Blockchain Technology: Implementing decentralized approaches could improve transparency and integrity in relevance scoring.

V. CONCLUSION

Page relevance computation is a vital component of web crawling that directly impacts search engine performance and user satisfaction. While various techniques exist, the challenges of dynamic content, scalability, and spam necessitate ongoing research and innovation. By addressing these challenges, future developments can enhance the efficacy of web crawlers in delivering relevant and high-quality information.

REFERENCES

[1]. Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems.
[2]. Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. Journal of the ACM.
[3]. Salton, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill.
[4]. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*.
[5]. Zhao, J., & Jiang, Z. (2019). Page Relevance Computation Based on Deep Learning. Journal of Information Science.