RESEARCH PAPER

Design and Implementation of Novel Techniques for Content-Based Ranking of Web Documents

Ritika Choudhary[1], Mukesh Rawat[2], Aarti Verma[3]

[1] Department of CSE, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh - 250005.
Email: Ritika.chaudhary.mtcs.2024@miet.ac.in

[2] Department of IT, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh - 250005.
Email: Mukesh.rawat@miet.ac.in

[3] Department of IT, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh - 250005.
Email: aarti.verma@miet.ac.in

Received: 2nd Mar, 2026 | **Revised:** 14th Mar, 2026 | **Accepted:** 4th Apr, 2026 | **Available Online:** 20th Apr, 2026

ABSTRACT

The Web content has exponentially grown making useful document ranking one of the essentials of a modern search engine. The classical hyperlink based ranking model like PageRank and HITS are largely based on link structure, and they do not always measure the true semantic relevance of Web documents. The content-based ranking methods, on the contrary, consider textual, semantic, and contextual data to deliver user-focused and most relevant output. The survey is based on a discussion of classical, machine learning, and emerging deep learning-based methods of ranking Web documents based on their content. It puts into emphasis their design principles, implementation strategies, evaluation metrics, limitations and real world applicability. The paper also covers the new hybrid and transformer-based approaches where semantic embeddings, learning-to-rank engines, and knowledge-enhanced presentations are combined to overcome existing challenges. The survey has ended by establishing the gaps in research and the development of future directions in the development of the advanced content-centric ranking systems.

Keywords: Web documents, Content-based ranking, Learning-to-Rank, Relevance scoring, Semantic ranking.

How to cite this article: Choudhary R, Rawat M, Verma A. Design and Implementation of Novel Techniques for Content-Based Ranking of Web Documents. Int J Drug Deliv Technol. 2026;16(34s):729-734. DOI: 10.25258/ijddt.16.34s.90

Source of support: Nil.

Conflict of interest: The authors declare no conflict of interest.

I. INTRODUCTION

The velocity of World Wide Web development has resulted in the excessive quantity of digital content, and the process of effective retrieval and prioritization of web pages has become a serious issue. The conventional search engines have been relying a lot on link-based ranking algorithms like PageRank and HITS that measure the authority of a web page with respect to its hyperlink network. Although they work well in the early phases of web, they do not usually portray the semantic relevance of content particularly when there are few or no links into a page, when the page was recently made or a web domain where hyperlinks are few and far between.

Alternative ranking methods have proved to be promising as a solution to these weaknesses considering that they have placed special attention on the textual, structural and semantic point of the contents of a web document. These methods break down term frequency, metadata, document structure, natural language semantics and user intent in order to calculate relevance scores. As machine learning (ML), deep learning (DL), natural language processing (NLP) and semantic web technologies have advanced, new models have been created that are superior to conventional ranking algorithms, delivering at least more context aware, personalized, and intelligent ranking.

The current survey paper gives an overall overview of new methods employed to rank web documents based on their content. It addresses traditional information retrieval (IR) methods, ranking methods boosted with NLP, relevance feedback methods, and the latest ML/DL-based ranking methods. The paper also reviews the current developments in the field, research gaps, and future research opportunities that can be used to develop more accurate, scalable and adaptive web document ranking systems.

II. LITERATURE SURVEY

Web document ranking on content basis is an important aspect in information retrieval systems since it allows users to get the relevant information out of the vast and heterogeneous pool of web information. The old methodology of search engines such as Google and Yahoo use a mix of the key word matching, linkage analysis and heuristic ranking functions. Nevertheless, the growing amount of data and complicated user queries has made researchers create new ranking methods to enhance accuracy, relevance, and contextual insights.

Prakash, Gupta and Rawat have made one of such contributions by suggesting a re-ranking method that receives the first results of the search on Google and re-orders them according to normalized frequency of query keywords terms in the contents of the documents [1]. Their approach is strictly content-based and it can be shown that the ranking of the keywords results in a substantial change in the rank of relevance in comparison with the default search engines.

Techniques based on the content can be vulnerable to noises and missing texts in web pages. In a bid to solve this challenge, scholars have attempted to use hybrid models whereby content features are merged with link structure. Calado et al. showed the effectiveness of the use of link information in enhancing web document classification by comparing several measures of link-derived similarity [2]. Their method involved a Bayesian network to integrate both the link features and content-based classifiers and demonstrated that link information alone can surpass the classical methods, with maximum improvements in F1 of 46 points. Their paper highlights that link structure provides good complementary information as well as cautions that noisy textual data has the potential to worsen joint performance.

The introduction of deep learning has seen the methods of ranking move towards sparse and dense neural networks. SPLADE is a model introduced by Formal et al., repurposed to build on the encoder framework of transformers to represent expanded and sparse representations of task ranking [3]. This method is also enhanced in SPLADE v2, which has a better sparsity control and training efficiency and is more feasible in large-scale retrieval conditions [4].

Similar to sparse models, dense passage retrievals have been of much interest. Karpukhin et al. introduce Dense Passage Retrieval (DPR): queries and passages are encoded in dense vector spaces, and the semantic matching is executed in dense space [5]. The performance and strength of DPR were confirmed by a replication study by Ma et al. in multiple training conditions with negative emphasis on the importance of negative sampling and quality of the dataset [6].

RocketQA which was created by Qu et al. enhances dense retrieval by adding refined sampling strategies and joint training approaches to the ranking and retrieval tasks [7]. This enables a superior semantic matching, and offers robust performance enhancements in diverse benchmarks. There is also retrieval robustness under real world query conditions. Zhuang and Zuccon investigated the behaviour of BERT-based retrieval with typographical noise, and found that contextual models stay consistent when the queries are noisy [8].

Modern IR systems are still affected by classical machine learning methods of ranking. Early studies including LambdaRank and LambdaMART by Donmez and others offer potent listwise ranking models which optimize ranking measures directly [9][10]. These methods are content, link and behavioural methods and laid the foundation of most neural ranking models.

Lastly, recent studies include the efficient sparse neural retrieval. A model suggested by Dudek et al. learns sparse lexical features in large vocabularies, gaining good retrieval accuracy and at a computational cost [11]. In the same vein, CSurF presented by Zhen and Callan incorporates lexical representations in the context, but remains sparse, that is, balancing scalability and semantic good [12].

Table1: survey table:

Authors / Year	**Technique**	**Type**	**Key Features**	**Limitations**
Prakash et al., 2022	Term-frequency re-ranking	Content-based	Reorders Google search results using normalized TF	No semantic understanding
Calado et al., 2003	Link + content combination	Hybrid	Link-based similarity +	Sensitive to noisy text

			Bayesian network	
Formal et al., 2021	SPLADE	Sparse neural	Sparse lexical expansions, transformer-based	Requires GPU resources
Formal et al., 2021	SPLADE v2	Sparse neural	Better sparsity control and efficiency	Complex training
Karpukhin et al., 2020	DPR	Dense retrieval	Semantic dual-encoder retrieval	Needs large training data
Ma et al., 2021	DPR Replication	Dense retrieval	Confirms robustness, analyzes performance factors	Dataset-dependent
Qu et al., 2020	RocketQA	Dense retrieval	Optimized sampling + joint training	Computationally heavy
Zhuang & Zuccon, 2021	BERT typo-robust ranking	Neural	Robust to noisy/typo queries	Slow inference
Donmez et al., 2008	Lambda Rank	Learning-to-rank	Direct optimization of ranking metrics	Requires feature engineering
Wu et al., 2010	Lambda MART	Learning-to-rank	Strong tree-based ranking model	Heavy model size
Dudek et al., 2023	Sparse lexical representations	Sparse neural	Efficient large-scale retrieval	Limited semantic depth
Zhen & Callan, 2023	CSurF	Sparse neural	Contextualized sparse surfaces	Still emerging model

III. Content-based ranking

The content-based ranking methods have gone a long way and their performance depends on how effective they are in terms of capturing the text relevance, semantic meaning and contextual relationship. The conventional methods of using key words like TF-IDF and BM25 primarily rely on frequency statistics of terms. These methods are effective with large sets of documents, but they fail frequently with documents containing synonyms of the query, or indirect expressions of it. As a result, they lose accuracy on semantically rich or ambiguous queries.

Such limitations are overcome by semantic-based methods, such as Latent Semantic Analysis and ontology-based ranking, which analyse conceptual relations between terms. They are more likely to be recalled and better understood within the context but they demand a lot of pre processing, special vocabulary storage and heavy calculation. When the domain specific ontologies are not available or are incomplete, their performance declines.

Content-based ranking is a fundamental approach in information retrieval (IR) that prioritizes documents based on the relevance of their **intrinsic textual content** to a user's query. Unlike link-based or behavior-driven ranking models, content-based ranking analyzes **what is written inside the document**, making it essential for domains where external metadata such as hyperlinks, click patterns, or citations are limited. This method is widely applied in search engines, digital libraries, semantic retrieval systems, and domain-specific indexing platforms.

The most balanced performance is provided by hybrid ranking schemes that use a combination of statistical and semantic and ML methods. They are more relevant, noisy data, and types of documents. The major hurdle is the complexity of the system, integration overheads and the ability to scale with the growth of document volume. On the whole, modern methods are characterized by the trade-off between accuracy, efficiency, and the complexity of implementation and the necessity to introduce new ranking mechanisms that are both semantically efficient and computationally efficient.

IV. Link-Based Ranking

Link-based ranking is a fundamental approach in information retrieval (IR) and web mining that determines the importance of a web page based on its **link structure**, rather than its textual content alone. This paradigm views the web as a directed graph where nodes represent pages and edges represent hyperlinks. The central assumption is that a page linked by many other pages—especially authoritative ones—is likely to be more important or trustworthy. Link-based ranking revolutionized web search, most notably through Google's PageRank algorithm, and continues to play a crucial role in modern search engines.

1. PageRank

PageRank evaluates page importance based on the number and quality of incoming links. Each link passes a fraction of its "rank" to the target page. The iterative computation uses the **random surfer model**, where rank stabilizes as a probability distribution of visits.

Key features:

- Considers both **quantity and quality** (importance) of links
- Uses damping factor (usually 0.85) to simulate random navigation.
- Robust against simple keyword stuffing.

$PR(A)=(1-d) +d (i=1\sum nC (Ti)PR(Ti))$

Where:

- **PR(A)** = PageRank of page A
- **d** = damping factor (usually **0.85**)
- **T_i** = pages linking to A
- **$C(T_i)$** = number of outgoing links from T_i

Damping Factor (d):

Simulates that a user keeps clicking links (85% chance) or jumps to a random page (15%).

2. HITS (Hyperlink-Induced Topic Search)

HITS classifies pages into:

- **Authorities**: pages providing valuable content
- **Hubs**: pages that link to good authorities

Each page receives two scores (authority and hub) through mutual reinforcement.

Authority update

$$a(p) = \sum_{q \in I(p)} h(q)$$

Hub update

$$h(p) = \sum_{q \in O(p)} a(q)$$

Where:

- **I(p)** = pages linking *to* p
- **O(p)** = pages p links *to*

V. Research Gaps as stated in the existing literature.

1 Stemming and Lemmatizer:

To minimize the inflection form, Stemming is the process to generate root words from inflected words; the generated words are derived form to deduce it to the common base form. The Lemmatizer generates the root words from the actual word but in stemming the stem might not be an actual word. In normalization process both has its own equal importance at the time of canonical representation for set of related words.

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Table 1 Difference between Stemming and Lemmatization.

VI. Conclusion

Ranking of web documents based on content remains a critical field of research because of the massive increase in online information and the demand of more people to get accurate and meaningful search results. The current literature shows that a lot of progress has been made by using key word based models, link based models, semantic model and the sparse/ dense retrieval frameworks. Nevertheless, there are still constraints in the contextual meaning awareness, dynamic content processing, trade-off considerations in computation and heterogeneous ranking signal integration. The survey of the existing approaches reveals that although some of them are better in terms of accuracy and relevance, none of them is effective in terms of the combination of content semantics, scale, and real-time flexibility. Thus, the research area still needs new combined methods that will be able to smartly combine several ranking parameters to provide more stable and user-friendly organization of documents.

overall common methodology and its improvement can be designed as per shown in the below block diagram.

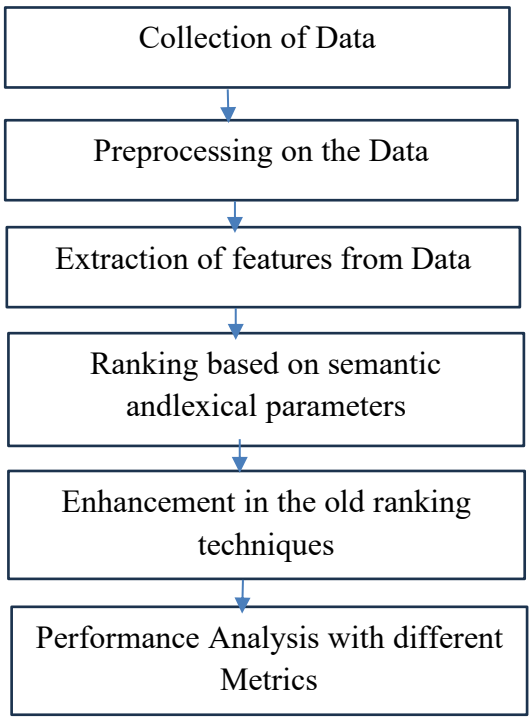

Fig. Overall Steps Included in proposed methods are shown

The common methodology steps are explained below as,

a) Collection of Data

The metadata, html pages etc web documents can be downloaded using scrapy.

b) Preprocessing on the Data

Preprocessing includes ads removal etc which is short make pages clean without noise after that web documents can be stored in format of indexable.

c) Extraction of features from Data

Multiple features can be extracted from the web documents such as metadata features, structural features, semantic features, lexical features etc.

d) Ranking based on semantic and lexical parameters

Different models can be designed and integrated for ranking such as lexical rankers TF-IDF, BM25, etc. semantic rankers such as SBERT / BERT etc.

e) Enhancement in the old ranking techniques

Additional features which include semantic expansion, keyword density, content similarity are considered. Graph based, query-based ranking also improves the results.

f) Performance Analysis with different Metrics

Performance metrics such as MAP (mean average precision), Precision@k, Recall@k can be evaluated to check the performance of different algorithms.

The overall structure of the steps included in content-based ranking for web documents is given in above steps.

REFERENCES

[1] Prakash, A., Gupta, S. K., & Rawat, M. (2022). *Design and Implementation of Novel Techniques for Content-Based Ranking of Web Documents*. In Process Mining Techniques for Pattern Recognition (pp. 35–45). CRC Press.

[2] Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., & Gonçalves, M. A. (2003). Combining link-based and content-based methods for web document classification. *Proceedings of CIKM*, 394–401.

[3] Formal, T., Lassance, C., Piwowarski, B., & Clinchant, S. (2021). SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. *arXiv*.

[4] Formal, T., Lassance, C., Piwowarski, B., & Clinchant, S. (2021). SPLADE v2: Improved Approaches for Sparse Lexical Expansion. *arXiv*.

[5] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP*.

[6] Ma, X., Sun, K., Pradeep, R., & Lin, J. (2021). A Replication Study of Dense Passage Retriever. *arXiv*.

[7] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.-X., Dong, D., Wu, H., & Wang, H. (2020). RocketQA: An Optimized Training Approach for Dense Retrieval. *arXiv*.

[8] Zhuang, S., & Zuccon, G. (2021). Dealing with Typos for BERT-based Passage Retrieval and Ranking. *arXiv*.

[9] Donmez, P., Svore, K. M., & Burges, C. J. C. (2008). On the Optimality of LambdaRank. *Microsoft Research Technical Report*.

[10] Wu, Q., Burges, C. J. C., Svore, K. M., & Gao, J. (2010). Learning to Rank using LambdaMART. *Microsoft Research*.

[11] Dudek, J., Kong, W., Li, C., Zhang, M., & Bendersky, M. (2023). Learning Sparse Lexical Representations over Expanded Vocabularies for Retrieval. *CIKM*.

[12] Zhen, F., & Callan, J. (2023). CSurF: Sparse Lexical Retrieval through Contextualized Surface Representation. *ICTIR*.