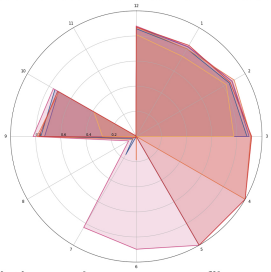
Graphical Abstract

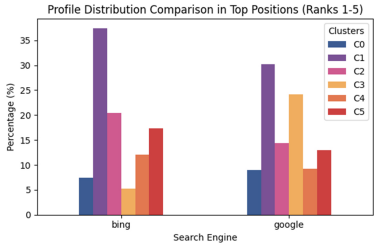
Characterizing Ranking Environments: An Empirical Study of Technical and Content Attributes for System Profiling

1. Diverse Resource Profiles



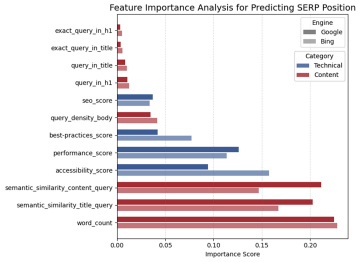
Six distinct web resource profiles emerged from clustering based on technical and content features (RQ1).

2. Engines Prefer Different Profiles



Google and Bing show significantly different profile distributions in top search results (RQ3).

3. Content Factors More Influential



Content attributes generally show greater aggregate importance for ranking visibility than technical factors (RQ4).

Highlights

Characterizing Ranking Environments: An Empirical Study of Technical and Content Attributes for System Profiling

- Content factors show stronger aggregate link to SERP visibility than technical scores for Google (70.1%) and Bing (61.8%).
- Google and Bing demonstrate significant differences in top-page attributes and inferred factor priorities.
- Counterintuitive Google finding: High title-query semantic similarity linked to worse ranking (multivariate models).
- Introduced a comparative profiling methodology for search engines using content-homogeneous data.

Characterizing Ranking Environments: An Empirical Study of Technical and Content Attributes for System Profiling

Abstract

This study empirically characterized and compared Google's and Bing's ranking environments, analyzing technical performance (Lighthouse scores) and content attributes (lexical/semantic features) within a homogeneous commercial discount domain. Using 14 465 SERP (Search Engine Results Page) items from 500 queries and 12 features, K-Means clustering identified six distinct web resource profiles, significantly associated with ranking tiers in both systems. Content-related attributes demonstrated a stronger aggregate association with visibility than technical scores for both Google (70.1% importance) and Bing (61.8%). Comparative analysis showed Bing's top-ranking results generally featured higher median values across numerous technical/content metrics than Google's. Notably for Google, high title-query semantic similarity unexpectedly associated with worse ranking odds in multivariate models, contrasting with positive main content-query similarity effects. Bing, conversely, prioritized content volume (word count) and explicit keyword signals more. These findings highlight system-specific nuances in factor weighting, contributing to IR system understanding and offering practical optimization insights. The dataset and analysis code will be made available upon publication.

Keywords: Ranking Factors, System Profiling, Technical Performance, Content Relevance, Semantic Similarity, Comparative Analysis

1. Introduction

1.1. Background

In contemporary digital society, large-scale information retrieval systems are primary gateways to vast online resources, fundamentally shaping how users discover and interact with information (Coghlan ., 2025). Achieving visibility within these complex ranking environments is crucial for diverse online entities, from commercial enterprises to educational platforms (Daly Ryan, 2024; Limongelli ., 2022). Ultimately, the success of these retrieval systems hinges on user satisfaction (Wang .,

2023), a critical metric that reflects how well they meet diverse user needs and expectations. However, the algorithms governing visibility are notoriously complex and opaque, often described as "black boxes," making it challenging for resource creators and researchers to understand the precise factors driving organic ranking outcomes (Coghlan ., 2025; Roumeliotis Tselikas, 2022). This opacity also poses challenges for evaluating search engine effectiveness, particularly in specialized domains like e-commerce, where traditional metrics may not fully capture user experience, prompting calls for more user-oriented evaluation approaches (Moffat, 2024). Consequently, there is a growing interest in empirical approaches to infer the implicit priorities and behavioral patterns of these influential systems – a form of "reverse engineering" based on observational data.

1.2. Problem Statement

A significant challenge in inferring system priorities is disentangling the influence of myriad factors. This is particularly true for the interplay between a resource's intrinsic content characteristics (e.g., relevance, quality, structure) and its technical performance attributes (e.g., loading speed, accessibility) (Nagpal Petersen, 2021; Srinivas Gowda, 2025). The sheer heterogeneity of content across the web often confounds attempts to isolate the specific impact of technical factors. Furthermore, while many studies investigate ranking factors within a single retrieval system (often implicitly Google) or focus on developing novel ranking algorithms to advance the state-of-the-art, large-scale comparative analyses characterizing the behavior of dominant, operational search engines remain relatively scarce. Such analyses are crucial for understanding how *different* major systems (e.g., Google vs. Bing) actually weigh various technical and content attributes in practice (Toms Taves, 2004).

This study addresses this specific gap. It is an empirical investigation aimed at characterizing and profiling existing systems, rather than proposing a new ranking algorithm to outperform state-of-the-art models. It employed a methodological approach designed to mitigate content variance by analyzing a large dataset of online resources from a relatively homogenous domain – specifically, pages related to commercial discounts and promotional offers. This content similarity provided a controlled environment, allowing for better isolation of technical performance effects and exploration of potential system-specific preferences for content presentation or semantic relevance, thereby offering insights into the "black-box" nature of these widely-used information retrieval systems. Thus, the 'baselines' in this study refer to the existing operational search engines themselves, whose behaviors we aim to profile and compare.

1.3. Research Objectives

This study aims to empirically characterize and compare the ranking environments of Google (System A) and Bing (System B) within a homogeneous domain. The specific research objectives designed to achieve this aim are:

1. To identify distinct profiles (clusters) of online resources based on a combination of their technical performance metrics (Lighthouse scores) and content characteristics (lexical and semantic features).
2. To determine if resources achieving higher SERP (Search Engine Results Page) visibility in Google and Bing are disproportionately associated with specific identified resource profiles.
3. To comparatively analyze the profiles and attribute distributions of top-ranking resources (e.g., top 5 positions) between Google and Bing for the same user queries.
4. To assess the relative strength of association between technical performance attributes versus content attributes and higher SERP visibility, comparing these patterns between Google and Bing.

1.4. Research Questions (RQs)

This study sought to answer the following research questions:

- **RQ1 (Profiling & Clustering of Online Resources):** Based on content relevance and Lighthouse metrics, what distinct profiles (clusters) emerged among the analyzed online resources? What were their defining characteristics?
- **RQ2 (Ranking Patterns and Profile Associations):** Did resources ranking higher (e.g., top 5 vs. 6-20) in System A and System B tend to belong to specific profiles/clusters? How did the profiles of top-ranked resources differ from lower-ranked ones?
- **RQ3 (System Differences):** How did the profiles of top-ranking resources differ between System A and System B for the same queries? Which content relevance or technical performance metrics appeared to be prioritized differently by these two retrieval systems within this domain?
- **RQ4 (Technical vs. Content Weight):** Within the analyzed resource set, did technical performance metrics (measured by Lighthouse) or content relevance metrics show a stronger association with higher visibility rankings or specific 'successful' profiles? Did this differ between System A and System B?

1.5. Contribution

This research offers several contributions primarily focused on the empirical characterization of existing large-scale information retrieval systems, rather than the development of new state-of-the-art ranking algorithms. Methodologically, it presents a large-scale approach for comparative system profiling using a content-homogeneous dataset as a control variable to better isolate the effects of technical factors and content presentation strategies (Jayaraman ., 2022; Çırakoğlu Koşaner, 2024). Empirically, the study provides robust, data-driven insights into the apparent ranking behaviors of Google (System A) and Bing (System B) concerning key technical and content attributes within a commercially relevant domain. Theoretically and practically, the findings deepen the understanding of the differing operational logics of these major information retrieval systems, particularly highlighting counter-intuitive factor weightings under multivariate conditions, and offer potential insights for creators seeking to optimize resource visibility across different digital environments. The dataset and analysis scripts will be made available to foster reproducibility and enable further research by the community upon acceptance of the manuscript.

1.6. Paper Structure

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the methodology. Section 4 presents the results. Section 5 discusses these findings. Section 6 outlines the availability of the data and code. Finally, Section 7 concludes the paper and suggests future research.

2. Literature Review and Related Work

2.1. Factors Influencing Organic Visibility

2.1.1. Technical Performance and Resource Structure

The technical underpinnings of online resources have long been recognized as crucial for visibility in retrieval systems. Beyond basic crawlability and indexability, factors like site speed, mobile-friendliness, security (HTTPS), and overall user experience related to performance are increasingly emphasized (Roumeliotis Tselikas, 2022). This scope includes not only server-side performance and accessibility but also structural elements such as URL design. URL design can influence both user perception and machine interpretability for SEO (Rastakhiz ., 2024). Retrieval systems also invest heavily in identifying and penalizing manipulative or low-quality content. This task shares challenges with areas like phishing detection, where hybrid feature models are employed for robustness (Zhu ., 2024). Tools like Google's Lighthouse provide standardized metrics for performance, accessibility, adherence to

best practices, and basic technical SEO checks. These metrics offer quantifiable indicators of technical health. Furthermore, concepts derived from link analysis, such as PageRank, remain foundational in understanding resource authority and trustworthiness (Srinivas Gowda, 2025; Toms Taves, 2004). Ongoing research continues to develop sophisticated models to assess website trust and combat misinformation (Niu ., 2022).

2.1.2. Content Signals and Semantic Relevance

While technical aspects form the foundation, content remains a primary driver of relevance. Early approaches focused heavily on keyword matching and density (Srinivas Gowda, 2025). Modern retrieval systems, however, employ sophisticated Natural Language Processing (NLP) techniques moving beyond lexical matching towards semantic understanding (Jindal ., 2014). This is crucial as users are often driven by a desire to reduce uncertainty or engage with interesting content (van der Sluis, 2025), motivations that are better served by semantically rich and relevant results rather than mere keyword presence. Techniques like Latent Semantic Analysis (LSA) (Nagpal Petersen, 2021), topic modeling, and more recently, vector embeddings from deep learning models like Sentence-BERT, allow for capturing semantic similarity even without exact keyword overlap (Pauzi Capiluppi, 2023). Understanding user intent and matching it with semantically relevant content is paramount (Jindal ., 2014; Limongelli ., 2022).

2.1.3. Measuring Content Quality and User Experience Signals

Defining and measuring "content quality" is inherently complex. Beyond relevance, factors such as readability, structure, originality, depth, and authority contribute to quality perceptions. Linguistic frameworks, like applying Grice's conversational maxims, offer structured ways to evaluate content clarity, informativeness, and appropriateness (Çırakoğlu Koşaner, 2024). Retrieval systems may also incorporate implicit user feedback signals (e.g., click-through rates, dwell time) as proxies for content quality and user satisfaction. However, these signals are harder for external researchers to measure directly.

2.1.4. Synthesizing Technical and Content Factors

Understanding the interplay between technical performance and content quality/relevance is crucial. Some studies suggest a trade-off or differing importance depending on context. Examples include the stage of the user journey (Nagpal Petersen, 2021) or the specific retrieval system being analyzed (Toms Taves, 2004). Modeling approaches sometimes attempt to integrate these diverse factors to predict

ranking outcomes or system behavior (Jayaraman ., 2022). The relative weighting assigned by different systems remains an active area of investigation.

2.2. Profiling and Characterizing Online Resources

Given the diversity of online resources, methods for grouping or classifying them based on shared characteristics are valuable. Clustering algorithms (e.g., K-Means, Hierarchical) can identify emergent profiles. These profiles are based on quantitative features like performance scores or content metrics (Yu ., 2024). Content analysis, including automated dictionary-based methods, can classify resources based on their topical focus or expressed attributes (Milei ., 2025). Such profiling helps in understanding the different strategies or archetypes present within a specific domain or result set.

2.3. Comparative Analysis of Information Retrieval Systems

While Google (System A) dominates market share, Bing (System B) and others represent significant alternative access points to information. Comparing their ranking behaviors is important for a comprehensive understanding of the information ecosystem. Studies have historically noted differences in how systems weigh factors like link structure versus content (Toms Taves, 2004). Understanding these differences is crucial for creators aiming for broad visibility. Furthermore, ethical considerations and potential biases embedded within the ranking logic of different systems are increasingly important topics (Coghlan ., 2025).

2.4. Reverse Engineering Approaches in Information Retrieval

Due to the proprietary nature of commercial ranking algorithms, researchers often resort to "black-box" analysis or reverse engineering. This involves systematically correlating observable input features (resource attributes) with observable output behavior (rankings). The goal is to infer potential algorithmic priorities. While unable to reveal exact algorithms, such empirical studies can provide valuable insights. They highlight factors strongly associated with higher visibility within specific system environments (Bardas ., 2025). Modern approaches increasingly explore sophisticated optimization objectives beyond simple relevance. One example is incorporating risk-sensitivity into deep learning ranking models to improve the robustness and fairness of outcomes (Silva Rodrigues ., 2025).

2.5. Research Gap

Despite extensive research on individual ranking factors and retrieval systems, a gap exists. Specifically, there is a need for large-scale comparative system profiling. This study aims to fill this gap by not only characterizing and contrasting

the ranking environments of System A and System B within a specific commercial context but also by employing a methodology that combines modern technical audits (Lighthouse) with advanced semantic feature analysis (Sentence-BERT) on a content-homogeneous dataset to reveal nuanced factor prioritizations and system-specific profiles. This can better isolate technical effects and system-specific content preferences in comparative system profiling.

3. Methodology

This section details the systematic approach undertaken to collect, process, and prepare the data. The goal was to characterize and compare the ranking environments of two major information retrieval systems.

3.1. Research Design

This study employed a quantitative, observational research design. Data pertaining to resource visibility (rankings) and associated attributes (technical performance, content features) were collected programmatically from Google (System A) and Bing (System B). Subsequent analysis involved data mining and statistical techniques to profile resources and infer system priorities.

3.2. Data Collection

Data collection occurred around April 17, 2025.

3.2.1. Query Set Preparation

The query set comprised 500 unique English search queries. These queries focused on online commercial promotions and discounts. They were curated to balance broad coverage with reduced semantic redundancy (e.g., '[Brand] promo code'). The full list is available in the accompanying dataset (`keywords.csv`). The dataset will be made publicly available upon acceptance of the manuscript. This domain homogeneity was a methodological choice to control for content variance.

3.2.2. Retrieval System Data Acquisition

For each query, the top $\sim$20 organic results were retrieved from Google (via Google Custom Search API) and Bing (via Bing Web Search API). Default API settings were used to minimize personalization. Raw JSON responses (URLs, titles, snippets, rank positions) were saved.

3.2.3. Resource Content Acquisition

Unique URLs were visited using a headless browser (Puppeteer via `pyppeteer`). This process saved full HTML content and captured screenshots. Access/rendering errors were logged.

3.3. Feature Engineering

3.3.1. Main Content Extraction

The `trafilatura` library (Barbaresi, 2021) extracted the main textual content from saved HTML, excluding boilerplate.

3.3.2. Technical Performance Features (Lighthouse)

The Google PageSpeed Insights API provided four core Lighthouse scores (Performance, Accessibility, Best Practices, SEO) for each URL (Roumeliotis Tselikas, 2022).

3.3.3. Content Relevance and Attribute Features

Eight content features were extracted for each query-URL pair:

1. **Lexical Features:** *query_in_title* (all query words in HTML title), *exact_query_in_title* (exact query string in title), *query_in_h1* (all query words in first `<h1>`), *exact_query_in_h1* (exact query string in first `<h1>`), *query_density_body* (percentage of query string in main text), and *word_count* (alphabetic words in main text).
2. **Semantic Similarity Features:** Using Sentence-BERT ('all-mpnet-base-v2' (Reimers Gurevych, 2019; Song ., 2020)), cosine similarity was calculated for *semantic_similarity_title_query* (query vs. HTML title) and *semantic_similarity_content_query* (query vs. main text). Long texts for the latter were chunked (Beltagy ., 2020; Devlin ., 2019).

All feature extraction was performed using custom Python scripts. These scripts leveraged libraries such as BeautifulSoup for HTML parsing, NLTK for tokenization, and Trafilatura for main content extraction. The analysis scripts will be made publicly available upon acceptance of the manuscript.

3.4. Dataset Description and Preprocessing

This study utilizes a custom-compiled dataset specifically designed for the empirical characterization and comparison of the ranking environments of Google (System

A) and Bing (System B). To ensure a controlled analytical environment by mitigating content variance, the dataset exclusively comprises web resources from a homogeneous domain: online commercial promotions and discount offers (as detailed in Section 3.2.1).

The data collection pipeline, which involved querying for 500 unique English keywords (focused on said domain) across both search engines, initially yielded approximately 19 950 raw SERP results. Following the programmatic acquisition of webpage content, comprehensive feature engineering (see Section 3.3 for details on the 12 technical and content features extracted), and necessary data cleaning and filtering (e.g., exclusion of ∼1.25% of entries with missing Lighthouse data), the **final dataset for analysis comprises** 14 465 **individual SERP result items**. Each item in this dataset represents a unique query-URL pair, annotated with its organic search rank position and the suite of extracted features. The structure of these entries and feature descriptions are detailed in Table 2.

The dataset encompasses 6929 **unique URLs** from 2238 **unique hostnames**. The distribution of SERP items is 5895 for Google and 8570 for Bing. It is pertinent to note that the *pwa_score* Lighthouse feature was entirely removed from the feature set due to 100% missing values across all collected data points.

Descriptive statistics for all numerical features in the final dataset, including an outlier analysis, are presented in Table 1. While some features exhibited outliers (e.g., $performance\_score$: 5.45%; $exact\_query\_in\_title$: 16.63%; $query\_density\_body$: 12.09%), these were generally retained to reflect real-world data variability. Robust statistical methods were employed in subsequent analyses where appropriate to account for these distributions. For multivariate analyses, including K-Means clustering and Ordinal Logistic Regression, all numerical features were Min-Max scaled to a [0,1] range to ensure equitable feature contribution and improve model performance.

3.5. Data Analysis Strategy

The preprocessed dataset was analyzed using Python (v3.9+). Key libraries included Pandas, NumPy, Scikit-learn, Statsmodels, and SciPy. The significance level α was set to 0.05. Numerical features were Min-Max scaled to [0,1] for multivariate analyses.

3.5.1. RQ1: Resource Profiling and Clustering

K-Means clustering was applied to 12 scaled technical and content features. The optimal number of clusters, $K = 6$, was determined using the Elbow method and Silhouette analysis (Figure 1). Clusters were characterized by their mean feature values (Table 3; Figure 2). Kruskal-Wallis H-tests (Table 4) and Random Forest Classifier feature importances identified discriminative features.

Table 1: Descriptive Statistics of the Final Analyzed Dataset (N=14 465)

Metric	Mean	Median	SD	Min	Max	Q1	Q3	IQR%
SERP Position								
Overall	10.15	10.00	5.67	1	20	5.00	15.00	–
Lighthouse Scores								
Perf.	87.83	95.00	15.04	12	100	82.00	99.00	5.45
Access.	87.36	90.00	8.45	34	100	83.00	94.00	2.96
Best Prac.	91.75	96.00	12.09	37	100	85.00	100.00	4.14
SEO	93.00	92.00	7.43	40	100	92.00	100.00	3.21
Content Relevance Metrics								
Q-Title	0.41	0.00	0.49	0.00	1.00	0.00	1.00	0.00
Q-H1	0.38	0.00	0.49	0.00	1.00	0.00	1.00	0.00
ExQ-Title	0.17	0.00	0.37	0.00	1.00	0.00	0.00	16.63
ExQ-H1	0.15	0.00	0.36	0.00	1.00	0.00	0.00	14.88
Q/B Density	0.19	0.00	0.46	0.00	7.73	0.00	0.20	12.09
Sim. Title	0.68	0.76	0.22	−0.02	1.00	0.65	0.81	12.22
Sim. Content	0.58	0.62	0.18	−0.04	0.88	0.49	0.72	2.24
Word Count	628.56	428.00	1294.32	0.00	132 115.00	168.00	795.00	6.24

Note: SD = Standard Deviation; Q1 = First Quartile; Q3 = Third Quartile. IQR% indicates percentage of outliers via IQR method. Metric abbreviations: Perf. (Performance Score), Access. (Accessibility Score), Best Prac. (Best Practices Score), SEO (SEO Score), Q-Title (Query in Title), Q-H1 (Query in H1), ExQ-Title (Exact Query in Title), ExQ-H1 (Exact Query in H1), Q/B Density (Query Density Body), Sim. Title (Semantic Similarity Title-Query), Sim. Content (Semantic Similarity Content-Query).

Table 2: Dataset Column Types and Descriptions

Category	Column Name	Description/Type
Search Engine	engine	Categorical (google/bing)
Search Engine	position	Integer (1-20)
Lighthouse	performance_score	Float (0-100)
Lighthouse	accessibility_score	Float (0-100)
Lighthouse	best-practices_score	Float (0-100)
Lighthouse	seo_score	Float (0-100)
Content	query_in_title	Integer (0/1)
Content	query_in_h1	Integer (0/1)
Content	exact_query_in_title	Integer (0/1)
Content	exact_query_in_h1	Integer (0/1)
Content	query_density_body	Float (%)
Content	semantic_similarity_title_query	Float (0-1, Cosine)
Content	semantic_similarity_content_query	Float (0-1, Cosine)
Content	word_count	Integer

3.5.2. RQ2: Visibility Patterns versus Profiles

Resources were grouped into High (1-5), Medium (6-10), and Low (11-20) rank tiers for Google and Bing separately.

- **Profile Distribution:** Pearson's χ^2 test assessed the association between profiles and rank tiers.
- **Feature Comparison:** Kruskal-Wallis H-test (with Dunn's post-hoc) compared median feature values across rank tiers (Table 5).

3.5.3. RQ3: Comparative System Analysis (Google vs. Bing)

Top-ranking (1-5) resources were compared between Google and Bing.

- **Profile Comparison:** Pearson's χ^2 test compared profile distributions.
- **Feature Comparison:** Mann-Whitney U test compared median feature values. Cohen's d estimated the effect size.

3.5.4. RQ4: Inferring Factor Priorities (Technical vs. Content)

Several methods were used:

- **Correlation Analysis:** Spearman's ρ correlated features with SERP *position*.

- **Ordinal Logistic Regression:** A Proportional Odds Logit model (‘statsmodels.miscmodels.ordinal_model.OrderedModel‘) predicted SERP quintiles (0=best to 4=worst). This was done using technical-only, content-only, and combined scaled feature sets. Model fit (Pseudo R^2, AIC, BIC) and coefficient significance/direction were examined (Nagpal Petersen, 2021; Toms Taves, 2004). A negative coefficient indicated a higher probability of a better rank. (Full models are in Appendix A.1).

- **Random Forest Feature Importance:** A Regressor predicted numerical SERP position. This was used to derive Gini importance for features, aggregated for technical versus content factors.

Non-parametric tests were used where parametric assumptions were violated.

4. Results

This section presents empirical findings from the dataset analysis.

4.1. RQ1: Derived Resource Profiles

K-Means clustering on 12 scaled features identified six distinct resource profiles. The optimal $K = 6$ (Silhouette score: 0.443) is shown in Figure 1. All features differed significantly across clusters (Kruskal-Wallis, $p < .001$; Table 4). Mean scaled feature values for each cluster are in Table 3 and visualized in Figure 2. Key profiles include:

- **Cluster 2 ("High Keyword & Semantic Relevance, Strong Technicals"; 16.3%):** Excelled across most metrics.

- **Cluster 3 ("Low Relevance, Low Technicals"; 16.0%):** Performed poorly (e.g., *semantic_similarity_title_query* mean scaled: 0.252; *word_count* median: 61).

Other clusters (0, 1, 4, 5) showed varied specializations. *Query_in_title*, *query_in_h*1, and *semantic_similarity_title_query* were most influential in cluster differentiation.

4.2. RQ2: Visibility Patterns and Profiles

Profile Distribution across Ranks: A significant association was found between cluster membership and ranking tiers for both engines (Pearson's $\chi^2, p < .001$; Figure 3). For Google, Profile 3 ("Low Relevance, Low Technicals") was more prevalent in lower ranks. For Bing, Profiles 1 and 2 were more prominent in higher ranks.

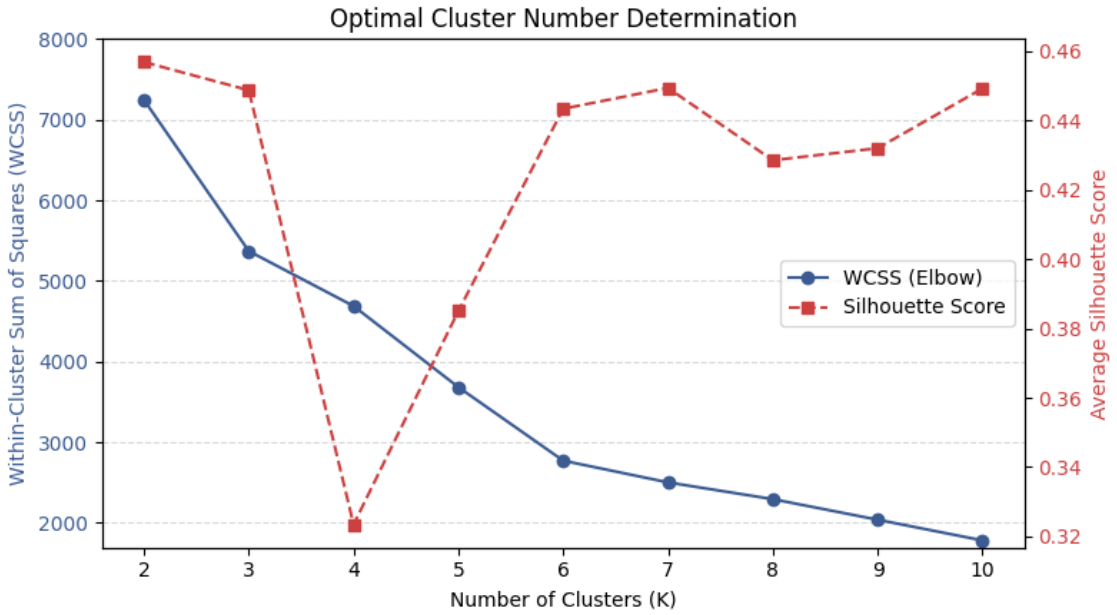


Figure 1: Determination of optimal cluster number (K) using Elbow method (left Y-axis, WCSS) and Silhouette analysis (right Y-axis, Average Silhouette Score). $K = 6$ was selected.

Table 3: Mean Scaled Feature Values for the Six Identified Cluster Profiles (RQ1)

Feature (Scaled)	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**
Perf.	0.858	0.870	0.879	0.811	0.870	0.872
Access.	0.806	0.823	0.827	0.736	0.805	0.838
Best Prac.	0.853	0.879	0.882	0.826	0.899	0.863
SEO	0.880	0.897	0.910	0.776	0.911	0.918
Q-Title	0.000	0.000	1.000	0.000	1.000	1.000
Q-H1	1.000	0.000	1.000	0.000	0.000	1.000
ExQ-Title	0.000	0.000	0.898	0.000	0.187	0.000
ExQ-H1	0.166	0.000	0.836	0.000	0.000	0.000
Q/B Density	0.031	0.010	0.070	0.000	0.046	0.012
Sim. Title	0.759	0.729	0.821	0.269	0.790	0.774
Sim. Content	0.747	0.716	0.762	0.389	0.711	0.722
Word Count	0.005	0.005	0.005	0.003	0.003	0.006
Cluster Size (N)	1054	5142	2364	2314	1522	2069
Percentage (%)	7.3	35.5	16.3	16.0	10.5	14.3

Note: C0-C5 represent Cluster 0 to Cluster 5. Features were Min-Max scaled to [0,1]. Metric abbreviations are defined in Table 1.

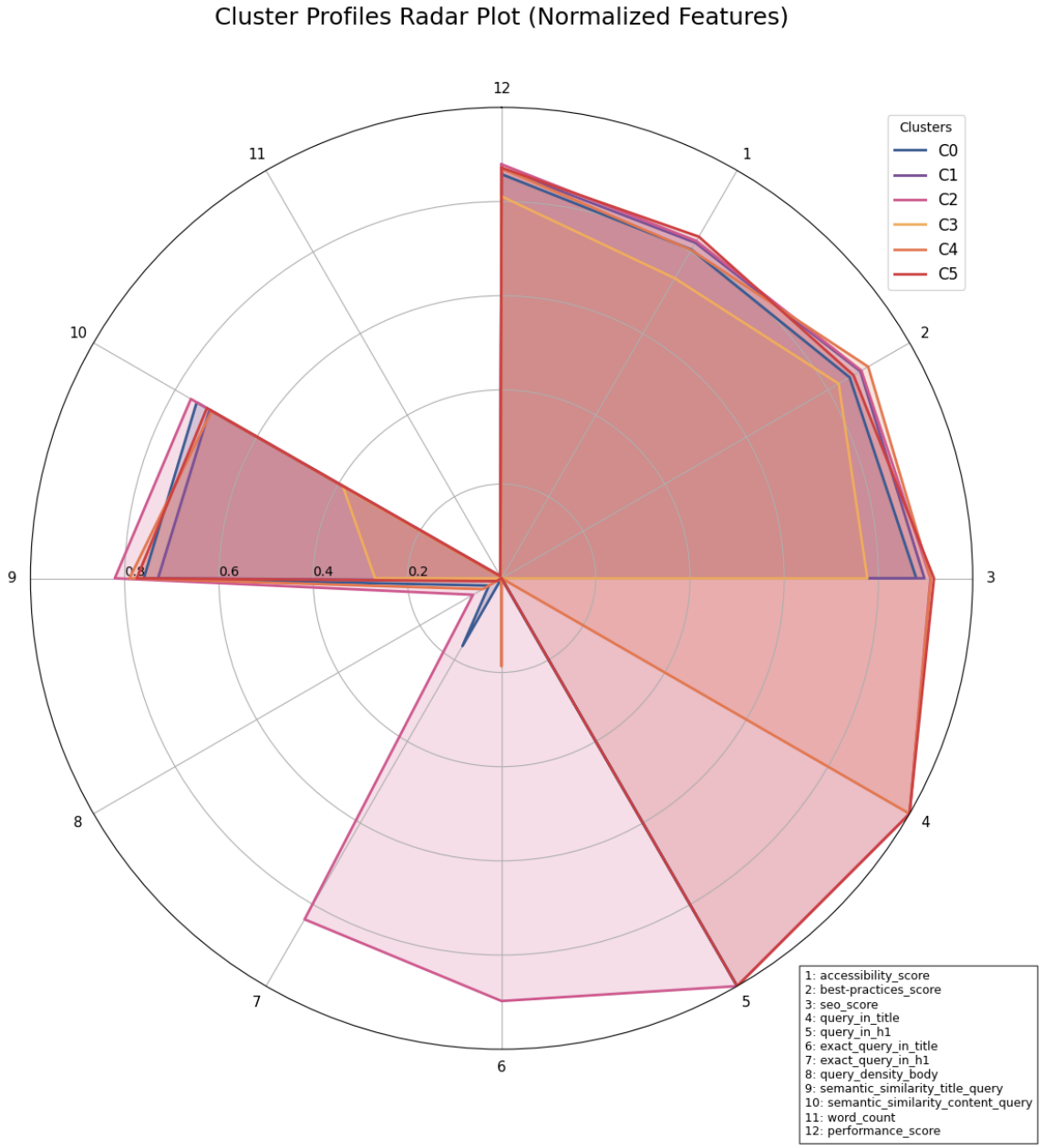


Figure 2: Radar plot of mean scaled feature values for the six cluster profiles (RQ1).

Table 4: Kruskal-Wallis H-Test Results for Feature Differentiation Across Clusters (RQ1)

Feature	**H-statistic**	**p-value**
Perf.	231.268	< 0.001
Access.	919.138	< 0.001
Best Prac.	172.419	< 0.001
SEO	1736.022	< 0.001
Q-Title	14 464.000	< 0.001
Q-H1	14 464.000	< 0.001
ExQ-Title	11 231.757	< 0.001
ExQ-H1	10 756.173	< 0.001
Q/B Density	4309.632	< 0.001
Sim. Title	7576.607	< 0.001
Sim. Content	4676.944	< 0.001
Word Count	2001.006	< 0.001

All p-values reported as $< .001$ were originally 0.0 in the source data.
Metric abbreviations defined in Table 1.

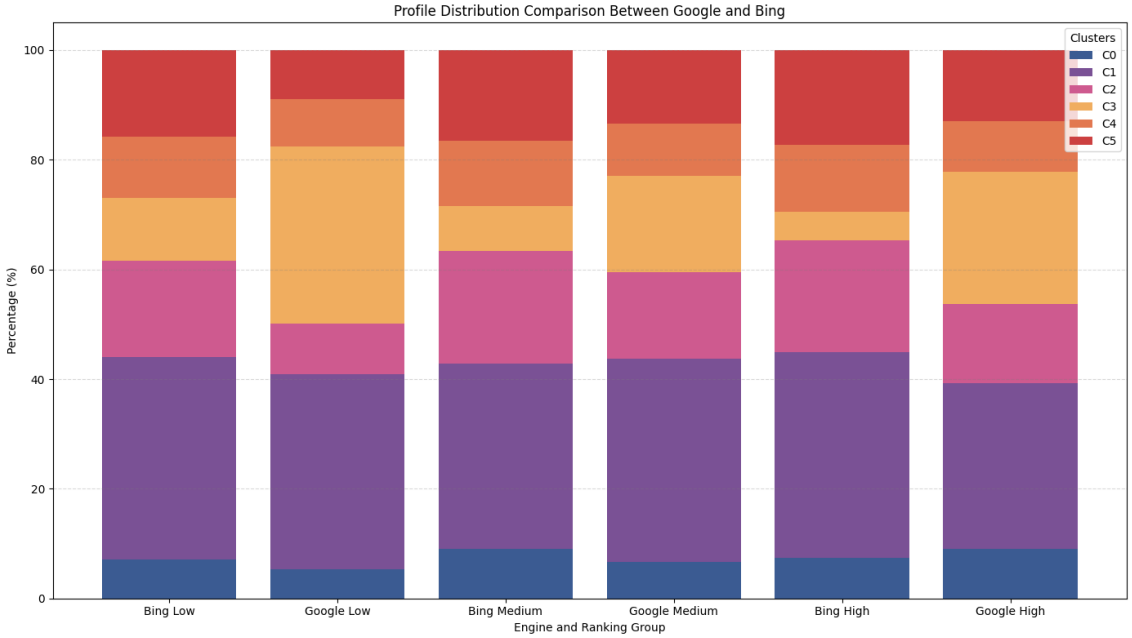


Figure 3: Profile Distribution Across Ranking Groups for Google and Bing (RQ2).

Feature Comparison across Ranks: Kruskal-Wallis H-tests (Table 5) indicated significant differences ($p < .05$) for all Lighthouse scores and most content metrics across ranking groups for both systems.

- For Google, higher median *semantic_similarity_content_query* (High: 0.606) and *accessibility_score* (High: 91.0) were associated with better ranks.
- For Bing, *accessibility_score* (High: 91.0), *semantic_similarity_title_query* (High: 0.785), and *word_count* (High: 630) showed clearer positive associations with better ranking tiers.

Table 5: Kruskal-Wallis H-Test Results for Features Across Ranking Tiers (High, Medium, Low) for Google and Bing (RQ2)

	Google (System A)		**Bing (System B)**	
Feature	**H-statistic**	**p-value**	**H-statistic**	**p-value**
Lighthouse Scores				
Perf.	35.896	< 0.001	86.829	< 0.001
Access.	29.247	< 0.001	359.950	< 0.001
Best Prac.	27.224	< 0.001	1.773	0.412
SEO	102.804	< 0.001	32.794	< 0.001
Content Relevance Metrics				
Q-Title	79.726	< 0.001	21.592	< 0.001
Q-H1	109.376	< 0.001	23.123	< 0.001
ExQ-Title	45.916	< 0.001	12.440	0.002
ExQ-H1	43.171	< 0.001	33.618	< 0.001
Q/B Density	91.336	< 0.001	127.544	< 0.001
Sim. Title	207.318	< 0.001	89.876	< 0.001
Sim. Content	242.241	< 0.001	9.394	0.009
Word Count	30.762	< 0.001	595.954	< 0.001

Ranking Tiers: High (1-5), Medium (6-10), Low (11-20). Dunn's test for post-hoc comparisons not shown here. Metric abbreviations defined in Table 1.

4.3. RQ3: Comparative System Analysis (Google vs. Bing)

Profile Comparison in Top Ranks: A significant difference ($\chi^2(5) \approx 302.26, p < .001$) was found in profile distribution within the top 5 ranks (Figure 4). Cluster

1 ("General Content, Average Technicals") was more prevalent in Bing's top 5. In contrast, Cluster 3 ("Low Relevance, Low Technicals") was more so in Google's.

Feature Comparison in Top Ranks: Bing's top-ranking pages generally had significantly higher median scores for most features (Mann-Whitney U, $p < .005$; Table 6). These included *semantic_similarity_title_query* (medium effect size) and *word_count* (small effect size).

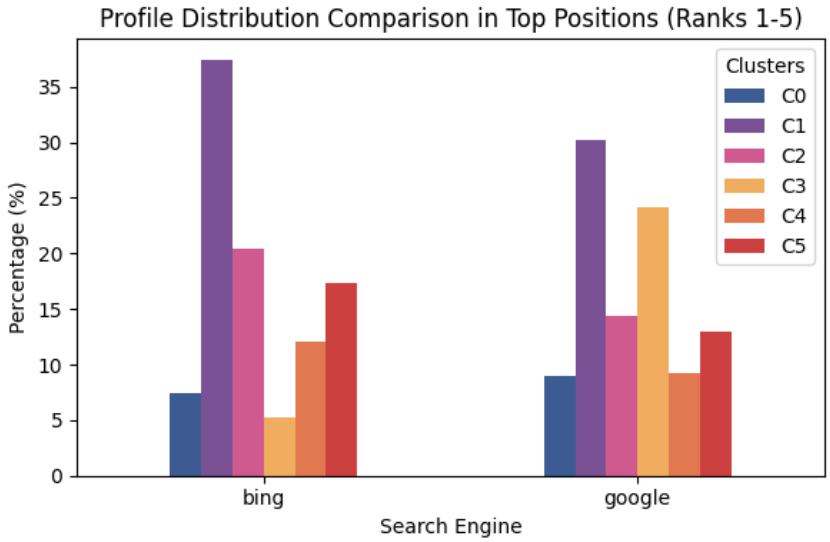


Figure 4: Comparison of Profile Distributions in Top 5 Ranks: Google vs. Bing (RQ3).

4.4. RQ4: Inferred Factor Priorities

Correlation Analysis: Spearman rank correlations are detailed in Appendix Appendix A.2 (Figure A.6).

- For Google, *semantic_similarity_content_query* ($\rho = -0.221$) and *semantic_similarity_title_query* ($\rho = -0.179$) showed the strongest significant negative correlations with *position*.
- For Bing, *word_count* ($\rho = -0.257$) and *query_density_body* ($\rho = -0.133$) were most notable. All these correlations had $p < .001$.

Ordinal Logistic Regression: (Full models are in Appendix A.1).

Table 6: Comparison of Median Feature Values for Top-Ranking (1-5) Pages between Google and Bing (RQ3)

Feature	**Median Google**	**Median Bing**	**p-value**	**Cohen's d**
Lighthouse Scores (0-100)				
Perf.	90.0	92.0	0.002**	−0.195
Access.	91.0	91.0	< 0.001***	−0.292
Best Prac.	96.0	96.0	0.001**	0.029
SEO	92.0	92.0	< 0.001***	−0.501
Content Relevance Metrics				
Q-Title (0/1)[a]	0.0	0.0	< 0.001***	−0.271
Q-H1 (0/1)[a]	0.0	0.0	< 0.001***	−0.180
ExQ-Title (0/1)[a]	0.0	0.0	< 0.001***	−0.140
ExQ-H1 (0/1)[a]	0.0	0.0	< 0.001***	−0.164
Q/B Density (%)	0.0	0.0	< 0.001***	−0.187
Sim. Title (0-1)	0.759	0.785	< 0.001***	−0.548
Sim. Content (0-1)	0.606	0.657	< 0.001***	−0.360
Word Count	567.0	630.0	< 0.001***	−0.304

Mann-Whitney U test used. p-values: * $p < .05$; ** $p < .01$; *** $p < .001$.
Cohen's d: negligible ($|d| < 0.2$), small ($0.2 \leq |d| < 0.5$), medium ($0.5 \leq |d| < 0.8$).
[a] For binary features, medians are shown; mean percentages discussed in text. Metric abbreviations defined in Table 1.

- For Google (Combined Model Pseudo $R^2 = 0.023$), significant predictors for better ranking included *accessibility_score* (coeff: -1.045), *query_in_h1* (coeff: -0.279), and *semantic_similarity_content_query* (coeff: -2.391). Conversely, *seo_score* (coeff: +1.590) and *semantic_similarity_title_query* (coeff: +0.730) were associated with worse ranking ($p < .05$).

- For Bing (Combined Model Pseudo $R^2 = 0.018$), significant predictors for better ranking included *accessibility_score* (coeff: -2.690), *exact_query_in_h1* (coeff: -0.299), *semantic_similarity_title_query* (coeff: -0.406), and *semantic_similarity_content_query* (coeff: -0.762). $Performance\_score$ was associated with worse ranking ($p < .05$).

Random Forest Feature Importance: Content factors had higher aggregate importance than technical factors for both Google (70.1%) and Bing (61.8%) (Figure 5). $Word\_count$ and semantic similarity metrics were top individual content predictors.

These analyses collectively indicate that content-related attributes generally exhibit a stronger association with ranking visibility than technical scores alone. However, specific influential factors and their impact direction vary between Google and Bing and across different analytical models.

5. Discussion

This section interprets the empirical findings from Section 4. It contextualizes them within existing literature and discusses their broader implications and limitations.

5.1. Interpretation of Resource Profiles (RQ1)

The K-Means clustering (RQ1) successfully delineated six distinct resource profiles within the commercial offers domain (Table 3, Figure 2). These profiles, such as "High Keyword & Semantic Relevance, Strong Technicals" (Cluster 2) and "Low Relevance, Low Technicals" (Cluster 3), illustrate the diverse optimization strategies and quality levels present. The differentiation was significantly driven by on-page lexical features (e.g., *query_in_title*) and title-focused semantic similarity. This underscores their role in characterizing resources in this niche (Yu ., 2024). This granular profiling moves beyond simplistic high/low quality categorizations, revealing a spectrum of resource archetypes.

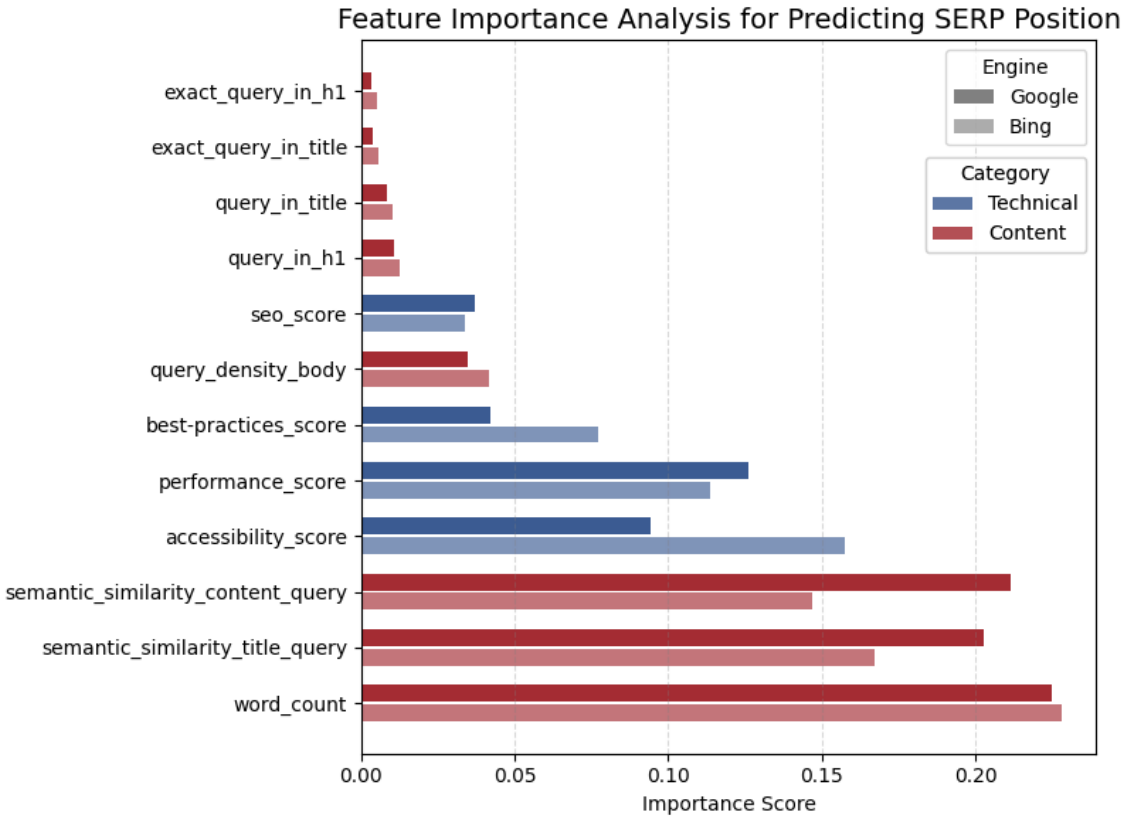


Figure 5: Relative feature importance from Random Forest for Google (left) and Bing (right) (RQ4).

5.2. Inferring System Ranking Logic (RQ2 & RQ4)

Profile Association with Ranking (RQ2): For both Google and Bing, resource profiles significantly associated with ranking tiers (Figure 3). Stronger content-relevant profiles were generally favored in higher ranks. This aligns with the established importance of content quality (Srinivas Gowda, 2025).

Individual Feature Association with Ranking (RQ2 & RQ4): Ordinal Logistic Regression (Appendix Appendix A.1) and Spearman correlations (Appendix Appendix A.2, Figure A.6) offered deeper insights. For **Google**, *semantic_similarity_content_query* was a robust predictor of better ranking. *Query_in_h*1 and *accessibility_score* also emerged as positively associated with better ranking in the combined regression model. A key counter-intuitive finding was the consistently positive coefficient for *semantic_similarity_title_query* in Google's regression models. This associated high title-query similarity with worse rank categories. This suggests that while title similarity correlates with better rank in isolation (as seen in RQ4 correlations), its effect changes in a multivariate context. When other factors like main content similarity are controlled, an extremely high title-query similarity might be less beneficial or even perceived negatively by Google. This could reflect a penalization for perceived over-optimization. Alternatively, it might indicate a preference for titles that offer broader context when main content relevance is already established (Nagpal Petersen, 2021). The non-significance of *query_density_body* for Google aligns with its shift towards semantic understanding (Jindal ., 2014).

For **Bing**, the Ordinal Logistic Regression (combined model) highlighted several significant predictors of better ranking. These included *accessibility_score*, *exact_query_in_h*1, *semantic_similarity_title_query*, and *semantic_similarity_content_query*. The positive coefficient for $performance\_score$ (worse ranking) was unexpected. This could be due to multicollinearity with other technical factors that more directly capture user-centric performance. It might also indicate that, for this dataset, pages ranking higher on Bing for other reasons coincidentally had slightly lower raw performance scores, despite the general expectation that good performance is favored. Bing's stronger reliance on *word_count* and *query_density_body* in correlational and Random Forest analyses suggests these explicit signals may still hold more weight for Bing compared to Google, even if their individual significance in the ordinal regression model varies.

Technical vs. Content Factors (RQ4): Random Forest models (Figure 5) indicated a greater aggregate importance for content factors over technical Lighthouse scores for both Google (70.1% content) and Bing (61.8% content). While specific technical metrics like *accessibility_score* proved significant in regression, content attributes collectively showed a stronger association with visibility. This supports the

"content is king" paradigm, while also affirming the non-negligible role of technical performance (Roumeliotis Tselikas, 2022). Combined regression models generally yielded better fit (Pseudo R^2) than single-factor-type models.

5.3. Profiling System Differences (RQ3)

Comparative analysis (RQ3) revealed significant differences in how Google and Bing rank resources. Bing's top results generally featured higher median values for a majority of the measured technical and content features (Table 6). This suggests Bing may more consistently reward a broader set of measured optimizations. The profile distribution in top ranks also differed (Figure 4). Google's top results included a higher proportion of Profile 3 ("Low Relevance, Low Technicals"). This could imply Google's reliance on unmeasured factors (e.g., domain authority, backlink profiles, user engagement signals) to a greater extent for some top positions. It might also reflect a different approach to result diversification (Coghlan ., 2025; Toms Taves, 2004). Bing's top results, conversely, more consistently featured profiles with stronger keyword relevance. These differences may stem from distinct interpretations of user intent or varying algorithmic philosophies. For instance, Google might prioritize satisfying a broader range of potential intents for ambiguous queries, even if some results are less optimized on measured features. Bing, on the other hand, might more directly reward explicit on-page relevance signals for the specific query type studied.

5.4. Synthesis and Overall Picture

Content relevance, especially semantic similarity and content volume (for Bing), appears dominant for visibility. However, the interplay with technical factors and specific feature weights differs between Google and Bing. The identified profiles show diverse success strategies. Google seems to prioritize deep semantic understanding of main content. It also has a complex view of title-query similarity in multivariate contexts, possibly to penalize over-optimization or favor titles with broader appeal when main content is highly relevant. Bing appears to reward a broader array of explicit on-page signals and technical performance more consistently. The counter-intuitive regression findings underscore that factors operate interdependently. An observed correlation for a single feature may not hold when controlling for others, highlighting the complexity of inferring ranking logic. The study emphasizes that correlational findings identify associations; they do not imply direct causation. Underlying causal mechanisms may involve unmeasured variables or complex interactions between features.

5.5. Implications

Practical Implications:

- For **Google**, content creators should prioritize high-quality, comprehensive content. This content should be semantically aligned with query intent (*semantic_similarity_content_query*) and ensure good *accessibility_score*. While title and H1 keyword presence is beneficial, overly narrow or repetitive title optimization might be counterproductive if not holistically supported by strong, distinct main content. A balanced approach is key.

- For **Bing**, a more direct approach appears more consistently rewarded. This involves explicit query terms in H1s (especially exact matches), ensuring substantial *word_count*, high *semantic_similarity_title_query*, and strong *accessibility_score*.

- The diverse "successful" profiles (RQ1) suggest that resource creators should not rely on a single checklist. Instead, they should identify their content's core strengths (e.g., deep semantic value vs. strong lexical targeting). Optimization should then be tailored to align with the nuanced preferences of each search engine.

Theoretical Implications: This research significantly advances the understanding of large-scale, operational "black-box" information retrieval systems, differentiating itself from prior work and contributing to IR theory in several key ways:

- **Enhanced Methodology for Comparative IR System Analysis:** This study introduces and validates a robust, large-scale empirical methodology for the *simultaneous comparative profiling* of major search engines (Google and Bing). Unlike many studies focusing on a single system or relying on heterogeneous web data, our use of a *content-homogeneous dataset* (commercial discount offers) acts as a crucial control variable. This novel approach allows for a clearer isolation and understanding of how technical performance (Lighthouse scores) and specific content attributes (including advanced semantic features from Sentence-BERT) are differentially weighted, thus revealing system-specific nuances in ranking logic with greater precision than previously possible. This work, therefore, updates and substantially expands upon earlier comparative IR studies (Toms Taves, 2004) by incorporating a more controlled design and modern feature sets.

- **Unveiling Algorithmic Complexity and System-Specific Nuances:** The empirical findings provide data-driven insights into the complex, often non-linear, and context-dependent nature of factor weighting within dominant commercial search engines. A key original contribution is the identification of counter-intuitive relationships, such as the negative association of high title-query semantic similarity with better Google rankings in multivariate models when other factors are controlled. Such findings challenge simplistic SEO heuristics and contribute to a more nuanced theoretical model of how search engines might perceive and penalize perceived over-optimization or balance diverse relevance signals. The distinct factor priorities identified for Google versus Bing (e.g., Bing's greater emphasis on content volume and explicit keyword signals) further underscores system-specific operational logics.

- **Empirically Grounding the Semantic Shift in Real-World Commercial Systems:** While the theoretical shift towards semantic understanding in IR is well-established, this study offers large-scale, contemporary empirical evidence of its manifestation, relative importance, and interplay with persistent lexical signals and content volume within today's leading commercial search platforms. This research, therefore, moves beyond theoretical postulations or smaller-scale academic model evaluations by examining these dynamics in operational, high-stakes environments. The differing sensitivities observed between Google and Bing also provide empirical grounding for discussions on varying philosophies or stages in the adoption and implementation of semantic technologies by major industry players.

- **Providing a Replicable Framework and Public Dataset for Future Profiling:** The detailed methodology, coupled with the commitment to make the comprehensive dataset (14 465 SERP items) and analysis scripts publicly available, offers a valuable and replicable framework. This directly contributes to the IR community by enabling further "reverse engineering" studies, longitudinal analyses of algorithmic changes, or extensions to different domains, languages, or search systems, thereby fostering greater transparency and understanding of these influential information access technologies.

5.6. Limitations

This study's findings should be considered within its limitations:

- **Data Snapshot and Domain Specificity:** Data are from May 2025 for 500 English commercial discount queries and may not generalize.

- **Feature Scope:** Excluded factors include backlinks, detailed Core Web Vitals, user engagement, and domain-level authority signals. Observed anomalies might be partially explained by these unmeasured variables.
- **Tool and API Limitations:** Accuracy depends on tools like Trafilatura, Sentence-BERT, and APIs.
- **Causation vs. Correlation:** The study identifies associations, not causal links. Inferred priorities are based on observed patterns. Actual causal mechanisms are opaque and might be influenced by unmeasured confounders or complex feature interactions not explicitly modeled.
- **Interaction Effects:** Regression models did not explicitly include interaction terms. These could explain some complex patterns (e.g., title similarity's effect varying with content similarity). Future work should explore these.
- **Search Engine Intent Interpretation:** Engines might interpret user intent for similar queries differently, influencing attribute prioritization. This was not directly modeled.
- **Personalization and Localization:** Efforts to minimize these using default API settings might not entirely eliminate their influence.
- **Scope of Baseline Comparison:** This study focused on characterizing existing search engine behaviors rather than proposing and evaluating a new ranking algorithm against state-of-the-art (SOTA) baselines. While Sentence-BERT (an LLM) was used for feature extraction, a direct comparison with LLM-based ranking models was beyond the current scope, which aimed to understand how incumbent systems respond to a defined set of features.

These limitations offer avenues for future research.

6. Data and Code Availability

The dataset generated and analyzed during the current study will be made publicly available upon acceptance of this manuscript. This includes the list of queries, raw SERP data, extracted features, and cluster labels. Similarly, the Python scripts used for data collection, feature engineering, and analysis ("SERP Profiler Kit") will also be openly shared at that time. This aims to ensure reproducibility and facilitate further research by the community. Specific repository links will be provided in the final published version.

7. Conclusion and Future Work

This study empirically characterized and compared the ranking environments of Google and Bing for commercial discount queries, based on an analysis of technical and content attributes from a large number of SERP entries. Six distinct resource profiles were identified, revealing varied optimization strategies. Content-related attributes, particularly semantic relevance and (for Bing) content volume, showed a stronger overall association with ranking visibility than technical Lighthouse scores for both search engines. However, specific influential factors and their impact direction varied significantly between Google and Bing and across different analytical models.

7.1. Principal Findings

Key findings include:

- **RQ1:** Six resource profiles emerged, differentiated primarily by query presence in titles/H1s and title-query semantic similarity.
- **RQ2:** Resource profiles significantly associated with ranking tiers in both Google and Bing. Stronger content-relevant profiles were generally more prevalent in higher ranks.
- **RQ3:** Bing's top-ranking pages generally exhibited higher median scores for most technical and content features compared to Google's. Profile distributions in top ranks also differed significantly.
- **RQ4:** Content factors showed greater aggregate importance (Random Forest) than technical factors for both systems. Ordinal Logistic Regression indicated specific predictors for each engine. For Google, *semantic_similarity_content_query*, *query_in_h*1, and *accessibility_score* were key for better rank, while *semantic_similarity_title_query* and *seo_score* surprisingly associated with worse rank. For Bing, *accessibility_score*, *exact_query_in_h*1, and semantic similarities predicted better rank, while *performance_score* associated with worse rank.

7.2. Concluding Remarks

This empirical study provides a data-driven characterization of search engine ranking environments. It highlights that visibility is shaped by a complex interplay of technical and content attributes, evaluated differently by Google and Bing. The identified resource profiles offer a granular view beyond simple quality distinctions. The nuanced, sometimes counter-intuitive, regression findings emphasize the

challenge of inferring modern IR algorithmic logic. They also suggest the importance of holistic, context-aware optimization rather than focusing on isolated metrics. The differing sensitivities to factors like title semantic similarity (Google) and word count/performance score (Bing) underscore the need for system-specific optimization strategies. We emphasize that these findings reveal correlations, not causation. Thus, inferred system priorities should be interpreted as empirical patterns, not definitive algorithmic rules.

7.3. Future Research Directions

Future work could include: longitudinal analysis to track algorithmic changes; expanding the feature set (e.g., backlinks, Core Web Vitals, user engagement signals); applying the methodology to diverse domains and languages. Further avenues are employing advanced predictive modeling, including the exploration of interaction effects between features, or causal inference techniques. Qualitative case studies could deepen understanding of specific ranking outcomes. Finally, further investigation into anomalous findings is warranted, particularly the role of title-query semantic similarity and SEO scores in Google's regression models and performance scores in Bing's.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the author(s) used ChatGPT (OpenAI) and Gemini (Google) in order to improve language, clarity, and assist with structuring the manuscript. After using these toolservices, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Appendix A. Detailed Statistical Test Results

Appendix A.1. RQ4 - Ordinal Logistic Regression Results

Table A.7: Ordinal Logistic Regression for Predicting SERP Quintiles - Google (System A) (RQ4)

	Technical Model		**Content Model**		**Combined Model**	
Predictor	**Coeff.**	**p-value**	**Coeff.**	**p-value**	**Coeff.**	**p-value**
Lighthouse Scores						
Perf.	−0.290	0.039*	–	–	0.213	0.140
Access.	−1.619	< 0.001***	–	–	−1.045	< 0.001***
Best Prac.	−0.105	0.427	–	–	−0.001	0.997
SEO	1.343	< 0.001***	–	–	1.590	< 0.001***
Content Metrics						
Q-Title	–	–	−0.175	0.016*	−0.146	0.045*
Q-H1	–	–	−0.250	< 0.001***	−0.279	< 0.001***
ExQ-Title	–	–	−0.110	0.322	−0.108	0.334
ExQ-H1	–	–	0.071	0.548	0.038	0.749
Q/B Density	–	–	0.088	0.803	0.054	0.880
Sim. Title	–	–	0.861	< 0.001***	0.730	< 0.001***
Sim. Content	–	–	−2.412	< 0.001***	−2.391	< 0.001***
Word Count	–	–	0.399	0.227	0.343	0.304
Pseudo R^2	0.0046		0.0193		0.0232	
AIC	18750.5		18482.9		18417.8	
BIC	18804.0		18563.1		18524.7	

Dependent Variable: SERP Quintile (0=best, 4=worst). Negative coefficients indicate increased likelihood of better ranking.
* $p < .05$; ** $p < .01$; *** $p < .001$. Thresholds (cut-points) for quintiles are omitted. Predictors Min-Max scaled.
Perf. = Performance Score; Access. = Accessibility Score; Best Prac. = Best Practices Score; SEO = SEO Score; Q-Title = Query in Title; Q-H1 = Query in H1; ExQ-Title = Exact Query in Title; ExQ-H1 = Exact Query in H1; Q/B Density = Query Density in Body Text; Sim. Title = Semantic Similarity of Title and Query; Sim. Content = Semantic Similarity of Content and Query; Word Count = Total Word Count in Content.

Table A.8: Ordinal Logistic Regression for Predicting SERP Quintiles - Bing (System B) (RQ4)

	Technical Model		Content Model		Combined Model	
Predictor	**Coeff.**	**p-value**	**Coeff.**	**p-value**	**Coeff.**	**p-value**
Lighthouse Scores						
Perf.	0.519	$< 0.001^{***}$	–	–	0.554	$< 0.001^{***}$
Access.	−2.671	$< 0.001^{***}$	–	–	−2.690	$< 0.001^{***}$
Best Prac.	0.202	0.078	–	–	0.211	0.067
SEO	−0.548	0.002^{**}	–	–	−0.051	0.785
Content Metrics						
Q-Title	–	–	−0.035	0.507	−0.062	0.245
Q-H1	–	–	−0.005	0.927	0.067	0.207
ExQ-Title	–	–	0.096	0.184	0.109	0.133
ExQ-H1	–	–	−0.234	0.002^{**}	−0.299	$< 0.001^{***}$
Q/B Density	–	–	−0.117	0.701	−0.243	0.422
Sim. Title	–	–	−0.725	$< 0.001^{***}$	−0.406	0.008^{**}
Sim. Content	–	–	−0.625	$< 0.001^{***}$	−0.762	$< 0.001^{***}$
Word Count	–	–	−0.157	0.936	−0.347	0.854
Pseudo R^2	0.0131		0.0060		0.0176	
AIC	27223.4		27426.7		27115.0	
BIC	27279.9		27511.4		27227.9	

Dependent Variable: SERP Quintile (0=best, 4=worst). Negative coefficients indicate increased likelihood of better ranking.

* $p < .05$; ** $p < .01$; *** $p < .001$. Thresholds (cut-points) for quintiles are omitted. Predictors Min-Max scaled.

Perf. = Performance Score; Access. = Accessibility Score; Best Prac. = Best Practices Score; SEO = SEO Score; Q-Title = Query in Title; Q-H1 = Query in H1; ExQ-Title = Exact Query in Title; ExQ-H1 = Exact Query in H1; Q/B Density = Query Density in Body Text; Sim. Title = Semantic Similarity of Title and Query; Sim. Content = Semantic Similarity of Content and Query; Word Count = Total Word Count in Content.

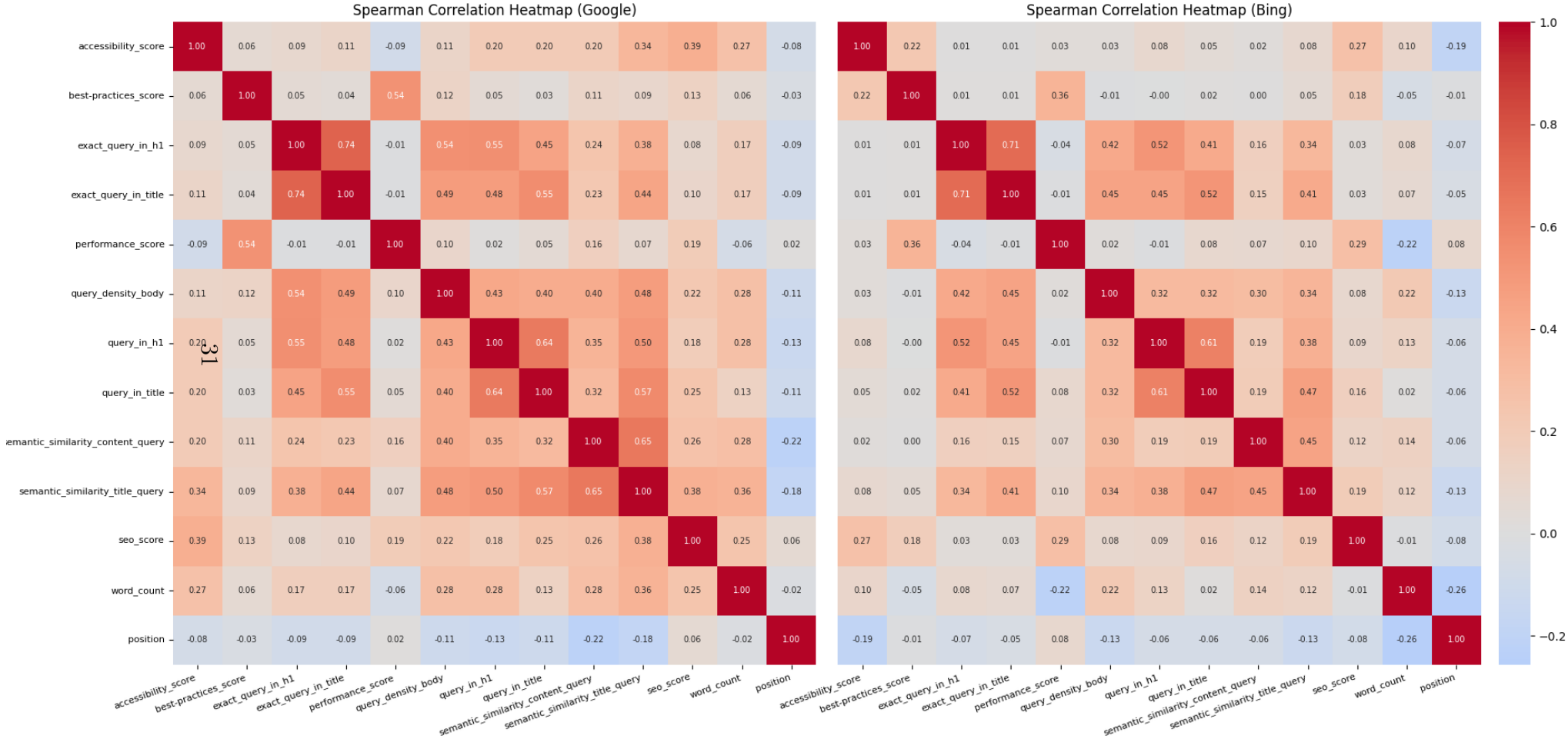


Figure A.6: Spearman Correlation Heatmaps with SERP Position for Google (left) and Bing (right) (RQ4).

References

Barbaresi2021Barbaresi, A. 202108. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. H. Ji, JC. Park R. Xia (), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations (122–131). OnlineAssociation for Computational Linguistics. https://aclanthology.org/2021.acl-demo.15/ 10.18653/v1/2021.acl-demo.15

Bardas2025Bardas, N., Mordo, T., Kurland, O., Tennenholtz, M. Zur, G. 202502. White Hat Search Engine Optimization using Large Language Models White Hat Search Engine Optimization using Large Language Models. arXiv preprint arXiv:2502.07315. 10.48550/arXiv.2502.07315

Beltagy2020Beltagy, I., Peters, ME. Cohan, A. 202004. Longformer: The Long-Document Transformer Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150. http://arxiv.org/abs/2004.05150

Sezerrakolu2024Çırakoğlu, FS. Koşaner, Ö. 202412. Linguistic insights into high-quality content for SEO: Optimizing high-quality content with Grice's conversational maxims Linguistic insights into high-quality content for SEO: Optimizing high-quality content with Grice's conversational maxims. Telematics and Informatics Reports16100-169. https://www.sciencedirect.com/science/article/pii/S2772503024000550 10.1016/j.teler.2024.100169

Coghlan2025Coghlan, S., Chia, HX., Scholer, F. Spina, D. 202503. Control search rankings, control the world: what is a good search engine? Control search rankings, control the world: what is a good search engine? AI and Ethics. https://link.springer.com/10.1007/s43681-025-00695-8 10.1007/s43681-025-00695-8

Daly2024Daly, TM. Ryan, JC. 202412. University 'Pay-for-grades': the bait and switch search engine optimization strategies of contract cheating websites in the United States University 'Pay-for-grades': the bait and switch search engine optimization strategies of contract cheating websites in the United States. International Journal for Educational Integrity20. 10.1007/s40979-023-00148-x

Devlin2019Devlin, J., Chang, MW., Lee, K. Toutanova, K. 201906. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. J. Burstein, C. Doran T. Solorio (), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (4171–4186). Minneapolis, MinnesotaAssociation for Computational Linguistics. https://aclanthology.org/N19-1423/ 10.18653/v1/N19-1423

Jayaraman2022Jayaraman, S., Ramachandran, M., Patan, R., Daneshmand, M. Gandomi, AH. 202202. Fuzzy Deep Neural Learning Based on Goodman and Kruskal's Gamma for Search Engine Optimization Fuzzy Deep Neural Learning Based on Goodman and Kruskal's Gamma for Search Engine Optimization. IEEE Transactions on Big Data8268–277. 10.1109/TBDATA.2020.2963982

Jindal2014Jindal, V., Bawa, S. Batra, S. 2014. A review of ranking approaches for semantic search on Web A review of ranking approaches for semantic search on Web. Information Processing and Management502416-425. https://www.sciencedirect.com/science/article/pii/S0306457313001106 10.1016/j.ipm.2013.10.004

Limongelli2022Limongelli, C., Lombardi, M., Marani, A. Taibi, D. 2022. A Semantic Approach to Ranking Techniques: Improving Web Page Searches for Educational Purposes A Semantic Approach to Ranking Techniques: Improving Web Page Searches for Educational Purposes. IEEE Access1068885–68896. 10.1109/ACCESS.2022.3186356

Milei2025Milei, P., Votintseva, N. Barajas, A. 202501. Automated Identification of Business Models Automated Identification of Business Models. Information Processing and Management62. 10.1016/j.ipm.2024.103893

Moffat2024Moffat, A. 202401. User-oriented metrics for search engine deterministic sort orders User-oriented metrics for search engine deterministic sort orders. Information Processing and Management61. 10.1016/j.ipm.2023.103547

Nagpal2021Nagpal, M. Petersen, JA. 202112. Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance? Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance? Journal of Retailing97746–763. 10.1016/j.jretai.2020.12.002

Niu2022Niu, X., Liu, G. Yang, Q. 202206. OpinionRank: Trustworthy Website Detection Using Three Valued Subjective Logic OpinionRank: Trustworthy Website Detection Using Three Valued Subjective Logic. IEEE Transactions on Big Data8855–866. 10.1109/TBDATA.2020.2994309

Pauzi2023Pauzi, Z. Capiluppi, A. 202304. Applications of natural language processing in software traceability: A systematic mapping study Applications of natural language processing in software traceability: A systematic mapping study. Journal of Systems and Software198111616. https://www.sciencedirect.com/science/article/pii/S0164121223000110 10.1016/j.jss.2023.111616

Rastakhiz2024Rastakhiz, F., Eftekhari, M. Vahdati, S. 2024. QuickCharNet: An Efficient URL Classification Framework for Enhanced Search Engine Optimization QuickCharNet: An Efficient URL Classification Framework for Enhanced Search Engine Optimization. IEEE Access. 10.1109/ACCESS.2024.3484578

Reimers2019Reimers, N. Gurevych, I. 201911. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks Sentence-BERT: Sentence embeddings using Siamese BERT-networks. K. Inui, J. Jiang, V. Ng X. Wan (), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp) (3982–3992). Hong Kong, ChinaAssociation for Computational Linguistics. https://aclanthology.org/D19-1410/ 10.18653/v1/D19-1410

Roumeliotis2022Roumeliotis, KI. Tselikas, ND. 2022. An Effective SEO Techniques and Technologies Guide-map An Effective SEO Techniques and Technologies Guide-map. Journal of Web Engineering211603-1650. 10.13052/jwe1540-9589.21510

SilvaRodrigues2025Silva Rodrigues, PH., de Sousa, DX., França, C., Rabbi, G., Couto Rosa, T. Gonçalves, MA. 2025. Risk-sensitive optimization of neural deep learning ranking models with applications in ad-hoc retrieval and recommender systems Risk-sensitive optimization of neural deep learning ranking models with applications in ad-hoc retrieval and recommender systems. Information Processing & Management624104126. https://www.sciencedirect.com/science/article/pii/S0306457325000688 https://doi.org/10.1016/j.ipm.2025.104126

Song2020Song, K., Tan, X., Qin, T., Lu, J. Liu, TY. 202004. MPNet: Masked and Permuted Pre-training for Language Understanding MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv preprint arXiv:2004.09297. http://arxiv.org/abs/2004.09297

Srinivas2025Srinivas, VM. Gowda, PMHC. 202502. A page rank-based analytical design of effective search engine optimization A page rank-based analytical design of effective search engine optimization. IAES International Journal of Artificial Intelligence1473–82. 10.11591/ijai.v14.i1.pp73-82

Toms2004Toms, EG. Taves, AR. 2004. Measuring user perceptions of Web site reputation Measuring user perceptions of Web site reputation. Information Processing and Management40291–317. 10.1016/j.ipm.2003.08.007

vanderSluis2025van der Sluis, F. 202503. Wanting information: Uncertainty and its reduction through search engagement Wanting information: Uncertainty and its reduction through search engagement. Information Processing and Management62. 10.1016/j.ipm.2024.103890

Wang2023Wang, P., Yang, H., Hou, J. Li, Q. 2023. A machine learning approach to primacy-peak-recency effect-based satisfaction prediction A machine learning approach to primacy-peak-recency effect-based satisfaction prediction. Information Processing & Management602103196. https://www.sciencedirect.com/science/article/pii/S0306457322002977 https://doi.org/10.1016/j.ipm.2022.103196

Yu2024Yu, B., Xu, R., Cai, M. Ding, W. 2024. A clustering method based on multi-positive–negative granularity and attenuation-diffusion pattern A clustering method based on multi-positive–negative granularity and attenuation-diffusion pattern. Information Fusion103102137. https://www.sciencedirect.com/science/article/pii/S1566253523004530 https://doi.org/10.1016/j.inffus.2023.102137

Zhu2024Zhu, E., Cheng, K., Zhang, Z. Wang, H. 2024. PDHF: Effective phishing detection model combining optimal artificial and automatic deep features Pdhf: Effective phishing detection model combining optimal artificial and automatic deep features. Computers & Security136103561. https://www.sciencedirect.com/science/article/pii/S0167404823004716 https://doi.org/10.1016/j.cose.2023.103561