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

ARTICLE INFO

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

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.

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 (**?**). Achieving visibility within these complex ranking environments is crucial for diverse online entities, from commercial enterprises to educational platforms (**??**). Ultimately, the success of these retrieval systems hinges on user satisfaction (**?**), 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 (**??**). 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 (**?**). 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) (**??**). 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 (**?**).

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

ORCID(s):

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 (**??**). 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 **??** reviews related work. Section **??** details the methodology. Section **??** presents the results. Section **??** discusses these findings. Section **??** outlines the availability of the data and code. Finally, Section **??** 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 (**?**). 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 (**?**). 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 (**?**). 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 (**??**). Ongoing research continues to develop sophisticated models to assess website trust and combat misinformation (**?**).

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 (**?**). Modern retrieval systems, however, employ sophisticated Natural Language Processing (NLP) techniques moving beyond lexical matching towards semantic understanding (**?**). This is crucial as users are often driven by a desire to reduce uncertainty or engage with interesting content (**?**), motivations that are better served by semantically rich and relevant results rather than mere keyword presence. Techniques like Latent Semantic Analysis (LSA) (**?**), topic modeling, and more recently, vector embeddings from deep learning models like Sentence-BERT, allow for capturing semantic similarity even without exact keyword overlap (**?**). Understanding user intent and matching it with semantically relevant content is paramount (**??**).

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 (**?**). 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 (**?**) or the specific retrieval system being analyzed (**?**). Modeling approaches sometimes attempt to integrate these diverse factors to predict ranking outcomes or system behavior (**?**). 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 (**?**). Content analysis, including automated dictionary-based methods, can classify resources based on their topical focus or expressed attributes (**?**). 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 (**?**). 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 (**?**).

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 (**?**). 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 (**?**).

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 ∼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 (**?**) 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 (**?**).

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_h*1 (all query words in first `<h1>`), *exact_query_in_h*1 (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' (**??**)), 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 (**??**).

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 **??**).

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 **??** 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 **??**.

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 **??**. 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 **??**). Clusters were characterized by their mean feature values (Table **??**; Figure **??**). Kruskal-Wallis H-tests (Table **??**) and Random Forest Classifier feature importances identified discriminative features.

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.

Table 1
Dataset Column Types and Descriptions

Category	Column Name	Type	Description (Metric)
Search Engine	engine	Categorical (google/bing)	Search Engine
Search Engine	position	Integer (1-20)	SERP Position
Lighthouse	performance_score	Float (0-100)	Perf.
Lighthouse	accessibility_score	Float (0-100)	Access.
Lighthouse	best-practices_score	Float (0-100)	Best Prac.
Lighthouse	seo_score	Float (0-100)	SEO
Content	query_in_title	Integer (0/1)	Q-Title
Content	query_in_h1	Integer (0/1)	Q-H1
Content	exact_query_in_title	Integer (0/1)	ExQ-Title
Content	exact_query_in_h1	Integer (0/1)	ExQ-H1
Content	query_density_body	Float (%)	Q/B Density
Content	semantic_similarity_title_query	Float (0-1, Cosine)	Sim. Title
Content	semantic_similarity_content_query	Float (0-1, Cosine)	Sim. Content
Content	word_count	Integer	Word Count

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
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	1.00	20.00	5.00	15.00	0.00
Lighthouse Scores								
Perf.	87.83	95.00	15.04	12.00	100.00	82.00	99.00	5.45
Access.	87.36	90.00	8.45	34.00	100.00	83.00	94.00	2.96
Best Prac.	91.75	96.00	12.09	37.00	100.00	85.00	100.00	4.14
SEO	93.00	92.00	7.43	40.00	100.00	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 are defined in Table **??**.

- **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 **??**).

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 (**??**). A negative coefficient indicated a higher probability of a better rank. (Full models are in **??**).
- **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: 0.442) is shown in Figure **??**. All features differed significantly across clusters (Kruskal-Wallis, $p < .001$; Table **??**). The internal cohesion and separation quality of these six profiles were quantitatively assessed using Silhouette and Calinski-Harabasz scores, as presented in Table **??**. These metrics provide a global and per-cluster evaluation of the clustering solution's robustness. As the table indicates, while the overall structure is reasonable (Average Silhouette Score: 0.442), the quality varies across profiles, with Cluster 2 showing the highest internal cohesion.

To better understand the structure of the clusters and their interrelationships, the results of a Principal Component Analysis (PCA) on the 12 features are visualized in Figure **??**. These first two components explain 43.0% of the total variance in the dataset. An analysis of the PCA loadings reveals that the First Principal Component (PC1) represents a strategic trade-off between Comprehensive Content & Technical Health (-) and Direct Keyword Targeting (+). The Second Principal Component (PC2) reflects a similar balance between Exact-Match Keyword Targeting (-) and Overall Technical Performance (+).

On this strategic map, the positions of the clusters corroborate the profiles presented in Table **??**. For instance, Cluster 3 (orange), located on the far left of the PC1 axis, represents resources with high word counts but weak keyword targeting, while Cluster 4 (pink), concentrated on the right, indicates a profile that adopts aggressive keyword optimization. The notable overlap observed between clusters suggests that many web resources employ hybrid strategies, a finding that is consistent with the moderate average Silhouette score of 0.442.

Mean scaled feature values for each cluster are in Table **??** and visualized in Figure **??**. 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: 0.269; *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 **??**). 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.

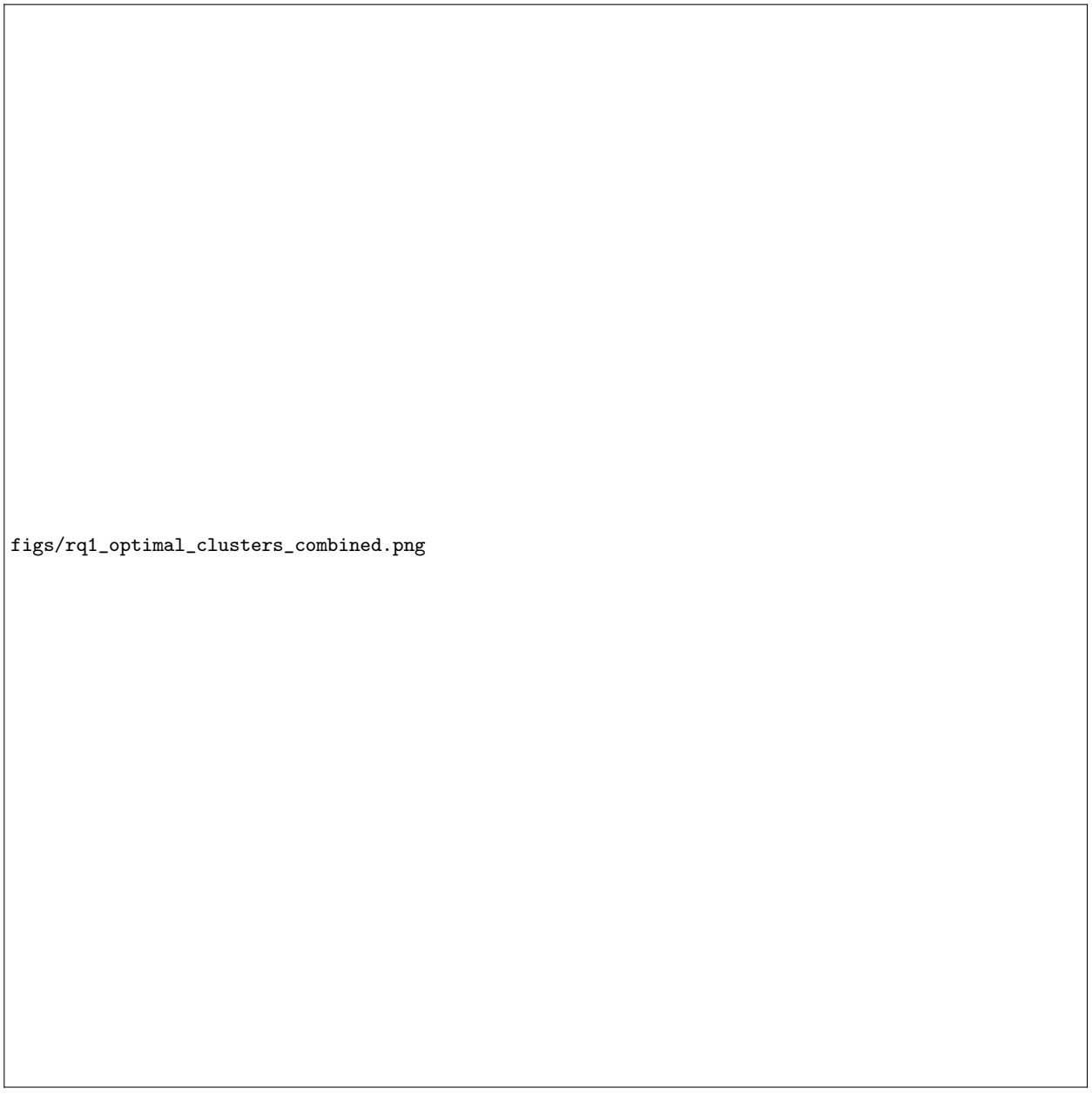

Figure 1: Determination of optimal cluster number (K) using Elbow method (left Y-axis, WCSS) and Silhouette analysis (right Y-axis, Average Silhouette Score). $K = 6$ was selected.

Feature Comparison across Ranks: Kruskal-Wallis H-tests (Table **??**) 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 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	0.881	0.811	0.870	0.871
Access.	0.806	0.823	0.826	0.736	0.805	0.837
Best Prac.	0.853	0.879	0.880	0.826	0.899	0.866
SEO	0.880	0.897	0.906	0.776	0.911	0.921
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	1.000	0.000	0.187	0.000
ExQ-H1	0.166	0.000	0.818	0.000	0.000	0.105
Q/B Density	0.031	0.010	0.072	0.000	0.046	0.016
Sim. Title	0.759	0.729	0.823	0.269	0.790	0.778
Sim. Content	0.747	0.716	0.758	0.389	0.711	0.730
Word Count	0.005	0.005	0.006	0.003	0.003	0.005
Cluster Size (N)	1054	5142	2122	2314	1522	2311
Percentage (%)	7.3	35.5	14.7	16.0	10.5	16.0

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

Table 4
Identified Cluster Quality and Cohesion Metrics (RQ1)

Cluster Profile	Cluster Size (N)	Percentage (%)	Mean Silhouette Score	WCSS
C0: Balanced Profile	1054	7.3	0.493	484.8
C1: Low Relevance	5142	35.5	0.451	532.5
C2: Content Focused	2122	14.7	0.544	439.2
C3: Low Relevance & Poor Technicals	2314	16.0	0.222	569.7
C4: High Relevance	1522	10.5	0.444	485.6
C5: Technical Excellence	2311	16.0	0.521	252.9
Overall / Average	**14465**	**100.0**	**0.442**	**2764.5**

Note: Higher Silhouette scores indicate better-defined clusters. Lower WCSS indicates tighter clusters. Cluster sizes are from Table **??**.

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 **??**). While Cluster 1 was the most common profile in Bing's top 5, Google's top results, though also led by Cluster 1 (Low Relevance), featured a significantly larger proportion of profiles like Cluster 3 (Low Relevance & Poor Technicals) and Cluster 4 (High Relevance).

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 **??**). These included *semantic_similarity_title_query* (medium effect size) and *word_count* (small effect size).

4.4. RQ4: Inferred Factor Priorities

Correlation Analysis: Spearman rank correlations are detailed in Appendix **??** (Figure **??**).

- For Google, *semantic_similarity_content_query* showed the strongest significant negative correlation with *position* ($\rho = -0.221$), followed by *query_in_h1* ($\rho \approx -0.19$) and *semantic_similarity_title_query* ($\rho = -0.179$).
- 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 **??**).

- For Google (Combined Model Pseudo $R^2 = 0.024$), several predictors were significantly associated with better ranking quintiles. These included *semantic_similarity_content_query* (coeff: -0.485), *accessibility_score*

Table 5
Kruskal-Wallis H-Test Results for Feature Differentiation Across Clusters (RQ1)

Feature	H-statistic	p-value
Perf.	232.613	< 0.001
Access.	920.863	< 0.001
Best Prac.	171.951	< 0.001
SEO	1756.283	< 0.001
Q-Title	14 464.000	< 0.001
Q-H1	14 464.000	< 0.001
ExQ-Title	12 798.195	< 0.001
ExQ-H1	9102.510	< 0.001
Q/B Density	4388.283	< 0.001
Sim. Title	7561.791	< 0.001
Sim. Content	4641.727	< 0.001
Word Count	2015.283	< 0.001

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

Table 6
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 are defined in Table **??**.

(coeff: -0.208), $query\_in\_h1$ (coeff: -0.121), $query\_in\_title$ (coeff: -0.070), and $performance\_score$ (coeff: -0.064). Conversely, a higher *semantic_similarity_title_query* (coeff: +0.202) and seo_score (coeff: +0.168) were significantly associated with worse ranking quintiles (all $p < .05$).

- For Bing (Combined Model Pseudo $R^2 = 0.032$), $word\_count$ emerged as the most powerful predictor of better ranking quintiles, with the largest significant negative coefficient (coeff: -1.133). Other factors significantly associated with better rankings included $accessibility\_score$ (coeff: -0.339), *semantic_similarity_content_query* (coeff: -0.093), and $exact\_query\_in\_h1$ (coeff: -0.085). In contrast, higher scores for $performance\_score$ (coeff: +0.093) and $exact\_query\_in\_title$ (coeff: +0.064) were unexpectedly associated with worse ranking quintiles (all $p < .05$).

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

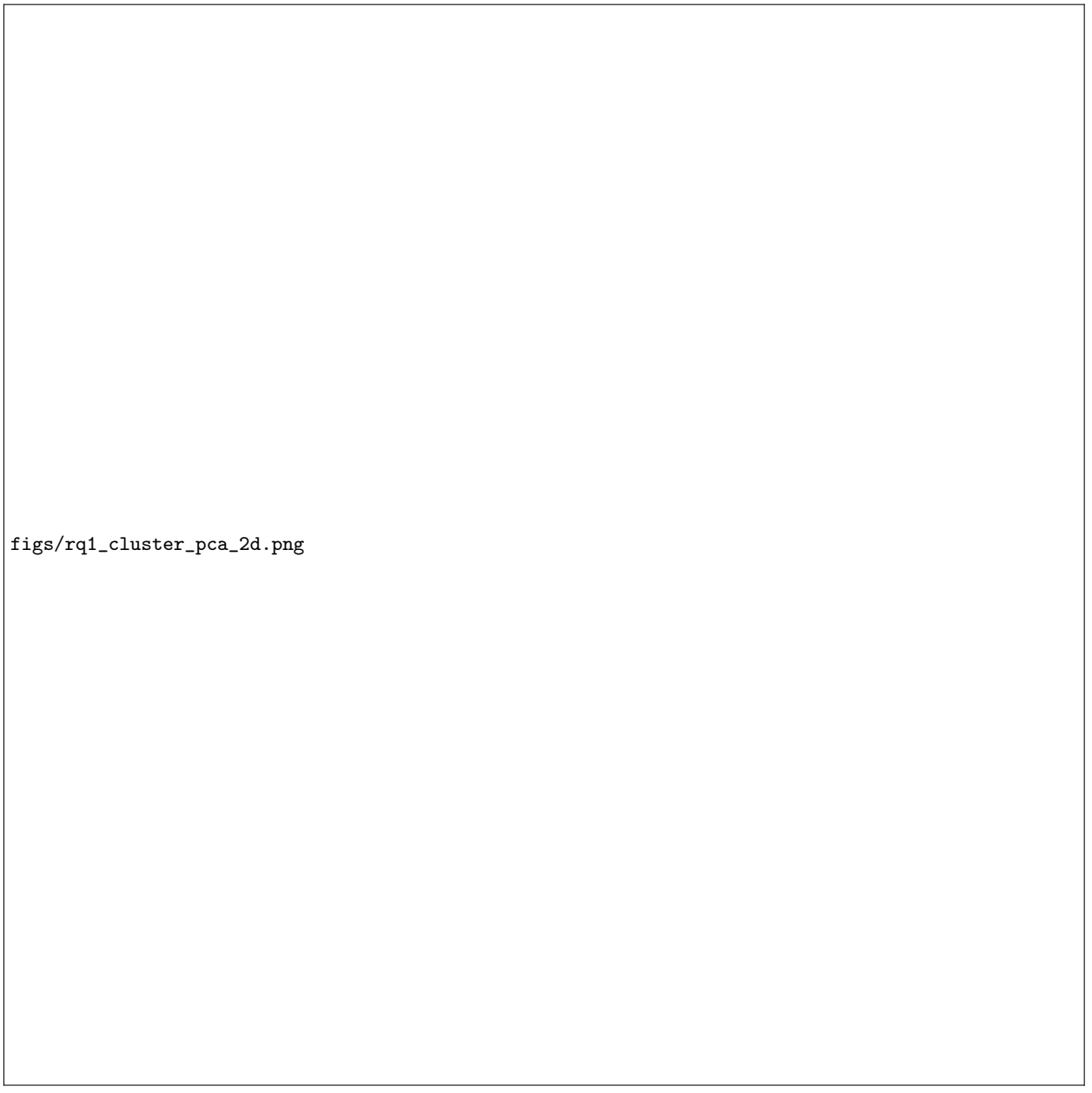

Figure 2: 2D PCA visualization of the six identified cluster profiles. The axes labels describe the main features contributing to each component, and the percentage of variance they explain.

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 **??**. It contextualizes them within existing literature and discusses their broader implications and limitations.

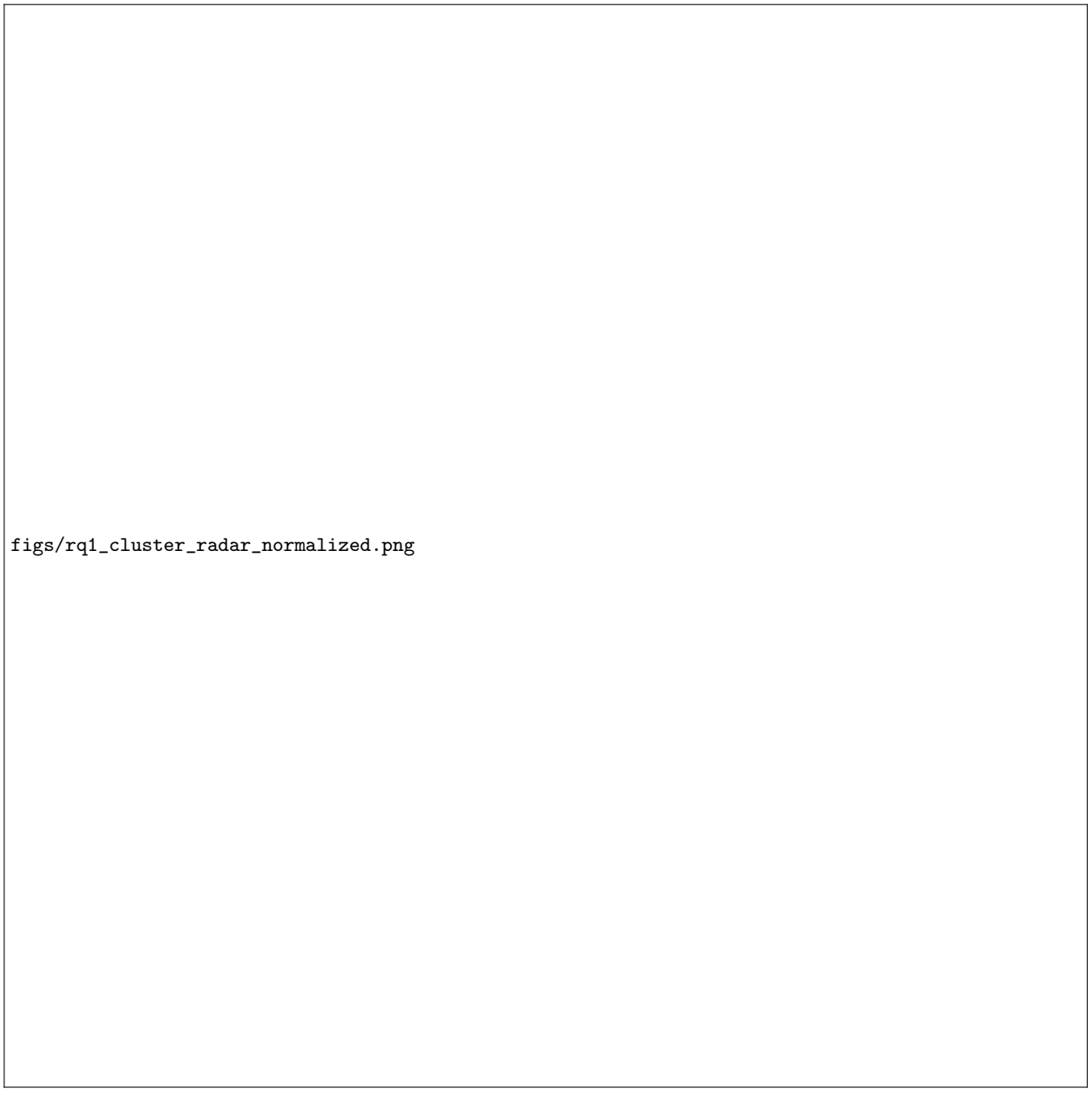

Figure 3: Radar plot of mean scaled feature values for the six cluster profiles (RQ1).

5.1. Interpretation of Resource Profiles (RQ1)

The K-Means clustering (RQ1) successfully delineated six distinct resource profiles within the commercial offers domain (Table **??**, Figure **??**). 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 (**?**). This granular profiling moves beyond simplistic high/low quality categorizations, revealing a spectrum of resource archetypes.

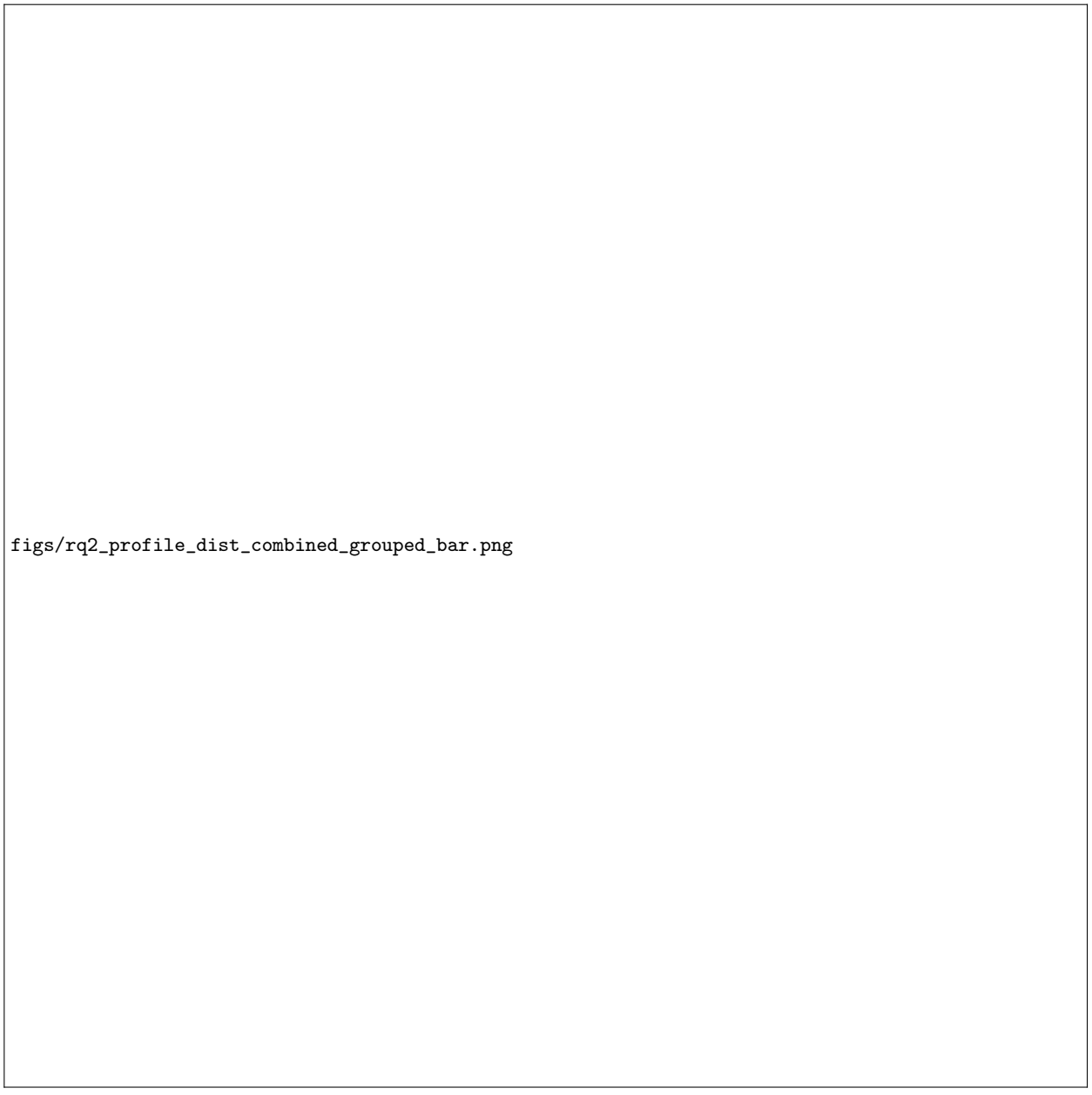

Figure 4: Profile Distribution Across Ranking Groups for Google and Bing (RQ2).

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 **??**). Stronger content-relevant profiles were generally favored in higher ranks. This aligns with the established importance of content quality (**?**).

Individual Feature Association with Ranking (RQ2 & RQ4): Ordinal Logistic Regression (Appendix **??**) and Spearman correlations (Appendix **??**, Figure **??**) offered deeper, and at times counter-intuitive, insights into system priorities.

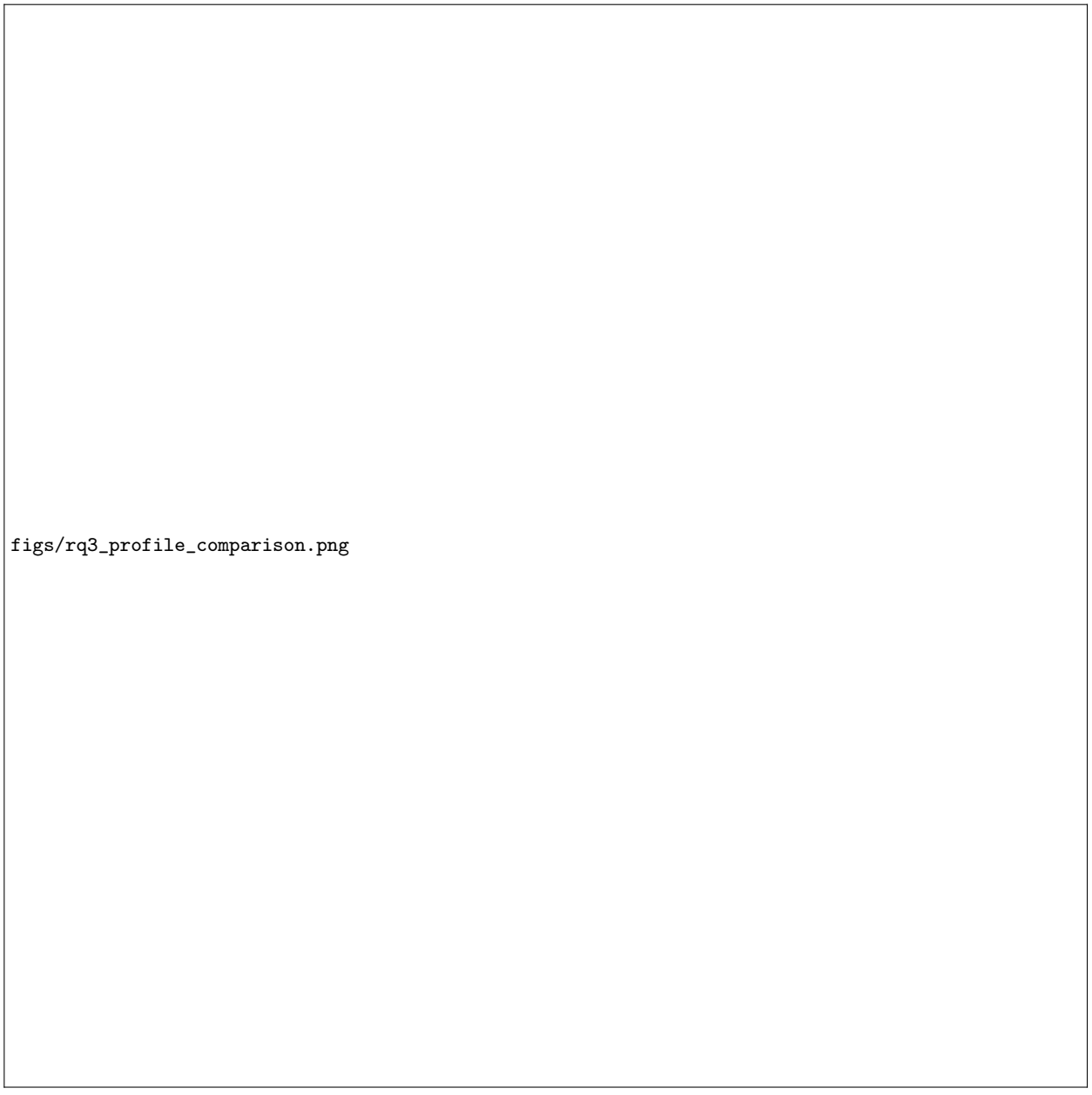

Figure 5: Comparison of Profile Distributions in Top 5 Ranks: Google vs. Bing (RQ3).

For **Google**, the combined regression model underscored a complex evaluation logic. The semantic relevance of the main content (*semantic_similarity_content_query*, coeff: -0.485) and technical accessibility (*accessibility_score*, coeff: -0.208) were strong, significant predictors of a better rank. This confirms Google's focus on deep, meaningful content and user-centric technical performance. However, two findings were particularly striking and counter-intuitive. Higher semantic similarity between the page title and the query (*semantic_similarity_title_query*) was associated with a *worse* ranking quintile (coeff: +0.202). Similarly, a higher Lighthouse *seo_score* also correlated with worse ranking outcomes (coeff: +0.168). These results suggest that, when other factors are controlled for, excessive or overly direct optimization targeting specific elements like titles might be perceived negatively, possibly as a signal of manipulation

Table 7
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**	−0.198
Access.	91.0	91.0	< 0.001***	−0.297
Best Prac.	96.0	96.0	< 0.001***	0.029
SEO	92.0	92.0	< 0.001***	−0.515
Content Relevance Metrics				
Q-Title [a]	0.0	0.0	< 0.001***	−0.270
Q-H1 [a]	0.0	0.0	< 0.001***	−0.179
ExQ-Title [a]	0.0	0.0	< 0.001***	−0.139
ExQ-H1 [a]	0.0	0.0	< 0.001***	−0.163
Q/B Density (%)	0.0	0.0	< 0.001***	−0.185
Sim. Title (0-1)	0.8	0.8	< 0.001***	−0.570
Sim. Content (0-1)	0.6	0.7	< 0.001***	−0.365
Word Count	567.0	630.0	< 0.001***	−0.302

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 are defined in Table **??**.

or low-quality optimization (**?**). Google's algorithm may favor titles that provide broader context or value beyond mere keyword matching, especially when the main content's relevance is already high.

For **Bing**, the regression analysis painted a picture of a system that, while also valuing relevance, heavily prioritizes content volume. *Word_count* emerged as the most powerful predictor of a better rank, with the largest significant negative coefficient by a substantial margin (coeff: -1.133). This implies a strong system-level preference for comprehensive, in-depth resources. Technical accessibility (*accessibility_score*, coeff: -0.339) and the presence of the exact query in H1 tags (*exact_query_in_h1*, coeff: -0.085) also significantly predicted better ranks, reinforcing the importance of fundamental technical health and clear, explicit relevance signals. Similar to Google, Bing's model also produced counter-intuitive results: higher scores for both $performance\_score$ (coeff: +0.093) and having the exact query in the title (*exact_query_in_title*, coeff: +0.064) were unexpectedly associated with worse ranking quintiles. The negative association with performance score could be due to multicollinearity or unmeasured factors, while the title result hints that both engines may share a complex, non-linear evaluation of title optimization.

Technical vs. Content Factors (RQ4): Random Forest models (Figure **??**) 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 (**?**). The improved fit of combined regression models over single-factor-type models further highlights the interdependent nature of these attribute categories.

5.3. Profiling System Differences (RQ3)

Comparative analysis (RQ3) revealed significant operational differences between Google and Bing. Bing's top results generally featured higher median values for a majority of the measured technical and content features (Table **??**), and its regression model rewarded a clear, volume-based content strategy ($word\_count$) most heavily. This suggests a more direct and perhaps more predictable ranking logic based on observable on-page factors.

In contrast, Google's logic appears more nuanced and complex. The negative coefficients for *semantic_similarity_title_query* and *seo_score* in its regression model suggest a system that may actively discount or penalize what it interprets as overt optimization attempts. Furthermore, the higher proportion of "Low Relevance" profiles (notably Cluster 1 and Cluster 3) in Google's top results (Figure **??**) could imply a greater reliance on unmeasured, off-page factors like domain authority, backlink profiles, or brand equity, which might allow certain resources to rank well despite weaker on-page metrics (**?**). This divergence suggests that Bing may currently reward a checklist-style

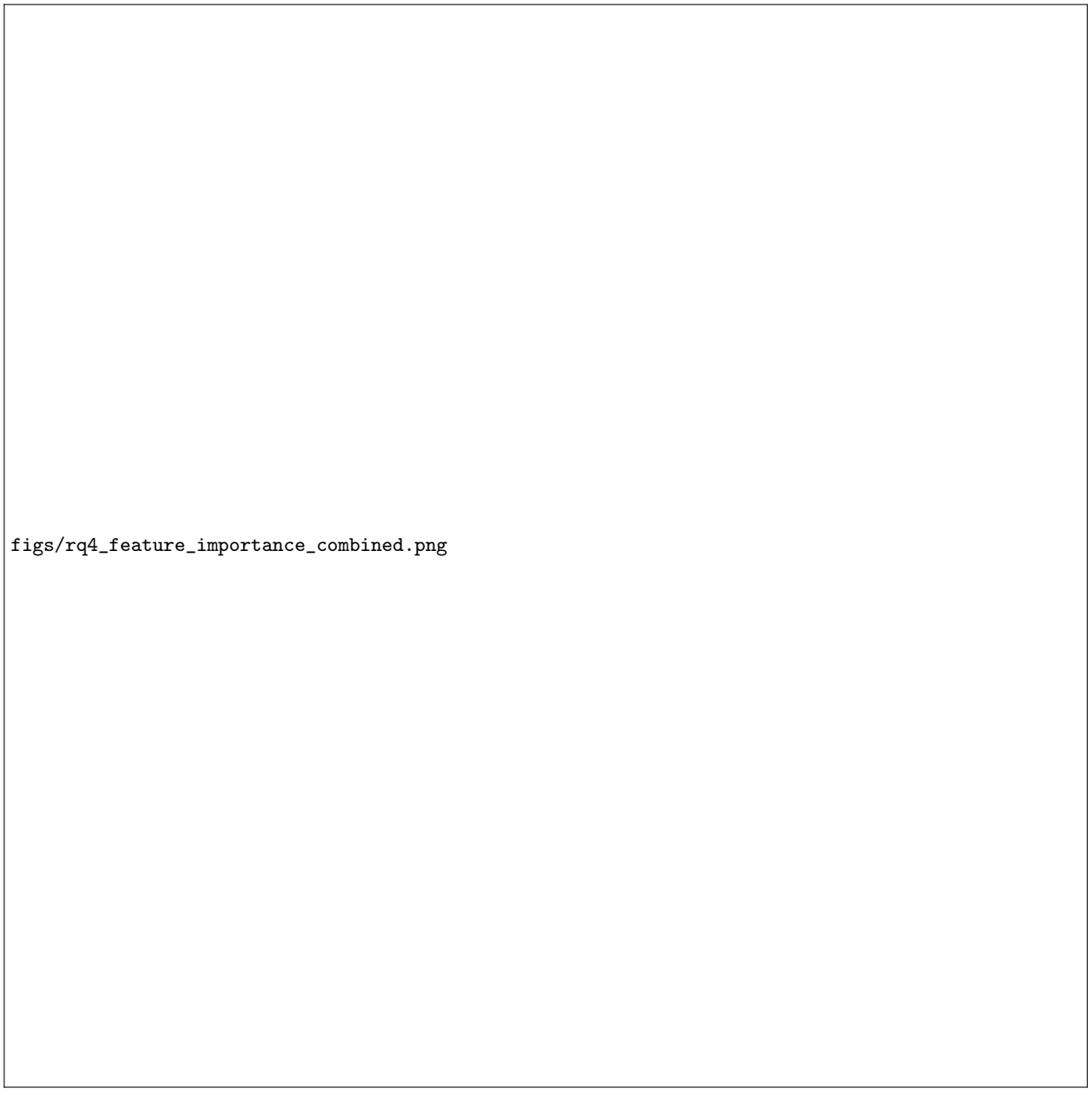

Figure 6: Relative feature importance from Random Forest for Google (left) and Bing (right) (RQ4).

optimization approach more reliably, whereas Google's algorithm requires a more holistic strategy that balances on-page relevance with a sophisticated, perhaps less direct, presentation.

5.4. Synthesis and Overall Picture

The findings paint a picture of two distinct, though overlapping, ranking ecosystems. Visibility in both is dominated by content, but the nature of that "dominant" content differs. **Bing's logic appears to be "bigger is better,"** strongly and clearly rewarding content volume (*word_count*) above all else, supplemented by explicit relevance signals.

Google's logic is more enigmatic, prioritizing the semantic depth of the main content while simultaneously showing a complex, and even punitive, response to aggressive optimization in titles and basic SEO metrics.

This study's most critical contribution is the unveiling of these counter-intuitive regression findings for both systems. The fact that certain "optimizations" correlate with worse rankings in a multivariate context powerfully illustrates that factors do not operate in a vacuum. A simple correlation may be misleading; the true impact of a feature is conditional on the presence and values of others. This highlights the inherent complexity of reverse-engineering modern IR systems and the pitfalls of relying on simplistic, single-factor analyses. The results strongly suggest that both algorithms have evolved beyond simple positive weighting of all "good" signals and now incorporate more sophisticated, context-dependent evaluations, possibly to combat manipulative practices and better approximate true user value.

5.5. Implications

Practical Implications: The updated results lead to more nuanced, actionable advice for content creators.

- For **Google**, the primary focus should be on creating high-quality, topically comprehensive content that is semantically rich and directly answers user intent (*semantic_similarity_content_query*). Technical accessibility must be flawless. However, creators should be wary of hyper-optimization in titles. A title that is too semantically close to the query may be less effective than one that offers additional context or a more natural phrasing, especially when the main content is already highly relevant. Blindly chasing a perfect Lighthouse *seo_score* may also be counterproductive.
- For **Bing**, the strategy is more straightforward: write longer, more detailed content. The strong, clear signal from the *word_count* coefficient suggests that content depth and volume are paramount. This should be combined with fundamental on-page practices like including exact-match keywords in H1 tags and ensuring high accessibility.
- The existence of multiple "successful" profiles (RQ1) and the complex regression results reinforce that a one-size-fits-all approach is obsolete. Optimization must be tailored not only to the target search engine but also to the resource's intrinsic strengths.

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:

- **Unveiling Algorithmic Complexity and Non-Linearity:** This study provides robust empirical evidence that the factor weighting in major search engines is complex, non-linear, and highly context-dependent. The identification of significant negative associations for intuitively "positive" factors (e.g., *semantic_similarity_title_query* for Google; *performance_score* for Bing) in multivariate models challenges simplistic, additive models of ranking. It contributes to a more nuanced theoretical framework where systems may penalize perceived over-optimization or balance signals in sophisticated ways.
- **System-Specific Ranking Philosophies:** The clear divergence between Google's complex, semantics-driven model and Bing's volume-centric approach provides strong empirical grounding for the theory that different IR systems develop distinct "ranking philosophies." This goes beyond surface-level differences and points to fundamental divergences in how relevance, quality, and user intent are modeled and operationalized.
- **Methodology for Isolating Nuanced Signals:** By using a content-homogeneous dataset and combining multiple analytical techniques (clustering, correlation, regression, feature importance), this study presents a powerful methodology for moving beyond broad-stroke observations to uncover subtle, system-specific factor interactions. This framework provides a template for future research aiming to "reverse-engineer" opaque systems.
- **Public Dataset for Reproducibility and Further Study:** The commitment to release the dataset and analysis scripts provides a valuable asset to the IR community, enabling replication, extension, and further exploration of these complex ranking dynamics, thereby fostering greater transparency and cumulative knowledge.

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 held greater aggregate importance than technical scores for both systems. Multivariate regression revealed distinct and complex priorities: For **Google**, strong predictors of better rank included main content semantic similarity and accessibility, while high title-query semantic similarity and a high SEO score were surprisingly associated with worse rank. For **Bing**, content volume ($word\_count$) was the most powerful predictor of better rank, while factors like performance score were unexpectedly associated with worse rank.

7.2. Concluding Remarks

This empirical study provides a data-driven characterization of two distinct search engine ranking philosophies. Visibility is not a simple sum of positive factors; it is shaped by a complex interplay of technical and content attributes, evaluated through different lenses by Google and Bing. The nuanced, and particularly the counter-intuitive, regression findings underscore the sophistication of modern IR algorithms and the folly of a one-size-fits-all optimization strategy. Google's logic appears to reward deep semantic relevance while penalizing overt optimization in specific elements like titles. Bing's logic seems more direct, placing a heavy premium on content volume as a proxy for authority and depth. These differing sensitivities emphasize the need for system-specific optimization strategies. We stress that these findings reveal correlations within a complex system, not simple causation. Inferred priorities should be interpreted as empirical patterns that reveal the distinct character of each search engine, not as immutable 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.

Table 8
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.	−0.064	0.016*	–	–	−0.064	0.016*
Access.	−0.208	< 0.001***	–	–	−0.208	< 0.001***
Best Prac.	−0.018	0.490	–	–	−0.018	0.490
SEO	0.168	< 0.001***	–	–	0.168	< 0.001***
Content Metrics						
Q-Title	–	–	−0.070	0.041*	−0.070	0.041*
Q-H1	–	–	−0.121	< 0.001***	−0.121	< 0.001***
ExQ-Title	–	–	−0.044	0.238	−0.044	0.238
ExQ-H1	–	–	0.005	0.896	0.005	0.896
Q/B Density	–	–	0.040	0.122	0.040	0.122
Sim. Title	–	–	0.202	< 0.001***	0.202	< 0.001***
Sim. Content	–	–	−0.485	< 0.001***	−0.485	< 0.001***
Word Count	–	–	0.045	0.055	0.045	0.055
Pseudo R^2	0.0046		0.0194		0.0240	
AIC	18750.1		18480.5		18480.5	
BIC	18560.7		18560.7		18560.7	

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.
Metric abbreviations are defined in Table **??**.

A. Detailed Statistical Test Results

A.1. RQ4 - Ordinal Logistic Regression Results

Table 9
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.	0.093	< 0.001***	–	–	0.093	< 0.001***
Access.	−0.339	< 0.001***	–	–	−0.339	< 0.001***
Best Prac.	0.026	0.223	–	–	0.026	0.223
SEO	−0.080	< 0.001***	–	–	−0.080	< 0.001***
Content Metrics						
Q-Title	–	–	−0.078	0.003**	−0.078	0.003**
Q-H1	–	–	0.022	0.394	0.022	0.394
ExQ-Title	–	–	0.064	0.024*	0.064	0.024*
ExQ-H1	–	–	−0.085	0.003**	−0.085	0.003**
Q/B Density	–	–	−0.036	0.079	−0.036	0.079
Sim. Title	–	–	−0.066	0.010**	−0.066	0.010**
Sim. Content	–	–	−0.093	< 0.001***	−0.093	< 0.001***
Word Count	–	–	−1.133	< 0.001***	−1.133	< 0.001***
Pseudo R^2	0.0131		0.0188		0.0319	
AIC	27222.5		27074.1		27074.1	
BIC	27158.8		27158.8		27158.8	

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.
Metric abbreviations are defined in Table **??**.

A.2. RQ4 - Spearman Correlation Heatmaps

Figure 7: Spearman Correlation Heatmaps with SERP Position for Google (left) and Bing (right) (RQ4).