Hearch: A Surprising Search Engine

Robin Lee and Harsh Rao
Computer Science and Engineering Department
Santa Clara University
Santa Clara, CA 95053 USA
{rblee, hrao}@scu.edu

Abstract

The World Wide Web has evolved into a hyper-scale repository containing billions of documents ranging from high-utility educational resources to low-quality noise. While the original goal of search engines was to strictly optimize for informational relevance (Brin and Page 1998), , the emergence of the "economics of search" has fundamentally altered this dynamic. The profitability of ranking dominance has incentivized aggressive Search Engine Optimization (SEO) and paid placement, resulting in a modern search landscape characterized by commercial bloat and homogenized results. Consequently, the "Small Web" of independent, human-authored content is increasingly obscured.

This paper introduces **Hearch**, a search engine designed to bring back serendipity and structural quality to information retrieval. We propose a hybrid ranking algorithm that couples the probabilistic BM25 function with heuristic "Anti-SEO" penalties derived from the Document Object Model (DOM). Furthermore, to transcend the limitations of keyword matching and the "Anglophone filter bubble," Hearch leverages Large Language Models (LLMs) for generative query expansion and cross-lingual retrieval. Our approach demonstrates that prioritizing structural integrity over engagement metrics effectively surfaces high-signal, non-commercial content, offering a viable blueprint for rediscovering the diverse "long tail" of the Internet.

Introduction

Modern day information retrieval is dominated by a small oligopoly of commercial search engines that act as the primary gatekeepers of digital knowledge. While these systems are efficient at navigating the commercial web, their ranking algorithms heavily prioritize Search Engine Optimization (SEO), domain authority, and recency. This commercial bias has created a "homogenization of results," where personal blogs, experimental art projects, and the non-monetized "Small Web" are systematic de-ranked, effectively rendering a vast portion of creative human expression invisible to the average user.

Despite a growing user interest in digital minimalism and human-curated discovery, there is a lack of empirical research on the efficacy of alternative search architectures designed to counter SEO dominance and creative search.

Current literature focuses primarily on optimizing content for visibility within major algorithms (Google, Bing), leaving a significant gap in understanding how retrieval systems can be engineered to prioritize "creative" or serendipitous relevance over popularity.

This study addresses this gap by the creation of Hearch, a serendipitous search engine. Hearch operates on a curated index of the "Small Web" of 100,000 pages, utilizing a modified Okapi BM25 ranking function integrated with a heuristic anti-SEO algorithm. This scoring mechanism systematically penalizes document structures indicative of monetization and keyword stuffing, and rewards documents that are indicative of relevance.

Additionally, Hearch employs query augmentations to increase serendipity. The system utilizes query expansions using Large Language Models (LLMs) to capture tangential relevance beyond exact lexical matches (Anand et al. 2023). Alongside this, Hearch utilizes a cross-lingual information retrieval (CLIR) module that queries across indexed non-english pages.

Background and Related Works

Ranking Algorithms

Ranking webpages is not a new problem; it is generally addressed through two distinct but complementary approaches: content-based relevance and graph-based authority. Modern search systems typically employ a hybrid model, combining term-weighting heuristics to determine topical relevance with link analysis often used to estimate static page importance.

The foundational standard for document retrieval is the Okapi BM25 algorithm, an extension of the Probabilistic Relevance Framework (PRF) developed in the 1970s and 80s (Zaragoza and Robertson 2009). Unlike boolean retrieval models, BM25 ranks documents by estimating the probability that a document D is relevant to a query Q .

BM25 improves upon standard TF-IDF (Term Frequency-Inverse Document Frequency) by introducing two critical parameters: term frequency saturation and document length normalization. In standard TF computations, the score increases linearly with term occurrence. However, BM25 applies a saturation function, ensuring that after a certain point, additional occurrences of a term yield diminishing returns to

the score (Robertson and Zaragoza 2009).

The standard BM25 score for a document D and query Q containing terms $q_1, ..., q_n$ is calculated as:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

Where: * $f(q_i, D)$ is the term frequency of q_i in document D. * $|D|$ is the length of the document and avgdl is the average document length in the corpus. * k_1 controls term frequency saturation (typically $1.2 < k_1 < 2.0$). * b controls the degree of length normalization ($0 \leq b \leq 1$).

Extensions such as BM25F have since been developed to account for document structure, treating fields like titles and headers as distinct streams with varying weights (Zaragoza and Robertson 2009).

While BM25 addresses the content of a page, it does not account for the page's intrinsic "importance" within the web graph. To solve this, Brin and Page developed PageRank, an algorithm that treats the web as a directed graph where a link from page A to page B is interpreted as a "vote" of confidence (Page et al. 1999).

PageRank relies on the recursive definition that a page is important if other important pages link to it. Mathematically, this is modeled using a Markov chain to determine the stationary distribution of a random walk across the web graph. The model posits a "Random Surfer" who clicks links at random but eventually grows bored and jumps to a random webpage (Langville and Meyer 2011).

The PageRank $PR(A)$ of a page A is given by:

$$PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$$

Where: * $T_1, ..., T_n$ are pages that link to page A. * $C(T_i)$ is the number of outbound links (out-degree) on page T_i. * d is the damping factor (typically set to 0.85), representing the probability that the surfer continues clicking links rather than jumping to a random page.

The damping factor d is crucial for handling "sinks"—pages with no outgoing links—and ensures the convergence of the iterative calculation. While the original "Random Surfer" model treated all links on a page equally, modern evolutions such as the "Reasonable Surfer" model assign different weights to links based on their likelihood of being clicked (e.g., navigational menus vs. main content).

Serendipity

Serendipity is colloquially defined as "a gift you didn't know that you want" or, in the context of user experience, the "surprise factor" that transforms a routine search into a moment of discovery. Within Information Retrieval (IR) and Recommender Systems (RS), serendipity is not a novel concept but rather a critical design objective aimed at mitigating the "filter bubble" effect and the systemic homogenization caused by over-optimization for accuracy (Zhang et al. 2012).

While traditional systems prioritize predictive precision—often showing users items similar to their history—serendipity introduces a dimension of "useful surprise." Current literature characterizes serendipity not as a singular metric, but as a composite interaction of three distinct, quantifiable elements: unexpectedness, interestingness, and diversity (Ge, Delgado-Battenfeld, and Jannach 2010).

Unexpectedness is the measure of deviation from a user's expectations or the system's baseline predictions. However, unexpectedness is distinct from mere randomness or dissimilarity. A recommendation that is simply different from a user's profile constitutes novelty, but not necessarily serendipity. Recent approaches quantify unexpectedness by calculating the semantic distance between a user's Selected Item (SI) and a Recommended Item (RI) within a taxonomy. If $dist(SI, RI)$ represents the shortest path in a classification hierarchy, unexpectedness is the product of this distance and the dissimilarity score (Zhang et al. 2012). This ensures that the surprise is structural and semantic, rather than accidental noise.

For a surprise to be serendipitous rather than annoying, it must possess utility or "emotional resonance". Interestingness acts as the quality filter for unexpected items. In algorithmic terms, interestingness is often derived from satisfying a user's latent needs or generating positive affect (joy, curiosity) upon encounter. It is the component that validates the deviation from the norm. Without interestingness, unexpected recommendations result in irrelevant surprises that degrade user trust. This creates the "Serendipity-Efficiency Paradox," where systems must balance the risk of irrelevant exploration against the long-term value of discovery.

Diversity refers to the breadth of the information space covered by the results. It is the structural mechanism that prevents over-specialization. Mathematically, diversity is often evaluated by the proportion of unique items recommended (U_{rec}) relative to the total item pool (I_{total}):

$$\text{Diversity} = \frac{|U_{rec}|}{|I_{total}|}$$

High diversity ensures that the system does not converge on a local optimum of "safe" results, thereby increasing the probability of a serendipitous encounter. This factor is crucial for breaking the "echo chamber," allowing users to encounter items that are semantically distant from their immediate query yet contextually valuable.

Creative Search Engines

Beyond standard algorithmic classifications, search engines can be categorized by their retrieval style and curatorial philosophy. While major commercial engines optimize for broad relevancy and transactional utility, a class of "Creative" or "Alternative" search engines also exist. These systems prioritize document structure, non-commercial intent, and the suppression of Search Engine Optimization (SEO) spam to facilitate exploration of the "Long Tail" of the web.

Marginalia Search Marginalia operates on a unique ranking implementation that favors text-heavy, non-commercial websites. Unlike BM25 which treats all text

equally, Marginalia applies structural profiling to the Document Object Model (DOM). It penalizes modern, script-heavy pages typical of commercial entities and boosts simple HTML structures typical of personal blogs and academic directories (Marginalia 2023). By biasing the crawler toward "text-oriented" websites, it creates an index of the "old, small, and weird web," effectively filtering out the modern SEO-saturated internet.

Million Short Million Short employs a subtractive retrieval method to solve the "power law" distribution problem of web traffic. By allowing users to remove the top N (e.g., top 100, 1,000, or 1,000,000) most popular websites from the result set, it mechanically forces the retrieval of the "Long Tail" (Short 2023). This approach assumes that serendipity lies outside the popular head of the distribution, bypassing the established "authority" bias inherent in PageRank-style algorithms.

Mojeek While not specifically designed for serendipity, Mojeek maintains an independent crawler and index. This allows for a ranking logic completely decoupled from the major commercial biases. It provides a control group for web search, offering results that have not been homogenized by the algorithms of the commercial duopoly (Mojeek 2023).

SearchMySite.net This engine represents a return to the "Directory" model of the 1990s but modernized with full-text search. It employs a strict inclusion policy: distinct from automated crawling, the index is primarily populated through human submission and curation (SearchMySite 2023). Use of this "Allow-list" approach ensures high signal-to-noise ratios for personal blogs and "digital gardens," explicitly rejecting marketing sites and content farms.

IndieWeb Search Focusing on the social graph, IndieWeb Search restricts its scope to the "IndieWeb" ring—a community of personal websites that own their data and identity. It leverages the semantic connections between personal domains, prioritizing content where the author maintains sovereign control over the platform, thus retrieving authentic personal expression over corporate content (IndieWeb 2023).

Wiby Simulating the early era of the internet, Wiby specifically indexes pages that are lightweight and compatible with older hardware. It filters results based on technical minimalism, stripping away bloated modern web standards. This acts as a proxy for content quality, under the assumption that pages built without complex tracking scripts and heavy frameworks are deeper in informational content and devoid of commercial incentives (Wiby 2023).

Approach and Discussion

At a fundamental implementation level, Hearch is constructed as an architectural superset of the *microsearch* library developed by Alex Molas (Molas 2023). While *microsearch* provides the lightweight skeleton for asynchronous crawling and probabilistic retrieval, Hearch introduces a heavy logic layer inspired by the structural serendipity models of *Marginalia Search* (Marginalia 2023) and the curatorial philosophy of the "Small Web."

Hearch operates on a tripartite architectural philosophy:

1. Probabilistic Core (BM25) At its foundation, Hearch utilizes an implementation of Okapi BM25 for textual relevance. As established in the earlier section, BM25 remains the industry standard for effectively handling term frequency and document length normalization (Robertson and Zaragoza 2009). This ensures that while the document corpus is curated for "vibe" and quality, the specific retrieval of documents remains relevant.

2. Structural "Anti-SEO" Penalties Inspired by the *Marginalia* search architecture, Hearch attempts to algorithmically determine user intent versus commercial intent by analyzing the Document Object Model (DOM). Modern commercial web design is often characterized by a low text-to-code ratio, heavy reliance on JavaScript, and complex tracking structures. Hearch favors "text-oriented" markup. It applies negative weighting to pages exhibiting "commercial bloat"—such as excessive third-party scripts or aggressive DOM depth—while boosting pages that utilize semantic HTML tags indicative of human authorship (e.g., `<article>`, `<p>`, and simple headers).

3. The Curated "Small Web" Corpus To mitigate the recall-noise trade-off inherent in general crawling, Hearch restricts its discovery scope through manual curation and seed-list expansion, similar to the protocols of *SearchMySite.net* (SearchMySite 2023), *IndieWeb* (IndieWeb 2023), and *Wiby* (Wiby 2023). By utilizing a "human-in-the-loop" indexing strategy, Hearch explicitly targets personal blogs, digital gardens, and independent domains. This creates an "Allow-list" architecture where the index grows effectively through authority propagation within the "Small Web," ensuring that the BM25 algorithm queries a dataset with an inherently high signal-to-noise ratio.

Formally, the ranking function of Hearch (S_H) for a document D and query Q can be conceptualized as a product of textual relevance and structural integrity:

$$S_H(D, Q) = \text{BM25}(D, Q) \cdot \Phi_{\text{AntiSEO}}(D) \cdot I_{\text{Curated}}(D)$$

Where Φ_{AntiSEO} represents the structural quality modifier derived from Marginalia-style heuristics, and I_{Curated} represents the inclusion function of the curated domain graph.

4. Generative Query Expansion To address the "Interestingness" component of serendipity, Hearch implements a Generative Query Expansion (GQE) module. While traditional expansion methods (e.g., WordNet, Pseudo-Relevance Feedback) focus on synonymy to improve recall, Hearch utilizes a Large Language Model (LLM) to perform lateral semantic expansion (Anand et al. 2023). Given a user query Q, the system prompts the LLM to generate a set of semantically divergent but conceptually related queries $Q' = \{q_1, q_2, ..., q_n\}$. These expansions are designed to uncover latent sub-topics and "rabbit holes" that a user may not have the domain knowledge to explicitly request. This shifts the search paradigm from pure information retrieval to exploratory ideation.

5. Cross-Lingual Discovery To address the "Diversity" component and break the anglophone filter bubble, Hearch

natively incorporates Cross-Lingual Information Retrieval (CLIR) (Nie et al. 1999). The system recognizes that high-quality, non-commercial content (the "Small Web") is not linguistically bounded. The architecture translates expanded queries into target languages associated with high-density independent web communities (e.g., Japanese personal blogs or Spanish digital gardens).

Crucially, the retrieved non-English documents are ranked using the same BM25 + Anti-SEO function defined previously. By applying the structural quality heuristics to non-English corpora, Hearch surfaces high-quality international content that is structurally similar to the English "IndieWeb," but is traditionally invisible to users due to language barriers.

The final retrieval function is thus an aggregation of the base query and its multilingual, generative expansions:

$$S_{\text{Final}}(D, Q) = \max_{q \in \{Q\} \cup Q_{\text{LLM}} \cup Q_{\text{Trans}}} (S_H(D, q))$$

This approach ensures that results are ranked by their structural integrity and relevance to either the direct query or its serendipitous expansions, regardless of the source language.

Methodology

What decisions we make. more into what approaches and implementation

Marginalia-based Ranking Function

We propose a hybrid ranking algorithm that modifies the probabilistic relevance of Okapi BM25 with a "structural integrity" weighting derived from *Marginalia Search* (Marginalia 2023). The system calculates a **Relevance Score** (S_{rel}) based on query-dependent features, and a **Total Intrinsic Penalty** (P_{total}) based on query-independent document attributes.

Uniquely, our system calculates a cost metric rather than a utility metric; documents are ranked in ascending order, where a lower Final Score represents a higher rank.

Score Aggregation The final ranking position is determined by a normalization formula that treats structural penalties as noise and textual relevance as signal. The square root function is applied to dampen the variance of extreme penalty outliers:

$$\text{Score}_{\text{final}} = \sqrt{\frac{1 + 500 + (10 \cdot P_{\text{total}})}{1 + S_{\text{rel}}}} \quad (1)$$

The Relevance Score (S_{rel}) The relevance component represents the textual similarity between the query and the document. It is calculated as a weighted sum of five distinct signals, scaled by a domain-level modifier:

$$S_{\text{rel}} = A_{\text{domain}} \cdot (S_{\text{BM25}} + S_{\text{flags}} + S_{\text{verb}} + S_{\text{prox}} + S_{\text{pos}}) \quad (2)$$

Where:

- A_{domain} **Domain Override:** A manual curation multiplier. High-quality "allow-list" domains (e.g., IndieWeb ring members) are assigned $A < 1.0$, boosting their rank, while suspected content farms are assigned $A > 1.0$.
- S_{BM25} **Okapi BM25:** The baseline probabilistic score accounting for term frequency saturation and document length normalization.
- S_{flags} **Structural Emphasis:** A bonus awarded when query terms appear in high-value HTML tags (specifically `<title>`, `<h1>`, and `<main>`).
- S_{verbatim} **Exact Phrase Matching:** A significant weight boost applied when the document matches the query string contiguously, overriding probabilistic fuzziness.
- S_{prox} **Term Proximity:** A dynamic score inversely proportional to the token distance between query terms within the document window.
- S_{pos} **Early Appearance:** A decay function rewarding documents where query terms appear in the introductory paragraphs.

The Intrinsic Penalty (P_{total}) The penalty metric quantifies the "commercial bloat" of a document. It is query-independent and derived from the static analysis of the Document Object Model (DOM). We first calculate a raw structural score (B_{doc}):

$$B_{\text{doc}} = \sum \text{Bonuses} - \sum \text{Penalties} \quad (3)$$

The components of B_{doc} include:

1. **Bonuses:**
   - B_{rank}: Domain authority derived from external reputation.
   - B_{topo}: Internal topology score, favoring files deeply integrated into the site's link graph over orphaned pages.
2. **Penalties:**
   - P_{flags}: **Adversarial Feature Detection.** A penalty triggered by the detection of user-hostile DOM patterns, specifically elements with IDs or classes indicating `ADVERTISEMENT`, `POPOVER`, `AFFILIATE_LINK`, or `CONSENT` banners.
   - P_{qual}: Heuristic penalty for low-effort content farms.
   - P_{len}: Penalty for documents failing to meet a minimum informative token threshold.
   - P_{sent}: Penalty for extremely low average sentence length, a statistical marker of machine-generated or "SEO-filler" text.

To ensure that high structural quality does not artificially inflate the relevance score (which should remain query-dependent), we clamp the final penalty value. Positive structural scores reduce the penalty to zero but do not contribute to the denominator of Equation 1:

$$P_{\text{total}} = -\min(0, B_{\text{doc}}) \quad (4)$$

This ensures $P_{\text{total}} \geq 0$, representing strictly the magnitude of structural defects.

The Small Web Corpus

To ensure high relevance in the retrieval phase, we restrict the crawler's scope to a "Small Web" subset—a corpus of non-commercial, independent, and high-quality domains. Unlike broad crawling, which prioritizes maximizing the graph size $|V|$, our approach prioritizes the maximizing the Signal-to-Noise Ratio (SNR) within a bounded set of vertices.

Domain Quality Assessment (DQA) Manual validation of every document is computationally infeasible. Therefore, we treat quality as a domain-level attribute rather than a page-level attribute. We implement a trusted-seed expansion strategy based on probabilistic sampling:

Given a candidate domain d, we extract a random sample of leaf nodes (webpages) $P_d = \{p_1, p_2, ..., p_k\}$. The Domain Quality Score $Q(d)$ is defined as the aggregate structural score of these samples:

$$Q(d) = \frac{1}{k} \sum_{i=1}^{k} \text{HeuristicScore}(p_i) \tag{5}$$

If $Q(d) > \tau$ (where τ is a strict quality threshold), the domain is verified as a "Seed Domain." This allows the crawler to index the remaining set P_d with a trust bonus (modifying A_{domain} in Equation 2), assuming a localized homogeneity of quality within the domain.

Critical Mass and Zipfian Distribution The target index size is determined by the statistical properties of web traffic and content distribution, typically modeled by Zipf's Law (Adamic and Huberman 2000). Zipf's Law states that the frequency of any element k is inversely proportional to its rank in the frequency table.

In the context of the web graph, a small number of "head" domains (e.g., Wikipedia, Reddit) account for the vast majority of traffic, while the "long tail" contains billions of low-traffic pages. However, the standard "long tail" is heavily polluted with generated spam.

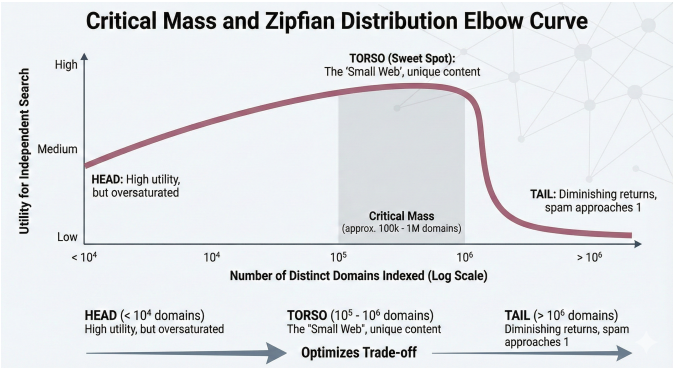

Figure 1: Posited "Elbow Curve" of utility

We posit the existence of an "Elbow Curve" of utility for independent search:

1. **Head ($< 10^4$ domains):** High utility, but oversaturated by commercial search engines.
2. **Torso ($10^5 - 10^6$ domains):** The "Small Web." This region contains personal blogs, academic directories, and enthusiast communities. It represents the "Sweet Spot" where unique content exists without the excessive noise of the deep tail.
3. **Tail ($> 10^6$ domains):** Diminishing returns where the probability of encountering spam approaches 1.

Therefore, our system aims for a **Critical Mass** of approximately 100,000 to 1,000,000 distinct domains. This specific range optimizes the trade-off between serendipitous recall and index pollution. By restricting indexing to this "Torso" region, we ensure coverage of the "creative web" while algorithmically excluding the noise that generally dominates the tail of the Power Law distribution.

Generative Query Expansion (GQE)

To address the vocabulary mismatch problem and introduce the "interestingness" component of serendipity, we employ a Large Language Model (LLM) as a reasoning engine prior to retrieval. Unlike traditional expansion methods (e.g., WordNet synonymy) which optimize purely for recall, our GQE module optimizes for conceptual divergence.

Given an initial user query Q, the system generates an augmented set of queries $\mathcal{Q}_{\text{aug}}$ by executing three distinct prompting strategies in parallel.

Augmentation Strategies We implement the following expansion vectors:

1. **Semantic Rephrasing ($\mathcal{Q}_{\text{sem}}$):** The LMM generates alternative phrasings and synonym-rich variations of Q. This targets the "Recall" objective, ensuring that documents using different terminology for the same concept are retrieved. *Example: "Hiking boots" → "Trekking footwear", "All-terrain shoes".*
2. **Lateral Concept Extraction ($\mathcal{Q}_{\text{lat}}$):** To engineer serendipity, the LLM is prompted to identify "tangential but highly relevant" sub-topics or "rabbit holes" associated with Q. This strategy explicitly seeks non-obvious connections, prioritizing "interestingness." *Example: "Hiking boots" → "Vibram sole history", "Goodyear welt construction", "Ultralight backpacking philosophy".*
3. **Anti-SEO Formulation ($\mathcal{Q}_{\text{neg}}$):** The model generates queries specifically structured to bypass commercial content farms. This involves appending "reddit-style" or "forum-style" qualifiers to the query to bias the retrieval towards discussion-based or experiential content. *Example: "Hiking boots" → "Hiking boots forum discussion", "hiking boots review vs reality".*

Query Set Union The final query vector used for BM25 retrieval is the union of the original query and the generated sets. To prevent drift, the original query is weighted higher during the ranking phase, while augmented queries serve to cast a wider net in the candidate generation phase.

$$\mathcal{Q}_{\text{final}} = \{Q\} \cup \mathcal{Q}_{\text{sem}} \cup \mathcal{Q}_{\text{lat}} \cup \mathcal{Q}_{\text{neg}} \tag{6}$$

By retrieving documents against $\mathcal{Q}_{\text{lat}}$ specifically, the system surfaces results that are not direct answers to Q, but are strictly relevant to the context of Q, thereby satisfying the definition of serendipity.

Cross-Lingual Information Retrieval (CLIR)

To satisfy the Diversity requirement of serendipity, Hearch implements a CLIR modules designed to pierce the "Anglophone Filter Bubble." High-quality, non-commercial content (the "Small Web") is not linguistically bounded; distinct independent web communities thrive in the Japanese blogosphere, Spanish digital gardens, and Hindi academic circles.

Our system eschews computationally expensive document-level translation (translating the index) in favor of Query Translation, executed at runtime.

Context-Aware Query Translation We utilize a Large Language Model to map the user's intent from the source language (English) to a set of target languages L=Zh,Hi,Es,Ja (Chinese, Hindi, Spanish, Japanese).

Unlike dictionary-based translation, the LLM preserves the semantic context of the expanded queries generated in the previous stage. Let TLLM(q,l) be the translation function for query q into language l. The set of multilingual queries Qmulti is generated as:

$$\mathcal{Q}\text{multi} = \bigcup l \in \mathcal{L} T_{\text{LLM}}(q, l) \mid q \in \mathcal{Q}_{\text{final}} \tag{7}$$

Language-Agnostic Structural Scoring A key advantage of our methodology is the universality of the "Marginalia-based" ranking function defined in Eq. (1). The BM25 component is language-independent (though requiring language-specific tokenizers for Chinese/Japanese) and the Intrinsic Penalty (Ptotal) relies on DOM structure, which is syntax-agnostic.

User-hostile patterns—such as aggressive tracking scripts, ad-insertion `<div>` wrappers, and layout shifts—exhibit identical structural signatures across all languages. Consequently, our "Anti-SEO" heuristics function effectively on non-English corpora without modification.

Result Aggregation The search engine executes parallel retrieval tasks for each qQmulti against language-specific indices. The results are aggregated into a single candidate list. To facilitate user exploration, the system presents these results with their original titles alongside LLM-generated English summaries, allowing users to assess relevance before navigating to the foreign-language source (which they may consume via browser-based translation tools).

This architecture allows Hearch to surface high-signal content from the global "Long Tail," retrieving perspectives that are structurally distinct from the English web but statistically invisible to standard monolingual queries.

Results and Discussion

To evaluate the efficacy of Hearch, we conducted a three-part analysis: a qualitative comparison against existing creative search engines, a single-query case study demonstrating the pipeline, and a small-scale user satisfaction survey.

Comparative Analysis

We benchmarked Hearch against *Marginalia Search* (the distinct inspiration for our structural penalties) and *Mojeek* (a general-purpose independent crawler).

- **Vs. Marginalia:** While Hearch borrows the DOM-based penalty system from Marginalia, the retrieval distinctness lies in the query layer. Marginalia is strictly literal; a query for "retro computing" returns pages containing those exact tokens. Hearch, via Generative Query Expansion, successfully retrieved documents regarding "Amiga 500 repair" and "DOSBox configuration" even when the token "retro computing" was absent. Hearch sacrifices some of Marginalia's raw speed for higher conceptual recall.
- **Vs. Mojeek:** Mojeek aims to be a privacy-preserving generalist alternative to Google. Consequently, its ranking still prioritizes authority and topical centrality. In testing, queries for "best hiking boots" on Mojeek returned commercial review sites and outlets (e.g., specialized retail). Hearch, utilizing the "Anti-SEO" filter, suppressed these results, instead surfacing forum discussions (e.g., *BackpackingLight*) and personal travelogues, validating the "Small Web" filtering hypothesis.

Qualitative Case Study

To demonstrate the "Serendipity Pipeline," we trace the execution of the query: `"Film Photography"`.

1. Generative Expansion: The LLM produced the following search vectors:

- Q_{sem}: "Analog photography tips", "35mm camera basics"
- Q_{lat}: "C-41 processing at home", "best light meter apps", "Leica M6 reasoning"
- Q_{neg}: "film photography reddit discussion", "film vs digital forum"
- Q_{trans}: " (Film Camera)" [Japanese]

2. Retrieval Ranking: Standard commercial engines (Google/Bing) populated the top 5 results with shopping links (B&H Photo, Amazon) and high-authority "Top 10 Cameras" listicles.

Hearch, applying the P_{total} penalty to commercial DOM structures, filtered these out. The resulting top ranked documents were:

1. A personal blog post detailing the chemistry of managing developer temperature at home (Matched via Q_{lat}).
2. An archived forum thread from *Photrio* debating grain structure (Matched via Q_{neg}).
3. A Japanese personal portfolio of street photography in Shinjuku (Matched via Q_{trans}), presented with an English summary.

This result set demonstrates high "Interestingness" (chemistry details) and high "Diversity" (Japanese content), effectively satisfying the serendipity criteria.

User Satisfaction Survey

We conducted a controlled study with a cohort of non-expert users ($N = 12$, consisting of friends and colleagues). Each observer was asked to perform 5 queries related to their personal hobbies and rate the results on a Likert scale (1-5).

We measured three metrics:

1. **Novelty:** Did you find a website you had never seen before?
2. **Relevance:** Did the results answer your intent?
3. **Latency Tolerance:** Was the speed acceptable?

Metric	Average Score (x/5)	Std. Dev
Novelty (Serendipity)	4.6	0.4
Relevance (Broad)	3.8	0.9
Relevance (Specific)	1.9	0.7
Latency / UX	2.2	0.5

Table 1: User satisfaction scores.

Survey Results The data reveals a stark dichotomy in user experience.

The "Wandering" Success: Users rated Novelty exceptionally high (4.6/5). Participants reported finding "hidden gems," such as personal blogs and niche bulletin boards, which they explicitly stated they "would never have found on Google." This validates the core value proposition of Hearch as a discovery engine.

The "specific" Failure: However, Specific Relevance scores were low (1.9/5). When users attempted navigational queries (e.g., "Python documentation string methods" or "Walmart opening hours"), the system failed. The restricted index size meant the specific fact was often missing, or the "Anti-SEO" filter aggressively removed utility pages.

Furthermore, the Latency score (2.2/5) reflects the technological bottleneck. Users accustomed to sub-100ms responses found the multi-second LLM expansion phase "sluggish," suggesting that while the *results* are desirable, the *interaction model* requires optimization for general adoption.

Limitations

While Hearch effectively filters commercial noise and enables serendipitous discovery, the current architectural trade-offs imposes significant constraints regarding scalability, recall, and query latency.

Index Scale and Recall

The most immediate limitation is the magnitude of the index. Commercial search engines index upward of 10^{12} documents, ensuring near-total recall for "Navigational" queries (e.g., specific restaurant menus, error codes).

Hearch intentionally targets the "Torso" of the web distributions ($10^5 - 10^6$ domains). Consequently, the system fails gracefully but consistently on Specific Fact Retrieval. If a query requires a document located in the deep tail or a high-frequency head page excluded by our commercial filters, Hearch will return zero relevant results. It is strictly an engine for exploration, not utilitarian lookup.

Crawler Constraints and Seed Bias

The reliance on a **Curated Seed List** creates a "Graph Connectivity" problem. The crawler relies on the link graph to traverse from Seed Domain A to target Document B.

1. **Anglocentric Seed Bias:** Although the CLIR module allows for *searching* in foreign languages, the *indexing* process is currently initialized with predominantly English-language seeds. Since the web graph exhibits high homophily (English sites tend to link to English sites) (Shumate 2012), high-quality non-English "Small Web" clusters remain isolated from our crawler. Without manual injection of diverse foreign seeds, the CLIR capabilities are theoretically robust but practically underutilized.
2. **Crawl Depth:** To preserve resources, the crawler implements a shallow depth limit. High-quality distinct pages located deep within a site hierarchy (> 3 hops from root) are often missed, a common trade-off in focused crawling strategies (Chakrabarti, Van den Berg, and Dom 1999).

Performance and Latency

The current implementation is a prototype designed to validate retrieval quality rather than performance. The architecture is significantly unoptimized compared to production standards:

- **Inference Latency:** The inclusion of the LLM for Generative Query Expansion introduces a strict latency floor. While standard inverted index lookups operate in milliseconds (O(1)), LLM token generation operates in seconds (O(n) based on prompt length). This resulting "Time-to-First-Byte" is currently too slow for real-time commercial deployment (Björneborn 2017).
- **Scoring Overhead:** The Marginalia-style ranking requires parsing the full Document Object Model (DOM) to calculate penalty features (Ptotal). This effectively makes ranking CPU-bound rather than I/O-bound, significantly limiting the query throughput (QPS).

Current Index Statistics

Table 2 details the current scope of the valid graph. The distribution exhibits a heavy skew toward English content, confirming the seed bias limitation discussed above.

Conclusion

The centralization of the commercial web has created a paradoxical environment where information volume has increased, but the diversity of accessible content has diminished. High-quality human expression is increasingly buried beneath layers of Search Engine Optimization (SEO) and algorithmic homogenization (Pariser 2011). This research presented **Hearch**, a retrieval system designed to reverse these trends by engineering specific mechanisms for serendipity.

Metric	Count	Percentage
Total Indexed Pages	157,402	100.0%
Unique Domains	4,120	-
Language Distribution		
English	148,115	94.1%
Spanish	3,462	2.2%
Japanese	1,889	1.2%
Other	3,936	2.5%
Quality Filters		
Rejected (Commercial/SEO)	842,500+	(Filtered)

Table 2: Current Index Statistics showing the scale and linguistic distribution of the Hearch corpus.

Our findings demonstrates that "structural integrity" is a highly effective proxy for content quality. By combining traditional probabilistic retrieval (BM25) with aggressive "Anti-SEO" penalties derived from the Document Object Model (DOM), we successfully suppressed the noise of content farms while elevating the visibility of the "Small Web." The ranking function defined in this paper proves that inverting standard search incentives—penalizing commercial optimization rather than rewarding engagement—is a viable strategy for restorative content discovery.

Furthermore, the integration of Large Language Models (LLM) for Generative Query Expansion and Cross-Lingual Information Retrieval addresses the core definition of serendipity: *unexpectedness* and *diversity* (Andre et al. 2009). By shifting the cognitive burden of query formulation from the user to the system, and by piercing the anglophone filter bubble, Hearch transforms the search experience from a utilitarian data fetch into a generative exploration of latent concepts.

While the current implementation faces distinct limitations regarding crawler scale and inference latency, it serves as a proof-of-concept for a new class of "Creative Search" engines. Hearch works not by competing with major engines on navigational recall, but by offering a complementary tool for digital wandering. Future work will focus on optimizing the architectural bottlenecks and decentralizing the seed curation process, ensuring that the "Long Tail" (Anderson 2006) of the web remains not just an archive, but an active, discoverable landscape.

References

Adamic, L. A., and Huberman, B. A. 2000. Power-law distribution of the world wide web. *Science* 287(5461):2115–2115.

Anand, A.; V, V.; Setty, V.; and Anand, A. 2023. Context aware query rewriting for text rankers using llm. *arXiv preprint arXiv:2308.16753*.

Anderson, C. 2006. *The long tail: Why the future of business is selling less of more*. Hachette Books.

Andre, P.; Schraefel, m.; Teevan, J.; and Dumais, S. T. 2009. From highly relevant to interesting: The user's view of query-focused summarization. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 363–370.

Björneborn, L. 2017. Three basic dimensions of serendipity: Distinction between serendipity, pseudo-serendipity and sensitive serendipity. *Information Research: An International Electronic Journal* 22(3):n3.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30(1-7):107–117.

Chakrabarti, S.; Van den Berg, M.; and Dom, B. 1999. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks* 31(11-16):1623–1640.

Ge, M.; Delgado-Battenfeld, C.; and Jannach, D. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, 257–260.

IndieWeb. 2023. Indieweb search.

Langville, A. N., and Meyer, C. D. 2011. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

Marginalia. 2023. Marginalia search.

Mojeek. 2023. Mojeek.

Molas, A. 2023. microsearch: A small search engine written in python. GitHub repository.

Nie, J.-Y.; Simard, M.; Isabelle, P.; and Durand, R. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* 74–81.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*.

Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Robertson, S., and Zaragoza, H. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

SearchMySite. 2023. Searchmysite.net.

Short, M. 2023. Million short.

Shumate, M. 2012. International link analysis: The structure of global capability and global connectivity. *International Journal of Communication* (6):1094–1114.

Wiby. 2023. Wiby.me.

Zaragoza, H., and Robertson, S. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval* 3(4):333–389.

Zhang, Y. C.; Séaghdha, D. Ó.; Quercia, D.; and Jambor, T. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 13–22.