Information Sources for Query Refinement: A Comprehensive Survey

Zahra Taheri[1], Mahdis Saeedi[1], Ziad Kobti[1] and Hossein Fani[1*]

[1]School of Computer Science, University of Windsor, Windsor, ON, Canada.

*Corresponding author(s). E-mail(s): hfani@uwindsor.ca;
Contributing authors: taherik@uwindsor.ca; msaeedi@uwindsor.ca; kobti@uwindsor.ca;

Abstract

The Web remains the largest and most dynamic information source available today. Web search plays a vital role in helping users locate relevant content for research, learning, and decision-making. A critical factor in search effectiveness is the system's ability to interpret user queries accurately. However, queries are often ambiguous, brief, or underspecified, leading to suboptimal retrieval results. Query refinement techniques address these issues by modifying the initial query through term addition, deletion, or substitution to better capture user intent. This paper presents a comprehensive survey of query refinement methods with a specific focus on the types of information sources leveraged in these techniques. We classify information sources into two primary categories: non-contextual (e.g., thesauri, WordNet, Wikipedia, retrieved documents) and contextual (e.g., search sessions, click-through data, temporal data, and social signals). For each category, we analyze recent methods that utilize these sources to support query expansion, suggestion, and reformulation through a systematic review of 67 papers published between 2016-2024. Our comparative analysis evaluates methods across effectiveness, scalability, robustness to query drift, and personalization capabilities, highlighting open research directions in this evolving field.

Keywords: Query refinement, Information retrieval, Query expansion, Search personalization, Contextual search

1 Introduction

Web search is a ubiquitous entry point for information-seeking activities across domains such as education, e-commerce, health, and enterprise. Users express their needs through short keyword queries, which are often ambiguous, incomplete, or semantically imprecise. These limitations hinder retrieval systems from correctly interpreting the user's intent and retrieving the most relevant results. Studies have shown that the average query length remains under three words [24], contributing to vagueness and under-specification in user input. To address this issue, query refinement techniques aim to improve search performance by reformulating the original query to better reflect the user's intent. These techniques include a range of strategies such as query expansion [1], suggestion [3], and reformulation. While the goals and mechanisms of these strategies differ slightly, they share a common objective: to produce queries that yield more relevant results. A key factor in the success of any query refinement method lies in the type and quality of the information it leverages. Traditionally, many approaches have relied on lexical resources, manually crafted thesauri, or term co-occurrence statistics. However, recent advances in machine learning and data availability have enabled the use of rich, heterogeneous sources of information from user interaction logs and session data to structured knowledge bases and temporal behavior patterns.

This paper focuses on categorizing and analyzing the types of information sources used in contemporary query refinement research through a systematic survey methodology. We divide these sources into two primary categories: **Non-contextual sources**, which do not depend on the user's behavior or history (e.g., lexical databases, Wikipedia, or initial retrieved documents), and **Contextual sources**, which rely on user-centric data such as session history, clicked documents, and social or temporal behavior.

Research Questions: This survey addresses three key research questions:

1. What are the primary categories of information sources used in modern query refinement techniques?
2. How do different information sources compare in terms of effectiveness, scalability, and personalization capabilities?
3. What are the current challenges and future research directions in leveraging diverse information sources for query refinement?

2 Survey Methodology and Information Source Definition

2.1 Survey Methodology

To ensure comprehensiveness and reproducibility, we followed a systematic survey methodology with clearly defined inclusion/exclusion criteria:

Database Sources: We searched the following academic databases: ACM Digital Library, IEEE Xplore, Springer, Elsevier (ScienceDirect), and arXiv.

Time Span: Papers published between January 2016 and December 2024, focusing on recent advances while ensuring contemporary relevance.
Search Terms: Our search strategy included combinations of: 'query refinement', 'query expansion', 'query suggestion', 'query reformulation', 'information retrieval', 'search personalization', and 'contextual search'.
Inclusion Criteria: a) Peer-reviewed conference papers and journal articles; b) papers focusing on query refinement techniques in IR; c) methods that explicitly leverage different information sources; d) papers with clear experimental validation.
Exclusion Criteria: a) Papers older than 2016 (to focus on recent advances); b) Non-English publications; c) Workshop papers without substantial technical contribution; d) Duplicate studies or minor variations of existing work. **Selection Process:** Initially, 156 papers were identified through database searches. After applying inclusion/exclusion criteria and removing duplicates, 67 papers were selected for detailed analysis. Two researchers independently reviewed abstracts and full texts, with disagreements resolved through discussion.

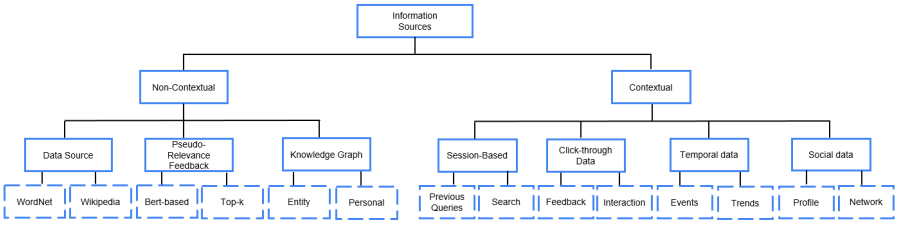

Fig. 1 Taxonomy of Information Sources for Query Refinement. The taxonomy categorizes sources into non-contextual and contextual types, highlighting subcategories such as external knowledge, retrieved documents, knowledge graphs, session data, click-through data, temporal data, and social data.

2.2 Information Source Definition

Recent advances in information retrieval have increasingly relied on diverse information sources to better capture users' needs and preferences. Effectively identifying and utilizing these sources is essential for improving retrieval performance. Figure 1 illustrates the proposed taxonomy of information sources for query refinement, categorized into two main groups: **Non-contextual Information** and **Contextual Information** This categorization provides a structured framework for understanding how different information types contribute to query refinement effectiveness.

3 Non-Contextual Information Sources

Non-contextual query refinement focuses exclusively on the content of the query itself, without incorporating additional contextual signals such as user

behavior or search session history. These approaches aim to enhance the effectiveness of the original query by leveraging external knowledge resources, rather than personalized or situational information.

3.1 External Data Sources

As previously discussed, query expansion techniques enhance an initial query by incorporating additional, semantically related terms. A notable example is the model proposed by Lucchese et al. [15], which employs a thesaurus to support query expansion through structured queries and efficiency-aware term selection strategies. This model is particularly well-suited for real-time applications, as it expands the original query into a Conjunctive Normal Form (CNF) by selecting candidate terms from a thesaurus. Beyond thesaurus-driven approaches, domain-specific keyword-based query languages have also been shown to improve retrieval precision. For instance, Rencis [36] introduced a configurable keywords-based query language tailored for the healthcare domain, enabling users to construct structured queries with precise control over term inclusion, synonyms, and logical operators. While similar in spirit to thesaurus-supported expansion, this approach emphasizes manual configurability over automatic term selection, making it particularly suited to specialized domains where terminology precision is critical.

To address these limitations, Azad et al. [4] proposed an advanced approach that employs Wikipedia and WordNet for query expansion. This method provides rich expansion terms for phrase queries using Wikipedia and for individual terms using WordNet. Their approach involves a two-level strategy for term selection from WordNet: initially extracting the synsets of the query terms, followed by retrieving the synsets of these synsets to capture a broader range of related terms.

Using resources like WordNet, Wikipedia, and thesaurus for query expansion has several advantages and disadvantages. On the positive side, these resources provide rich semantic frameworks, enabling the identification of synonyms, hyponyms, and related terms, which enhance the retrieval of relevant documents. They offer domain-specific knowledge, particularly useful in specialized fields, and their structured nature helps in precise and contextually appropriate term expansion. However, there are notable drawbacks. The static nature of these resources means they might not capture the latest terminology or emerging trends, potentially leading to outdated expansions. Furthermore, they may not fully capture context-specific nuances, leading to the inclusion of contextually irrelevant terms, and their polysemous terms can cause ambiguous expansions.

3.2 Pseudo-Relevance Feedback (PRF) Methods

Another important source of information for query refinement is the set of documents retrieved in response to the initial query. Methods that leverage this feedback are referred to as pseudo-relevance feedback (PRF) techniques. In

PRF, the top-ranked retrieved documents are assumed to be at least partially relevant and are used to refine or expand the original query. Recent PRF methods [12, 13] utilize entire feedback documents for expansion. However, this often leads to overly verbose and noisy expansions that may degrade performance. To address this issue, Zheng et al. [5] propose a BERT-based query expansion method that selectively incorporates relevant content. Their approach begins by identifying the top-ranked documents using a fine-tuned BERT model.

Beyond query expansion, PRF documents can also be leveraged for query suggestion, particularly in domains such as scientific literature search, where users may lack the domain expertise to judge the relevance of query suggestions. In this context, Medlar et al. [9] introduce a method designed for exploratory search, which generates alternative queries that are independent of the original query but capable of retrieving similar content to that currently displayed. Their approach works by summarizing the content of search result documents and generating new queries based on those summaries. In terms of query representation, most existing models produce a single representation for each query. However, this can be limiting when a query expresses multiple underlying intents. To address this, Hashemi et al. [8] propose a model based on BART that learns multiple distributed representations for each original query. Their method uses the top-ranked documents from the retrieval results to uncover different semantic aspects of the query, allowing the model to capture diverse user intents and produce more accurate, intent-aware representations.

In conclusion, leveraging retrieved documents for query expansion—commonly referred to as pseudo-relevance feedback (PRF)—offers both notable advantages and inherent limitations. On the positive side, PRF enhances the contextual richness of the query by incorporating terms from top-ranked documents, which can improve retrieval precision. It also increases recall by adding synonyms, semantically related terms, and additional relevant concepts, thereby enabling the system to retrieve a broader and more relevant set of documents. However, PRF is not without its challenges. One significant drawback is the risk of query drift, where the expanded query includes terms that diverge from the user's original intent, potentially retrieving irrelevant results. Moreover, the process of analyzing top-ranked documents to identify candidate terms can be computationally expensive, especially in large-scale retrieval systems. This can also result in overly long or verbose queries, which may negatively impact system efficiency. Additionally, the presence of irrelevant or noisy terms in feedback documents may reduce the overall precision of search results, highlighting the need for more robust selection and filtering mechanisms in PRF-based approaches.

3.3 Knowledge Graph Integration

In this section, we examine methods that incorporate knowledge information into query reformulation and related tasks. Utilizing knowledge sources in query suggestion helps improve the interpretation of ambiguous queries

by providing additional semantic context. Moreover, the integration of such knowledge enables personalization, allowing systems to tailor suggestions based on individual user preferences and search histories. Broadly, knowledge sources offer two types of information: 1- the relationships between entities, typically represented in the form of knowledge graphs. 2- detailed descriptions of entities, as found in structured encyclopedic sources.

Li et al. [6] propose a knowledge-enhanced pipeline that addresses limitations of traditional approaches. To find the knowledge contained in the initial query, they built an entity linker that identifies the query entities present in a large-scale knowledge repository. This approach enhances the understanding of the query context by leveraging structured knowledge bases like knowledge graphs, which provide relationships between entities and help in identifying relevant expansion terms. In the proposed pipeline, query reformulation and neural retrieval modules are optimized alternately from the feedback of each other, creating a cooperative system that continuously refines query suggestions based on both user intent and contextual knowledge. Additionally, incorporating detailed descriptions from sources such as Wikipedia ensures that the expanded queries are semantically rich and contextually accurate. By integrating these knowledge-based techniques, the pipeline significantly enhances the overall retrieval performance, making the search process more intuitive and aligned with user needs. Another application of knowledge entities is personalization. A personal knowledge repository composed of entities derived from search queries and visited web pages is constructed by Baek et al. [7] for the query suggestion task. The captured knowledge is used as context for a large language model input. This repository leverages structured knowledge graphs and ontologies to provide rich, semantically relevant context for user queries. The proposed knowledge-augmented model is capable of lightweight personalization through retrieval from the knowledge store, without the need for explicit user profiling. This approach enhances query expansion by incorporating user-specific context and preferences directly into the LLM, making the suggestions more relevant and tailored.

In summary, knowledge-based query expansion enriches semantic context by linking terms through synonyms, hypernyms, and related concepts, thereby improving query understanding and relevance. These methods help prevent query drift by preserving original intent and can be adapted to domain-specific needs, often with greater efficiency than pseudo-relevance feedback since they rely on curated knowledge sources. However, static knowledge bases may miss new terminology or trends, leading to outdated expansions, and they typically lack personalization, producing results that may not align with individual user intent. To overcome this, newer approaches integrate user interaction data—such as search logs and click behavior—to generate more contextually relevant and personalized query reformulations.

4 Contextual Information Source

Contextual query refinement involves leveraging information about the user's current context, previous interactions, or ongoing search session to improve the relevance and specificity of retrieved results.

4.1 Session-Based Information

Users interact with search engines by submitting sequences of related queries, often grouped into sessions defined by consecutive queries and user actions (e.g., clicks). Within a session, users iteratively adjust their queries—adding, removing, or substituting terms—to refine their information needs. Interaction signals such as click-through rates and dwell time provide valuable clues about evolving intent. By analyzing these reformulation and interaction patterns, systems can disambiguate terms, capture user preferences, and generate more accurate and personalized query reformulations.

4.1.1 With Previous Queries

This subsection focuses on approaches that incorporate a user's initial query along with the sequence of previously issued queries within a search session to better capture the user's evolving intent. By leveraging the full context of the session, systems can generate more relevant reformulations that align closely with users' evolving goals.

One notable approach is proposed by Dehghani et al. [10], who introduced a query-aware attention mechanism based on a sequence-to-sequence (seq2seq) architecture for query suggestion. Their model incorporates three key components: 1- an attention mechanism that selectively focuses on relevant parts of the current and previous queries to infer intent, 2- a copying mechanism that enables the reuse of useful terms from earlier queries in the session, and 3- a generation mechanism that introduces new terms or phrases aligned with the user's evolving information need. By considering multiple queries within a session, the model dynamically adapts its suggestions to the user's current context while maintaining coherence with previous query formulations. In addition to general purpose search, session-based query reformulation has been explored in domain-specific search environments. For instance, Cao et al. [11] constructed a large-scale query reformulation corpus using search logs collected from Stack Overflow. Their process involved identifying original queries and their corresponding refinements from search sessions, which were then used to train a transformer based model. The trained model was capable of automatically generating reformulated queries tailored to the information needs of developers conducting programming related searches. In the context of conversational search, Yang et al. [14] proposed a zero-shot query reformulation framework using a transformer model that leverages conversational context. Unlike traditional approaches that require large-scale training data with explicit query–reformulation pairs, their method uses prior conversational history—including earlier user queries and system responses—to

generate reformulated queries that remain coherent with the ongoing dialogue. This approach demonstrates the potential of context-aware models to generalize across unseen tasks without the need for extensive supervision.

One key application is to encode a user's query history to enhance the representation of the current query, thereby making the user's information need more explicit and easier to satisfy. In this context, Zhou et al. [16] treat the current query as a direct expression of the user's intent and propose a method for encoding prior queries in the session to construct a context-aware representation of the current query. By incorporating historical search behavior, their approach enriches the semantic representation of the query, enabling more accurate interpretation and downstream retrieval performance. In the domain of conversational search, systems aim to further refine query understanding by posing clarifying questions that resolve ambiguity in user intent. Aliannejadi et al. [23] introduced a model for selecting clarifying questions that considers both the user's initial query and the sequence of prior question–answer interactions. Their approach leverages the entire dialogue history to identify information gaps or ambiguities and select follow-up questions that are most likely to elicit informative user responses. By dynamically adapting to conversational context, the system improves its ability to understand the user's true intent, especially in complex or under-specified queries.

Another line of research in this domain focuses on enhancing retrieval effectiveness in conversational search through query expansion, also known as query rewriting. To this end, Mo et al. [17] proposed a query selection framework that selectively expands the current query using only the most relevant historical queries, rather than incorporating all previous queries in the session. Their method employs a pseudo labeling strategy, in which historical queries are automatically labeled based on their relevance to the current search intent. The model leverages contextual signals from earlier conversational turns including the user's initial query and subsequent interactions to determine which past queries are most likely to contribute useful context. In the e-commerce domain, Hirsch et al. [18] conducted a large-scale analysis of search logs to investigate user query reformulation behavior within a session. Their study revealed that reformulated queries tend to be longer and are associated with higher click-through and purchase rates, particularly for the final query in a session—regardless of how many reformulations occurred. Based on these findings, the authors addressed a novel challenge in e-commerce search: predicting whether a user will reformulate their query before any results are shown. This insight enables the development of pre-retrieval strategies aimed at anticipating user behavior and improving both search relevance and commercial outcomes.

4.1.2 With Query and Click-through Data

This section examines approaches that combine previously issued queries with click-through data to better capture user intent. By analyzing both query sequences and user interactions, these methods enhance tasks such as query

suggestion, reformulation, and representation learning. Click-through data serves as implicit feedback on document relevance, and when paired with prior queries, it reveals behavioral patterns that refine query interpretation. Empirical studies show that incorporating session-level clicks improves precision and relevance by moving beyond surface-level term matching to infer deeper semantic intent—particularly useful when queries are ambiguous or underspecified.

Learning how to represent search session information plays a crucial role in capturing the semantic and syntactic relationships between user queries and interactions. In this context, Jiang et al. [19] proposed a query suggestion model that integrates query and click-through data from search logs into term embeddings using a heterogeneous network embedding framework. Their approach leverages a large-scale search log dataset (AOL) to identify meaningful patterns between user queries and the links they clicked. The core of their method is a bidirectional recurrent neural network (BiRNN), which encodes sequences of previous reformulations to infer the next likely reformulation in the embedding space. By embedding query terms and associated click data into a shared semantic space, the model effectively captures user intent and the semantic relationships among queries and clicked documents. Search activities within a session—such as query reformulations and result clicks—often exhibit strong interdependencies. These dependencies offer rich contextual signals that can be leveraged to enhance retrieval models. To exploit these interactions, Ahmad et al. [22] introduced a two-level hierarchical recurrent neural network that learns both the search session representation and the structure of query–click dependencies. Their model explicitly incorporates prior queries and associated clicks from the session to improve both document ranking and query suggestion performance. Another complementary direction is to treat clicked suggested queries as feedback for improving query suggestion systems. In this line of work, Li et al. [20] presented a feedback-aware query suggestion model, which uses user feedback—specifically, the act of clicking on suggested queries—to adaptively refine future suggestions. By interpreting user interactions as implicit feedback, the model can better align suggestions with the user's evolving intent. To more effectively capture users' underlying search intent, Li et al. [20] model the sequence of issued search queries alongside clicks on previously suggested queries, treating these interactions as implicit user feedback. Their proposed framework employs adversarial training techniques to improve the model's robustness and maintain performance even under challenging conditions, such as noisy inputs or ambiguous user behavior. This approach enables the system to dynamically refine its understanding of user intent, thereby improving the accuracy of subsequent query suggestions and delivering a more tailored and effective search experience. Building on this idea, Chen et al. [21] proposed a query suggestion model that incorporates both short-term context (i.e., queries and click behavior from the current session) and long-term context (i.e., interactions from previous sessions). Their model is built on an attention-based hierarchical neural network, which is designed to

capture user preferences toward specific query types. The attention mechanism allows the model to selectively focus on the most informative elements of both current and past sessions, resulting in more accurate and contextually aware suggestions. This hierarchical design ensures that the system takes into account not only the user's immediate behavior but also long-term patterns, thereby enhancing the overall effectiveness of query refinement. In a related direction, Wu et al. [25] introduced a feedback memory network that encodes a variety of user interactions including issued queries, clicked documents, and skipped items as positive and negative feedback. This feedback-aware representation is then integrated with a sequence-to-sequence model to drive a query suggestion system capable of adapting to nuanced user behavior. By explicitly modeling feedback signals, the system is better equipped to distinguish between relevant and irrelevant content, leading to more refined and personalized query suggestions.

An important application of integrating users' query and click data is personalized search. Natural language is inherently ambiguous—terms such as "Python" can refer to either a programming language or a species of snake, depending on the user's background and intent. As such, these terms should ideally have user-specific semantic representations. Motivated by this observation, Yao et al. [27] and Zhou et al. [26] proposed personalized search approaches that leverage users' queries and click behaviors within search sessions to model individual user interests. Yao et al. [27] addressed this by first training a global word2vec model on the full AOL query log to capture general semantic relationships between terms. While this unsupervised model learns word co-occurrences effectively, it does not account for user-specific preferences embedded in click behavior. To overcome this limitation, the authors developed a supervised personalization model that fine-tunes the word embeddings using users' click data. The result is a set of personalized word vectors that more accurately reflect each user's unique interests and search patterns. Similarly, Zhou et al. [26] introduced a self-supervised learning framework for personalized search that utilizes a contrastive sampling technique. Their approach extracts informative pairs from user behavior sequences recorded in query logs, allowing the model to refine the representations of queries, documents, and users. By learning from implicit feedback in user interactions, their method enhances the alignment between search results and individual user preferences, leading to more effective and personalized retrieval outcomes.

While many approaches rely on users' historical queries and click behavior to personalize and disambiguate queries, recent research has also explored interactive strategies, such as posing clarifying questions to users. For example, Erbacher et al. [35] introduced a simulated query clarification framework that enables multi-turn interactions between information retrieval (IR) systems and user agents. Their framework simulates user behavior based on click data, allowing the system to iteratively clarify ambiguous queries through sequential interactions, thereby improving retrieval effectiveness in real-time.

Beyond traditional search, search session information is also increasingly utilized in the search advertising domain, where ad relevance and placement are critical. Typically, ad selection models in this domain are trained using large-scale ad click data. To address challenges associated with rare or long-tail queries, Lee et al. [29] proposed a method that uses a sequence-to-sequence model to generate relevant keywords based on the user's input query. These generated keywords are then used to expand the query, thereby improving ad retrieval and selection. This approach ensures that relevant ads are surfaced even for underrepresented queries, enhancing both click-through rates and overall advertising performance.

4.2 Temporal data

This section examines temporal data - including query timestamps and event information - for query refinement. Understanding the time aspect can help improve the relevance of search results by considering seasonal trends, recent events, or time-sensitive information.

As we mentioned before, query auto-completion (QAC) methods suggest queries to search engine users as they begin typing. Most current QAC methods rank these suggestions based on their past popularity, measured by the number of times they have been previously submitted. However, query popularity changes over time and can differ significantly among users. To address this problem, Cai et al. [30] proposed a time-sensitive query auto-completion approach. This approach ranks query completions by predicting query popularity based on periodicity and recent trends in query frequency. They incorporate several key elements, including a time-sensitive popularity prediction mechanism that analyzes historical query logs to identify periodic patterns and recent trends, and a prefix-adaptive mechanism that adjusts suggestions based on different lengths and forms of user-typed prefixes. By capturing the dynamic nature of query popularity over time, the method ensures that auto-completion suggestions remain relevant and up-to-date, which is crucial for handling trending topics and seasonal searches.

Another form of temporal data used in information retrieval tasks is event information. Rosin et al. [34] proposed an event-aware query expansion approach that incorporates temporal context by identifying events related to the user's initial query. In this method, events are first detected, and terms that are semantically associated with both the query and the identified events are selected as expansion candidates. To facilitate this, both words and events are embedded into a shared vector space, allowing for the measurement of semantic similarity between query terms and event-related concepts. Unlike approaches that rely heavily on search log data, they utilize structured external knowledge sources—specifically Wikipedia and DBpedia—to extract relevant event information. By incorporating event-based expansions, the model enhances the contextual relevance of queries, particularly in cases where time-sensitive or event-driven topics are involved. This leads to more accurate and timely search results that reflect the user's current informational needs.

While search log–based methods offer valuable insights into user behavior and enable highly personalized refinements by capturing recurring patterns within sessions, they also raise privacy concerns since they rely on sensitive user histories. Moreover, their scope is limited to individual behavior on a single platform, which may not reflect broader interests or external factors. In contrast, social information–based approaches provide diversity by leveraging shared interests, collaborative behavior, and community-level signals from social networks. These extend beyond isolated sessions and support refinements informed by collective perspectives. Together, session-based methods emphasize precision and personalization, while social-based techniques offer breadth and diversity complementary strategies that enhance the adaptability and relevance of retrieval systems.

4.3 Social Information

Approaches leveraging users' search sessions on social platforms enhance search relevance by building personalized models that generate diverse and contextually appropriate query suggestions. These methods combine signals from user-generated content, social connections, and trending topics to align recommendations with both individual interests and broader social dynamics. By integrating personal and collective activity, they provide more personalized, trend-aware, and timely query refinements, improving overall search effectiveness.

One prominent example of a social network platform used for contextual search enhancement is LinkedIn. Zhou et al. [31] proposed a query suggestion model aimed at improving job search experiences on LinkedIn by leveraging a combination of user profiles, prior search history, and job descriptions. Their model employs semantic matching techniques to ensure that the suggested queries align closely with both the users' qualifications and the roles they are exploring. Further advancements in this domain were made by Zhong et al. [32], who introduced a query suggestion model that incorporates user-specific professional features to enhance personalization. Their approach extracts career-related attributes—such as job titles, skills, and industry experience—from LinkedIn member profiles and integrates these features into the suggestion model. This results in more personalized and contextually appropriate recommendations, ultimately improving the efficiency and relevance of job-related search on professional platforms.

User modeling can also be enhanced by incorporating information from users' social connections, particularly in scenarios where individual historical data is limited. Zhou et al. [33] proposed a personalized search model that integrates both individual and group profile information to improve the relevance of search results. Their approach combines a user's personal search history with data derived from their social network, forming user groups based on similar interests or behavioral patterns.

Incorporating social information into query refinement enhances personalization by aligning suggestions with users' interests, professional roles, and

social ties. It improves relevance when individual histories are sparse by leveraging group behavior and structured data such as resumes or job descriptions. However, this approach faces challenges: privacy and ethical concerns around handling personal data, reliance on platform-specific structures, and risks from outdated or noisy social signals. Large-scale graph processing is also computationally costly, and there is potential for reinforcing existing social biases, which may reduce diversity in query suggestions.

5 Comparative Analysis

To provide a structured comparison of the surveyed methods, we present comparative analyses across multiple dimensions based on our systematic review in Tabel 1, 2 and 3. Our comparative analysis evaluates methods across four key dimensions: *Effectiveness:* Based on reported performance improvements in original studies. *Scalability:* Assessed by computational requirements and real-time applicability. textitRobustness to Query Drift: Evaluated by semantic preservation mechanisms. *Personalization:* Measured by user-specific adaptation capabilities. Ratings (High/Medium/Low) were assigned based on quantitative results reported in the original papers and qualitative assessment of method characteristics.

Table 1 Comparative Analysis of Non-Contextual Methods

Method	Information Source	Effectiveness	Scalability	Robustness to Query Drift	Personalization	Year	Key Innovation
Lucchese et al. [15]	Thesaurus	Medium	High	Medium	Low	2018	Conjunctive Normal Form expansion
Azad et al. [4]	WordNet + Wikipedia	High	Medium	High	Low	2019	Two-level synset extraction
Zheng et al. [5]	Retrieved Docs (BERT)	High	Medium	Medium	Low	2021	Selective chunk-based expansion
Hashemi et al. [8]	Retrieved Docs (BART)	High	Medium	High	Low	2021	Multiple query representations
Li et al. [6]	Knowledge Graph	High	Low	High	Medium	2022	Entity linking with neural retrieval
Baek et al. [7]	Personal Knowledge	High	Medium	High	High	2024	Lightweight personalization
Li et al. [12] [7]	Retrieved Documents	Medium	Medium	Low	Low	2018	Neural PRF approach
Wang et al. [13] [7]	Retrieved Documents	High	Low	Medium	Low	2020	End-to-end neural framework

Beyond the tabulated summary, a closer look at key evaluation criteria highlights notable differences among the surveyed methods. In terms of robustness to query drift, neural architectures—particularly transformer-based and reinforcement learning–driven refiners—demonstrate stronger resilience, as they can model context across longer query sequences and correct deviations from the original intent. In contrast, purely lexical or heuristic approaches tend to be more susceptible to drift, especially in longer sessions. Regarding personalization capability, models that explicitly incorporate user profiles, session histories, or social information (e.g., friend networks, professional profiles) consistently outperform generic refinement methods, delivering more

Table 2 Comparative Analysis of Contextual Methods

Method	Information Source	Effectiveness	Scalability	Robustness to Query Drift	Personalization	Year	Key Innovation
Dehghani et al. [10]	Session Queries	High	Medium	High	High	2017	Copy-generate mechanism
Jiang et al. [19]	Query + Click Data	High	Medium	Medium	High	2018	Heterogeneous network embedding
Ahmad et al. [22]	Session + Clicks	High	Low	Medium	High	2019	Two-level hierarchical learning
Li et al. [20]	Feedback Clicks	High	Medium	High	High	2019	Adversarial training approach
Cai et al. [30]	Temporal Data	Medium	High	Medium	Medium	2016	Periodicity and trend prediction
Zhou et al. [31]	Social + Professional	High	Low	High	High	2022	Professional feature integration
Zhong et al. [32] [31]	Career Attributes	High	Low	High	High	2020	Skills and experience modeling
Zhou et al. [33] [31]	Social Networks	High	Low	Medium	High	2021	Individual + group profiling
Rosin et al. [34] [31]	Event Information	Medium	Medium	High	Low	2021	Event-query semantic embedding
Wu et al. [25] [31]	Multi-modal Feedback	High	Low	Medium	High	2018	Comprehensive feedback encoding
Chen et al. [21] [31]	Short/Long-term Context	High	Medium	High	High	2020	Multi-temporal context modeling

Table 3 Information Source Taxonomy

Category	Subcategory	Examples	Key Characteristics	Applications
Non-Contextual	External Knowledge	WordNet, Wikipedia, Thesauri	Static, General, Semantic	Query Expansion, Synonym Detection
	Retrieved Documents	PRF, Top-k Results	Dynamic, Query-specific	Contextual Expansion
	Knowledge Graphs	Entity Linking, Relations	Structured, Semantic	Entity-based Refinement
Contextual	Session Data	Query Sequences, User History	Personal, Sequential	Personalized Suggestion
	Click-through	User Interactions, Feedback	Behavioral, Implicit	Intent Modeling
	Temporal	Time, Events, Trends	Time-sensitive, Dynamic	Trending Topics, Seasonality
	Social	Network Data, Profiles	Collaborative, Social	Social Search, Professional

relevant and user-aligned suggestions. However, this advantage often comes at the cost of higher data requirements and potential privacy concerns, whereas non-personalized approaches remain easier to deploy in data-constrained or privacy-sensitive contexts.

6 Future Directions and Open Challenges

Query refinement research continues to evolve, yet several technical, methodological, and ethical challenges remain, alongside emerging opportunities that could shape the next generation of methods.

Technical Challenges: One persistent issue is query drift mitigation. Many expansion methods inadvertently introduce terms that diverge from the original user intent, reducing retrieval relevance. Future work should prioritize robust filtering strategies and semantic coherence measures to ensure that expanded queries remain aligned with user needs. Scalability is another critical hurdle—real-time query refinement in large-scale systems is particularly demanding for knowledge graph–based and complex contextual approaches. This calls for more efficient indexing schemes, approximation algorithms, and resource-aware processing pipelines. Furthermore, multi-modal integration remains underexplored: most current approaches rely on a single information source. Harnessing the complementary strengths of multiple sources (e.g., combining contextual signals with structured knowledge) requires sophisticated fusion and weighting mechanisms.
Evaluation and Benchmarking: The field lacks standardized evaluation benchmarks that enable consistent comparisons across different information source types. A unified, comprehensive evaluation framework—covering both quantitative and qualitative dimensions—would be instrumental in advancing the field. Moreover, current evaluations often emphasize immediate retrieval improvements, neglecting long-term effectiveness measures such as sustained user satisfaction, learning effects, and behavioral changes over time. Addressing these gaps would yield more reliable assessments of method impact in real-world use. **Privacy and Ethical Considerations:** Contextual methods frequently rely on sensitive user data, raising important privacy concerns. Developing privacy-preserving refinement techniques that maintain retrieval effectiveness without compromising user confidentiality is a pressing need. Additionally, both social and contextual methods can inadvertently reinforce existing biases present in training data or interaction logs. Research on bias detection, mitigation, and fairness-aware refinement strategies is essential to ensure equitable and trustworthy systems. **Emerging Opportunities:** Large language models (LLMs) offer new possibilities for query understanding, reformulation, and expansion, but their integration must be tempered by careful evaluation of hallucination tendencies and factual accuracy. Real-time adaptation—dynamically adjusting refinement strategies to evolving user contexts and information landscapes—remains a promising yet unsolved challenge. Finally, cross-domain transfer represents an important research frontier: developing models and techniques that generalize effectively across diverse domains, languages, and application scenarios would substantially increase the applicability and robustness of query refinement systems.

7 Conclusion

This survey provided a comprehensive analysis of query refinement methods through the lens of information sources, systematically reviewing 67 papers published between 2016-2024. We established a clear taxonomy distinguishing between non-contextual sources (external knowledge, retrieved documents,

knowledge graphs) and contextual sources (session data, click-through behavior, temporal information, social signals). Our comparative analysis reveals that contextual methods generally provide superior personalization capabilities but at the cost of increased complexity and privacy concerns. Non-contextual methods offer better generalization and scalability but may lack user-specific adaptation. The most promising approaches combine multiple information sources, suggesting that hybrid models integrating structured knowledge with behavioral context hold significant potential. **Key Contributions:** a) Systematic methodology for surveying query refinement literature; b) Novel taxonomy based on information source types; c) Comprehensive comparative analysis across effectiveness, scalability, and personalization dimensions; d) Identification of critical challenges and future research directions.
Future Work: We recommend focusing on: (1) developing privacy-preserving contextual methods, (2) creating standardized evaluation benchmarks, (3) investigating effective multi-source integration techniques. The development of explainable refinement systems that provide transparency in the refinement process remains an important open challenge. This structured understanding of information sources in query refinement facilitates more informed decisions when selecting or designing techniques and provides a foundation for advancing the field toward more adaptive, accurate, and user-centered retrieval systems.

References

[1] H.K. Azad and A. Deepak, Query expansion techniques for information retrieval: a survey, Information Processing & Management, Elsevier, 2019.

[2] B. Vélez, R. Weiss, M.A. Sheldon, and D.K. Gifford, Fast and effective query refinement, SIGIR, 1997.

[3] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, Context-Aware Query Suggestion by Mining Click-Through and Session Data, Association for Computing Machinery, 2008.

[4] H.K. Azad and A. Deepak, A new approach for query expansion using Wikipedia and WordNet, Information Sciences, Elsevier, 2019.

[5] Z. Zheng, K. Hui, B. He, X. Han, L. Sun, and A. Yates, Contextualized query expansion via unsupervised chunk selection for text retrieval, CIKM, 2021.

[6] X. Li, J. Mao, W. Ma, Z. Wu, Y. Liu, M. Zhang, S. Ma, Z. Wang, and X. He, A cooperative neural information retrieval pipeline with knowledge enhanced automatic query reformulation, WSDM, 2022.

[7] J. Baek, N. Chandrasekaran, S. Cucerzan, A. Herring, and S. K. Jauhar, Knowledge-augmented large language models for personalized contextual query suggestion, Proceedings of the ACM Web Conference, 2024.

[8] H. Hashemi, H. Zamani, and W. B. Croft, Learning multiple intent representations for search queries, CIKM, 2021.

[9] A. Medlar, J. Li, and D. Głowacka, Query suggestions as summarization in exploratory search, CHIIR, 2021.

[10] M. Dehghani, S. Rothe, E. Alfonseca, and P. Fleury, Learning to attend, copy, and generate for session-based query suggestion, CIKM, 2017.

[11] K. Cao, C. Chen, S. Baltes, C. Treude, and X. Chen, Automated query reformulation for efficient search based on query logs from Stack Overflow, ICSE,2021.

[12] C. Li, Y. Sun, B. He, L. Wang, K. Hui, A. Yates, L. Sun, and J. Xu, NPRF: A Neural Pseudo Relevance Feedback Framework for Ad-hoc Information Retrieval, EMNLP, 2018.

[13] L. Wang, Z. Luo, C. Li, B. He, L. Sun, H. Yu, and Y. Sun, An end-to-end pseudo relevance feedback framework for neural document retrieval, CIKM, 2020.

[14] D. Yang, Y. Zhang, and H. Fang, Zero-shot query reformulation for conversational search, SIGIR, 2023.

[15] C. Lucchese, F.M. Nardini, R. Perego, R. Trani, and R. Venturini, Efficient and effective query expansion for web search, cikm, 2018.

[16] Y. Zhou, Z. Dou, and J.-R. Wen, Encoding history with context-aware representation learning for personalized search, SIGIR , 2020.

[17] F. Mo, J.-Y. Nie, K. Huang, K. Mao, Y. Zhu, P. Li, and Y. Liu, Learning to relate to previous turns in conversational search, SIGKDD, 2023.

[18] S. Hirsch, I. Guy, A. Nus, A. Dagan, and O. Kurland, Query reformulation in E-commerce search, SIGIR, 2020.

[19] J.-Y. Jiang and W. Wang, RIN: Reformulation inference network for context-aware query suggestion, CIKM, 2018.

[20] R. Li, L. Li, X. Wu, Y. Zhou, and W. Wang, Click feedback-aware query recommendation using adversarial examples, WWW, 2019.

[21] W. Chen, F. Cai, H. Chen, and M. de Rijke, Hierarchical neural query suggestion with an attention mechanism, CIKM, 2020.

[22] W. U. Ahmad, K.-W. Chang, and H. Wang, Context attentive document ranking and query suggestion, SIGIR, 2019.

[23] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, SIGIR, 2019.

[24] Statista Research Department. Number of search terms used in internet research in the United States as of January 2022. *Statista*. https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/

[25] B. Wu, C. Xiong, M. Sun, and Z. Liu, Query suggestion with feedback memory network, WWW, 2018.

[26] Y. Zhou, Z. Dou, Y. Zhu, and J.-R. Wen, PSSL: self-supervised learning for personalized search with contrastive sampling, CIKM, 2021.

[27] J. Yao, Z. Dou, and J.-R. Wen, Employing personal word embeddings for personalized search,SIGIR, 2020.

[28] L. Wang, N. Yang, and F. Wei, Query2doc: Query expansion with large language models, CoRR, 2023.

[29] M.-C. Lee, B. Gao, and R. Zhang, Rare query expansion through generative adversarial networks in search advertising, SIGKDD, 2018.

[30] F. Cai, S. Liang, and M. de Rijke, Prefix-adaptive and time-sensitive personalized query auto completion, IEEE Transactions on Knowledge and Data Engineering, 2016.

[31] Z. Zhou, X. Zhou, M. Li, Y. Song, T. Zhang, and R. Yan, Personalized query suggestion with searching dynamic flow for online recruitment, CIKM, 2022.

[32] J. Zhong, W. Guo, H. Gao, and B. Long, Personalized query suggestions, SIGIR, 2020.

[33] Y. Zhou, Z. Dou, B. Wei, R. Xie, and J.-R. Wen, Group based personalized search by integrating search behaviour and friend network, SIGIR, 2021.

[34] G. D. Rosin, I. Guy, and K. Radinsky, Event-driven query expansion, WSDM, 2021.

[35] P. Erbacher, L. Denoyer, and L. Soulier, Interactive query clarification and refinement via user simulation, SIGIR, 2022.

[36] E. Rencis, Application of a configurable keywords-based query language to the healthcare domain, Journal of Advances in Information Technology, 2021.