The Impact of Query Specificity on Retrieval Effectiveness: Postprint

Authors: Ren Ke, Lu Wei, Ding Heng

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] To conduct a comprehensive analysis of retrieval effectiveness for queries with different specificity levels, providing insights for improving search engine performance and enhancing user retrieval experience. [Method] Based on TREC Web Track queries, we manually constructed a query specificity annotation set, selected three models including language model with Dirichlet smoothing, language model with linear interpolation smoothing, and BM25, and used commonly employed information retrieval evaluation metrics as benchmarks to investigate the impact of query specificity strength on retrieval effectiveness at different levels. [Results] The difference in retrieval effectiveness between strong and weak specificity queries is most pronounced in the top-ranked results, with strong specificity queries achieving significantly better performance than weak specificity queries. [Limitations] Experiments were conducted only on the TREC dataset, and further validation on other datasets is needed. [Conclusion] Search engines should prioritize the accuracy of top-ranked results under the dimension of specificity as a starting point for improving retrieval models.

Full Text

The Impact of Query Specificity on Retrieval Effectiveness

Ren Ke[1], Lu Wei[1,2], Ding Heng[1]

[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Study of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract

[Objective] This study provides a comprehensive analysis of retrieval effectiveness across queries with varying specificity levels, offering insights for improving

search engine performance and user experience. [**Methods**] We manually constructed a query specificity annotation set based on TREC Web Track queries and employed three retrieval models—language model with Dirichlet smoothing, language model with linear interpolation smoothing, and BM25. Using standard information retrieval evaluation metrics, we examined how query specificity affects retrieval effectiveness at different levels. [**Results**] The most significant differences in retrieval effectiveness between strong and weak specificity queries appear among the top-ranked results, with strong specificity queries demonstrating notably better performance. [**Limitations**] Our experiments were conducted solely on the TREC dataset and require further validation on other datasets. [**Conclusions**] From the perspective of query specificity, search engines should prioritize improving the accuracy of top-ranked results as a key strategy for performance enhancement.

Keywords: Query intention, Query specificity, Retrieval effectiveness
Classification: G353.1

Introduction

The Internet provides users with rich and diverse information resources. While this abundance offers convenience, it also leads to information overload [1], increasing the difficulty of information seeking. Search engines such as Google and Baidu serve as interfaces between users and web resources, returning result lists based on user queries. However, queries that are either too broad or too narrow may produce biased results, forcing users to manually filter results or conduct secondary searches. This wastes time and effort and may result in information loss. Therefore, developing different retrieval strategies for queries with varying specificity represents a promising direction for improving retrieval models.

Query specificity constitutes a semantic feature of search queries [2] and influences information retrieval effectiveness [3]. It reflects the breadth of concepts expressed in a query and captures users' requirements for detail and certainty in retrieved information [4], thereby indicating user intent to some extent. Enhanced research on query specificity can help search engines better understand users' latent needs and provide more relevant result lists.

Current research on "how query specificity strength affects retrieval effectiveness" has focused primarily on the entire result list. However, due to individual differences in user search and click behaviors, analyzing the complete list fails to capture diverse user needs adequately. Based on the definition of query specificity, this paper systematically categorizes classification criteria and manually classifies all queries from the TREC Web Track 2009–2012. Using standard information retrieval evaluation metrics, we comprehensively analyze the impact of query specificity on retrieval effectiveness to identify improvement opportunities along this dimension.

Query intention represents an intermediate form between user queries and actual information needs, expressing users' search purposes [5]. Deep analysis of

query intention facilitates constructing users' information need space and clarifying their search goals, enabling the provision of more direct, relevant, and comprehensive information. In 2002, Broder [6] classified queries into three categories from a user purpose perspective: navigational, informational, and transactional. Subsequent scholars refined this framework [7-8], but the basic structure remains. However, this classification system is too simplistic to handle complex information needs. Consequently, González-Caro et al. [9] proposed ten dimensions—including genre, topic, specificity, task, objective, and scope—to reveal users' complex information needs and rich information spaces. This paper focuses on the specificity dimension for query intention analysis.

Query specificity represents an important aspect of search queries, but its semantic nature makes it difficult to measure. Current research on query specificity falls into three main areas:

1. **Classification of query specificity**: This area focuses on the specificity dimension itself, exploring its 内涵 and characteristics. Approaches include classifying queries by specificity strength (e.g., Hafernik [10] used nine semantic attributes to categorize queries as strong or weak specificity, finding that query length and part-of-speech could improve identification accuracy; Tang et al. [4] analyzed query features and used machine learning to automatically identify specificity strength) and classifying by representation (e.g., specificity represented by number of retrieved documents, morphological variations, or domain-specific terms, then examining correlations among these representations [11]).

2. **Relationship between query specificity and other attributes**: This area positions specificity within a network of retrieval attributes to enhance consistency and completeness. Phan et al. [12] found that query specificity correlates with query length, with shorter queries having lower specificity and three words typically serving as the boundary between strong and weak specificity. Kim [3] examined the relationship between query specificity and document relevance, returning to the essence of retrieval.

3. **Impact of query specificity on retrieval effectiveness**: Research aims to improve retrieval effectiveness. Mu et al. [13] used query specificity and length as two approaches for query expansion, studying how increasing or decreasing specificity and length affects overall retrieval effectiveness in health information retrieval. Heine [14] investigated how database informativeness, query length, and average query specificity affect retrieval effectiveness in MEDLINE, finding that specificity had limited impact in this context.

These studies exhibit several limitations: (1) Research on "query specificity' s impact on retrieval effectiveness" concentrates on medical information retrieval without analyzing open-web search scenarios; (2) Studies only examine the relationship between query specificity and overall retrieval effectiveness without

analyzing results across different metrics and levels.

Addressing these gaps, this paper refines the classification system for query specificity features, constructs an annotation set for query specificity strength on the TREC Web Track dataset, and employs BM25 and language models with multiple evaluation metrics to comprehensively analyze retrieval effectiveness. We also examine how different models perform across specificity levels.

Query Specificity Classification

3.1 Classification of Query Specificity Strength

Different specificity strengths may produce varying retrieval effectiveness across contexts. Our objective is to classify queries in the dataset and compare retrieval effectiveness across different specificity levels.

Currently, no clear classification standards exist for query specificity strength. Queries are typically categorized into two classes (strong/narrow vs. weak/broad) or three classes (specific, medium, broad). Given that English queries are typically short and concise, this paper adopts an information need perspective [15-16] and uses a two-class system:

1. **Strong specificity queries**: Users express clear information needs with minimal or no ambiguity, clearly defining their purpose and search scope, often involving domain-specific knowledge. Examples include queries seeking exact answers, comparisons between topics, or specific dates, such as "who invented music" or "mothers day songs."
2. **Weak specificity queries**: Users express ambiguous information needs with high ambiguity, or their search purpose and scope are broad, belonging to general domains that cannot be precisely located. These queries often require secondary searches or manual filtering, such as "cell phones,""korean language," or "dieting."

3.2 Feature Analysis of Query Specificity Strength

Query features consist of basic features (length, term count) and content features (meaning). As a semantic feature, query specificity focuses on content features. Specificity typically examines what constraints users employ to clarify their needs, such as quantity, name, time, or location constraints. Building on Hafernik [10] and Tang et al. [4], we selected nine attribute features to characterize specificity strength, as shown in Table 1 .

Table 1 Query Attribute Features with Examples

Query Attribute Feature	Example
Query includes URL, website name, or IP	yahoo

Query Attribute Feature	Example
Query contains specific location names and other terms	map of the united states
Query compares different things or different aspects of the same thing	butter and margarine
Query is a question containing an exact answer	who invented music
Query contains information needs regarding directions, advice, or guidance	how to build a fence
Query contains object names and other terms	Obama family tree

Since user query intention is crucial for understanding specificity, selected attributes should indicate both user needs and strong specificity characteristics based on prior research. For instance, a query containing a URL suggests the user seeks a specific website (strong specificity), while a query seeking guidance indicates the user wants specific steps to achieve a goal. Based on these attributes, we classify queries as strong specificity if they contain one or more such features, and weak specificity if they contain none.

Research Design

4.1 Research Subjects

Our dataset comprises queries from TREC Web Track 2009–2012, with 50 queries published annually (200 total). These web-focused queries represent open-web search scenarios commonly encountered by the general public, maximizing practical relevance. We indexed the collection using Indri 5.7, applied the standard Indri stopword list, and performed stemming with Krovetz. Two graduate students from Wuhan University' s School of Information Management annotated queries for specificity strength using the attributes in Table 1. To assess inter-annotator reliability, we calculated the Kappa statistic [17], obtaining a value of 0.91 (>0.8), indicating excellent agreement and validating our annotations.

For large-scale corpora, listing all relevant documents per query is impractical. Therefore, we employed the pooling method [18], combining the top K documents from multiple retrieval systems into a subset for relevance judgment. We used TREC evaluation benchmarks to assess the top 1000 documents for each query. Among 200 queries, two lacked relevance judgments in the Track' s provided qrels, leaving 198 queries for analysis.

4.2 Research Methods

Language models [19] and the Okapi BM25 model [20] are the most widely used information retrieval models. We employed both to reduce potential bias from single-model evaluation. To avoid zero-probability problems, we applied two smoothing methods in language models: Dirichlet smoothing and Jelinek-Mercer

linear interpolation smoothing. We used trec_{eval} to compute evaluation metrics and assess result differences.

Traditional recall and precision measurements are set-based and poorly suited for ranked retrieval results from search engines. Therefore, we employed standard ranked retrieval metrics: Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), R-Precision (R-Prec), Binary Preference-based measures (Bpref), Reciprocal Rank (Recip_{Rank}), and Precision at K (P@5, P@10, P@20).

1. **MAP**: A single-value metric reflecting system performance across all relevant documents.
2. **NDCG**: Measures ranking quality across all retrieved documents.
3. **R-Prec**: Precision among the top R results, where R is the total number of relevant documents for the query.
4. **Bpref**: Focuses on how often non-relevant documents appear before relevant ones.
5. **Recip_{Rank}**: Indicates the system's ability to return the first relevant document.
6. **P@K**: Reflects accuracy of the top K retrieved results.

These metrics serve different purposes: MAP and NDCG evaluate overall effectiveness, while Recip_{Rank} and P@K assess top-ranked results, enabling analysis of specificity' s impact at different retrieval levels.

4.3 Research Framework

We manually classified 198 queries by specificity strength and evaluated the top 1000 retrieved documents across three models (Dirichlet language model, linear interpolation language model, and BM25) using TREC relevance judgments. This yielded relevant document counts and metric values (MAP, NDCG, R-Prec, Bpref, Recip_{Rank}, P@5, P@10, P@20). We then conducted comprehensive analysis from two perspectives: (1) "Impact of specificity strength within the same model" to examine differences between specificity levels, and (2) "Comparison of different models under the same specificity level" to examine model performance differences. This dual approach provides a 立体 and comprehensive description of how query specificity affects retrieval effectiveness.

Impact of Specificity Strength on Retrieval Effectiveness

We used descriptive statistics (mean, standard deviation, median, min, max) and boxplots to visualize the impact of specificity strength, applying the Mann-Whitney U Test to assess significant differences ($p < 0.05$).

5.1 Within-Model Comparison of Strong vs. Weak Specificity

We compared retrieval effectiveness across specificity levels using multiple metrics within each model. Table 2 presents Mann-Whitney U test results.

Table 2 Comparison of Evaluation Metrics Across Retrieval Models

Metric	Dirichlet LM	Linear Interpolation LM	BM25
MAP	-	0.040	-
Recip_{Rank}	-	0.029	0.012
P@5	0.036	0.047	-

Table 2 shows that under Dirichlet smoothing, P@5 = 0.036 (<0.05), indicating significant differences. Under linear interpolation smoothing, MAP (0.040), Recip_{Rank} (0.029), and P@5 (0.047) all show significant differences. Under BM25, Recip_{Rank} and P@5 also show significant differences. These six significant metric pairs (25% of 24 total comparisons) demonstrate that specificity strength affects retrieval results in each model. Boxplots for these significant differences appear in Figure 1 [Figure 1: see original paper].

Specifically, P@5 shows significant differences across all three models, indicating that accuracy of the top 5 results varies markedly between specificity levels. As shown in Figure 1(a), (d), (f), strong specificity queries achieve higher maxima, means, and fewer outliers than weak specificity queries, demonstrating superior performance. Weak specificity queries have broader scopes without constraints, causing top results to cover diverse aspects (e.g., "apple" returning both Apple Inc. and the fruit), leading to intent mismatches. Strong specificity queries, with clearer constraints and less ambiguity, produce better top-5 results.

Recip_{Rank} also shows significant differences under linear interpolation smoothing and BM25 (p = 0.029 and 0.012). As a metric for first relevant document retrieval, Recip_{Rank} indicates the relevance strength of the first result. Figure 1(c), (e) show strong specificity queries outperform weak ones in maxima and means, as strong specificity queries typically have clear answers or focused topics, making the first relevant result appear earlier. Compared to P@5, Recip_{Rank} shows greater instability between specificity levels.

Finally, MAP shows marginal significance under linear interpolation smoothing (p = 0.040). Since MAP averages precision across all relevant documents, differences between specificity levels diminish. Figure 1(b) shows minimal differences except in outlier patterns.

Overall, strong specificity queries outperform weak ones when fewer results are considered, with differences diminishing as more results are retrieved. Even in non-significant metrics, strong specificity averages slightly outperform weak specificity. From a model perspective, Dirichlet smoothing shows better adaptability and sensitivity to both specificity levels, with minimal differences beyond P@5, while the other two models perform poorly on weak specificity queries.

5.2 Within-Specificity Comparison Across Models

We compared statistical measures across models for each specificity level using pairwise comparisons. Tables 3 through 5 present Mann-Whitney U test results.

Table 3 Dirichlet LM vs. Linear Interpolation LM

Metric	Strong Specificity	Weak Specificity
MAP	<0.05	<0.05
NDCG	<0.05	<0.05
R-Prec	<0.05	<0.05
P@20	<0.05	<0.05

Table 4 Dirichlet LM vs. BM25

Metric	Strong Specificity	Weak Specificity
All metrics	>0.05	>0.05

Table 5 Linear Interpolation LM vs. BM25

Metric	Strong Specificity	Weak Specificity
MAP	<0.05	<0.05
NDCG	<0.05	<0.05
R-Prec	<0.05	<0.05
Recip_{Rank}	<0.05	<0.05
P@10	<0.05	<0.05
P@20	<0.05	<0.05

Table 3 shows significant differences for MAP, NDCG, R-Prec, and P@20 under both specificity levels when comparing Dirichlet and linear interpolation smoothing. However, Table 4 shows no significant differences between Dirichlet smoothing and BM25 across any metric, indicating similar performance. Table 5 reveals numerous significant differences between linear interpolation smoothing and BM25 (MAP, NDCG, R-Prec, Recip_{Rank}, P@10, P@20) for both specificity levels. Notably, when strong specificity shows significant model differences, weak specificity typically shows similar patterns. Figure 2 [Figure 2: see original paper] visualizes these significant comparisons.

As Figure 2 and Table 4 demonstrate, Dirichlet smoothing and BM25 produce similar maxima, minima, and means across metrics, with all Mann-Whitney U test values >0.05, indicating equivalent performance for queries of the same specificity level. Both models outperform linear interpolation smoothing in

quartile and median values. Therefore, ranking model performance for same-specificity queries: Dirichlet smoothing = BM25 > Linear interpolation smoothing. Search engines should employ Dirichlet smoothing or BM25 for better alignment with user needs when considering specificity.

Discussion and Conclusion

Our findings reveal that due to differences in query clarity, strong and weak specificity queries show the largest performance gaps in top-ranked results, with strong specificity queries achieving significantly better effectiveness. As more documents are retrieved, these differences diminish, and performance converges. Since users typically examine only the top 10–20 results, search engines should prioritize improving top-result accuracy along the specificity dimension. Additionally, Dirichlet smoothing and BM25 outperform linear interpolation smoothing for queries of the same specificity level. Search engines could leverage machine learning to automatically identify specificity strength and employ Dirichlet smoothing or BM25 to return personalized results matching users' information need constraints, thereby improving performance and user experience.

We also observed that: (1) Bpref is the only metric showing no significant differences in either within-model specificity comparisons or across-model comparisons, likely because TREC queries have relatively few relevant documents (max: 167, mean: 37), limiting Bpref' s discriminative power; (2) P@5 shows significant differences between specificity levels within models but not across models within the same specificity level, indicating all models retrieve better results for strong specificity queries but struggle with weak specificity queries.

This study has limitations, as experiments were conducted only on the TREC dataset and require validation on other datasets.

References

[1] Wang Na, Chen Huimin. Investigation on the Harm and Cause of Information Overload in Ubiquitous Network [J]. Information Studies: Theory & Application, 2014, 37(11): 20-25.

[2] Jones K S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval [J]. Journal of Documentation, 1972, 28(1): 11-21.

[3] Kim G. Relationship Between Index Term Specificity and Relevance Judgment [J]. Information Processing & Management, 2006, 42(5): 1218-1229.

[4] Tang Xiangbin, Lu Wei, Zhang Xiaojuan, et al. Feature Analysis and Automatic Identification of Query Specificity [J]. New Technology of Library and Information Service, 2015(2): 15-23.

[5] Song Wei. Research on Topic Based Query Intent Identification [D]. Harbin: Harbin Institute of Technology, 2013.

[6] Broder A. A Taxonomy of Web Search [J]. ACM SIGIR Forum, 2002, 36(2): 3-10.

[7] Rose D E, Levinson D. Understanding User Goals in Web Search [C]. In: Proceedings of the 13th International Conference on World Wide Web. New York, NY, USA: ACM, 2004: 13-19.

[8] Baeza-Yates R, Calderón-Benavides L, González-Caro C. The Intention Behind Web Queries [C]. In: Proceedings of the 13th International Conference on String Processing and Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2006: 98-109.

[9] González-Caro C, Calderón-Benavides L, Baeza-Yates R, et al. Web Queries: The Tip of the Iceberg of the User' s Intent [C]. In: Proceedings of the 4th ACM WSDM Conference, Hong Kong, China. 2011.

[10] Hafernik C T. The Relationship Between Query Length, Parts of Speech Usage and Web Search Query Specificity [D]. The Pennsylvania State University, 2013.

[11] Tamine L, Chouquet C, Palmer T. Analysis of Biomedical and Health Queries: Lessons Learned from TREC and CLEF Evaluation Benchmarks [J]. Journal of the Association for Information Science and Technology, 2015, 66(12): 2645-2663.

[12] Phan N, Bailey P, Wilkinson R. Understanding the Relationship of Information Need Specificity to Search Query Length [C]. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 07). New York: ACM, 2007: 395-402.

[13] Mu X, Lu K. Improving UMLS Metathesaurus Query Expansion Based on the Query Specificity and Length [C]. In: Proceedings of the ACM SIGKDD Workshop on Health Informatics. 2012.

[14] Heine M H. An Investigation of the Relative Influences of Database Informativeness, Query Size and Query Term Specificity on the Effectiveness of Medline Searching [J]. Journal of Information Science, 1995, 21(3): 173-185.

[15] Ingwersen P, Jarvelin K. The Turn: Integration of Information Seeking and Retrieval in Context [M]. Springer, 2005.

[16] Ramírez G, de Vries A P. Relevant Contextual Features in XML Retrieval [C]. In: Proceedings of the 1st International Conference on Information Interaction in Context. New York: ACM, 2006: 56-65.

[17] Carletta J. Assessing Agreement on Classification Tasks: The Kappa Statistic [J]. Computational Linguistics, 1996, 22(2): 249-254.

[18] Siegel S, Castellan N J. Non-parametric Statistics for the Behavioral Sciences [J]. American Catholic Sociological Review, 1957, 18(2). DOI: 10.2307/3708383.

[19] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval [C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998: 275-281.

[20] Robertson S E, Jones K S. Relevance Weighting of Search Terms [J]. Journal of the American Society for Information Science, 1976, 27(3): 129-146.

Author Contributions

Ren Ke: Literature review, data processing and analysis, manuscript drafting, revision, and final version.
Lu Wei: Research design, manuscript revision and final version.
Ding Heng: Data analysis, manuscript revision.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn:
[1] Ren Ke, Lu Wei, Ding Heng. eval-adhoc-dir.xlsx. Retrieval results under Dirichlet language model across different metrics.
[2] Ren Ke, Lu Wei, Ding Heng. eval-adhoc-jm.xlsx. Retrieval results under linear interpolation language model across different metrics.
[3] Ren Ke, Lu Wei, Ding Heng. eval-adhoc-bm25.xlsx. Retrieval results under BM25 model across different metrics.
[4] Ren Ke, Lu Wei, Ding Heng. 数据标注结果集.xlsx. Data annotation results.

Received: 2016-07-18
Revised: 2016-09-01

Note: Figure translations are in progress. See original paper for figures.