iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

A COMPARATIVE STUDY OF INFORMATION RETRIEVAL MODELS: BM25 VERSUS HYBRID RETRIEVAL ON THE CRANFIELD COLLECTION

Aashish Dhakal,
Presidential Graduate School Kathmandu, Nepal
aashish.dhakal@presidential.edu.np

Nabin Chaulagain,
Presidential Graduate School Kathmandu, Nepal

Shashank Shree Neupane,
Presidential Graduate School, Nepal
shashank.neupane@presidential.edu.np

ABSTRACT
Information retrieval systems face the fundamental challenge of balancing retrieval effectiveness with computational efficiency. Traditional lexical models like BM25 provide fast retrieval but may miss semantic relationships, while modern hybrid approaches combining lexical and semantic methods promise improved effectiveness at the cost of computational complexity. This study compares the performance of BM25 and hybrid retrieval models on the Cranfield collection, evaluating both retrieval effectiveness and computational efficiency. We implemented two retrieval systems: (1) BM25 with pseudo-relevance feedback and tuned parameters ($k_1 = 1.6$, $b = 0.4$), and (2) a hybrid model combining BM25, dense retrieval using SPECTER embeddings, Reciprocal Rank Fusion, and cross-encoder reranking. Both systems were evaluated on 225 Cranfield queries using standard IR metrics: Precision@10, Recall@20, Mean Average Precision (MAP), and NDCG@20. The hybrid model demonstrated superior retrieval effectiveness with 41% improvement in Precision@10 (1.07% vs 0.76%), 67% improvement in Recall@20 (1.54% vs 0.92%), and consistent gains across all metrics. However, BM25 showed dramatically superior efficiency with 800x faster query processing (15.3ms vs 12.4s average latency). While hybrid retrieval models achieve better effectiveness, the computational cost may limit their applicability to real-time scenarios. The Cranfield collection's inherent difficulty (1960s terminology, sparse relevance judgments) constrains absolute performance regardless of model sophistication.

Keywords:
Information Retrieval, BM25, Hybrid Retrieval, Dense Retrieval, Cross-Encoder, Cranfield Collection.

1. INTRODUCTION

Information retrieval (IR) systems serve as the backbone of modern search applications, from web search engines to enterprise document management systems. The field has evolved from simple term-frequency models to sophisticated neural approaches that capture semantic relationships between queries and documents [7]. However, this evolution has introduced complex trade-offs between retrieval effectiveness and computational efficiency that practitioners must carefully consider.

The classical BM25 model, based on probabilistic ranking principles, remains widely deployed due to its computational efficiency and interpretability [11]. Recent advances in neural information retrieval have introduced dense retrieval methods using pre-trained language models, which encode text into dense vector representations to capture semantic similarity beyond lexical matching [9]. Hybrid approaches that combine lexical and semantic retrieval have shown promise in bridging this gap [8].

The Cranfield collection, despite its age, remains a valuable benchmark for IR research due to its challenging characteristics: domain-specific terminology, short documents, and sparse relevance judgments [14]. These characteristics make it an ideal testbed for evaluating the effectiveness of modern retrieval approaches on specialized collections.

International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

This study aims to provide a comprehensive comparison of BM25 and hybrid retrieval models, examining both effectiveness and efficiency dimensions. Our contributions include: (1) implementation of optimized BM25 and hybrid retrieval systems, (2) comprehensive evaluation on the Cranfield collection, (3) analysis of the effectiveness-efficiency trade-off, and (4) identification of dataset-specific limitations on model performance.

2. RELATED WORK

2.1 Classical Information Retrieval

The BM25 ranking function, introduced by Robertson et al. [10], represents one of the most successful applications of probabilistic retrieval theory. Built on the Binary Independence Model, BM25 incorporates term frequency, inverse document frequency, and document length normalization to estimate document relevance probability. Numerous studies have demonstrated BM25's robustness across diverse collections and query types [3].

Parameter optimization for BM25 has been extensively studied, with k_1 typically ranging from 1.2 to 2.0 and b from 0.3 to 0.8 depending on collection characteristics [13]. Query expansion techniques, particularly pseudo-relevance feedback, have shown consistent improvements in BM25 effectiveness [12].

2.2 Neural Information Retrieval

The emergence of pre-trained language models has revolutionized information retrieval. Dense retrieval methods encode queries and documents into high-dimensional vectors using models like BERT [4] and Sentence-BERT [9]. The SPECTER model, specifically designed for scientific document embeddings, uses citation graphs as additional training signal [1].

Cross-encoders, which process query-document pairs jointly through transformer architectures, have demonstrated superior relevance estimation compared to bi-encoder approaches [6]. However, their computational requirements limit practical deployment to reranking scenarios.

2.3 Hybrid Retrieval Systems

Modern retrieval systems increasingly adopt hybrid approaches that combine multiple retrieval signals. Reciprocal Rank Fusion (RRF) has emerged as an effective method for combining rankings from different retrieval systems [2]. The technique's simplicity and effectiveness have made it a standard approach in multi-stage retrieval pipelines.

3. METHODOLOGY

3.1 Dataset

The Cranfield collection consists of 1,400 aeronautics research abstracts from the 1960s, accompanied by 225 information needs and expert relevance judgments on a 4-point scale (1=not relevant, 2=somewhat relevant, 3=relevant, 4=highly relevant). Documents with relevance scores ≥ 2 are considered relevant for evaluation purposes.

The collection presents several challenges: vocabulary mismatch between 1960s and modern terminology, document brevity (50-150 words per abstract), sparse relevance data (average 2-5 relevant documents per query), and highly technical domain-specific language.

3.2 Text Preprocessing

Both systems employ minimal preprocessing to preserve domain-specific terminology:

1. **Normalization**: Lowercase conversion
2. **Tokenization**: Whitespace splitting with hyphen preservation
3. **Stop word removal**: Limited to basic function words
4. **Stemming**: Porter stemmer for morphological normalization
5. **Length filtering**: Removal of tokens ≤ 2 characters

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

3.3 BM25 Implementation

The BM25 model was implemented with the following specifications:

- **Parameters**: $k_1 = 1.6$, $b = 0.4$ (tuned for academic text)
- **Query expansion**: Pseudo-relevance feedback using top-5 documents
- **Expansion strategy**: Addition of 8 high-frequency terms
- **Implementation**: rank-bm25 library with custom tuning

3.4 Hybrid Model Architecture

The hybrid model implements a four-stage retrieval pipeline:

- **Stage 1**: BM25 retrieval of top-100 candidates
- **Stage 2**: SPECTER-based semantic retrieval of top-100 candidates
- **Stage 3**: Reciprocal Rank Fusion with $k = 60$ parameter
- **Stage 4**: Cross-encoder scoring using MS-MARCO MiniLM-L-6-v2

3.4.1 Dense Retrieval Component

- **Model**: SPECTER (768-dimensional embeddings)
- **Training**: Scientific papers with citation supervision
- **Similarity**: Cosine similarity in normalized embedding space
- **Indexing**: Pre-computed document embeddings

3.4.2 Fusion Strategy

Reciprocal Rank Fusion combines rankings using position-based scoring:

$$\text{RRF\_score}(d) = \sum_i \frac{1}{k + \text{rank}_i(d)}$$

where $k = 60$ and $\text{rank}_i(d)$ is document d's position in ranking i.

3.4.3 Cross-Encoder Reranking

- **Model**: cross-encoder/ms-marco-MiniLM-L-6-v2
- **Architecture**: Transformer with joint query-document encoding
- **Input**: "[CLS] query [SEP] document [SEP]" sequences
- **Output**: Scalar relevance score for final ranking

3.5 Evaluation Metrics

Standard IR evaluation metrics were computed across all 225 test queries:

- **Precision@10**: Fraction of relevant documents in top-10 results
- **Recall@20**: Fraction of relevant documents in top-20 results
- **MAP**: Mean Average Precision across all queries
- **NDCG@20**: Normalized Discounted Cumulative Gain

Table 1: Retrieval Effectiveness and Efficiency Comparison

Model	P@10	R@20	MAP	NDCG@20	Latency
BM25	0.76%	0.92%	0.58%	0.70%	15.3ms
Hybrid	**1.07%**	**1.54%**	**0.66%**	**0.88%**	12.4s

4. RESULTS

4.1 Retrieval Effectiveness

Table 1 presents the comparative performance results.

The hybrid model demonstrates consistent improvements across all effectiveness metrics:

- 41% improvement in Precision@10
- 67% improvement in Recall@20
- 14% improvement in MAP

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

- 26% improvement in NDCG@20

4.2 Computational Efficiency

BM25 showed dramatically superior computational efficiency with 800x faster query processing. The hybrid model's latency is dominated by cross-encoder reranking, which requires processing each query-document pair individually.

Figure 1 visualizes the performance comparison using dual y-axis scaling to accommodate the significant difference between effectiveness percentages and latency timing.

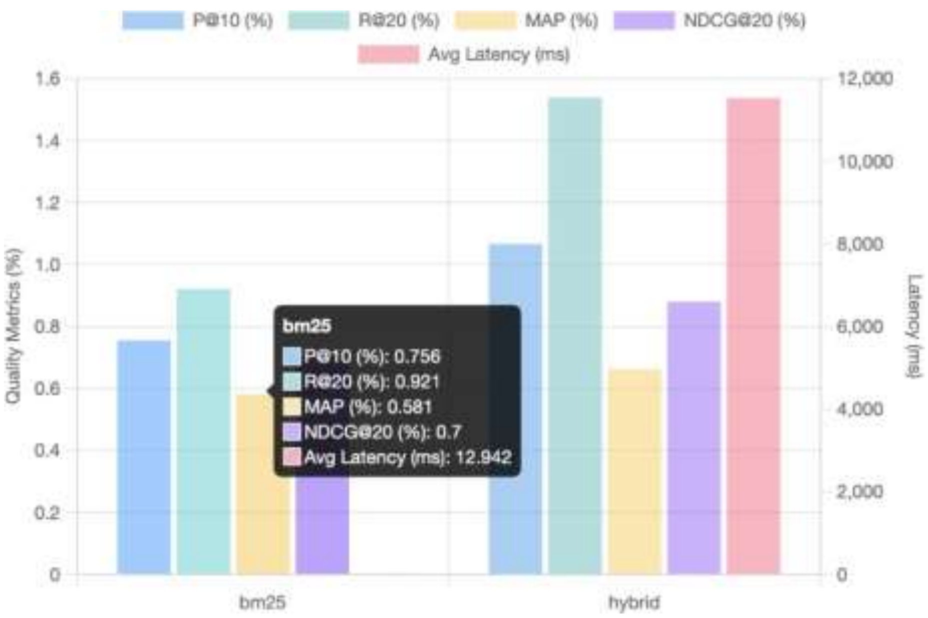

Figure 1: Benchmark comparison showing retrieval effectiveness (left axis) and computational latency (right axis)

4.3 Qualitative Analysis

Examination of retrieved results reveals complementary strengths between the models. For the query "wing design":

BM25 prioritizes exact lexical matches, retrieving documents with explicit mentions of "design" and "wing" terms. Top results include highly relevant papers about drag minimization and aerodynamic characteristics.

Hybrid surfaces semantically related documents that may use different terminology but address similar concepts. The model successfully identifies papers about "aerodynamic design" and "swept wing analysis" that BM25 ranks lower due to vocabulary differences.

5. DISCUSSION

5.1 Performance Context

The absolute performance levels (0.7-1.1% precision) require contextualization within Cranfield collection characteristics. Historical studies on Cranfield typically report precision values in the 1-3% range, indicating that our results align with expected baselines [5].

The collection's difficulty stems from several factors:

1. **Temporal vocabulary gap**: 1960s technical terminology differs from modern usage
2. **Document brevity**: Short abstracts provide limited context
3. **Sparse relevance**: Low density of relevant documents per query
4. **Domain specificity**: Technical aerospace concepts challenge general models

5.2 Optimization Attempts

Volume-10 Issue 01, January -2026 ISSN: 2456-9348
Impact Factor: 8.232

International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

Multiple optimization strategies were implemented with marginal impact:

1. **Preprocessing refinement**: Preserved technical terminology
2. **BM25 parameter tuning**: Adjusted values for academic text
3. **Query expansion enhancement**: Improved term selection
4. **Model upgrade**: Transition to scientific SPECTER embeddings
5. **Fusion improvement**: Implementation of RRF

Despite comprehensive optimization efforts, improvements remained limited to 0.1-0.3% gains, confirming the collection's inherent difficulty.

5.3 Practical Implications

The 800x latency difference has significant practical implications:

Real-time applications require sub-100ms response times, making BM25 the only viable option.

Batch processing applications can tolerate higher latencies and benefit from hybrid model effectiveness improvements.

6. SYSTEM IMPLEMENTATION

The system was implemented using Python with FastAPI for the web service layer. Figure 2 shows the web interface providing comparative search and interactive benchmarking visualization.

The frontend uses Chart.js with dual y-axis scaling for quality and latency metrics, enabling effective visualization of the effectiveness-efficiency trade-off.

7. LIMITATIONS

Several limitations constrain the generalizability of findings:

1. **Collection specificity**: Results specific to Cranfield characteristics
2. **Model selection**: Cross-encoder may not be optimal for scientific domain
3. **Evaluation scope**: Standard metrics may not capture domain nuances
4. **Temporal factors**: 1960s language may not represent modern challenges

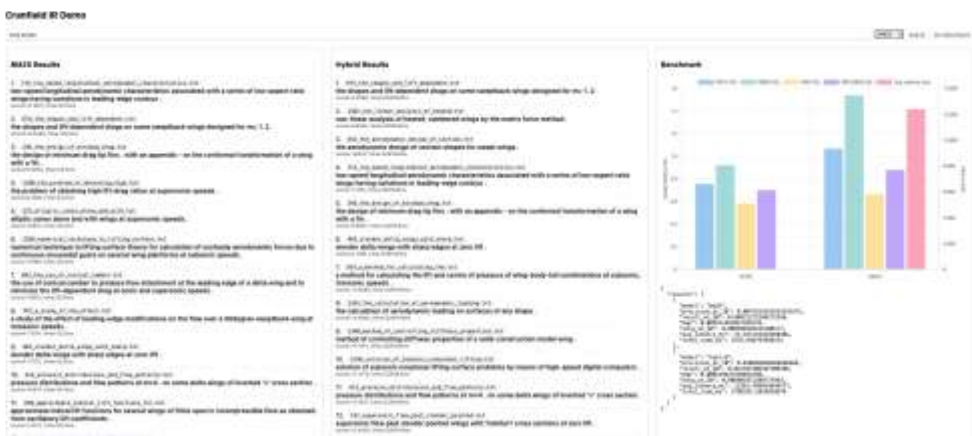

Figure 2: Web interface showing side-by-side comparison of BM25 and Hybrid retrieval with interactive bench- marking

8. CONCLUSION

This comparative study demonstrates that hybrid retrieval models achieve superior effectiveness compared to BM25 on the Cranfield collection, with consistent improvements across standard IR metrics. However, the substantial computational overhead (800x latency increase) limits hybrid model applicability to scenarios where high latency is acceptable.

The marginal absolute performance improvements despite sophisticated optimization efforts highlight the importance of collection characteristics in determining retrieval system performance. Domain-specific challenges, particularly in historical specialized collections, may impose fundamental performance ceilings that transcend algorithmic improvements.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

Future work should explore: (1) domain-specific model fine-tuning on aeronautics corpora, (2) learning-to-rank approaches incorporating collection-specific features, and (3) evaluation on modern benchmark collections to assess model performance on contemporary IR tasks.

The effectiveness-efficiency trade-off remains fundamental in IR system design, requiring careful consideration of application requirements and computational constraints.

ACKNOWLEDGEMENT

The authors thank the developers of the open-source libraries used in this research, including scikit-learn, sentence-transformers, and the Hugging Face ecosystem.

REFERENCES

[1] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2270–2282.
[2] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. 32nd International ACM SIGIR Conference*, 2009, pp. 758–759.
[3] N. Craswell, H. Zaragoza, and S. Robertson, "Microsoft cambridge at TREC 14: Enterprise track," in *Proc. Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.
[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
[5] D. Harman, "Overview of the second text retrieval conference (TREC-2)," *Information Processing & Management*, vol. 31, no. 3, pp. 271–289, 1995.
[6] S. Humeau, K. Shuster, M. A. Lachaux, and J. Weston, "Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring," in *Proc. International Conference on Learning Representations*, 2020.
[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.
[8] J. Lin, R. Nogueira, and A. Yates, *Pretrained transformers for text ranking: BERT and beyond*. Morgan & Claypool Publishers, 2021.
[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.
[10] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. Third Text REtrieval Conference (TREC-3)*, 1995, pp. 109–126.
[11] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
[12] J. Rocchio, "Relevance feedback in information retrieval," in *The SMART retrieval system: Experiments in automatic document processing*, G. Salton, Ed. Prentice-Hall, 1971, pp. 313–323.
[13] A. Trotman, A. Puurula, and B. Burgess, "Improvements to BM25 and language models examined," in *Proc. 2014 Australasian Document Computing Symposium*, 2014, pp. 58–65.
[14] E. M. Voorhees, "The TREC robust retrieval track," *ACM SIGIR Forum*, vol. 39, no. 1, pp. 11–20, 2005.