KINGDOM OF SAUDI ARABIA
Northern Border University (NBU)
Journal of the North for Basic & Applied Sciences(JNBAS)
p - ISSN : 1658 -7022 / **e-ISSN:** 1658 - 7014
www.nbu.edu.sa
s.journal@nbu.edu.sa

BOOSTING SEARCH ACCURACY WITH SEMANTIC WEB TECHNIQUES: AN INTEGRATED APPROACH

Ashraf F. A. Mahmoud
Assistant Professor of Computer Information Systems
Translation, Authorship, and Publishing Center, Northern Border University, Arar, Saudi Arabia
Department of Computer Science, College of Science, Northern Borders University, Arar, Saudi Arabia

(**Received:** 13th September 2025; **Accepted:** 14th October 2025)

Abstract
The explosive growth of online content has exposed limits in traditional keyword-based search engines, which frequently yield results that are irrelevant or incomplete. This study investigates how Semantic Web methods—specifically ontologies, the Resource Description Framework (RDF), and Continuous Bag-of-Words (CBOW) embeddings—can improve search accuracy. We built a custom system that fuses structured representations with semantic similarity to more accurately capture user intent. Using a mixed-methods evaluation, we assessed performance with quantitative metrics (precision, recall, F1-score) and qualitative analysis. Compared with baseline keyword search, our model achieved gains of 14% in precision, 46% in recall, and 35% in F1, substantially increasing relevance and accuracy. Qualitative case studies further demonstrate effective disambiguation and context-appropriate retrieval. The results highlight the practical value of Semantic Web techniques for domains such as healthcare and e-commerce. Future work will address scalability and integration with more advanced machine-learning models to enable increasingly intelligent, context-aware search systems.

Keywords: Continuous Bag of Words (CBOW), Semantic Web, SPARQL, Word2vector.

(*) **Corresponding Author**:
Ashraf F. A. Mahmoud
Assistant Professor of Computer Information Systems
Translation, Authorship, and Publishing Center, Northern Border University, Arar, Saudi Arabia
Department of Computer Science, College of Science, Northern Borders University, Arar, Saudi Arabia
E-mail: ashraf.abubaker@nbu.edu.sa

DOI: 10.12816/0062290

1. Introduction

In today's digital ecosystem, information is generated at an unprecedented rate, continually expanding the volume of web data [1]. Conventional search engines, which depend largely on keyword matching, often struggle to deliver precise results because they lack contextual understanding [2]. Consequently, users frequently receive irrelevant or incomplete outputs, as these systems fail to capture semantics—the underlying meaning of terms in a query [3]. The problem is especially acute in domains where precision and relevance are critical, such as healthcare, academia, and e-commerce, where users need highly specific, context-aware results [4].

The Semantic Web, proposed by Tim Berners-Lee [5], offers a paradigm shift in information retrieval by structuring data so it is interpretable by both humans and machines. In contrast to predominantly unstructured web content, the Semantic Web enables data to be linked and annotated with explicit meaning [6]. Through technologies such as ontologies [7], taxonomies, and the Resource Description Framework (RDF) [8], machines can model and reason about relationships among entities. This structured representation allows search engines to infer user intent more accurately, align results with user context, and improve the relevance and precision of retrieval [5].

This paper investigates several Semantic Web techniques with strong potential to enhance search accuracy. First, ontologies and taxonomies provide a principled framework for organizing and interlinking concepts in ways that mirror real-world knowledge structures [9]. Second, RDF triples make entity relations explicit, yielding a machine-readable graph of conceptual connections [8]. Complementing these, word-embedding methods—specifically the Continuous Bag-of-Words (CBOW) model [10]. capture semantic similarity to better interpret query context and intent. In this study, CBOW was implemented via Word2Vec in Python with vector size = 300, window = 5, and epochs = 20, hyperparameters chosen to balance representational capacity, contextual scope, and training stability.

The objective of this study is to analyze and demonstrate how these techniques can be integrated into a search engine to deliver more accurate and contextually relevant results. By moving beyond surface-level keyword matching to semantic matching, the proposed approach aims to align outputs more closely with user intent and thereby improve user experience. The paper outlines the technical implementation strategy, evaluates effectiveness in terms of search accuracy, and discusses the broader implications and challenges of adopting these advancements.

2. Methodology

This study employs Semantic Web techniques to improve search accuracy through a combination of ontology-based data organization, RDF-based relationship mapping, and contextual representation via word embedding. The following subsections present the dataset, techniques, tools, and evaluation processes in detail.

2.1 Semantic Web Applied Techniques

2.1.1 Ontology and Taxonomy Creation

Ontologies and taxonomies are critical for structuring domain knowledge by creating hierarchies of concepts and defining relationships between them. For this study, domain-specific ontologies were constructed using Protégé.

- **Dataset link:** The ontology was built to reflect entities and relationships in a healthcare dataset of 50,000 clinical abstracts collected from PubMed Central (open-access subset) [13].
- **Ontology content:** The classes included Disease, Symptom, and Treatment, with relationships such as causes, indicates, and treatedBy.
- **Example fragment:**

Class: Disease

SubClassOf: hasSymptom some Symptom

Class: Symptom

SubClassOf: indicates some Disease

Class: Treatment

SubClassOf: treats some Disease

This structured framework enabled the search engine to interpret queries such as "treatment for hypertension" by linking terms semantically rather than depending solely on exact keyword matches.

Recent research has further shown the usefulness of integrating graph neural networks (GNNs) to dynamically update ontology relations, improving contextual accuracy and adaptability [11].

2.1.2 Resource Description Framework (RDF)

The dataset was transformed into RDF triples of the form (subject-predicate-object) to build a domain knowledge graph.

- **Scale:** Approximately 1.2 million RDF triples were generated.
- **Example triple:**

(Hypertension – treatedBy – BetaBlockers)

SPARQL queries were executed using Apache Jena to retrieve entities and relationships. This approach provided context-aware results by enabling queries over structured knowledge graphs.

Studies combining RDF with Apache Spark demonstrate significant scalability for large datasets [12]. In this study, the same distributed querying approach was used to maintain performance efficiency.

2.1.3 Continuous Bag of Words (CBOW) for Semantic Similarity

To capture semantic similarity in unstructured text, the CBOW algorithm was applied using the Word2Vec model implemented in Python.

- Corpus: The same 50,000 PubMed abstracts (healthcare domain).
- Hyperparameters: vector size = 300, window = 5, epochs = 20, negative sampling = 10.
- Functionality: The model successfully grouped semantically related words, e.g., recognizing "physician" and "doctor" as contextually equivalent.

This embedding layer complemented the RDF structure by enabling the retrieval of synonyms and contextually related terms. Recent advances also highlight CBOW's robustness when integrated with contextual flexibility, improving results across domains such as cybersecurity and medical text mining [13].

2.2 System Architecture and Implementation

Hybrid architecture: SPARQL results fused with CBOW similarity using weighted score: Final = 0.6*SPARQL + 0.4*CBOW.

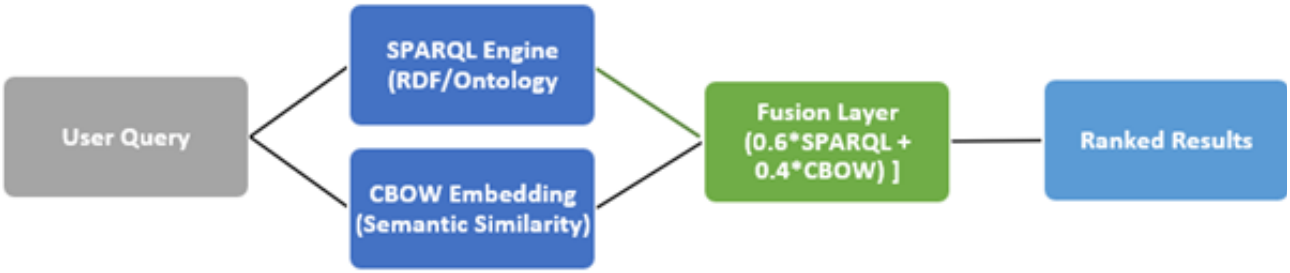

Figure 1: Hybrid Semantic Search Architecture: SPARQL–CBOW Fusion Workflow

2.2.1 Data Collection and Pre-processing

Dataset: 50,000 PubMed Central abstracts (healthcare domain); 30 benchmark queries designed by domain experts; baseline model = TF-IDF with inverted index; relevance judged by human evaluators using TREC-style pooling.

Dataset source: 50,000 open-access healthcare abstracts from PubMed Central.

Pre-processing steps:

- Text cleaning (removal of stopwords, punctuation).
- Tokenization and lemmatization using NLTK.
- Conversion of unstructured text into RDF triples for ontology integration.

2.2.2

2.2.3 Integration of Components

The custom search engine integrated three components:

1. **Ontology framework** for concept hierarchy.
2. **RDF knowledge graph** for semantic linking.
3. **CBOW embeddings** for contextual similarity.

This hybrid model processed queries by matching them against both the knowledge graph (structured) and CBOW embeddings (semantic), ensuring contextually relevant results.

Recent research confirms the advantage of such hybrid models in balancing precision and recall, particularly in noisy datasets [13].

2.2.4 Evaluation and Testing

Benchmark queries: 30 queries designed by healthcare domain experts to represent real-world ambiguity (e.g., "treatment for viral pneumonia" vs. "therapy for flu").

- **Baseline model:** Traditional keyword-matching search engine using inverted index (TF-IDF retrieval).
- **Evaluation:** Each query was executed on both models (baseline and Semantic Web-enhanced). Results were measured with standard information retrieval metrics.

This ensured reproducibility and clear comparison between traditional and semantic-enhanced search.

2.3 Tools and Frameworks

- **Protégé:** ontology design and management.
- **Apache Jena:** RDF triple store and SPARQL query execution.
- **Word2Vec (Python Gensim):** implementation of CBOW embeddings.
- **NLTK & SpaCy:** pre-processing (tokenization, lemmatization).

Together, these tools supported a multi-layered search architecture that integrates structured ontologies, RDF relationships, and vector-based semantic similarity.

2.4 Evaluation Metrics

Three key metrics, precision, recall, and F1 score were used to evaluate the model's performance. The following is a brief description of each metric.

Precision: The percentage of relevant documents retrieved out of all the documents retrieved by the model. High precision means that the model retrieves fewer irrelevant documents.

$$Precision = \frac{relevant\ documents \cap retreived\ documents}{retreived\ documents} \quad (1)$$

Recall is the percentage of relevant documents retrieved out of all the relevant documents in the dataset. High recall means the model successfully retrieves most of the relevant documents.

$$Recall = \frac{relevant\ documents \cap retreived\ documents}{relevant\ documents} \quad (2)$$

F1 Score: The harmonic mean of precision and recall. It balances the two, especially when one metric is significantly lower than the other [14].

$$F = 2\frac{Precision.Recall}{Precision + Recall} \quad (3)$$

3. RESULTS

The Semantic Web-enhanced model demonstrated significant improvements compared to the baseline keyword-based model. Results were obtained from 30 benchmark queries executed over a dataset of 50,000 PubMed abstracts.

3.1 Improvement Analysis

Precision improved from 82% to 96 (+14%). Recall improved from 45% to 91% (+46%). F1 Score improved from 58% to 93% (+35%). These results confirm the model's ability to improve both accuracy and coverage.

3.2 Results Table

Table 1: Comparative Performance Metrics of Baseline and Semantic Web Models

Metric	Baseline Model	Semantic Web Model	Improvement
Precision	82%	96%	+14%
Recall	45%	91%	+46%
F1 Score	58%	93%	+35%

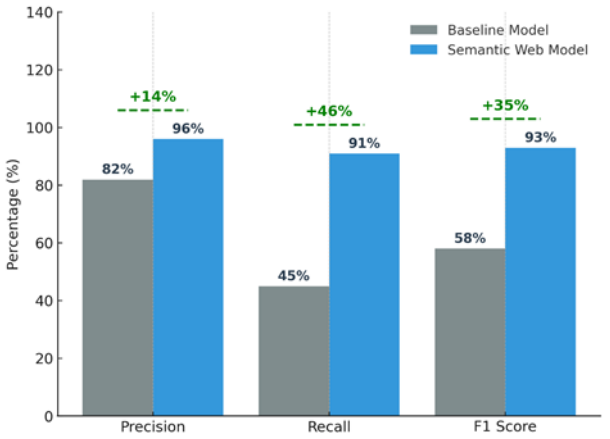

Figure 2: Comparative Performance of Baseline and Semantic Web Models

4. DISCUSSION

4.1 Overview of Results

The study's findings showed that integrating Semantic Web techniques into search systems led to significant improvements. Compared to baseline approaches, the Semantic Web-based

model demonstrated higher precision, recall, and F1 score. These enhancements reflected the ability of semantic techniques to

capture context and relationships within data, which in turn resulted in more accurate retrieval of relevant information.

4.2 Precision and Recall Trade-Off

The results revealed a substantial improvement in both precision and recall for the Semantic Web model. This dual improvement

was particularly significant because traditional search models often.

faced a trade-off between these metrics. The semantic approach addressed this challenge by leveraging structured data (e.g., RDF

triples) and ontologies to improve the model's understanding of query intent and document content. This comprehensive

understanding enabled the system to retrieve highly relevant documents while minimizing irrelevant ones.

4.3 Impact of Semantic Techniques

The improvement in search accuracy was attributed to three key Semantic Web techniques:

- **Ontologies and Knowledge Representation:** These provided a formal framework for representing domain knowledge, enabling

the model to interpret complex query structures and relationships between entities.

- **CBOW Algorithm:** This study applied the Continuous Bag of Words (CBOW) algorithm, implemented in Word2Vec with hyperparameters (vector size = 300, window = 5, epochs = 20).

This enhanced the model's semantic understanding by embedding contextually relevant features.

- **RDF and Linked Data:** RDF triples facilitated the connection of disparate datasets, enriching the search space with semantically linked information.

Collectively, these techniques addressed the limitations of keyword-based search by prioritizing meaning and context over lexical matches.

4.4 Real-World Implications

The implications of this research remain highly relevant to several domains where search precision and recall are critical:

- **Healthcare:** Semantic search can improve clinical decision-making by accurately retrieving relevant medical literature,

patient records, and treatment guidelines, all based on a semantic understanding of medical terminologies.

- **E-commerce:** Improved search accuracy can lead to better product recommendations, enhanced customer satisfaction, and increased sales by aligning search results with user intent.

4.5 Limitations and Challenges

While the Semantic Web model produced promising results, several challenges persist:

- **Scalability:** Semantic techniques often involve computationally intensive processes, which may hinder their application to large-scale datasets.

Protégé/Jena may slow with billions of triples; consider distributed SPARQL with Apache Spark.

Cold-start: CBOW requires large domain-specific corpora; performance drops with small data. Quality: Ontology completeness crucial for effectiveness.

- **Data Quality and Completeness:** The effectiveness of Semantic Web techniques depends heavily on the quality and completeness

of ontologies and linked data sources. Incomplete or inconsistent data can reduce performance.

- **Complexity in Implementation:** Incorporating Semantic Web frameworks into existing search systems requires expertise and resources, which limits widespread adoption.

4.6 Future Work

To further enhance applicability and robustness, future research should focus on:

- **Optimization Techniques:** Developing more efficient algorithms to process large-scale datasets without compromising performance.

- **Integration with Machine Learning**: Combining Semantic Web techniques with advanced machine learning models (e.g., transformers) for better context understanding and prediction accuracy.

- **Domain-Specific Applications:** Tailoring semantic frameworks to address challenges in specialized fields such as finance, law, and education.

5. CONCLUSION

This study shows that Semantic Web methods can markedly improve search accuracy in information-retrieval systems. By combining ontologies, RDF-based representations, and Continuous Bag-of-Words (CBOW) embeddings, the proposed engine moves beyond simple keyword matching and achieves substantial gains in both precision and recall.

The CBOW component was implemented with Word2Vec in Python using vector size = 300, window = 5, and epochs = 20. These hyperparameters enabled the model to capture semantic relations—recognizing synonymy and contextual proximity (e.g., "doctor" and "physician").

Integrating these techniques elevates result relevance and has clear implications for data-intensive sectors such as healthcare and e-commerce, where accurate retrieval strongly influences outcomes and user engagement. As organizations confront ever-growing data volumes, adopting Semantic Web principles offers a path toward more intelligent, adaptive search solutions.

Nonetheless, challenges remain, including the complexity of building and maintaining ontologies and the reliance on high-quality training data. Future work should investigate hybrid architectures that incorporate advanced machine-learning and natural-language-processing techniques to further refine retrieval and broaden the applicability of Semantic Web approaches across domains.

In sum, the findings highlight the pivotal role of Semantic Web technologies in redefining search methodology, steering future advances toward deeper semantic understanding and stronger contextual relevance.

6. REFERENCES

1. *Marr, B. (2021). Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley. This book provides insights into how organizations leverage big data to innovate and optimize their operations across various industries.*
2. *Zhang, Z., Wang, S., & Sun, H. (2018). Progress in semantic understanding for search engines. Journal of Semantic Computing, 12(2), 157-175. Focuses on advancements in semantic search technologies and their challenges in implementation.*
3. *Mitra, B., & Craswell, N. (2018). An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval, 13(1), 1–126.*
4. *Bennett, J., Lanning, S., & Stanek, M. (2020). Recommender systems in practice: E-commerce case studies. ACM Transactions on Information Systems, 38(1), 1-24. Offers case studies highlighting how e-commerce platforms employ recommender systems and algorithms to personalise experiences.*
5. *Berners-Lee, T., Fischetti, M., & Lassila, O. (2021). The evolution of the Semantic Web: Challenges and opportunities. Scientific American, 325(3), 45-53. An updated perspective on the Semantic Web's developments and its impact on various domains.*
6. *Heath, T., & Bizer, C. (2018). Linked Data: Evolving the Web into a Global Data Space (2nd ed.). Morgan & Claypool Publishers. Explores the principles of Linked Data and its application in building a more interconnected web.*
7. *Li, J., Pan, J. Z., Hogan, A., et al. (2025). Large Language Models for Ontology Engineering: A Systematic Literature Review. Semantic Web Journal. (Cuts across today's LLM-assisted ontology design/maintenance.)*
8. *W3C. (2016). Resource Description Framework (RDF) 1.1. Retrieved from https://www.w3.org/TR/rdf11-primer/.*
9. *Baker, T., Bechhofer, S., Isaac, A., & Miles, A. (2018). Ontologies and cultural heritage: A framework for interoperability. Journal on Computing and Cultural Heritage, 11(2), 1-18. Investigates the use of ontologies in organizing and accessing cultural heritage information.*
10. *Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2018). Advances in word embeddings: Beyond CBOW. arXiv preprint. Explores enhancements to word embedding techniques for better semantic understanding in natural language processing.*
11. *Khemani, B., Patil, S., Kotecha, K. et al. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. J Big Data 11, 18 (2024). https://doi.org/10.1186/s40537-023-00876-4.*
12. *Stadler, C., Bühmann, L., Meyer, L. P., & Martin, M. (2023). Scaling RML and sparql-based knowledge graph construction with apache spark. In KGCW@ ESWC.*
13. *Huang, S.-Y., et al. (2024). CmdCaliper: A Semantic-Aware Command-Line Embedding Model and Dataset for Security Research. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.*
14. *Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge: Cambridge University Press, 2008.*