Enhancing Semantic Search Precision through the CBOW Algorithm in the Semantic Web

Ashraf F.A. Mahmoud
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
ashraf.abubaker@nbu.edu.sa (corresponding author)

Zakariya M. S. Mohammed
Center for Scientific Research and Entrepreneurship, Northern Border University, Arar, Saudi Arabia
zakariya.mohammed@gmail.com

Mohamed Ben Ammar
Department of Information Systems, College of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia
mohammed.Ammar@nbu.edu.sa

Ali Satty
Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia
alisatty1981@gmail.com

Faroug A. Abdalla
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
faroug.abdalla@nbu.edu.sa

Gamal Saad Mohamed Khamis
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
jamal.khamis@nbu.edu.sa

Mohyaldein Salih
Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia
mohyyassin@gmail.com

Abdelnasser Saber Mohamed
Computer Science Department, Applied College, Northern Border University, Arar, Saudi Arabia
abdelnasser.mohammed@nbu.edu.sa

Received: 30 October 2024 | Revised: 18 November 2024 | Accepted: 29 November 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.9450*

ABSTRACT

The Semantic Web enhances data interoperability and enables intelligent information retrieval through structured data representation. However, challenges remain in achieving high precision in semantic search. This paper uses the Continuous Bag of Words (CBOW) model to enhance semantic search precision. By generating rich word embeddings, CBOW enables a better understanding of contextual relationships among terms within semantic queries. Our approach has been evaluated using the websites intended to be used as a sample for testing the efficiency of semantic information retrieval, demonstrating significant improvements in search precision compared to traditional methods. The findings indicate that integrating CBOW into semantic search frameworks can lead to more relevant and accurate search results, paving the way for future advancements in Semantic Web technologies.

Keywords-semantic search; continuous bag of words; semantic information retrieval; Word2vec

I. INTRODUCTION

The Semantic Web represents a significant evolution of the internet, enabling machines to process and understand data in a meaningful way. Linking structured data across various domains enhances the potential for intelligent information retrieval and data integration. The Semantic Web facilitates data sharing and reuse across applications, enterprises, and communities by improving the web with semantic annotations [1, 2]. However, despite these advancements, achieving high precision in semantic search remains a critical challenge [3, 4]. Current semantic search systems often struggle to accurately interpret user queries and provide relevant results, leading to inefficiency and user frustration [5]. Addressing this challenge is crucial to realizing the full potential of the Semantic Web. One promising approach to improve semantic search precision is the integration of Natural Language Processing (NLP) techniques. The Continuous Bag of Words (CBOW) model, a prominent NLP technique, offers a promising solution to this problem. By generating dense vector representations of words based on their context, CBOW captures semantic relationships and nuances that traditional keyword-based search methods may overlook [6].

This study reviews existing work on semantic search methodologies [7], highlights the limitations of current approaches [2], and discusses the role of word embedding models, particularly CBOW, in enhancing semantic search [6]. Despite progress in semantic search, several challenges need to be addressed to improve performance. Semantic search systems aim to improve the retrieval of relevant information by understanding user intent and the context of queries. Traditional keyword-based search methods often fail to address human language's complexities and web data's rich semantics [8]. Recent advancements in semantic search have focused on leveraging structured data and ontologies to improve precision and relevance. For instance, an ontology-based semantic search framework was proposed that uses ontological hierarchies to enhance query understanding and result ranking. However, such approaches may still struggle with ambiguities and the semantic variance of natural language. To further improve the effectiveness of semantic search systems, several key challenges must be overcome:

- Ambiguity in Queries: Users often formulate queries with ambiguous terms, which can lead to irrelevant results. This issue necessitates robust disambiguation techniques to enhance search precision.
- Heterogeneous Data Sources: Integrating data from diverse sources can introduce inconsistencies and reduce the overall precision of search results.
- Scalability: As the volume of web data grows, maintaining search efficiency and precision becomes increasingly tricky.

Recent studies have shown that word embedding models, including Word2Vec, GloVe, and FastText, can significantly enhance semantic search performance by capturing contextual word relationships [6, 9, 10]. These models generate dense vector representations of words that reflect their meaning with respect to each other. The success of these models highlights the potential of word embeddings to improve semantic search capabilities. The CBOW model, a variant of the Word2Vec algorithm, leverages context to predict target words, making it well suited for capturing nuanced word semantics. Research has demonstrated the potential of CBOW to enhance various NLP tasks, such as sentiment analysis and document classification [11]. Despite the success of word embedding models, the integration of CBOW to semantic search frameworks is an underexplored area. Most current approaches either stick to traditional keyword-based methods or fail to fully exploit the potential of embeddings for semantic search. This paper addresses this research gap by showcasing how CBOW can be effectively incorporated into semantic search systems to significantly improve the precision and relevance of results.

II. METHODOLOGY

This section describes the methodology used in this study to integrate the CBOW model into the semantic search framework. The process consists of two main components: the first stage is the data preparation stage, and the second stage is the use of CBOW word embedding structures for information retrieval. In addition, the evaluation of the methodology is presented.

A. Data Preparation

At this stage, randomly selected words from websites are collected and used as a sample to measure the efficiency of semantic information retrieval. The data were collected using a crawler and stored in the database (https://drive.google.com/file/d/1o8sXYbg-BmkdOs1umDyFOVAut1K_xVWL/view?usp=drive_link). Figure 1 demonstrates the steps of this stage.

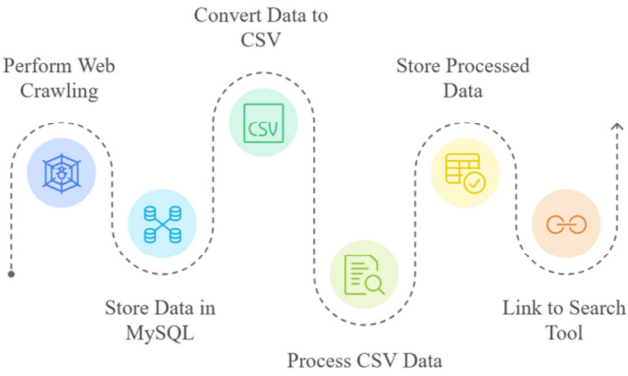

Fig. 1. Data preparation steps.

B. Use of CBOW for information retrieval

At this stage, the CBOW method is used to represent words as vectors to obtain the similarity between the search query word and the results retrieved from the database, and to evaluate the relevance of the results to the word. The stage is divided into the following steps, as shown in Figure 2.

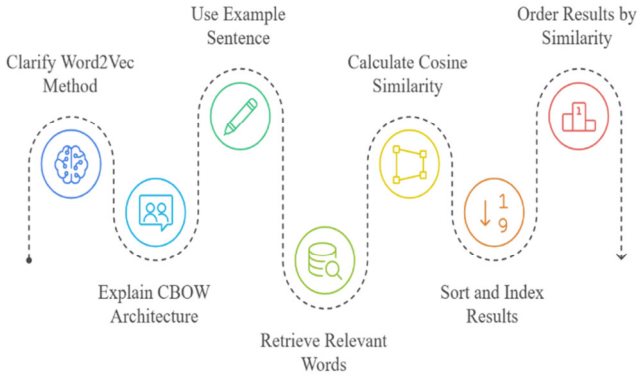

Fig. 2. Steps of word embedding (CBOW) for information retrieval.

1) First Step

In this step, the method for building word embeddings (Word2vec) is clarified and the structure of word representation as vectors to retrieve related words (CBOW) is explained. A vector is formed from the words in the sentence, excluding all special symbols, identifiers, and conjunctions. A word is selected from the words in the vector, and the word 'comprehensive' has been selected. Figure 3 shows the structure of the representation of words as vectors.

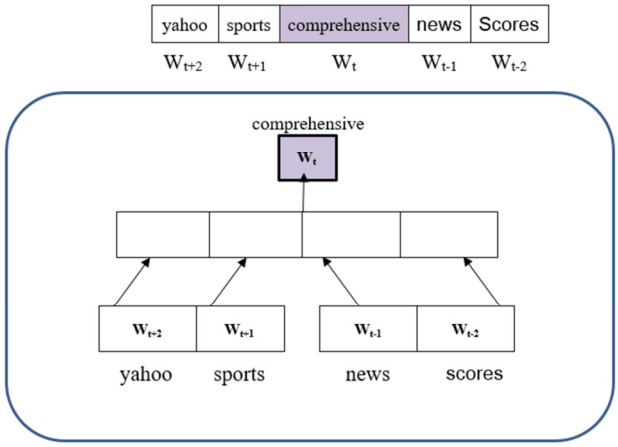

Fig. 3. Structure of the representation of words as vectors (CBOW).

The query consists of multiple words. First, the feature vector of each word is calculated by multiplying the vector of the word X_i by the embedding matrix W_w, which is retrieved from the previous word embedding stage (Word2vec), as shown in (1). Then, the average of the vectors is calculated to obtain a single vector representing the user's query.

$$V_{query}\ (w) = \frac{1}{n}\sum W_w X_i \quad (1)$$

where W_w $(w \leq w_1, w_2, \ldots w_n >)$ is the word embedding for the word X_i, which can be learned by the CBOW architecture [12].

2) Second Step

This step explains the mechanism for retrieving the most relevant words from the database and the method for finding the similarity ratios between words. Cosine similarity is generally easier to compute than other distance metrics and is widely used in Word2vec. It is a normalized dot product of two vectors and this ratio defines the angle between them as shown in (2). A cosine similarity of 1 means that the two vectors have the same direction, while a cosine similarity of 0 means that the vectors are at a 90-degree angle to each other. A cosine similarity of -1 means that the vectors are opposite to each other, regardless of their magnitude.

$$sim(A,B) = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} \quad (2)$$

All search queries are converted into semantic vectors using word embedding techniques. The cosine similarity between the vectors is then calculated to find the most relevant topic to the query [12]. Then, using a code written in Python, the similarity results are sorted and indexed the from the highest to the lowest similarity score. After that, the existing data in the database are accessed to retrieve the appropriate words related to the topic of the query.

3) Third Step

In this step, the mechanism for ordering the results with a similarity percentage greater than 0.04 is clarified, followed by the presentation of the results according to the similarity percentage. In this study, a minimum similarity threshold of 0.04 was adopted to achieve the highest accuracy in retrieving related words and the highest recall rate. Therefore, any similarity results lower than 0.04 are excluded from the query results and the results are ranked from highest to lowest.

C. Evaluation

The performance is evaluated in terms of precision, recall, and F score. High recall indicates high coverage of the system. Precision represents the correctness of the results, as shown in (3), which is the ratio of relevant retrieved words to the total number of relevant as well as irrelevant words retrieved [12].

$$\text{Precision} = \frac{\text{relevant words} \cap \text{retreived words}}{\text{retreived words}} \quad (3)$$

Recall represents the completeness of coverage as shown in (4), meaning the ratio of the number of relevant words retrieved to the total number of all possible relevant words [12].

$$\text{Recall} = \frac{\text{relevant words} \cap \text{retreived words}}{\text{relevant words}} \quad (4)$$

The F1 score is the harmonic mean of precision and recall, as shown in (5) [12].

$$\text{F} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

III. RESULTS

The dataset, which initially contained 194 queries, was processed to remove duplicates and irrelevant queries. Ten random queries were then selected to test the proposed system. Word embedding was used to transform the user queries and

words into vectors, and the cosine similarity between the vectors was calculated. This enabled the retrieval of related words associated with the presented concept as results. Table I shows the retrieved words, including their index and similarity ratio.

TABLE I. RESULTS OF THE HIGHEST SIMILARITY RATIO FOR RETRIEVED WORDS RELATED TO THE QUERY

No	The retrieved word relevant to the query	Similarity	The index specific to the word
1	comprehensive	0.06599839	140
2	comprehensive	0.06599839	131
3	comprehensive	0.06599839	122
4	comprehensive	0.06599839	113
5	standings	0.040465258	143

The performance of the semantic search tool is measured by precision, recall, and F-score. The retrieved words are used to calculate precision and recall, as well as the F-score, when the similarity ratio is greater than or equal to 0.04, as shown in Table II and Figure 4.

TABLE II. RESULTS OF PRECISION, RECALL, AND F FOR RELEVANT QUERIES AFTER SEARCHING TEN QUERIES AS A SAMPLE WITH A SIMILARITY THRESHOLD GREATER THAN OR EQUAL TO 0.04 (>= 0.04)

No	Words	Number of retrieved words	Number of relevant words	Precision (%)	Recall (%)	F1 score (%)
1	comprehensive	4	4	100	100	100
2	operator	7	6	85.7	100	92
3	comparison	4	4	100	100	100
4	elements	4	4	100	100	100
5	parameter	11	8	72.7	100	84
6	sports	4	4	100	100	100
7	scores	4	4	100	100	100
8	news	25	10	40	100	57
9	technology	16	3	18.7	100	32
10	media	3	3	100	100	100
Average				82	100	87

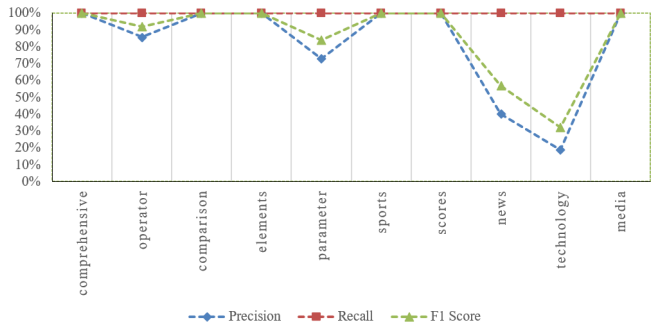

Fig. 4. Precision, recall, and F value for testing a similarity ratio (>=0.04).

IV. DISCUSSION

The results of testing the proposed semantic search system on a dataset of 194 queries, with 10 random queries selected for evaluation, provide valuable insights into the system's performance in terms of similarity, precision, recall, and the F1 score. Here's a breakdown of the process and key findings from the results:

A. Data Preprocessing and Word Embedding

After preprocessing the dataset by removing duplicates and irrelevant queries, word embeddings were employed to transform user queries and their corresponding words into vectors. The cosine similarity between these vectors was calculated to determine the semantic closeness between the queries and retrieved words. Cosine similarity is a standard measure used in semantic search to assess how similar two-word vectors are, based on the angle between them, with values ranging from -1 to 1, where 1 indicates identical vectors.

For example, in Table 1, the highest similarity word retrieved was "comprehensive," with a similarity score of 0.06599839 across multiple indices. This low similarity ratio suggests a potential limitation in the model's ability to clearly distinguish between closely related words, which is expected in semantic searches that rely on general word embeddings.

B. Precision and Recall Calculation

The system's performance in retrieving relevant words for the selected queries was evaluated using standard information retrieval metrics:

- Precision: The proportion of retrieved words that are relevant.
- Recall: The proportion of relevant words retrieved out of all relevant words in the dataset.
- F1 score: The harmonic mean of precision and recall, providing a balanced performance measure [7, 13].

Table II presents the precision, recall, and F1 scores at a similarity threshold of 0.04, meaning that only words with a cosine similarity score of 0.04 or higher were considered relevant for the given queries. Let's delve into these metrics:

1) High Precision and Recall in Some Queries

Several queries showed high precision and recall values, reflecting the system's ability to retrieve highly relevant words in those instances. For words such as "comprehensive," "comparison," "elements," "sports," and "media," the precision and recall were both 100%, resulting in an F1 score of 1.0. This indicates that every retrieved word was relevant, and that all relevant words were retrieved. This is a positive outcome, indicating that the model effectively captures the semantic meaning in specific queries and retrieves appropriate results with high confidence.

2) Performance Challenges in Some Queries

However, the system's performance was weaker for other queries, such as "news", "technology," and "parameter", where the precision dropped significantly:

- For the query "news," while the recall was 100%, the precision was 40%, indicating that many retrieved words were irrelevant to the query. The F1 score of 0.57 reflects the imbalanced nature of these results.

- For "technology," the precision was even lower, at 18.7%, with an F1 score of 0.32, showing significant noise in the retrieved words despite the high recall.

This suggests that the model struggles with more complex or ambiguous queries, where the semantic relationships between words may be more nuanced or context-dependent. Relying on a fixed similarity threshold may also result in retrieving words that share surface-level similarities but are semantically distant in context.

3) Implications of Similarity Threshold

The similarity threshold of 0.04 seems relatively low and possibly contributes to the retrieval of irrelevant words. For instance, words with similarity scores just above this threshold may not be contextually relevant, even if their vector representations show minimal differences. Increasing the threshold could improve precision by filtering out less relevant words, although this may come at the cost of lower recall, as fewer words are retrieved overall. A balance between precision and recall must be considered when choosing the optimal similarity threshold for the search system.

C. Overall Semantic Search Tool Performance

The semantic search tool performs well in contexts where there are clear and strong semantic relationships between query terms and retrieved words. However, for more ambiguous or diverse queries, such as "technology" and "news", the system's performance degrades and precision drops significantly. This suggests that while the system can effectively handle straightforward or well-defined queries, it may struggle with queries that require deeper contextual understanding or involve multiple possible interpretations.

D. Potential Areas for Improvement

Several enhancements can be considered to improve the system's overall performance:

- Dynamic similarity threshold: Adjusting the threshold based on the query's complexity instead of a fixed similarity threshold could improve the balance between precision and recall.
- Contextual word embeddings: Leveraging more advanced contextual models such as BERT or GPT-based embeddings could help better capture the nuanced meanings of words in different contexts, leading to improved accuracy for ambiguous queries.
- Query expansion: Incorporating query expansion techniques could retrieve semantically related terms, enhancing precision by considering a broader set of contextually relevant terms [14].
- Ontological integration: Integrating domain-specific ontologies could improve the system's understanding of complex queries in specific fields, leading to more targeted and relevant search results.

The test results highlight both the strengths and limitations of the current semantic search tool. While it successfully retrieves relevant results for some queries, it struggles with more complex or ambiguous queries. By refining the similarity threshold and incorporating more advanced contextual understanding, the system could further enhance its precision and provide more reliable results across a broader range of queries in the future.

V. IMPLICATIONS OF FINDINGS AND LIMITATIONS

The improved precision suggests that semantic search systems can benefit from adopting CBOW-based approaches. This would lead to better user satisfaction and more effective information retrieval. Additionally, this methodology can be extended to other NLP applications within the Semantic Web, such as entity linking and recommendation systems.

Some limitations have been made when applying our approach:

- Domain Dependency: CBOW's performance may vary across different domains, potentially affecting the generalizability of the results.
- Scalability: The computational cost of training CBOW on large datasets can be a challenge in real-time applications.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel approach to enhance semantic search precision by integrating the Continuous Bag of Words (CBOW) model. Our experiments demonstrated significant improvements in search results, highlighting the effectiveness of CBOW in capturing semantic relationships within the data. In this paper, we reached the following conclusions:

- Using the CBOW architecture has increased the efficiency of semantic search.
- Presenting the search results verified that more relevant results can be obtained for the research topic.
- Using open-source resources in software design and development facilitates the software development process, especially in the long run.
- It greatly reduced time and cost by presenting results related to the search terms and excluding results not related to the meaning of the words.

By achieving a recall of 100%, the approach adopted in this paper outperformed the algorithms used in [15], where the maximum recall achieved was 98%. However, the proposed approach falls short in precision and F1 score compared to the results reported in [15]. Additionally, the approach in this paper surpasses the algorithm used in [12] in terms of precision, recall, and F1 score. Overall, our research contributes to the ongoing efforts to improve semantic search capabilities within the Semantic Web, paving the way for more intelligent and effective information retrieval systems.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work under the project number "NBU-FFR-2024-1635-01".

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 29–37, May 01, 2001, https://doi.org/10.1038/scientificamerican0501-34.

[2] N. Shadbolt, T. Berners-Lee, and W. Hall, "The Semantic Web Revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, Jan. 2006, https://doi.org/10.1109/MIS.2006.62.

[3] P. Hitzler, M. Krotzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*. New York, NY, USA: Chapman and Hall - CRC Press, 2009, https://doi.org/10.1201/9781420090512.

[4] L. Ding *et al.*, "Swoogle: a search and metadata engine for the semantic web," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, Washington, D.C., USA, 2004, pp. 652–659, https://doi.org/10.1145/1031171.1031289.

[5] J. D. Ullman and J. Widom, *A First Course in Database Systems*, 3rd ed. New York, NY, USA: Pearson, 2008.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," presented at the *International Conference on Learning Representations*, Scottsdale, AZ, USA, May 2-4, 2013.

[7] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003, pp. 700–709, https://doi.org/10.1145/775152.775250.

[8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[9] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Jun. 2017, https://doi.org/10.1162/tacl_a_00051.

[11] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014, pp. 1188–1196.

[12] E. H. Mohamed and E. M. Shokry, "QSST: A Quranic Semantic Search Tool based on word embedding," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 934–945, Mar. 2022, https://doi.org/10.1016/j.jksuci.2020.01.004.

[13] A. Singhal and I. Google, "Modern Information Retrieval: A Brief Overview," *IEEE Data Engineering Bulletin*, vol. 24, no. 4, pp. 35–43, 2001.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[15] M. M. El-Gayar, N. E. Mekky, A. Atwan, and H. Soliman, "Enhanced Search Engine Using Proposed Framework and Ranking Algorithm Based on Semantic Relations," *IEEE Access*, vol. 7, pp. 139337–139349, 2019, https://doi.org/10.1109/ACCESS.2019.2941937.