Data and Metadata. 2025; 4:443
doi: 10.56294/dm2025443

ORIGINAL

Optimizing Query Using the FOAF Relation and Graph Neural Networks to Enhance Information Gathering and Retrieval

Optimización de la consulta mediante la relación FOAF y redes neuronales de grafos para mejorar la recopilación y recuperación de información

Ahmed Mahdi Abdulkadium[1] ✉, Asaad Sabah Hadi[1]

[1]Software Department, College of Information Technology, University of Babylon. Babylon, Iraq.

Cite as: Mahdi Abdulkadium A, Sabah Hadi A. Optimizing Query Using the FOAF Relation and Graph Neural Networks to Enhance Information Gathering and Retrieval. Data and Metadata. 2025; 4:443. https://doi.org/10.56294/dm2025443

Submitted: 15-02-2024 **Revised:** 09-06-2024 **Accepted:** 12-10-2024 **Published:** 01-01-2025

Editor: Adrián Alejandro Vitón Castillo

Corresponding Author: Ahmed Mahdi Abdulkadium ✉

ABSTRACT

A lot of students suffer expressing their desired enquiry about to a search engine (SE), and this, in turn, can lead to ambiguit and insufficient results. A poor expression requires expanding a previous user query and refining it by adding more vocabularies that make a query more understandable through the searching process. This research aims at adding vocabulary to an enquiry by embedding features related to each keyword, and representing a feature of each query keyword as graphs and node visualization based on graph convolution network (GCN). This is achieved following two approaches. The first is by mapping between vertices, adding a negative link, and training a graph after embedding. This can help check whether new information reaches for retrieving data from the predicted link. Another approach is based on adding link and node embedding that can create the shortest path to reaching a specific (target) node, . Particularly, poor data retrieval can lead to a new concept named graph expansion network (GEN). Query expansion (QE) techniques can obtain all documents related to expanding and refining query. On the other hand, such documents are represented as knowledge graphs for mapping and checking the similarity between the connection of a graph based on two authors who have similar interst in a particular field, or who collaborate in a research publications. This can create paths or edges between them as link embedding, thereby increasing the accuracy of document or paper retrieval based on user typing.

Keywords: Query Expansion; Graph Neural Networks (GNNs); FOAF; Link Predicted; Node Embedding.

RESUMEN

Muchos estudiantes sufren al expresar su consulta deseada sobre un motor de búsqueda (SE), y esto, a su vez, puede dar lugar a resultados ambiguos e insuficientes. Una expresión deficiente requiere ampliar una consulta previa del usuario y refinarla añadiendo más vocabulario que haga más comprensible la consulta a través del proceso de búsqueda. El objetivo de esta investigación es añadir vocabulario a una consulta incorporando características relacionadas con cada palabra clave, y representando una característica de cada palabra clave de la consulta como gráficos y visualización de nodos basada en redes de convolución de grafos (GCN). Esto se consigue siguiendo dos enfoques. El primero consiste en mapear los vértices, añadir un enlace negativo y entrenar un gráfico después de la incrustación. Esto puede ayudar a comprobar si llega nueva información para recuperar datos del enlace predicho. Otro enfoque se basa en añadir enlaces e incrustaciones de nodos que pueden crear el camino más corto para llegar a un nodo específico (objetivo). En particular, una recuperación de datos deficiente puede dar lugar a un nuevo concepto denominado red de expansión de grafos (GEN). Las técnicas de expansión de consultas (QE) pueden obtener todos los documentos relacionados con la consulta ampliada y refinada. Por otro lado, dichos documentos se representan como grafos de conocimiento para mapear y comprobar la similitud entre la conexión de un grafo basado en

dos autores que tienen un interés similar en un campo concreto, o que colaboran en una publicación de investigación. Esto puede crear caminos o aristas entre ellos como incrustación de enlaces, aumentando así la precisión de la recuperación de documentos o artículos basada en la escritura del usuario.

Palabras clave: Expansión de Consultas; Redes Neuronales Gráficas (GNNs); FOAF; Predicción de Enlaces; Incrustación de Nodos.

INTRODUCTION

The Web comprises enormous amounts of data, which are gathered, examined, and used by a huge number of users every day. Unstructured content such as Web pages, books, journals, and files make up a sizable portion of the data on the Internet. Ac-cordiglly, gathering the right information from such enormous data becomes diffi-cult and time-consuming. This is because trivial keyword-based information retriev-al systems rely heavily on data statistics. Such systems frequently encounter word mismatch issues brought on by a term's inescapable meaning and context fluctua-tions. Moreover, most information retrieval algorithms find documents by matching keywords. This method is undoubtedly ineffective for finding, for example, papers with similar meanings and different syntactic forms. Query expansion (QE) is one of the widely used strategies to address this restriction.[1]

Based on the above discussion, it becomes urgent to arrange such vast amounts of data and this, in turn, can lead to quickly processing information in a broad context while taking data semantics into account. Although semantic Web is widely used to store unstructured data in an orderly and organized manner, ontologies have signifi-cantly improved the performance of various information retrieval techniques. Hence, data can be retrieved by ontological information retrieval systems based on semantic comparison among user query and the indexed data.[2] In ontologies, data can be described as a semantic context. As such, ontology mapping has emerged as a key component in addressing the heterogeneity issues of semantically described data. As a result, alignments across ontologies must be applied during various run-time pro-cedures, most likely during the design time. A collection of linkages between the source and target ontologies are described by such alignments. Thus, the mappings can demonstrate how instance data from one ontology could be formulated in terms of another one.[3]

This research aims to enhance query and information retrieval. In comparison to previous literature, the present study adds three contributions:

- Predicting negative link (PNL): there are many nodes within a graph which have a poor connection due to the few number of relationships. We aim to map between those nodes and link them with nodes that have similar features and stronger connection so as to get the entire document related to our query.
- Heterogeneity Node Embedding: a graph can have heterogeneous attributes, whereby the node may have different types. In this case, it is a challenge to learn the embeddings that express the smallest differences of each node type.
- Author Relational Module: some nodes of an author graph are not widely connected due to fewer citations or event limitations and complexity.

Related Work

It can be challenging for many users to express their information demands in language. Sometimes a user may simply type one or two words, which do not accurately indicate the type of information required. This leads to a list of articles that are only marginally pertinent to the user's wants. Query expansion (QE) is a method for expanding the user's word list to make it more representative. Pseudo Relevance Feedback is a method that can be used to broaden a query. The findings of work that has been done to expand query using Pseudo Relevance Feedback on an information retrieval system (IRS) is discussed based on the Indonesian Wikipedia dataset, which consisted of roughly 450 thousand documents. The cosine similarity is used to calculate how similar the query and the list of tourism news documents whereas TF-IDF is used to determine how much weight should be given to each term. The quality of the new query has a significant impact on the choice of the increase in the baseline dataset, so datasets from an identical domain are encouraged. As for the use on a QE basis, the QE reference dataset should be domain-specific and filtered.[4]

Zhang[5] used embedding (i.e. term, user, and subject embedding) and improved individualized query reformulation by more efficiently exploiting contextual information and user preferences. They completed two main tasks. In the former, candidate questions are created by changing or adding one term from the initial search query, and contextual similarity between terms is calculated using term embedding that why enhanced with user customization. The word, user, and topic embedding in a graphical model were used to evaluate and grade (re-rank) the candidate queries created in the first task based on their consistency with the semantic meaning of users and their preferences. Experiments demonstrate that the suggested model outperforms state-of-the-art techniques by a wide margin.

https://doi.org/10.56294/dm2025443

Zhang et al.[6] used graph neural networks (GNNs) for text categorization tasks. In addition to having a more straightforward data structure than other types of graph data, trees can also provide rich hierarchical information known as (HINT) for text classification. They created the coding tree by reducing structural entropy after being inspired by the graph, so HINTwas offered. HINT, attempts to fully utilize the hierarchical information included in the text for the purpose of text categorization. Precisely, a dependency parsing graph was first created for each text, This was followed by develop a structural entropy minimization technique to decode the crucial data in the graph and turn each graph into the relevant coding tree. By upgrading the graphical representation of non-leaf nodes in the coding tree layer by layer, it is possible to derive a representation of the whole network based on the coding tree's hierarchical structure. Finally, they demonstrated how well hierarchical information classified texts. According to experimental findings, HINT performs better than cutting-edge techniques on well-known benchmarks while having an easy-to-understand structure with a few number of parameters.

Liu et al.[7] proposed a new graph-based model named DADGNN, for text classification. It successfully separated the essential processes of GNNs, namely representation transformation and propagation, in order to train a deep neural network. Additionally, they introduced an attention diffusion mechanism in a single layer to effectively collect non-direct-neighbor context information. The concept of node-level attention was proposed as a means to acquire more accurate representations at the document level. By employing the aforementioned procedures, the receptive field is expanded and the depth was augmented, and this leads to, mitigating the adverse effects associated with excessive smoothing. The theoretical analysis demonstrated that DADGNN has the ability to gather appropriate filters for adapting the dataset. Specifically, it was capable for retaining a greater amount of relevant high-frequency information from nodes, hence enhancing the model's expressiveness. Numerous experiments have demonstrated that their model outperforms other competing approaches. Significantly, the research did not only establish a robust foundational model for text classification, but also added a valuable contribution to the field of graph representation learning.

Blagec et al.[8] created the Intelligence task ontology and knowledge graph (ITO), which is a comprehensive, meticulously organized, and manually maintained database including information on AI tasks, benchmark outcomes, and performance indicators. The ITO framework encompasses a total of 685 560 edges, 1 100 classes dedicated to AI operations, and 1 995 features specifically designed for performance indicators. The primary goal of the international telecommunication union (ITU) was to facilitate the comprehensive analysis of global artificial intelligence (AI) initiatives and capacities. ITO was constructed using technologies that provide smooth automatic deduction, ongoing expert supervision of the fundamental ontological models, and effortless integration and enhancement with external data.

Jain et al.[9] stated that query expansion is widely recognised as the most crucial approach for attaining precise outcomes in the field of information retrieval. In order to overcome the limitations of the existing web system and use the advantages of query expansion, a unique framework was proposed for information retrieval. This framework is built upon fuzzy ontology, aiming to enhance the effectiveness of the retrieval process. The recommended approach utilises domain-specific information to facilitate the construction of ontologies. ConceptNet was utilised for constructing a fuzzy ontology, which served as a foundational structure for both domain-specific ontologies and a comprehensive global ontology. The identification of the most semantically related words for the given query is accomplished by utilising a constructed fuzzy ontology. Subsequently, the query was enlarged based on this information.

Information Retrieval Data

The Citeseer, Cora and Pumbed datasets are scientific digital libraries which consist of scientific publications that are classified into classes for the literature in computer and information sciences (Agents, AI, DB, IR, ML and HCI). Publishing papers can be represented as a network of nodes, and the relations of (cited by /citing) as links between them. Each publication is described by a [0/1] vector element indicating the absence/presence of the corresponding word from the dictionary of unique words.

Table 1. The citation network publication datasets

Dataset	Node	Edges	Classes	Features (word dictionary)
Citeseer	3,327	4 732	8	3,703
Cora	2,708	5,429	7	1,433
Pumped	19,717	44,338	3	500

Academic Relationship Approach

In figure 1, the interconnection of Author and Paper leads to the analysis of data to help render scientific research publication, thus making the academic relationship more cooperative. Authors could have many Friend of a Friend's relationship (FOAF) to cite each other and this could, expand the relationship between them to gather academic papers from different communities.

https://doi.org/10.56294/dm2025443

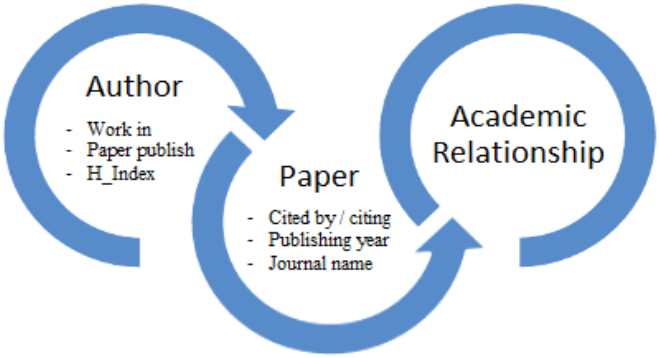

Figure 1. Academic Relationship Approach

METHOD

In this paper, the proposed methodology deals with the following tasks:

1. Extracting tokens (keywords) from a query and labeling them.
2. Estimating a user's desire from the extracted token related to a specific keyword and converting it to a graph.
3. Taking the behavior of expression and interpreting the behind meaning, through which the knowledge can be obtained about whom will cite each other.
4. Drawing the graph and relation between node features to create an obvious appearance, resulting in a graph called Graph Neural Network Author Relations (GNNAR), as shown in figure 2.

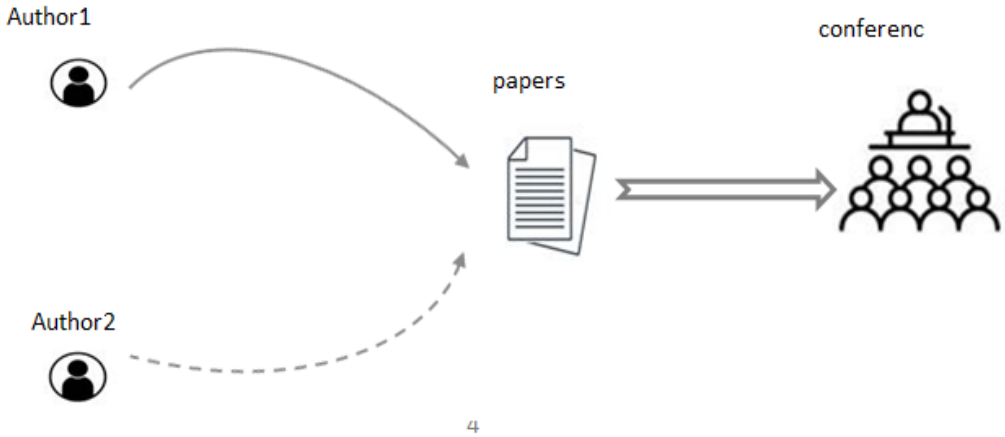

Figure 2. GNNAR Methodology

Graph Neural Network (GNN)

GNN and Hierarchical Graph Transformer Network

Users' lack of query expression typing and mismatching may retrieve text or documents far away beyond queries. This research handles an author search for a specific paper's or journal ISSN numbers. The retrieved results can include all details relating to a target author. The graph-based approach will help in context to learn representations of the authors in the graph, It can be used for a variety of tasks such as predicting which authors are likely to collaborate in the future, identifying clusters of authors with a similar interests, and predicting the impact of new publications within graph neural network (GNN).

Each author shares his/her common information such as name, scientific degree, organization, DOI, and previous papers publishing, that is represented in a graph and RDF format. To train a graph neural network on this type of information, the first step is to construct the graph by identifying co-authorship relationships between authors. Then, each node (author) is represented with a feature vector that captures relevant information about the author, such as the number of publications, the number of citations, and the FOAF relation for making a link predicting for the nearest author who shares same information (a same concept similarity).Thus a relationship will occur and curriculum learning to train the connecting nodes.

A general framework-knowledge-enhanced hierarchical graph transformer network (KHGT).[10] It includes the following steps:

- The KHGT framework explicitly achieves high-order relation learning in the knowledge-aware multi-behavior collaborative graph under the hierarchically structured graph transformer network.
- Two stages are used to jointly integrate user- and item-wise collaborative similarities under the multi-behavior modeling paradigm of KHGT.
- The graph-structured transformer module captures the type-specific user-item interactive patterns in a time-aware environment.
- The attentive fusion network encodes the cross-type behavior hierarchical dependencies and discriminates the type-specific contribution in forecasting the target behaviors.
- The Proposed KHGT framework was applied to three real-world datasets of movie, venue, and product recommendations.

Anchor-Aware Entropy Graph Neural Network Fused (IEA-GNN)

Graph neural networks have the ability to address the discussed constraints, making them an attractive option for various machine learning applications that involve graph data. A possiple application is the link prediction problem.[11]

The suggested anchor-aware graph neural network fused with information entropy (IEA-GNN) framework has three distinct modules namely, anchor point selection, node-to-anchor path calculation, and node embedding computation.

The module responsible for anchor point selection by employs the concept of information entropy to quantify the informational significance of nodes within a graph. Subsequently, anchor points are determined by using the obtained information entropy values. The graph contains dispersed anchors that exhibit extensive information exchange with neighboring nodes.

The module is responsible for determining the route between a node and anchor points based on a function to determine the distance of the path. This calculation is performed for each anchor point in a set, resulting in the retrieval of the node's relative position information and the remaining shared nodes linked to each anchor point. The IEA-GNN module employs a GNN encoder to integrate attributes, neighborhood structure, and global position information in order to derive the ultimate embedding representation of the node.[12]

The Motivation of GNN

The purpose of using GNN is to address the limits of traditional machine learning (ML) approaches. These constraints can be categorized as follows:

- One downside of standard machine learning algorithms is their inability to differentiate between an input graph G and any other graph that is isomorphic to G. This means such these algorithms will provide a same output for both G and any other graph that is obtained by re-labeling the nodes of G in a random way.
- Real-life graphs exhibit a high degree of irregularity. This means that it is common to come across several nodes with few or no connections, whereas a small number of nodes may has a significantly larger number of connections. Therefore, within our dataset, several nodes possess a dense network of connections, offering a substantial amount of valuable information that can be utilized for classification purposes. Conversely, the majority of nodes within the dataset possess only limited and scattered data, which represent a challenge when attempting to do accurate prediction tasks.
- Finally, the edges able to collect correlations between connected nodes. The inclusion of such correlations has the potential to enhance the forecasting precision of machine learning algorithms. To exemplify, a social network a social relationships between pairs of authors are chosen. Social relationships influence individual behaviors where an ML algorithm is more effective in deciding whether an author *"a"* is interested in publishing paper *"p"* if the authors who are close to *"a"* in the social graph were interested in citing *"p"*. The relational structure of a graph-based dataset constitutes an asset that cannot be ignored in the learning process.

ABERT (Author DistilBERT) Model

For classifiying, the main focus lies on he hidden state associated with the initial token [CLS]. This in somehow can capture the semantics of the entire sentence better than others. Thuerefore, this embedding can be used as input for a classifier that is built on its top.

A lighter version of BERT called Author DistilBERT is used in the present work. This model is similar to the original DistilBERT and still maintains about 97 % performance on various Natural Language Processing (NLP) tasks. Moreover, BERT's input is tokens in which it associates a tokenizer that preprocess text. Words are often split into sub-words, so special tokens called CLS are added to indicate the beginning of the sentence, whereas SEP is added to separate multiple sentences.Finally, PAD is added to make all sentences with the same number of tokens.

https://doi.org/10.56294/dm2025443

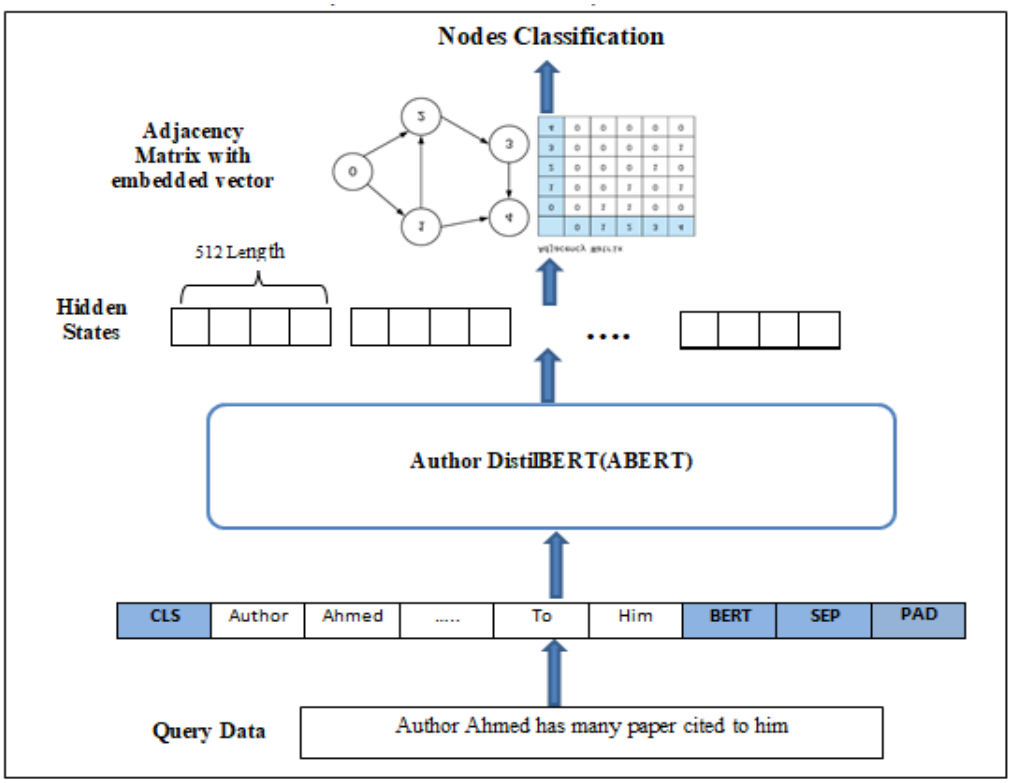

Figure 3. Author DistilBERT model

Friend-of-a-Friend Graph Prediction

Recommendation Systems (RSs) have the ability to establish a connection between users and products, enabling them to suggest items that may appear irrelevant but are actually of interest to users based on their behaviour. The performance of RSs is significantly impacted by the challenge of data sparsity. This is due to the limited quantity of information available to every user, which becomes increasingly apparent as the number of users and objects still continues to expand. In order to address the problem of data sparsity in recommendation systems, a unique approach to collaborative filtering prediction is introduced. This approach is referred to as multi-order nearest neighbour prediction (MNNP). The notion of multi-order neighbours is a further development of the friends-of-a-friend (FOAF) concept.[13]

Link prediction is a crucial and foundational area of study in sophisticated network research, which has been explored by numerous scientists through diverse investigations. Nevertheless, the majority of current link prediction techniques fail to take into account the directionality of network edges or fully utilize the information available from network nodes. The utilization of the topological similarity method in a directed network is proposed as a potential solution for addressing this particular challenge. Firstly, this study introduces enhancements to the Sorensen index in the context of directed networks, along with the proposal of other related variants. Additionally, the matrix representation of each fundamental index is denoted using matrix algebra. The topological nearest-neighbors similarity index is derived by applying the concept of the global leicht holme hewman (GLHN) similarity index to each fundamental index.[14]

Negative Link Predicted

This work proposes a Negative Link Predicted to Graph Neural Network (NLPGNN) by enhancing the previous approach, The proposed approach includes three key steps.

First, P1 is to check the closest neighbors who do not have a connection between vertices. This is to make a predicted link by taking the advantage of concept similarity without depending on a direct path or normal similarity for adding vertices.

Second, P2 is to check the farthest neighbors in a knowledge graph, Moreover, it is used to make a graph convolution network (GCN) one-by-one for nodes that have a high score of similarity to connect an author to another vertex within a graph.

Finally, P3 is to keep trying and perform convolution for all other nodes to add a link between unconnected nodes that have no matches or do not have features for link predicted and embedding nodes. This is to add links that have a strong relationship, so the added path is cycled and tracked for accurate data retrieval.

Negative Link Predicted Author Relation (NLPAR)

The prediction of relationship is performed according to the following interconnection:

- Supposing that an author named “Ahmed” has many details about his paper title, published journal, and area of interest.
- Since there are mutual relations, there is a chance to predict links with other authors who have the same interest or maybe have the same concept in a different area that can make both work on a similar concept. This justifies the link connection to retrieve more data related to the author “Ahmed”.
- Checking if there are any previous collaboration or FOAF relations can help in predict a link connection to nodes that have mutual features.

Figure 4 shows that there is a similarity between Author 1 and Author 2 graphs in a feature that both work , for example, at the University of Babylon. This can provide a reason to predict a link and make a further interaction such as sharing interest, working on a joint paper and publishing an article together. After connecting the two graphs, there was a need for node embedding techniques to reduce the dimensionality of the network and capture important features. Node embedding can improve the performance of downstream machine learning tasks, such as node classification. It can also be visualized and interpreted, allowing to gain insights into the structure and properties of the graphs.

Based on similarity, the first and second-order similarity measurements can be defined when node A1 and node A2 in the same level of neighbors , so a similarity coefficient R=0 will be set. Otherwise it will be set to R=1. In other words, if two authors have collaborated on a paper, their FOAF profiles can be linked by defining work property. Similarly, if two authors have a shared interest in a topic, the future interest property can be used to link them and their profiles together.

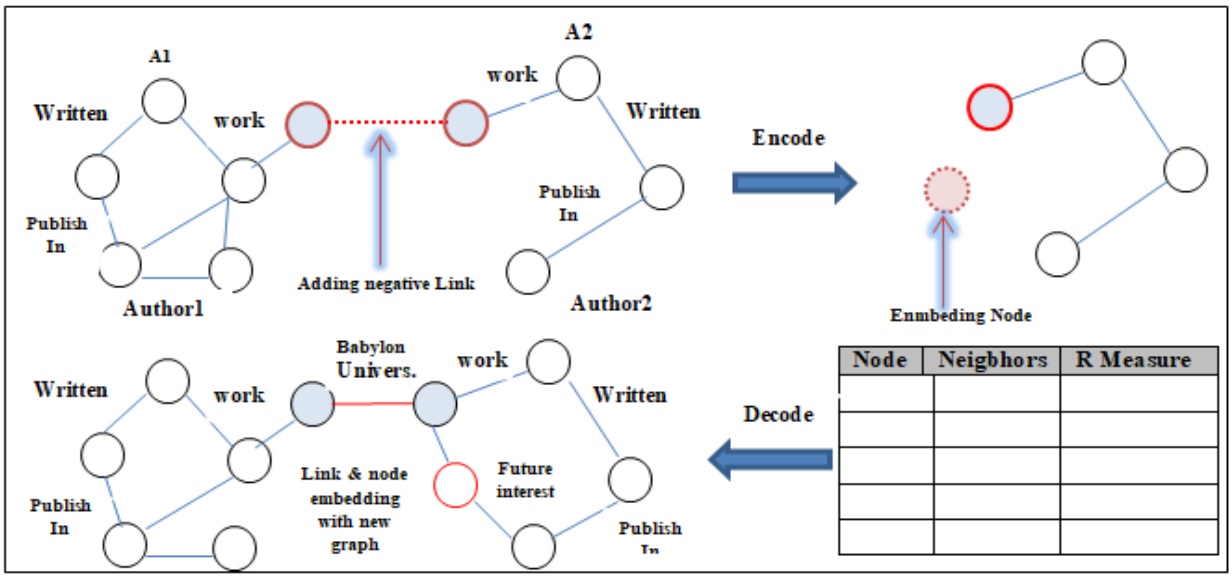

Figure 4. Node and Link Predicted Based on FOAF Author Relationships

By creating a network of authors using FOAF, it is possible to identify potential collaborators or experts in a particular field as presented in table 2. This is useful in academic research and corporation, where finding the right collaborators can lead to more impactful and innovative papers publishing.

Table 2. A1 and A2 collaboration features

Source	Target	Features	R Coefficient
A1	A2	work	1
A1	A2	publishing	0
A2	A1	written	0
A2	A1	collaborating	1
A1	A2	future interest	1

https://doi.org/10.56294/dm2025443

Node and Graph Embedding

Node embedding is a technique used to represent each node in a graph as a low-dimensional vector. It is used to capture the structural and semantic properties of a node and its relations with other nodes in the graph.

Graph embedding, on the other hand, is a technique to represent the entire graph as a low-dimensional vector. It is used to capture the global structural properties of the graph, such as its topology and connectivity.

Embedding Nodes in an Author Graph

To embed nodes in an author graph, five steps should be followed. This includes:

1. Constructing the author graph: the author graph can be constructed by representing each author as a node, whereas the co-authorship relations between authors can be represented as edges.
2. Choosing an embedding technique: a node embedding technique is chosen based on the size of the author graph, the computational resources available, and the quality of embedding required.
3. Generating node embedding: the node embedding technique is chosen to generate embedding for each author node in the graph.
4. Evaluating the node embedding: the quality of the nodes embedding is evaluated by measuring their performance on downstream tasks such as authorship or clustering.
5. Fine-tune the embedding: the node embedding is fine-tuned by using techniques such as transfer learning to improve their quality on specific tasks.

Similarity Concept Embedding

Node and link embedding can be used to capture the similarity between nodes and links in a graph. In this context, similarity refers to the degree of similarity or closeness between nodes or links in terms of their attributes or structural properties. A process of creating node and link embedding with a similarity concept after graph construction can be perforemed by two steps:

1. Measuring similarity: once the node and link embedding are generated, the similarity measure between nodes or links can be made based on their embedding. Various similarity metrics such as cosine similarity, Euclidean distance, or Pearson correlation coefficient are used to measure the similarity.
2. Using the similarity for downstream tasks: the similarity between nodes or links is used for various downstream tasks such as node classification, link prediction, or clustering to predict the class label of a new node based on its similarity with the existing nodes in the graph.[15]

The meta-path aside cast and a novel method to learn the low-dimensional vector space perserve both structural and semantics information in heterogeneous information network (HIN). Specifically, take advantage of GNN to conduct the structural information of HIN and train the model by the task-guided objective function (node classification loss in this paper). To handle the challenges of HIN mentioned above, a dedicated Type-aware Attention Layer was designed instead of the convolutional layer in the conventional GNN. For each type-aware attention layer, a transformation operation that projects vertices from different entity spaces to the same low-dimensional target space is defined for the interaction between heterogeneous nodes (C1), and the attention strategies focusing on different types of edges are applied for the aggregation of neighboring vertices with different semantics (C2).

The Neighborhood Aggregation is preserve the semantic of different types of relationship between nodes, the |R| attention scoring functions are utilized to match different relation patterns, i.e., $F^{(l+1,\,m)} = \{f_r^{(l+1,\,m)} \mid r \in R\}$. For a vertex j, an attention coefficient is computed for each link edge $e = (i, j, r) \in E_j$ in the form as in equation 1:[16]

$$o_e^{(l+1,m)} = \sigma\left(\mathrm{f}_{\mathrm{r}}^{(\mathrm{l+1,\,m})}\left(\mathbf{h}_{\varphi(j),j}^{(l+1,,m)}, \mathbf{h}_{\varphi(j),i}^{(l+1,,m)}\right)\right) \quad (1)$$

Where σ is an activation function implemented by LeakyReLU. The attention coefficient o_e indicates the importance of edge e to the target vertex j. $\varphi(j)$ refer to the hidden state of nodes.[16]

A graph convolution networks model MS-GCNs with the motif-structure information integration is presented to improve the expression ability of the model by integrating the following steps:

- The MS-GCNs model combines the node's first order neighborhood of edge information and motif-structure information to improve the convolution operation.
- Same label rates of several real datasets are calculated and analyze the degree of assimilation of nodes in the same motif from the perspective of data, which indicates the validity of higher-order information of the motif.
- Node classification experiments are carried out on different types of real networks. By comparing the effects according to relevant indicators, their baseline models are perforemed correspondingly.[17]

https://doi.org/10.56294/dm2025443

Information Retrieval Algorithm

Input: Cora, Citeseer and Pubmed Dataset
Output: Features with High accuracy ($feature_{best}$)
Step 1: Set i attribute, j Record

Dataset Preprocessing and Cleaning Step

Step 2: For each i in a dataset
Step 3: For each j in a dataset
Step 4: Call Stop Word Removal Function
Step 5: Call Stemming Function
Step 6: Call Tokenization Function
Step 7: End For
Step 8: End For

Feature Extraction and Selection Step

Step 10: For each i in a dataset
Step 11: For each j in a dataset
Step 12: Call TF-IDF
Step 13: Call Correlation Function
Step 14: Call ADistilBERT Function
Step 15: End For
Step 16: End For

Represent selected Feature as a Graph Step

Step 17: For each i in a dataset
Step 18: For each j in a dataset
Step 19: Call GNN Function (Convert Data to Graph)
Step 20: Call PNL Function (Add negative Link)
Step 21: SPARQL endpoint for Query a Graph after enrichment
Step 22: End For
Step 23: End For

Graph Represented as a data store step

Step 24: For each node in a Graph
Step 25: For each vertex in a Graph
Step 26: Extract Data from a Graph
Step 27: Call Index Storage
Step 28: End For
Step 29: End For

Retrieving and Ranking Step

Step 30: For each i in a dataset
Step 31: For each j in a dataset
Step 32: return more data from an embedding graph
Step 33: Call Ranking
Step 34: Call Relevant Feedback
Step 35: End For
Step 36: End For

Evaluation and Comparison Step

Step 37: For each i in a dataset
Step 38: For each j in a dataset
Step 39: Call F-measure
Step 40: Call Precision
Step 41: Call Recall
Step 42: $feature_{best}$ call Best Result measure
Step 42: End For
Step 43: End For
Step 44: return $feature_{best}$

https://doi.org/10.56294/dm2025443

Table 3. The result and comparison of Fusion embedding for citeseer graph prediction				
Model	**5 %**	**10 %**	**30 %**	**50 %**
GAT - SBERT (study in[18])	0,61 ± 0,01	0,65 ± 0,01	0,71 ± 0,01	0,73 ± 0,01
GraphSAGE - TF-IDF[18]	0,66 ± 0,01	0,67 ± 0,01	0,73 ± 0,01	0,78 ± 0,01
GraphSAGE - Sent2Vec[18]	0,64 ± 0,01	0,66 ± 0,01	0,73 ± 0,01	0,77 ± 0,01
AsK-LS (study in[19])	The result applied to the entire graph is 0,748±0,013			
LS-SVM (study in[19])	The result applied to the entire graph is 0,724±0,011			
NLPGNN-AuthorDistilBERT	0,61 ± 0,01	0,68 ± 0,01	0,71 ± 0,01	0,79 ± 0,01

Table 4. The result and comparison of Fusion embedding for cora graph prediction				
Model	**5 %**	**10 %**	**30 %**	**50 %**
GAT - SBERT (study in[18])	0,63 ± 0,01	0,65 ± 0,01	0,71 ± 0,01	0,77 ± 0,01
GraphSAGE - TF-IDF[18]	0,62 ± 0,01	0,63 ± 0,01	0,68 ± 0,01	0,71 ± 0,01
GraphSAGE - Sent2Vec[18]	0,64 ± 0,01	0,66 ± 0,01	0,77 ± 0,01	0,78 ± 0,01
AsK-LS (study in[19])	The result applied to the entire graph is 0,748±0,013			
NLPGNN-AuthorDistilBERT	0,61 ± 0,01	0,68 ± 0,01	0,71 ± 0,01	0,79 ± 0,01

The results of table 2 and table 3 show the graph prediction results for the training edges (5 %, 10 %, 30 %, 50 %) of metric lies between 0 and 1 , where a higher value means better results, It is clear from the findings that combining SBERT with GCN was less effective than using it alone according to the obtaining results above. However, GraphSAGE with TF-IDF models yields results comparable to plain text embedding. The overall graph structure increases distortion in fine document classification embedding. Effective use of AuthorDistilBERT properties requires careful selection of local neighbors. At the same time sufficient results were obtained (highlighted in bold) regardless the number of train edges, unlike other approaches that decrease significantly when the percentage decreases.

CONCLUSIONS

The utilization of the FOAF relation and the notion of graph neural network (GNN) has the potential to enhance the efficiency of querying outcomes and information gathering and retrieval. The utilization of the FOAF relation, a well-established method for representing social relationships, enabled the extraction of significant information from social networks. This information was then employed to improve the relevance and precision of search results. The FOAF relation facilitated the identification of connections and similarities among distinct individuals, enabling the generation of more focused and customized search outcomes.

The use of the GNN paradigm can proficiently record intricate connections and interdependencies among diverse nodes within a graph. GNN possessed the capability of analyzing both the local and global properties of nodes. This ability allowed to acquire knowledge and discern patterns within the data. Consequently, GNN can be leveraged to enhance the optimization of querying outcomes.

In general, the integration of the FOAF relation and the GNN concept presented a robust methodology for enhancing the effectiveness and productivity of information acquisition and retrieval. By employing these methodologies, its possible to effectively obtain more pertinent and precise data from social networks, yielding advantages across diverse sectors such as social media, e-commerce, and medical treatment. In the future , the performance of a link prediction model can be expaneded further to retrieve more information from the dataset. Moreover, otology can be built and enriched with the prediction model to enhance the querying outcomes.

BIBLIOGRAPHIC REFERENCES

1. ALMarwi, H.; Ghurab, M.; Al-Baltah, I. A hybrid semantic query expansion approach for Arabic information retrieval. J Big Data 7, 39 (2020). https://doi.org/10.1186/s40537-020-00310-z.

2. Alvarado MAG. Gentrification and Community Development: An analysis of the main lines of research. Gentrification 2023;1:2–2. https://doi.org/10.62486/gen20232.

3. B. Wang; L. Cheng; J. Sheng; Z. Hou; Y. Chang, “Graph convolutional networks fusing motif - structure information,” Sci. Rep., pp. 1–12, 2022, doi: 10.1038/s41598-022-13277-z.

4. Castillo VS. Gentrification as a field of study in the last decade: a bibliometric analysis in Scopus. Gentrification 2023;1:5–5. https://doi.org/10.62486/gen20235.

5. Dinkar AK, Haque MA, Choudhary AK. Enhancing IoT Data Analysis with Machine Learning: A Comprehensive Overview. LatIA 2024;2:9–9. https://doi.org/10.62486/latia20249.

6. F. Chen; G. Yin; Y. Dong; G. Li; W. Zhang, "KHGCN: Knowledge-Enhanced Recommendation with Hierarchical Graph Capsule Network," Entropy, vol. 25, no. 4, p. 697, Apr. 2023, doi: 10.3390/e25040697.

7. Fahad, M.. Ontology-based Mediation with Quality Criteria. In International Conference on Business Intelligence, pp. 74-88, (2023, July). Cham: Springer Nature Switzerland.

8. Genes APC. Theoretical foundations and methodological guidelines for the appropriation of ICT in the pedagogical practice of teachers. Multidisciplinar (Montevideo) 2024;2:104–104. https://doi.org/10.62486/agmu2024104.

9. Gonzalez-Argote J, Maldonado EJ. Indicators of scientific production on Health Policy. Management (Montevideo) 2024;2:107–107. https://doi.org/10.62486/agma2024107.

10. Guo, F.; Zhou, W.; Wang, Z.; Ju, C.; Ji, S.; Lu, Q. A link prediction method based on topological nearest-neighbors similarity in directed networks. Journal of Computational Science, 69, 102002 (2023).

11. H. Bu; J. Xia; Q. Wu, "A Dyna H. Bu, J. Xia, Q. Wu, and L. Chen, "A Dynamic Heterogeneous Information Network Embedding Method Based on Meta-Path and Improved Rotate Model," Applied Sciences, vol. 12, no. 21, p. 10898, Oct. 2022, doi: 10.3390/app122110898.

12. H. Y. Husni; Yeni Kustiyahningsih; Fika Hastarita Rachman; Eka Mala Sari Rochman, "Query expansion using pseudo relevance feedback based on the bahasa version of the wikipedia dataset," AIP Conf. Proc., vol. 2679, no. 1, 2023.

13. He, Mingzhen; He, Fan; Shi, Lei; Huang, Xiaolin ; Suykens, Johan. (2022). Learning with Asymmetric Kernels: Least Squares and Feature Interpretation.

14. Hernández-Lugo M de la C. Artificial Intelligence as a tool for analysis in Social Sciences: methods and applications. LatIA 2024;2:11–11. https://doi.org/10.62486/latia202411.

15. Hong; Huiting; et al. "An attention-based graph neural network for heterogeneous structural learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 04. 2020.

16. Jain; Shivani; K. R. Seeja; Rajni Jindal. "A fuzzy ontology framework in information retrieval using semantic query expansion." International Journal of Information Management Data Insights 1.1 (2021): 100009.

17. K. Blagec; A. Barbosa-Silva; S. O. ; M. Samwald, "a curated, ontology-based, large- scale knowledge graph of artificial intelligence tasks and benchmarks." 2022.

18. León MP. The impact of gentrification policies on urban development. Gentrification 2023;1:4–4. https://doi.org/10.62486/gen20234.

19. M. N. Asim, M.; Wasim, M.; Usman, G.; Khan, N. Mahmood; W. Mahmood, "The Use of Ontology in Retrieval : A Study on Textual , Multilingual , and Multimedia Retrieval," IEEE Access, vol. 7, pp. 21662–21686, 2019, doi: 10.1109/ACCESS.2019.2897849.

20. Madariaga FJD. Pedagogical model for the integration of ICTs into teaching practices in official educational institutions in rural Monteria. Multidisciplinar (Montevideo) 2024;2:105. https://doi.org/10.62486/agmu2024105

21. Makarov I; Makarov M; Kiselev D. Fusion of text and graph information for machine learning problems on networks. PeerJ Comput Sci. (2021,May)11;7:e526. doi: 10.7717/peerj-cs.526.

22. Muthusundari M, Velpoorani A, Kusuma SV, L T, Rohini O k. Optical character recognition system using artificial intelligence. LatIA 2024;2:98–98. https://doi.org/10.62486/latia202498.

23. Navarro WS, Duque NEA, Ramirez FMB, Chaparro AMT. Strategic Analysis from a consulting context for the Super Kinder School Institution. Management (Montevideo) 2024;2:30–30. https://doi.org/10.62486/agma202430.

24. P. Zhang; J. Chen; C. Che; L. Zhang; B. Jin; Y. Zhu, "IEA-GNN : Anchor-aware graph neural network fused with information entropy for node classification and link prediction," Inf. Sci. (Ny)., vol. 634, no. November 2022, pp. 665–676, 2023, doi: 10.1016/j.ins.2023.03.022.

25. Pérez GAJ, Cruz JMH de la. Applications of Artificial Intelligence in Contemporary Sociology. LatIA 2024;2:12–12. https://doi.org/10.62486/latia202412.

26. Pirela CV, Plata AO, Hernandez GL. Strategic thinking as a potential factor in the growth of companies in the dairy sector. Management (Montevideo) 2024;2:40–40. https://doi.org/10.62486/agma202440.

27. Sonal D, Mishra K, Haque A, Uddin F. A Practical Approach to Increase Crop Production Using Wireless Sensor Technology. LatIA 2024;2:10–10. https://doi.org/10.62486/latia202410.

28. Sun; Xiaohan; Li Zhang. "Multi-order nearest neighbor prediction for recommendation systems." Digital Signal Processing, 127 (2022). https://doi.org/10.1016/j.dsp.2022.103540.

29. Vargas OLT, Agredo IAR. Active packaging technology: cassava starch/orange essential oil for antimicrobial food packaging. Multidisciplinar (Montevideo) 2024;2:102–102. https://doi.org/10.62486/agmu2024102.

30. Velazquez MDCR, Chririnos AAN, Brito AV. Decision-making styles developed by commercial enterprises in the municipality of Barrancas. Management (Montevideo) 2024;2:35–35. https://doi.org/10.62486/agma202435.

31. X. F. Yonghao Liu; Renchu Guan; Fausto Giunchiglia; Yanchun Liang, "Deep Attention Diffusion Graph Neural Networks for Text Classificatio." pp. 8142–8152, 2021.

32. X. Liu; X. Li; G. Fiumara; P. De Meo, "Link prediction approach combined graph neural network with capsule network," Expert Syst. Appl., vol. 212, no. August 2021, p. 118737, 2023, doi: 10.1016/j.eswa.2022.118737.

33. Zhang, C.; Zhu, H.; Peng, X.Q.; Wu, J.; Xu, K. (2021). Hierarchical Information Matters: Text Classification via Tree Based Graph Neural Network. International Conference on Computational Linguistics.

34. Zhang, X. "Improving personalised query reformulation with embeddings. Journal of Information Science, 48(4), 503-523(2022).

FINANCING

Currently, there are no available financing sources designated for this project. This absence of financial support underscores the need for strategic planning to identify potential funding avenues that could facilitate the successful implementation and advancement of the initiative.

CONFLICT OF INTEREST

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Ahmed Mahdi Abdulkadium, Asaad Sabah Hadi.
Investigation: Ahmed Mahdi Abdulkadium, Asaad Sabah Hadi.
Methodology: Ahmed Mahdi Abdulkadium, Asaad Sabah Hadi.
Writing - original draft: Ahmed Mahdi Abdulkadium, Asaad Sabah Hadi.
Writing - review and editing: Ahmed Mahdi Abdulkadium, Asaad Sabah Hadi.