International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 5, July-August 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Automatic Query Expansion Using Word Embedding Based on Fuzzy Graph Connectivity Measures

Tarun Goyal[1], Ms. Shalini Bhadola[2], Ms. Kirti Bhatia[3]

[1]M Tech Student, [2]Assistant Professor, [3]HOD,
[1, 2, 3]Computer Science & Engineering, Sat Kabir Institute of Technology and Management
Bahadurgarh (HR) Affiliated by Maharshi Dayanand University (Rohtak), Haryana, India

ABSTRACT

The aim of information retrieval systems is to retrieve relevant information according to the query provided. The queries are often vague and uncertain. Thus, to improve the system, we propose an Automatic Query Expansion technique, to expand the query by adding new terms to the user‟s initial query so as to minimize query mismatch and thereby improving retrieval performance. Most of the existing techniques for expanding queries do not take into account the degree of semantic relationship among words. In this paper, the query is expanded by exploring terms which are semantically similar to the initial query terms as well as considering the degree of relationship, that is, "fuzzy membership" between them. The terms which seemed most relevant are used in expanded query and improve the information retrieval process. The experiments conducted on the queries set show that the proposed Automatic query expansion approach gave a higher precision, recall, and F- measure then non-fuzzy edge weights.

KEYWORDS: *Information Retrieval; Fuzzy Log; WordNet; Centrality*

How to cite this paper: Tarun Goyal | Ms. Shalini Bhadola | Ms. Kirti Bhatia "Automatic Query Expansion Using Word Embedding Based on Fuzzy Graph Connectivity Measures" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-5, August 2021, pp.1429-1442, URL: www.ijtsrd.com/papers/ijtsrd45074.pdf

IJTSRD45074

Copyright © 2021 by author (s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (http://creativecommons.org/licenses/by/4.0)

1. INTRODUCTION

Information retrieval (IR) is defined as searching for particular information in one or more documents or searching for a document itself [1]. Documents can be any multimedia, text or anything that resides in the concerned database. The performance of an IR system is often ineffective due to imprecise queries. To retrieve high-quality results, information retrieval systems generally require exact keywords [2]. This is a problem since the user is often unsure about the nature of the content they need. Furthermore, several other challenges are like the constantly evolving nature of data, high volatility, intense growth leads to lower precision. Thus, to improve the information retrieval process, query expansion techniques are used [9].

Query Expansion (QE), is defined as the set of methodologies, techniques or algorithms for refining search queries, for retrieving of more relevant results [4].QE is performed by expanding the initial query through the addition of new related words and phrases to the original query. There are two types of QE techniques. Manual Query Expansion, the user manually reformulates by adding more terms to the original query. Automatic Query Expansion, the system automatically reformulates by adding more terms to the original query without any user intervention.

Recently, Zhang, Wang, Si, and Gao used four corpora as data sources (one industry and three academic corpora) and presented a Two Stage Feature Selection framework (TFS) for query expansion known as the Supervised Query Expansion (SQE)[8]. The first stage is an Adaptive Expansion Decision (AED), which predicts whether a query is suitable for SQE or not. For queries which is unsuitable, SQE is skipped with no term features being extracted at all, so that the computation time is reduced. For suitable queries, the second stage conducts Cost Constrained

Feature Selection (CCFS), which chooses a subset of effective yet inexpensive features for supervised learning. A drawback of corpus specific QE is that they fail to establish a relationship between a word in the corpus and those which are used in different communities, e.g., “senior citizen” and “elderly”

Maryamah Maryamah, Agus Zainal Arifin proposed a query expansion method based on BabelNet search and Word Embedding (BabelNet Embedding). Query expansion method focuses on developing queries based on semantic relationships on queries to understand the context of the query [3]. User queries were developed by finding synonyms which, measure similarity using WordNet, Word Embedding on all articles of Wikipedia, and BableNet Embedding on articles Wikipedia Online. Successful Query expansion retrieves relevance data from the Arabic documents which are searched. A drawback is that it does not take degree of centrality into consideration while creating query expansion.

There exist several graph-based query expansion methods which use semantic relationships between query terms to get more relevant terms [11][12][14]. Jain et al used semantic relationships and various graph centrality measures for query expansion [10] where words or phrases are represented as nodes and the semantic relationship between them represented as edges connecting the graph. In this paper, for semantic relations, a lexicon i.e. WordNet has been used. Once the weights for the edges are computed, then various graph centrality measures were applied on nodes of graph. Finally nodes which were having score above average were selected and incorporated with the original query. But this method had a drawback that it considered the edge weights as crisp values (0 or 1). Therefore, not capturing the degree of relationship between words.

Balaneshin-kordan, Saeid, and Alexander Kotov explored sequential query expansion using concept graph where nodes are words or phrases and the edges represent the semantic relationship between them which achieved improvement in retrieval accuracy over other methods [11].

But in concept graph, finding candidate concepts which are related to the query is a laborious problem. Jain, Amita, Sonakshi Vij, and Oscar Castillo used Hindi WordNet relations between words and fuzzy graph connectivity measures for query expansion [12]. But here the edge weight was defined by experts and was not automatically generated. The edge weight was defined by the type of relationship that exists between words like hypernymy, hyponymy etc. Moreover, this method was based on Hindi language and did not provide a comparison against other state of the art methods.

In this paper, we proposed a novel fuzzy graph-based method to automatically expand a query and deal with the limitations occurring in other methods. Semantic relations like hypernymy, hyponymy, meronymy, holonymy, entailment and troponymy between query terms are extracted from WorldNet and base graph is constructed. Next a subgraph from base graph is extracted. After subgraph construction, the edge weights between words is not considered as crisp but used as a fuzzy edge weight that describes a degree of relationship that exists between words. We used different word embedding models for automatic generation of fuzzy logic value of edges. Next, various fuzzy graph centrality measures are evaluated on the sub-graph for each node. A top ten words list is prepared for each measure. The nodes that are present in at least three of the measures are chosen since they are likely to have higher relevance to the original query terms. This helps in query expansion as the nodes that rank high in centrality measures have more degree of relevance attached to them and thus used to extract more relevant terms.

The paper is organized as follows, Section 2 discusses the related work; Section 3 examines the preliminaries followed by Section 4 that deliberates the proposed method, Section 5 illustrating the method using an example, Section 6 describes the Experimentation, result and discussion and Section 7 is the conclusion and future work.

2. Related Work

There are various approaches to Query Expansion. They are separated into two categories namely local and global where local methods use relevance feedback while global methods use WordNet, automatic thesaurus etc. [13].

Recently word embedding techniques are widely used for Query Expansion. An embedding framework based on distributed neural language model word2vec [5]. Based on the embedding framework, it extracted terms similar to a query provided by the user using the K-nearest neighbor approach. The experimental study was performed on standard TREC ad-hoc data. It showed considerable improved result as compared to the classic term overlapping-based retrieval approach. It is also noticed that word2vec framework based query expansion perform more or less the same with and without any feedback information.

Some other works using word embedding techniques are based on locally-trained word embedding (such as word2vec and GloVe) for ad hoc IR . They also used local embeddings that capture the nuances of topicspecific languages and are better than global embeddings. They also suggested that embeddings be learned on topically-constrained corpora, instead of large topically-unconstrained corpora[6]. In a query-specific manner, their experimental results suggest towards adopting local embeddings instead of global embedding because of formers potentially superior representation.

Similarly, a QE technique based on word embeddings that uses Word2Vec"s Continuous Bag-of-Words(CBOW) approach[7]. CBOW represents terms in a vector space based on their co-occurrence intext windows. It also presents a technique for integrating the terms selected using word embeddings with an effective pseudo relevance feedback method.

Query expansion using graph-based methods work by determining the significance of each vertex. The graph is made by using various semantic relations that exist between words using WordNet. The importance of each node is determined by using graph centrality measures [14]-[15]. Semantic relations considered are associative(cause-effect), equivalence (synonymy), hierarchical (e.g. hypernymy, hyponymy) etc. [17]. Query Expansion approaches were categorized as extensional, intentional or collaborative by Grootjen [17]. Extensional materializes information in terms of documents. Intentional use semantics of keywords which are primarily ontology-based. The third category uses user"s behaviors like mining query logs as an additional approach with previous ones.

In graph and network analysis, identifying the significant nodes of the network is a major task[18]. Centrality measures namely Key Player Problem, Degree Centrality, Betweenness, PageRank and HITS are used to measure how important is a node in the network. Among these, two of the measure namely PageRank [19] and HITS [20] are extensively studied and have been extremely influential for information retrieval. PageRank defines how important a page is by counting links to it by other web pages. HITS have two components for rating web pages namely authority and hub values. A page has good authority if it is pointed by several pages and it has a good hub if it points to several others and thus, they both have a mutually supplementing relationship. The difference between the two is that PageRank works on sub-graphs of relevant pages while HITS uses the entire graph.

Sinha and Mihalcea [14] used in-degree, closeness, betweenness, PageRank of vertices to identify the significant nodes of the graph. Geodesic distance between two nodes was used by Free-man [21] to ascertain the importance of nodes relative to each other. According to him, how "in-between" a node is among other nodes determines its betweenness. Key player Problem (KPP) was proposed by Borgatti which used relative closeness of a vertex with all other vertices to determine its importance computed by "reciprocal of total shortest distance from a given node to all other nodes" [18].

Jain, Amita, Sonakshi Vij and Oscar Castillo used fuzzy weights in edges in their graph-based query expansion method by assigning specific values to semantic relationships that exists between words [12].

In this paper, we computed the fuzzy relationship value between words of fuzzy graph by using different word embeddings models and using the graph centrality measures on the fuzzy graph to include more relevant terms and expand user"s query.

3. Preliminaries

This section describes the fuzzy logic system, fuzzy graph, WordNet and the word embedding models used in our method to automatically generate the fuzzy weight for all the edges.

A. Fuzzy Logic system

Zadeh (1965) described fuzzy logic using a mathematical framework, principle behind it was that "everything is a matter of degree" [22]. Thus, a fuzzy set is a matter of degree and not of denial or affirmation mapping each element in it to (0, 1) in which (0, 1). This includes words since two or more words need not have only a binary relation between them but be corresponded with a degree. Thus, to deduce a relation between two pairs of words, fuzzy relations can be used.

B. Fuzzy Graph

A fuzzy graph is a symmetric binary fuzzy relation on a fuzzy subset. A Fuzzy Graph [Rosenfeld et al. 1975; Sunitha 2001; Zadeh et al. 1975; Mathew and Sunitha 2009; Bhattacharya 1987] defined as $G = (\sigma, \mu)$ is a pair of functions $\sigma: S \rightarrow [0,1]$ and $\mu: S \times S \rightarrow [0,1]$, where for all x, y in S we have $\mu(x, y) \leq \sigma(x) \cap \sigma(y)$. Any relation $R \subseteq S \times S$ on a set S can be represented as a graph with node set S and arc set R. Similarly, any fuzzy relation $\mu: S \times S \rightarrow [0,1]$ can be regarded as defining a fuzzy graph, where the arc $(x, y) \in S \times S$ has weight $\mu(x, y) \in [0,1]$ [10].

C. WordNet

WordNet is a large lexical database of English. Nouns, adverbs, adjectives and verbs are grouped into various sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are always interlinked by means of lexical relations and conceptual-semantic [4]. WordNet apparently resembles a thesaurus, in thesaurus it groups words together on the basis of their meanings. However, there are some major differences both of them.

Firstly, WordNet interlinks based on specific senses of words, not just on word forms - strings of letters. As a result, those words which are in close proximity with one another in the network are always semantically disambiguated. Secondly, WordNet labels all the semantic relations among words, whereas the groupings of words in a thesaurus do not follow any kind of explicit pattern other than meaning is similar.

D. Word Embedding Models

Word embedding are the methods which map words into real valued vectors in a vector space . Word embedding is the collective name given to a set of feature learning techniques and language modeling in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers [23] . It increases the ability of networks to learn from the data by reducing the number of dimensions of text data. They are essential in text processing and machine learning tasks where it helps to derive relations between different textual data. Word embedding models represent different techniques to generate vectors from text data and deduce relations in it. We used the following pre-trained Word embedding models:-

1. Word2Vec

To use textual data in algorithms, some technique was needed to change text into numerical representation. Mikolov, Tomas, et al.[24] proposed the CBOW (Continuous Bag of Words) and Skip-gram models to convert words to vectors. This model was open sourced by Google and is now known as Word2Vec. This model uses neural network of two layers to map text into vectors. Vector representation allows various operations like subtraction, addition etc. to determine relationships among the words.

Similar word embeddings are generated for words in similar contexts. For example, the sentences "Alex is a generous person " and "Alex is a kind person" has similar word embeddings for words „generous‟ and „kind‟ due to similar context.

The model used in this paper is trained on 3 million words by Google using the google news data. This model has 300 dimensions [25]. Negative sampling and skip gram were used to construct this model.

2. GloVe

GloVe or Global Vectors for words representation was made by team at Stanford using word to word co-occurrence [26]. It has various models varying in dimensions and tokens. It is an unsupervised algorithm that combines the advantage of local context window method and global matrix factorization.

GloVe is based on the idea that words appearing next to each other have more impact than considering individual words since group of words have more information than a single word. For example, bank can be a financial institute or land alongside a body of water. But if we say river bank the meaning is much clear. The model used here was trained on dataset of Wikipedia and had 300 dimensional vectors.

3. Fast Text:

Fast Text model was proposed by Joulin, Armand, et al.[27]. It uses neural nets to generate word embeddings. In this model, words are represented using bag of characters n-grams. It is developed by Facebook and has good performance since it uses character level representations.

This model first takes the average of n-gram embeddings to obtain a sentence vector, and then uses multinomial logistic regression for classification and finally softmax to fasten the computation process. This model allows both supervised and unsupervised learning algorithms to convert words into vectors. Facebook has pretrained this model for several languages. The model used here has 1 million vectors trained on Wikipedia dataset.

Proposed work of Automatic Query Expansion using Fuzzy Graph Centrality Measures

In this section, we describe the steps involved in the proposed method to perform query expansion. Terms used in the algorithm are defined below:

- G: graph constructed around the original query terms, i.e., Fuzzy WordNet graph
- G‟: Fuzzy sub-graph
- V‟: nodes of Fuzzy sub-graph
- Tj : query term under consideration
- Tk : linked query term to Tj
- L‟: maximum distance allowed between query terms
- N : Maximum length of original query

A. Construction of Fuzzy Graph

First the user enters a query having i terms where 1≤i≤N.Then pre-processing of query is performed. All the stop words are removed and terms are Part of Speech (POS) tagged to identify adjective, adverb, verb, and noun. Only these POS tagged words are considered for the original query. Using WordNet, WordNet graph G is constructed by exploring the semantic relations. Subgraph G‟ is constructed by depth first search performing for every on WordNet graph G. Then, for every Tk in G where ! = j, if there is a path, ..., Tk of length <= L‟, the intermediate nodes and edges are added to the subgraph G‟.

IJTSRD International Journal of Trend in Scientific Research and Development ISSN: 2456-6470

B. Automatic assignment of fuzzy weights to edges

After the construction of subgraph G‟, fuzzy weight for each edge is calculated automatically that quantifies the degree of relationship between the two vertices forming the edge. This is achieved by converting each vertex to a vector and calculating the similarity score of the vectors. This score can be calculated using different techniques like cosine similarity, Euclidean distance, Jaccard distance, word mover‟s distance etc. [28]. Cosine similarity is the most widely used membership function and is used in the word embedding models namely Word2Vec, GloVe and FastText. Cosine similarity metric models a text document as a vector of terms. By this model, the similarity between two documents can be derived by calculating cosine value between two documents‟ term vectors [39]. Implementation of this metric can be applied to any two texts (sentence, paragraph, or whole document). In search engine case, similarity value between user query and documents are sorted from the highest one to the lowest one. The higher similarity score between document‟s term vector and query‟s term vector means more relevancy between document and query. Cosine similarity for similarity measurement between document and user query should accommodate to the word‟s meaning.

Cosine Similarity: It is computed by multiplying cosine angles of the two vectors. [29]

$$\cos\alpha = \frac{A \; X \; B}{|A| \; X \; |B|} = \frac{\sum_{i=1}^{n} A_i \; X \; B_i}{\sqrt{\sum_{i=1}^{n} ()^2} \; x \; \sqrt{\sum_{i=1}^{n} ()^2}} \qquad (1)$$

Where Ai and Bi are the components of vector A and B respectively.

To generate vectors and calculate the similarity score we have used three-word embedding models namely Word2Vec, GloVe, and FastText. Comparison of the different models is shown in result section (Section 6).

C. Centrality measures to select expansion terms

Centrality Measures are heavily used in wide applications in network analysis. In our method we use centrality measures to identify and compare importance of nodes which are relative to each other to select the most important nodes and thus the most relevant terms.

In this section, various centrality measures are being defined for a fuzzy graph.

Fuzzy Degree Centrality: It considers the degree that a vertex has which is defined as the number of connections it has to other vertices. For a fuzzy graph, the degree for a vertex v ($degv$ (v)) is [30]

$$degf\ (v) = \sum u{\neq}v\ re\ (u,\ v)\ belongs\ to\ t5e\ set\ of\ edges\ E \qquad (2)$$

For a vertex v, Fuzzy degree centrality ($Cf\ (v)$) is its degree standardized by the maximum degree

$$C_f\ (v) = \frac{deg\ f\ (v)}{|v|-1} \qquad (3)$$

Fuzzy Eigenvector Centrality: This centrality has the idea that “connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes” [31]. It has two variants namely Fuzzy PageRank and Fuzzy HITS. These measures have been proposed for a fuzzy graph in [30].

Fuzzy PageRank for vertex ( ) is calculated as

$$v_a = \frac{(1-d)}{|V|} + d \sum_{(v_a\ ,v_b\ )\in E} \frac{\mu\, ba}{\sum_{(v\ b\ ,\ v_c)\in E} \mu\, bc} (v_b\ ) \qquad (4)$$

where μba is the weight an edge connecting two vertices namely (va) and (vb), E is the set of edges and d is the damping factor where d ∈ [0, 1]. In this article, we took d = 0.85 as advocated by Luca and Carlos [31]. The initial fuzzy PageRank score for every vertex is taken as 1 The second variant, fuzzy HITS is similar to fuzzy PageRank but makes the distinction between hubs and authorities.

For fuzzy graphs, hubs □f (v)and authorities af (v)have been proposed by Jain, Amita, and D. K. Lobiyal [24]

as

$$□f\ ^{(}v^{)} = {}^{\sum}(u,\ v)\in E\ \mu uv\ af\ ^{(}u^{)} \qquad (5)$$

$$af\ (v) = \sum_{(u,v)\in E} \mu uv\ □f\ (u) \qquad (6)$$

where μuv is the edge strength between two vertices, u and v. In our method, we take it as a single measure HITS by adding the authority and hub values.

Fuzzy Closeness Centrality: It is based on the principle that a vertex is considered important by its relative closeness to all the other vertices.

Fuzzy Closeness centrality ( ) for a fuzzy graph has been defined [30] as

$$u_n = \frac{\sum_{u_0\ \in V, u_0\ \neq u_n} \frac{1}{Min_{all\ paths\ u_0\ to\ u_n}[\delta(u_0,u_n)]}}{|v|-1} \qquad (7)$$

($u0$,) is the length of path P: ($u0$, $u1$,….un) that exists between two vertices ($u0$, un) and is computed as

$$(u_0, u_n) = \sum^{n}_{i=1} \frac{1}{\mu_{(i-1)(i)}} \tag{8}$$

If a node is disconnected, its closeness centrality is given by $\frac{1}{|V|}$

Fuzzy Betweenness Centrality: This centrality is based on the idea of shortest paths. For a vertex v, it is calculated as

$$fuzzy\ betweenness_f\ (v) = \sum_{s,t\,\in V, s\neq v\neq t} \frac{|d_v(x,y)|}{|d(x,y)|} \tag{10}$$

Where $|dv$ (x, y) | represents the number of shortest paths from vertex x to vertex y, |d (x, y) | represents the number of all the paths that pass through the vertex v.

In a fuzzy graph, for a vertex v (fuzzy $betweennesscentf\ (v)$) it is given as a fraction of fuzzy betweenness(v) to the maximum pair of vertices that do not have v [24].

$$fuzzy\ betweennesscentf\ \ (v) = \frac{betweenness\ \ f\,(v)}{(|V|-1)(|V|-2)} \tag{11}$$

4. Query Expansion

The technique proposed improves the process of information retrieval by improving the original query. In this method, the query is expanded by appending more terms relevant to the original query. After constructing the entire subgraph G‟, graph centrality measures are calculated for every vertex. Then each measure is sorted in descending order and top ten words are taken from each measure. Then, the vertices V‟ that doesn‟t exist in the original query and exists as top ten in at least three or more centrality measures are selected. Finally, the selected terms are chosen to expand the original query. The proposed algorithm is as following:

Algorithm 1: Algorithm for Fuzzy Query Expansion:

//Considering the initial query entered by user having terms I where 1≤i≤n .and value of maximum distance „L‟ considered as L=3

Step 1: Query is entered by the user having the terms 1≤i≤n.

Step 2: Remove Stop words and Tag POS to identify adjective, adverb, verb, and noun. Only these POS tagged words are considered for the original query.

Step 3: Using WordNet, the two semantic relations of terms are explored to get more relevant terms and fuzzy WordNet Graph G is constructed.

Step 4: To Create Sub Graph from WordNet Graph G Initialize the SubGraph G‟ to NULL

For j=1 to n(Looping through the query terms), Repeat

1. For every perform Depth First Search (DFS) of the WordNetGraph (G) exploring semantic relations for T .
2. For every Tk where k!=j is between j+1 to n, if there is path Tj, ..., Tk of length <= L‟ add all the intermediate nodes and edges of the is path Tj, ..., Tk to the WordNet Sub Graph G‟.

Step 5: In G‟, assign every edge a weight using word embedding models on the nodes.

Step 6: Calculate the **Graph Centrality Measures**

For each nodes of the Subgraph G‟, calculate the below Fuzzy Centrality Measures for the graph.

- Fuzzy Degree Centrality
- Fuzzy Page Rank
- Fuzzy Betweenness
- Fuzzy Closeness
- Fuzzy HITS

Step 7: Sort each fuzzy centrality measure and take the Top 10 nodes from each measures.

Step 8: Select the nodes V‟ from G‟ such that

1. Nodes V‟ should not be equal to any Query Terms which existed in the original query.
2. Nodes V‟ exists as Top 10 in the three or more centrality Measures.

Step 9: Add the above-selected terms to the Original Query for the query expansion.

5. Explanation through an Example

The proposed technique is explained using an example query. Suppose user has entered query: **"creative work in art and history"**. Here Stop Words „in‟, „and‟ are removed. POS tagged to each terms of the query:

„Creative‟ – Adjective, „work‟ – Noun, „art‟ - Noun, „history‟ – Noun. The final terms selected are „creative‟,

„work‟, „art‟, „history‟. Using WordNet, the two semantic relations of the terms are explored to get more relevant terms and a graph is constructed around all the terms. The Fuzzy WordNet graph is given in Figure 1 shows a part of fuzzy WordNet Graph for query terms (Creative and Art)

Next, a fuzzy subgraph is extracted from Fuzzy WordNet graph by selecting only those nodes where the distance between any two linked nodes is less than or equal to a maximum distance „L‟. Next, all the edge weights are assigned by looking at edge‟s nodes and finding out the degree of fuzzy relationship between them using the Word2Vec word embedding model .We have explained the example using Word2Vec word embedding model, similarly other word embedding models (GloVe,FastText) are used for assigning the fuzzy edge values. A comparison of all word embedding models is shown in Table II. The fuzzy subgraph is shown in Figure 2. Edge weight is taken as unity if the relation is not defined between the two words in word embedding model, represented by "?" in Figure 2.

After constructing the fuzzy subgraph, fuzzy graph centrality measures are calculated and for each centrality measure Top 10 words were found out and result obtained are shown in Table I. The words that occur in at least three or more measures were selected to obtain the final expanded query. The words which were occurring in at least three or more measures are highlighted in the Table I.

From the words obtained in Table I, we see that terms *product, creation, book, movie, publication, activity, written record* excluding the original query terms are present in at least three or more measures. Thus, the final query is "*creative work art history product creation book movie publication activity written record*".

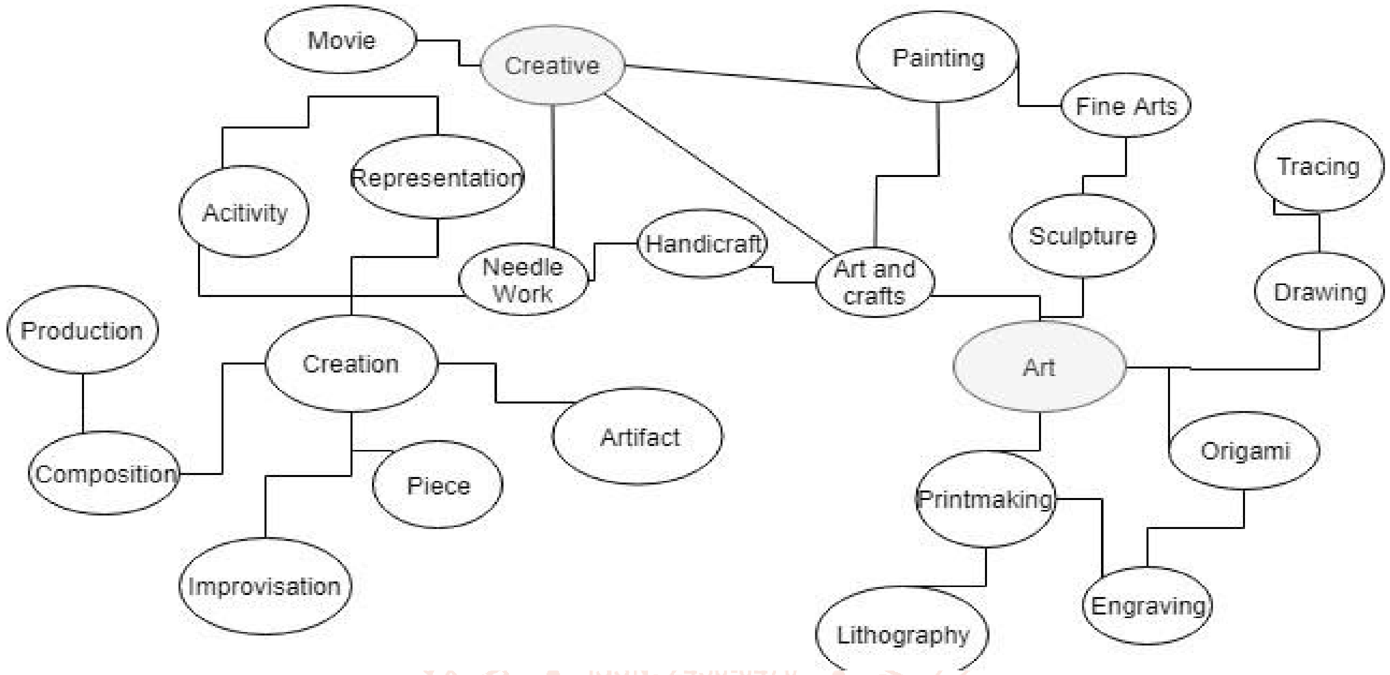

Fig. 1: Part of the Fuzzy WordNet graph around the query terms (Creative and Art)

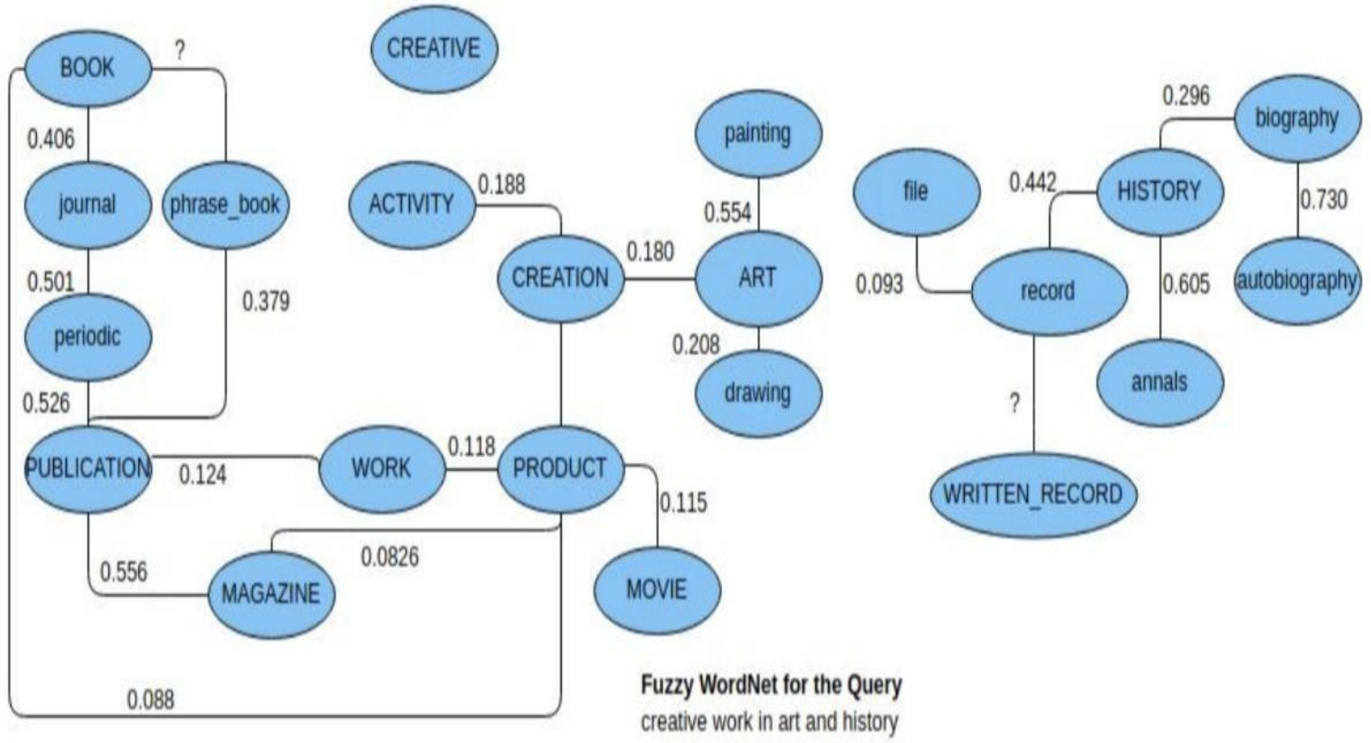

Fig. 2: Fuzzy Subgraph constructed for the query

Table I. Top 10 words in descending order of their value of fuzzy centrality measure

Fuzzy Degree Centrality	Fuzzy Betweenness	Fuzzy PageRank	Fuzzy Closeness	Fuzzy HITS
Activity (0.2268370607028754)	Creation (0.39768165806504463)	Activity (0.09426675405500479)	Creation (1.0514663964324533)	Written Record (1.7717315686760078)
Book (0.11182108626198083)	Product (0.39571557303186694)	Book (0.05419606375461807)	Product (1.0460735777596928)	Working Papers (0.10570855479920042)
Movie (0.06389776357827476)	Activity (0.31524944703858443)	Movie (0.031101163361487185)	Work (1.0326228336534522)	History (0.05875651697458283)
Creation (0.06070287539936102)	Work (0.23984189399524863)	Written record (0.03079878863915959)	Inspiration (1.0209787209958276)	Evidence (0.02063332923870081)
Written Record (0.05750798722044728)	Book (0.16543786352093062)	Work (0.02286601849445781)	Newspaper (1.0106193340179628)	Memorabilia (0.017670365347016017)
Work (0.05750798722044728)	Art (0.1414352420742197)	Art (0.019726866534460007)	Classic (1.0076748109549492)	File (0.011091495227915315)
Product (0.04472843450479233)	Publication (0.11692061931678545)	Creating by mental acts (0.018340932125813496)	Activity (0.9969985902040367)	Stub (0.005107890044318016)
Art (0.04472843450479233)	Movie (0.09805849102973703)	Creation (0.018019251592849794)	Turn (0.9937735016715603)	Medical History (0.0036543204529104513)
Publication (0.04153354632587859)	Handicraft (0.046018677807815184)	Creating from raw materials (0.0164188117426234 2)	Puncture (0.989475387152774)	Family History (0.0036543204529104426)
Handicraft (0.03514376996805112)	Magazine (0.0369869746866552)	Publication (0.01622771772785263)	Solo (0.9892784064422213)	Autobiography (0.0007703669391399912)

6. Experimental Result and Analysis

The proposed query expansion approach was examined on the standard TREC dataset [31]. TREC (Text REtrieval Conference) is a yearly forum/conference, which aims at the evaluation of large-scale Information Retrieval systems and approaches. TREC gives a set of text datasets, called test collections, against which any information retrieval system can be evaluated. The TREC dataset is a dataset for question classification consisting of open-domain, fact-based questions divided into broad semantic categories. It has both a six-class (TREC-6) and a fifty-class (TREC-50) version[31]. Each of them have 5,452 training examples and 500 test examples, but TREC-50 has finer-grained labels. Using the NLTK library in python, the query was stemmed and all the stop words were removed. The fuzzy WordNet graph was created around the query words and a suitable fuzzy subgraph was extracted. Using different word embeddings namely Word2Vec, FastText, and GloVe similarity between words was calculated and edge weights were assigned. Then the various fuzzy graph centrality measures were calculated to pick out the most relevant terms. The performance of our proposed work was evaluated on three parameters: Precision, Recall, F Score.

In Information Retrieval System, Precision (P) is defined as the number of true positives Tp over the number of true positives plus the number of false positives ( ) [30].

$$P = \frac{Tp}{Tp + Fp} \quad (11)$$

In Information Retrieval System, **Recall** (R) is defined as the number of true positives ( ) over the number of true positives (T) plus the number of false negatives (Fn) [30].

$$R = \frac{Tp}{Tp + Fn} \quad (12)$$

In Information Retrieval System, **F-Score** (F) is the harmonic mean (average) of the precision and recall [30].

$$F = 2 * \frac{P * R}{P + R} \quad (13)$$

After expanding the input query, the terrier was used to get precision and recall for different methods. Table II shows the Precision, Recall, and F-Measure calculated for the graph based IR methods where Non-Fuzzy edge weight represents the method where edge weights are considered as crisp, and the rest three methods have fuzzy edge weights calculated using different word embedding models namely Word2Vec, FastText and GloVe. We used three sets of random samples of sizes 10, 25 and 50 into consideration. From the table, it can be seen that the graph-based IR based on fuzzy edge weights outperformed the crisp edge weight (Non- Fuzzy Edge Weight). Moreover, method based on edge weight given by GloVe model had the best results. Figures 3-5 shows the comparison of the results.

Next, Three Baseline methods namely OkapiBM25 [32], Weighted Word Pair (WWP) [32], and Kullback-Leibler Divergence Based Query Expansion (KLDBQE) [33] were used for comparison. Table III shows the Precision, Recall, and F-Measure calculated for the methods. Comparison of the baseline methods with our method where edge weights were calculated using GloVe model showed that our method had lower precision, comparable F score, but much better recall. Figures 6-8 shows the comparison of the results.

From the graphs, we infer that the fuzzy methods gave higher recall and lower precision as compared to KLDBQE, OkapiBM25, and WWP. In F Score, fuzzy methods outperformed for lower sample space but as the number of samples increase the existing methods like OkapiBM25, WPP and KLDBQE start giving higher F Score.

Table II. Comparison of Non-Fuzzy edge weights and fuzzy edge weights by different models for graph connectivity based IR

Methods	10 Retrieved documents			25 Retrieved documents			50 Retrieved documents		
	Precision	Recall	F score	Precision	Recall	F score	Precision	Recall	F score
Non- Fuzzy Edge Weight	0.1191	0.2452	0.1603	0.0974	0.3215	0.1495	0.0857	0.4325	0.1412
Word2Vec Fuzzy Edge Weight	0.1468	0.2214	0.1765	0.1040	0.4214	0.1667	0.0868	0.5237	0.1489
FastText Fuzzy Edge Weight	0.1404	0.2335	0.1754	0.1017	0.43347	0.1648	0.0874	0.5025	0.1489
GloVe Fuzzy Edge Weight	0.1404	0.2022	0.1657	0.2035	0.4283	0.1657	0.0878	0.5016	0.1494

Table III. Comparison of baseline query expansion methods and fuzzy graph-based IR using GloVe model

Methods	10 Retrieved documents			25 Retrieved documents			50 Retrieved documents		
	Precision	Recall	Fscore	Precision	Recall	F score	Precision	Recall	F score
OkapiBM25	0.2276	0.1176	0.1551	0.2197	0.1967	0.2076	0.1934	0.2976	0.2344
WWP	0.2445	0.1211	0.162	0.2256	0.2167	0.2211	0.2114	0.3011	0.2484
KLDBQE	0.2423	0.1203	0.1608	0.2234	0.2145	0.2189	0.2099	0.3001	0.247
GloVe Fuzzy Edge Weight	0.1404	0.2022	0.1657	0.2035	0.4283	0.1657	0.0878	0.5016	0.1494

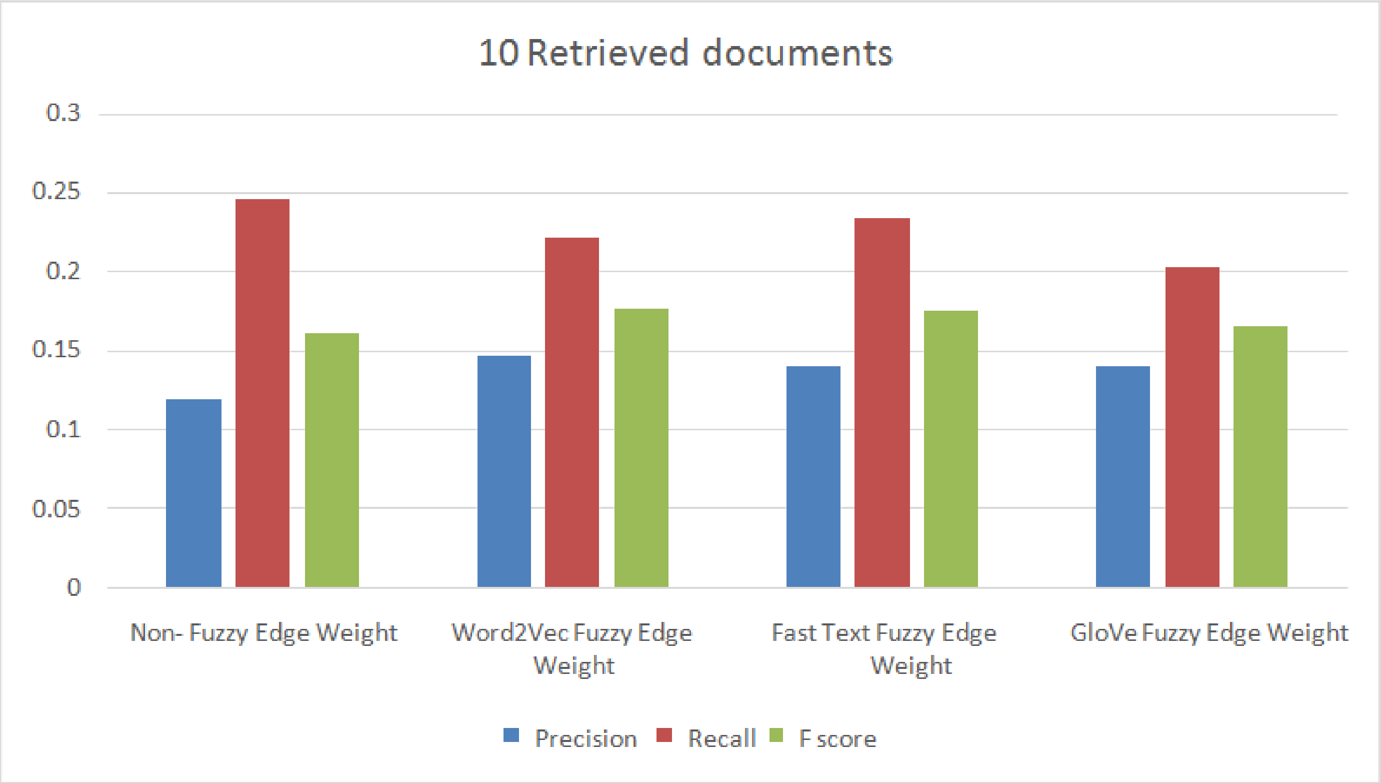

Fig. 3: Precision, Recall, F score comparison for graph based IR methods on 10 retrieved documents

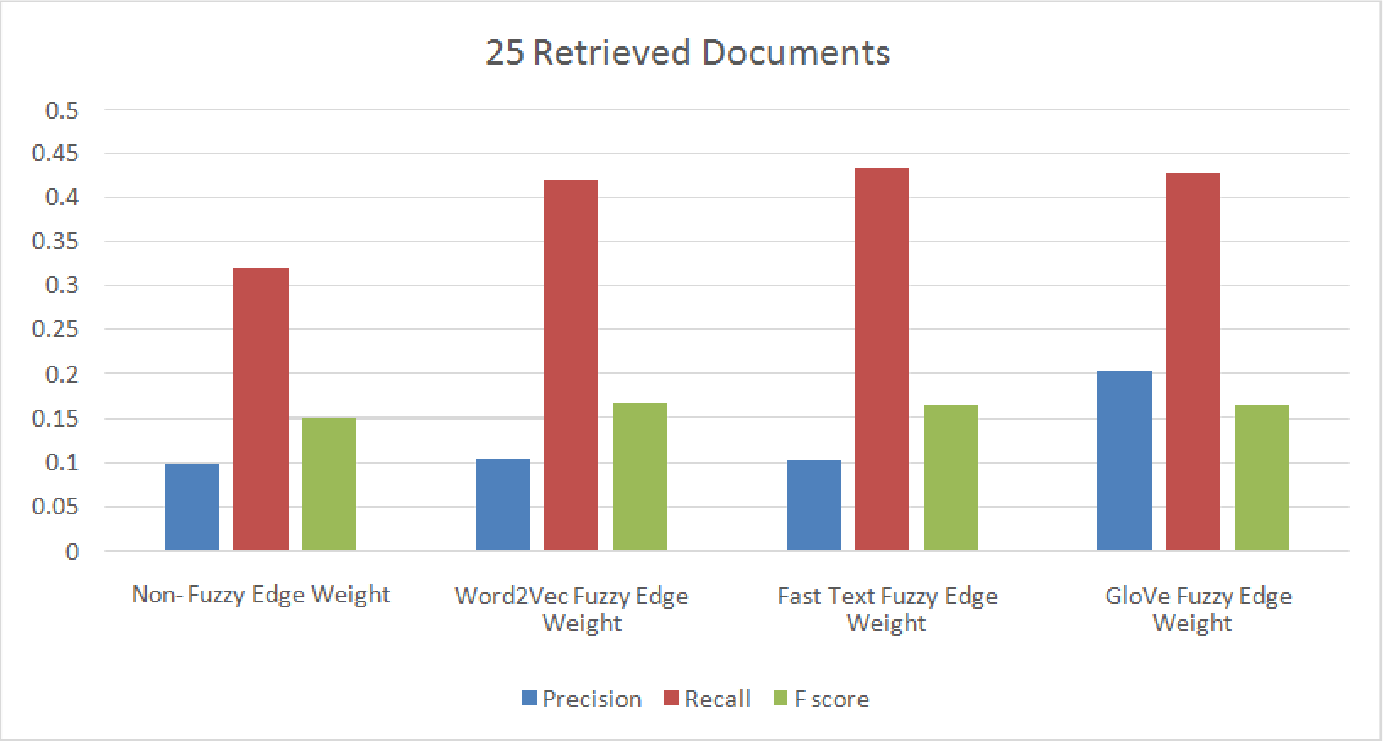

Fig. 4: Precision, Recall, F score comparison for graph based IR methods on 25 retrieved documents

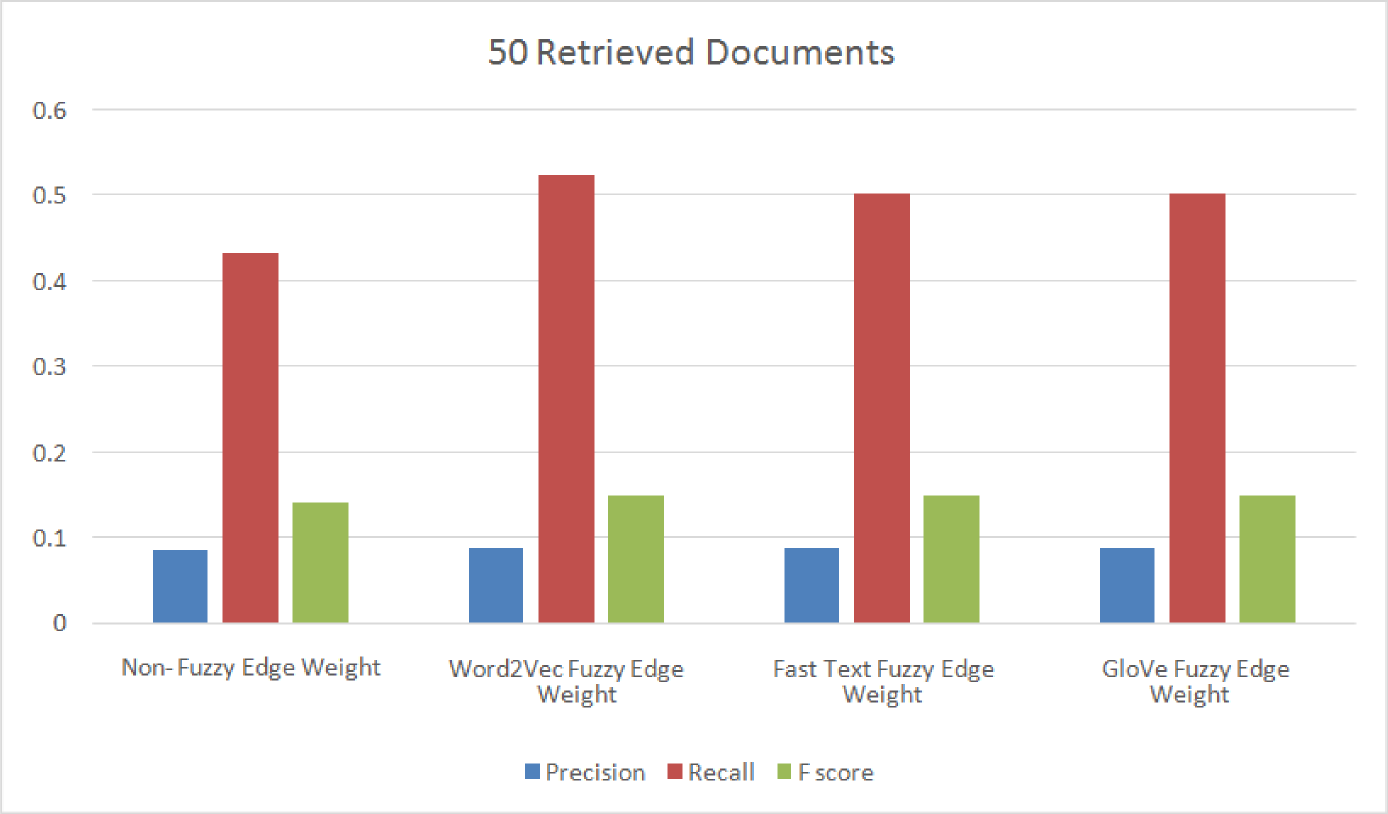

Fig. 5: Precision, Recall, F score comparison for graph based IR methods on 50 retrieved documents

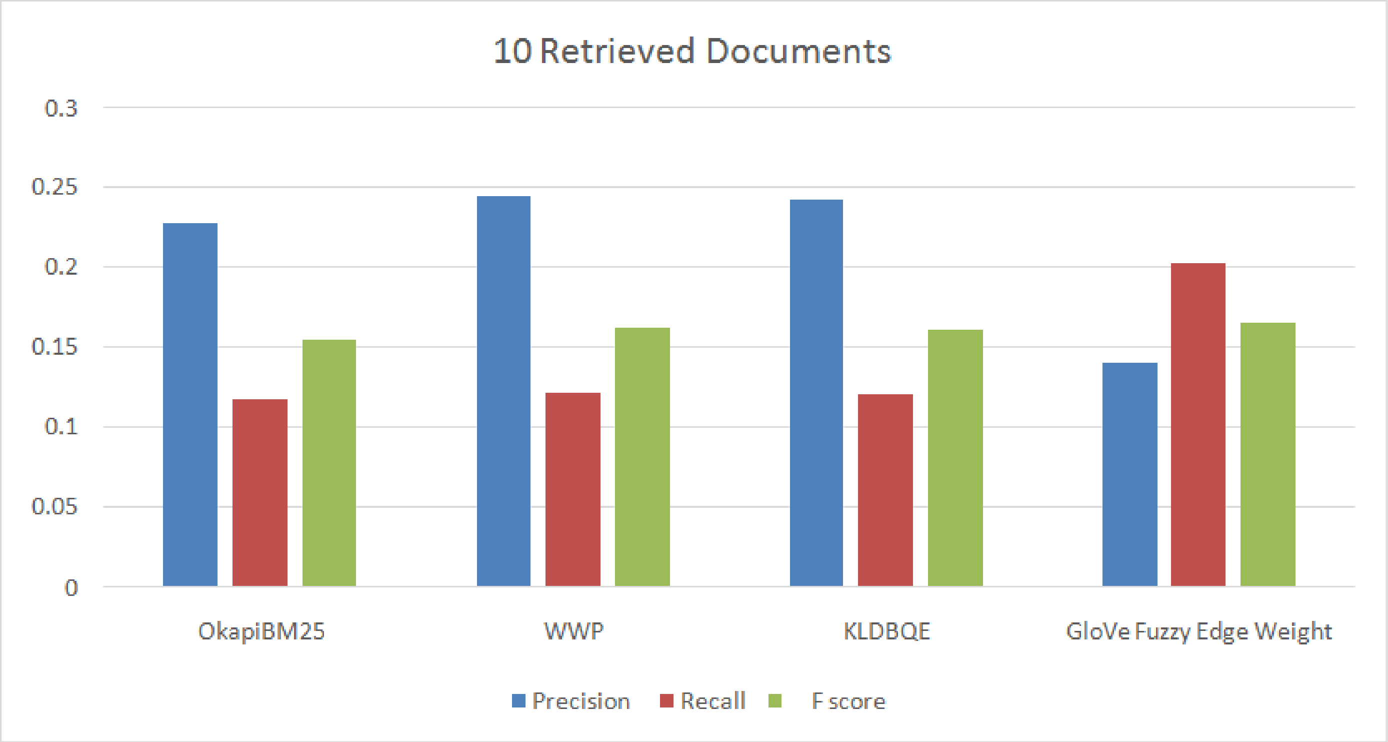

Fig. 6: Precision, Recall, F score comparison of baseline methods with GloVe method on 10 retrieved documents

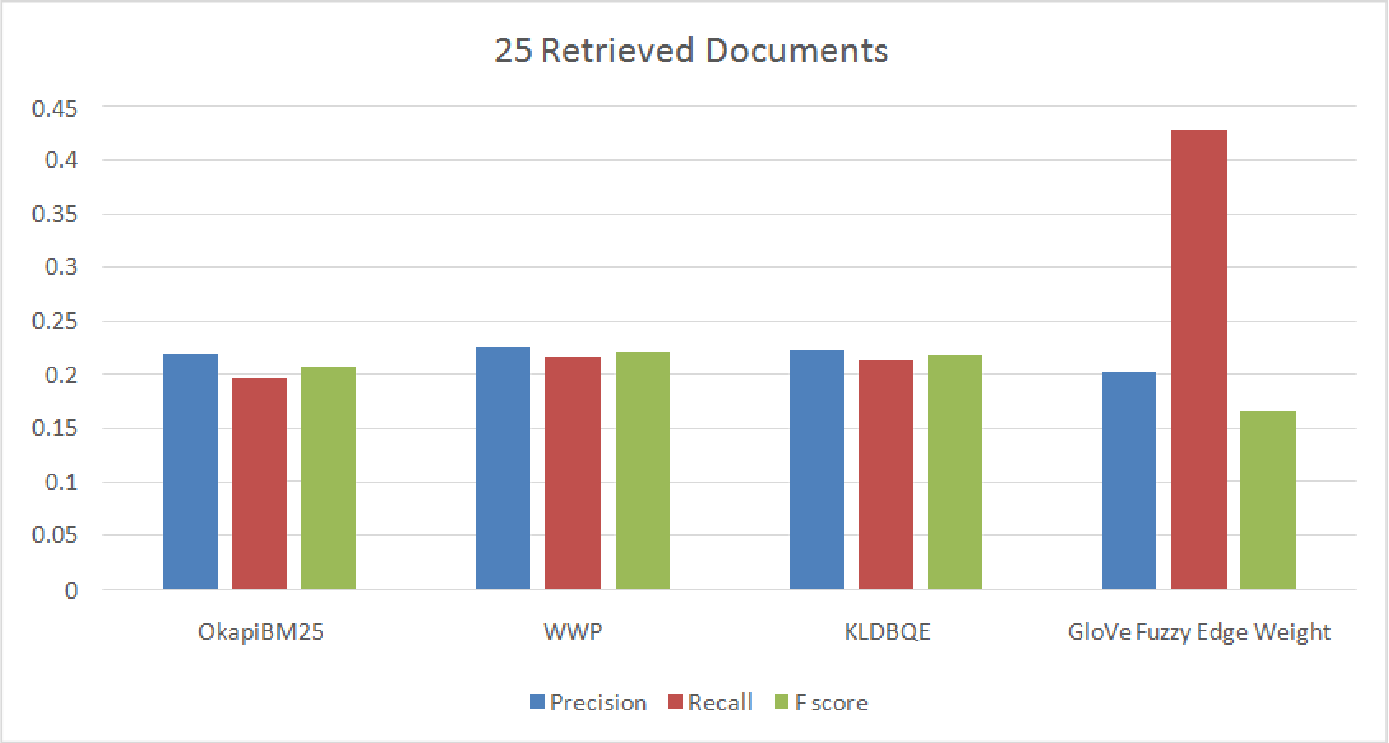

Fig. 7: Precision, Recall, F score comparison of baseline methods with GloVe method on 25 retrieved documents

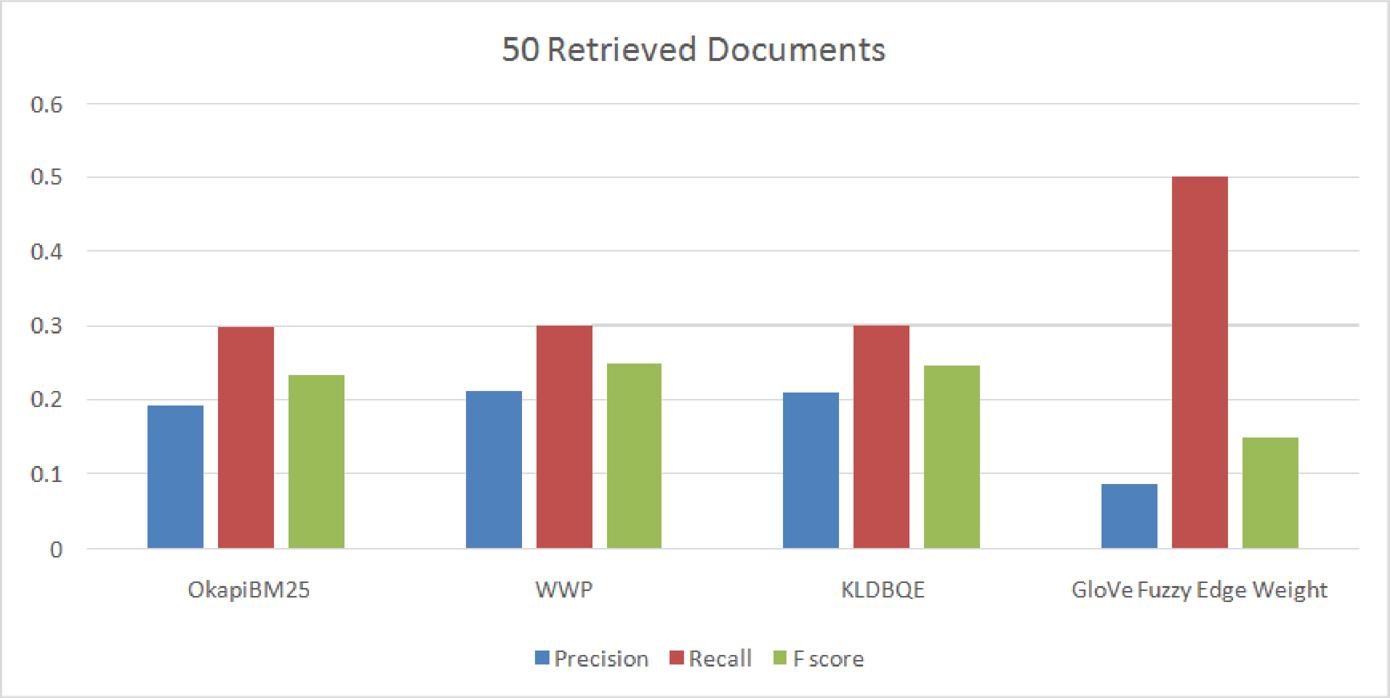

Fig. 8: Precision, Recall, F score comparison of baseline methods with GloVe method on 50 retrieved documents

7. Conclusion and Future Work

The method proposed here to expand a query using fuzzy graph connectivity has shown improved performance over graph connectivity-based IR without fuzzy edges. Moreover, it has shown comparable performance against other state of art methods with a much better recall. Semantic relations that exists between words are explored using WordNet and is found that these relations play a big role in including relevant terms to the query. Since the technique proposed in this paper accounts for the degree of relationship between words, it provided better results than comparable approaches used in the past. In this way, a query is improved for better information retrieval. In future, new methods to compute fuzzy relationship values between the query words can be explored. Instead of type-1 fuzzy logic, type-2 fuzzy logic or interval type 2 fuzzy logic can be explored.

References

[1] Deo, Arpit, Jayesh Gangrade, and Shweta Gangrade., A survey paper on information retrieval system, International Journal of Advanced Research in Computer Science 9.1, 2018.

[2] Ooi, Jessie, et al., A survey of query expansion, query suggestion and query refinement techniques, 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS). IEEE, 2015.

[3] Maryamah Maryamah, Agus Zainal Arifin1, Riyanarto Sarno, Yasuhiko Morimoto, Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents, International Journal of Intelligent Engineering and Systems, Vol.12, No.5, 2019

[4] B. He and I. Ounis, Studying query expansion effectiveness, in European Conference on Information Retrieval, 2009, pp. 611-619: Springer

[5] Roy, D., Paul, D., Mitra, M., & Garain, Using word embeddings for automatic query expansion, arXiv preprintarXiv:1606.07608, 2016.

[6] Diaz, F., Mitra, B., & Craswell, N., Query expansion with locally-trained word embeddings, arXiv preprintarXiv:1605.07891, 2016.

[7] Kuzi, S., Shtok, A., & Kurland, O., Query expansion using word embeddings, Proceedings of the 25th ACM international on conference on information and knowledge management, 2016.

[8] Zingla, M. A., Chiraz, L., & Slimani, Y., Short query expansion for microblog retrieval, Procedia Computer Science, 96, 225–234(2016).

[9] Vechtomova, Olga, Query Expansion for Information Retrieval, Springer Science + Business Media, LLC, 2009.

[10] Jain, Amita, Kanika Mittal, and Devendra K. Tayal, Automatically incorporating context meaning for query expansion using graph connectivity measures, Progress in Artificial Intelligence 2.2-3 (2014): 129-139.

[11] Balaneshin-kordan, Saeid, and Alexander Kotov, Sequential query expansion using concept graph, Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.

[12] Jain, Amita, Sonakshi Vij, and Oscar Castillo, Hindi Query Expansion based on Semantic Importance of Hindi WordNet Relations and Fuzzy Graph Connectivity Measures, Computación y Sistemas 23.4, 2019

[13] Mogotsi, I. C., Christopher d. manning, prabhakarraghavan, and hinrichschütze: Introduction to information retrieval., 2010 : 192-195.

[14] Sinha, Ravi, and Rada Mihalcea, Unsupervised graph-based word sense disambiguation using measures of word semantic similarity, International conference on semantic computing (ICSC 2007). IEEE, 2007.

[15] Navigli, Roberto, and Mirella Lapata. "An experimental study of graph connectivity for unsupervised word sense disambiguation." IEEE transactions on pattern analysis and machine intelligence 32.4, 2009.

[16] Miller, George A, WordNet: a lexical database for English, Communications of the ACM 38.11 : 39-41, 1995.

[17] Grootjen, Franciscus Alexander, and Th P. Van Der Weide, Conceptual query expansion, Data & Knowledge Engineering 56.2: 174-193, 2006.

[18] Borgatti, Stephen P, Identifying sets of key players in a network, IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No. 03CH37502). IEEE, 2003.

[19] Brin, Sergey, and Lawrence Page, The anatomy of a large-scale hypertextual web search engine, 1998 .

[20] Kleinberg, Jon M, Authoritative sources in a hyperlinked environment, Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, 1998.

[21] Freeman, Linton C, Centrality in social networks conceptual clarification, Social networks 1.3: 215- 239, 1978

[22] Zadeh, Lotfi A, Fuzzy sets, Information and control 8.3 : 338-353, 1965.

[23] Mandelbaum, Amit, and Adi Shalev, Word embeddings and their use in sentence classification tasks, arXiv preprint arXiv:1610.08229, 2016 .

[24] Mikolov, Tomas, et al., Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 2013.

[25] Word2Vec, https://code.google.com/archive/p/word2vec/, Accessed on : 25th January 2020

[26] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

[27] Joulin, Armand, et al, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759, 2016.

[28] Comparison of different word embeddings on text similarity – A use case in NLP https://medium.com/@Intellica.AI/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c,Oct 4 2019

[29] Lahitani, AlfirnaRizqi, Adhistya Erna Permanasari, and Noor AkhmadSetiawan, Cosine similarity to determine similarity measure: Study case in online essay assessment, 2016 4th International Conference on Cyber and IT Service Management. IEEE, 2016.

[30] Jain, Amita, and D. K. Lobiyal, Fuzzy Hindi WordNet and word sense disambiguation using fuzzy graph connectivity measures, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 15.2 : 1-31, 2015.

[31] Eigen vector Centrality, https://en.wikipedia.org/wiki/Eigenvector_centrality, Accessed on : 20th January 2020

[32] Becchetti, Luca, and Carlos Castillo, The distribution of PageRank follows a power-law only for particular values of the damping factor, Proceedings of the 15th international conference on World Wide Web. 2006.

[33] Colace, Francesco, et al, A query expansion method based on a weighted word pairs approach, Proceedings of the 3rd Italian Information Retrieval (IIR), 17-28 964, 2013.

[34] Singh, Jagendra, and Aditi Sharan, Relevance feedback based query expansion model using Borda count and semantic similarity approach, Computational intelligence and neuroscience, 2015.

[35] WordNets in the World, Global WordNet Association, Retrieved 19 January 2020.

[36] Ch Anwar ul Hassan, Muhammad Sufyan Khan, Munam Ali Shah, Comparison of Machine Learning Algorithms in Data classification, Proceedings of 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018.

[37] Introduction to fuzzy graph : Basic definitions and notations, https://shodhganga.inflibnet.ac.in/bitstream/10603/191172/3/chapter-1.pdf

[38] Princeton. University. About WordNet, https://wordnet.princeton.edu/, 2010.

[39] G. Salton and C. Buckley, Term-weighting Approaches in Automatic Text Retrieval, Information Processing and Management, vol.24, no.5, pp.513–523, 1988.

[40] Introduction to Word Embedding and Word2Vec, https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec, Sep 1, 2018

[41] Terrier, http://terrier.org/, 18th November 2004

[42] TREC Dataset, https://trec.nist.gov/data.html, Accessed On: 25th January 2020