Prog Artif Intell (2014) 2:129–139
DOI 10.1007/s13748-014-0041-x

REGULAR PAPER

Automatically incorporating context meaning for query expansion using graph connectivity measures

Amita Jain · Kanika Mittal · Devendra K. Tayal

Received: 7 June 2013 / Accepted: 14 January 2014 / Published online: 4 February 2014

Abstract In order to improve the retrieval performance, the query is reformulated by the process of Query expansion (QE). Most of the existing query expansion techniques do not consider the context of the terms present in the user's query which can result in low precision and recall. Through this paper, the query consisting of ambiguous terms (polysemy words) is expanded by selecting the terms, which are in close proximity to the query terms while context meaning of the terms is automatically incorporated. The basis of this query expansion method is to investigate the role of graph structure (which is being created for the query) and determining the importance of each node in the graph using WordNet. The relevant nodes representing word senses are identified from the graph and can be chosen as additional terms to be added to the query for improving the retrieval of web pages. The experiments, conducted on data sets of ambiguous queries show that proposed approach outperforms other query expansion methodologies by enhancing precision and recall.

Keywords Query expansion · Natural language processing · Information retrieval · PageRank · Hypertext induced topic selection (HITS) · Key player problem (KPP) · Centrality

A. Jain (✉)
Department of CSE, Ambedkar Institute of Advanced Communication, Tech and Research, Delhi, India
e-mail: amitajain@aiactr.ac.in; amita_jain_17@yahoo.com

K. Mittal
Department of CSE, Bhagwan Parshuram Institute of Technology, New Delhi, India
e-mail: Kkanika_virgo@yahoo.com

D. K. Tayal
Department of CSE, Indira Gandhi Delhi Technical University for Women, Delhi, India
e-mail: dev_tayal2001@yahoo.com

1 Introduction

Information retrieval [1] is a process of retrieving the documents from the document database when the user enters his query in the search engine. The main aim of information retrieval system is to evaluate the degrees of relevance of the collected documents with respect to a user's query and retrieve the documents with a high degree of satisfaction to the user. But sometimes, it results in the retrieval of irrelevant documents along with relevant documents, as the user is unclear about the information actually needed. The uncertainty in the user's query induces ambiguity due to which inappropriate documents are retrieved. In addition, heterogeneous and dynamically changing information are the major challenges of web data [2], which results in low precision. In order to improve the retrieval efficiency, QE technique is used in which user's query is modified by addition of certain terms into the original query The expansion of the initial query is done by finding and adding the relevant terms from the retrieved documents to the initial query, and weighing of the terms is done using an appropriate weighing technique [3,4]. Through query expansion, the ambiguity of terms can be dealt and the effects of the word mismatch problem are reduced which is a result of different terms being used in reference to a single concept, both in the documents and in the user queries. The process of adding terms to the query can either be manual, automatic or user-assisted.

In literature there are several techniques for query expansion, such as relevance feedback technique [5] in which the user is presented with list of answers to the query and the user can then mark as relevant or irrelevant to the information need. A variation of relevance feedback, namely, pseudo relevance feedback was proposed by Buckley et al. [6]. A term cluster query expansion [7] in which classification information is generated based on which term clusters are made

which are then selected by user and additional query terms are selected accordingly. Manning et al. [1] categorized query expansion into two classes: global methods based on expanding the query independent of query terms so that new query matches other semantically related terms and local method used the documents retrieved using unmodified query. Query can also be expanded by using ontology, which provides vocabulary and word representations for clear communication within a particular domain such as WordNet, Euro WordNet [8] etc.

The methods proposed for query expansion in the literature are not deprived from drawbacks. Lioma and Ounis [9] attempted two approaches for query expansion technique that is based firstly, a purely shallow syntactic-based query expansion (SSQE) technique and second, a combination of the SSQE method and the probabilistic pseudo-relevance feedback approach. However, this assumption was not accurate as frequently occurring part-of-speech blocks are merely a result of sentence construction in natural language documents. In addition their approach was computationally intensive as it required parsing of documents. Another way of expanding the query which was based on co-occurrence of terms has a drawback that two terms which co-occur in the same sentence seem more correlated than two terms which occur distantly within a document, but the simple co-occurrence does not necessarily mean that the terms are correlated. Moreover, this approach gave more importance to rare terms than to common terms.

However, there are several problems with ontology-based approach too like issues related to vocabulary mismatch between the query terms and the concepts in the ontology. Secondly, a lot of effort is required if an ontology for a particular domain does not exist to construct ontology from scratch. The design and construction of domain ontology is labor intensive, time consuming and difficult.

In this paper, the focus is on graph-based methods for query expansion and investigating the role of graph structure to determine the importance of a node in the graph. The graph is analyzed to find the additional and relevant nodes out of all the candidate nodes to be added to the original query nodes (terms) so as to expand the query. The various graph connectivity measures, namely, degree centrality, betweenness centrality, key player problem, PageRank & HITS have been analyzed which will assess the relative importance of the node within the graph.

Through this paper, a method is derived to improve the performance of query expansion and overcome the limitations of other approaches proposed for query expansion. In this approach, the semantic relations between all the query terms are explored by constructing a WordNet graph as WordNet interlinks, not just word forms but specific senses of words. The reason behind choosing WordNet is that all the concepts/word senses are related with each other through various relations defined in WordNet. These relations play a significant role in representing the concepts/word senses in a semantically enriched way. This helps in extracting the terms for query expansion. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. In addition, WordNet labels the semantic relations among words; whereas the grouping of words in a thesaurus does not follow any explicit pattern other than meaning similarity. Then based on the minimum distance between the query terms (selecting a particular value for minimum distance), a sub-graph is extracted from WordNet graph consisting of all the neighboring and candidate terms for expansion including the original query terms. This technique is regardless of the ambiguous terms present in the query. After sub-graph construction, the graph connectivity measures are calculated to find out the respective value of each connectivity measure. If the average score is found out for all the measures and the candidate nodes having the respective value of any three graph connectivity measures greater than the average score calculated, then those nodes will be selected as additional and relevant nodes to be added to original query nodes. In this way, query is expanded by considering the value of graph connectivity measures contributing to higher accuracy as the terms having higher value will only be added to the query.

In Sect. 2, the related work done in this field is mentioned. In next section, we will discuss about the evaluation parameters used to select the expansion terms; in Sect. 4, the proposed method for query expansion is explained. Going further, in Sect. 5, we will explain the proposed method with the help of an example along with the results, finally in last section, we conclude our research and future work.

2 Related work

In the literature, different QE approaches are studied in different ways. For instance, Manning et al. [1] provided a classification of QE approaches into global and local methods, where global methods are query-independent since all documents are examined for all queries. Global methods include QE using WordNet, reformulation using automatic thesaurus generation and local methods include query expansion using relevance feedback, pseudo-relevance feedback or indirect relevance feedback. Cao et al. [10] and Collins-Thompson and Callan [11] captured both direct and indirect term relationships for query expansion through external knowledge sources such as ontology and statistical processing of the document corpus, respectively, as independent usage of the sources showed minimal improvement in retrieval performance. But there is minimal improvement as several important factors have not been examined and utilized extensively; e.g., the query structure, length, and linguistic characteris-

tics. One of the noticeable limitations of using the WordNet Ontology given by Mihalcea [12] for query expansion is the limited coverage of concepts and phrases within the ontology. There are graph-based methods for query expansion which determined the importance of each node. The graph can be constructed by exploring semantic relations between different concepts using WordNet. The types of relations considered are hierarchical (e.g., IS-A or hypernym-hyponym, part-whole, etc.,), associative (e.g., cause-effect), equivalence (synonymy), etc., [13] and the degree of importance of each node can be found out using certain graph connectivity measures [12,14]. Kim et al. in [15] proposed a query term expansion and reweighting method which considers the term co-occurrence within the feedbacked documents. The further categorization of QE approaches was given by Grootjen and van der Weide [16] as extensional, intentional, or collaborative ones. The first approach materializes information needed in terms of documents; for instance, relevance feedback and local analysis methods. The second category, i.e., intentional approach which takes advantage of the semantics of keywords, is primarily thesauri/ ontology-based. Collaborative approaches are focused towards exploiting users' behavior, e.g., mining query logs, as a complement to previous approaches. Sanasam et al. [17] proposed a method for query expansion based on real-time implicit feedback from user. Voorhees in his work [18] used WordNet for query expansion by adding synonyms to the original query for expansion. Many approaches have been used for expanding queries using automatically derived thesaurus which was basically used in domain-specific search engines. Gong et al. [19] used the combination of WordNet and Term Semantic Network (TSN) for query expansion. Salton and McGill [5] proposed a method for query expansion based on relevance feedback in which the user is presented with a list of answers to the query, and the user can then mark as relevant or irrelevant to the information need. A variation of relevance feedback, namely, pseudo-relevance feedback was proposed by Buckley et al. [6] in which the relevant terms are extracted from top ten documents that are returned in response to the original query. The additional terms are selected based on statistical heuristics and added to the original query, and the expanded query is run again to return a fresh set of documents. Certain graph-based query expansion methods have been proposed which disambiguate the ambiguous terms and identify the importance of each node in the graph.

Graph connectivity measures have been studied extensively in the social sciences, especially within the field of Social Network Analysis (SNA) [20]. A social network is basically a network consisting of groups of people with some pattern of contacts or interactions between them. Examples include the patterns of friendship between individuals or the business relationships between companies. To determine which individuals are most central or important in the network (by being most connected or having most influence) and how they are connected to one another is one of the fundamental problems in network analysis is. There are certain measures such as centrality and connectivity, which allow us to characterize the structure and properties of large networks and make predictions about their behavior. Among these measures, PageRank [21] and HITS [22] have been extremely influential and are widely studied link analysis algorithms for information retrieval. PageRank, has the purpose of measuring the relative importance of each element within the set and assigns a numerical weighting to each element of a hyperlinked set of documents. It is based on the idea of "voting" or "recommendation" i.e., when one vertex links to another one; it is casting a vote for that vertex [21]. The vertex with highest number of votes casted will have higher importance or relevance, whereas HITS rates Web pages for their authority and hub values. Hubs and authorities exhibit a mutually reinforcing relationship: a good hub is a document that points to many others, and a good authority is a document that many documents point to. The difference between PageRank and HITS is that former is computed on a subgraph of relevant pages and later takes the entire graph into account.

Rada and Mihalcea in their work [12] identify the importance of a particular node in a graph by finding graph centrality. An unsupervised graph-based method for WSD proposed by Sinha and Mihalcea, was based on an algorithm that computes graph centrality of nodes in the constructed semantic graph, they made use of the in-degree, the closeness, and the betweenness of the vertices in the graph, as well as PageRank to measure the centrality of the nodes. PageRank and HITS are variants of the another graph connectivity measure, namely, eigenvector centrality measure which assigns relative scores to all nodes in the graph based on the recursive principle that connections to the nodes having a high score contribute more to the score of the node [23]. Freeman in his work [24], determined the closeness of a vertex by calculating the shortest geodesic distance between two nodes which in turn determine the relative importance of node. According to Freeman, the betweenness of a vertex is defined in the terms of how "in-between" a vertex is among all the other vertices present in the graph. Borgatti [25] in his work proposed a measure, namely, Key player Problem (KPP) and used it to determine the importance of a vertex by its relative closeness with all the other vertices. It is calculated as reciprocal of total shortest distance from a given node to all other nodes. Barathi and Valli [27] in their work proposed an ontology-based query expansion for retrieving information to capture the context of particular concept(s) and discover semantic relationships between them. In [28], Marco and Navigli proposed a method for improving web results by acquiring the various senses (i.e., meanings) of an

ambiguous query and then cluster the search results based on their semantic similarity to the word senses induced.

3 Evaluation measures to select expansion terms

There are certain local measures and global, which determine the degree of relevance of a vertex 'v' in graph G and the influence of a node over the network. They are helpful in determining the graph connectivity and can be used for both directed and undirected graphs. In this paper, we will discuss about only local measures. We can define a local measure l as:

$$l : V \rightarrow [0, 1]$$

A value close to 1 indicates that a vertex is important, whereas a value close to zero indicates that the vertex is peripheral. In the literature [20], there are several local methods for determining the graph connectivity and importance of a particular vertex or node in the graph, namely, key player problem (KPP), PageRank, HITS, centrality i.e., degree centrality, betweenness centrality, and other variants.

The measures are discussed briefly below:

3.1 Centrality

The basic idea behind the graph centrality is to determine the importance of a node in the graph taking into account the relation of the node with other nodes in the graph [12].

The variants of centrality are:

3.1.1 Degree centrality

It is the simplest way to determine a vertex importance by its degree [14]. The degree of a vertex refers to the number of edges incident on that vertex. For an undirected graph, the number of outgoing edges and number of incoming edges are same; i.e., in-degree is equal to out-degree. However, for directed graphs it is different. The degree of a vertex is given by:

$$\deg(v) = |\{(u, v) \in E : u \in V|$$

If a vertex is present in the center of the graph, it has high degree. The degree centrality is the degree of a vertex normalized by the maximum degree and calculated as [14].

$$C_D(v) = \frac{\deg(v)}{|V| - 1} \tag{1}$$

3.1.2 Betweenness centrality

It is defined in terms of how "in between" a vertex is among the other vertices in the graph [12]. The betweenness centrality of a node 'v' is the ratio of number of shortest paths from one node to another that are passing through 'v' and the number of shortest path between two nodes.

It is a computationally expensive method owing to the number of shortest paths that needs to be calculated. Betweenness centrality is calculated as:

$$\text{Betweenness(v)} = \sum \frac{\sigma_{ij(v)}}{\sigma_{ij}} \tag{2}$$

where σ_{ij} is the number of shortest paths between node i and j, and $\sigma_{ij}\ (v)$ is the number of shortest paths between node i and j passing through vertex v.

The node is considered to be important if that node is involved in large number of paths as compared to the total number of paths.

3.1.3 Key player problem (KPP)

KPP considers the importance of a vertex by its relative closeness with all the other vertices [25]. It is calculated as reciprocal of total shortest distance from a given node to all other nodes.

It is calculated as:

$$\text{KPP}(v) = \frac{\sum_{u \in V : u \neq v} \frac{1}{d(u,v)}}{|V| - 1} \tag{3}$$

where, the inverse of the shortest distance between v and all other nodes is the numerator, and denominator is the nodes in the graph.

3.1.4 PageRank

PageRank is one of the popular algorithms to rank the nodes or find the importance of a node in a network. It is based on the idea of "voting" or "recommendation" i.e., when one vertex links to another one; it is casting a vote for that vertex [21]. The vertex with highest number of votes casted will have higher importance or relevance. Moreover, the importance of vertex casting a vote determines how important a vote is. All the nodes that link to 'v' contribute towards determining its relevance. The PageRank algorithm was initially proposed for directed graphs, but it can be applied on undirected graphs also.

The PageRank of a node 'v' for an undirected graph is calculated using a recursive function as:

$$\text{PageRank}\,(v) = \frac{1 - d}{|V|} + d \sum_{u,v \in E} \frac{\text{PR}(u)}{\text{outdegree}(u)} \tag{4}$$

where d is the damping factor introduced, which has the role of integrating into the model the probability of jumping

given vertex to another random vertex [26] and its value is set between zero and 1. A value for zero means that the ranking of the page does not depend on its outgoing links, and 1 indicates that the score is exclusively determined by the links with neighboring pages. The typical value of d is 0.85.

3.1.5 Hypertext induced topic selection (HITS)

HITS is similar to PageRank but the only main difference is, it makes a distinction between authority and hubs; i.e, in this method two values are determined for a node 'v' i.e., authority (a(v)) and hub value (h(v)). The authority corresponds to the pages that are good and reliable sources and have numerous incoming links, whereas hub value corresponds to the pages having many outgoing links [22].

Another difference between PageRank and HITS is that the former is computed on a sub-graph of relevant pages and later takes the entire graph into account.

For every vertex, HITS produce two set of scores—*authority score* and *hub score.* They are found out using below equations:

$$\mathrm{HITS}_A(Vi) = \sum_{V_j \in In(V_i)} \mathrm{HITS}_H(Vj) \tag{5}$$

$$\mathrm{HITS}_H(Vi) = \sum_{V_j \in Out(V_i)} \mathrm{HITS}_{HA}(Vj) \tag{6}$$

For each iteration, these scores are normalized, so that the authority scores for all vertices add up to 1. HITS can also be applied to undirected graphs.

4 Proposed method

To improve the retrieval of documents, the user query is expanded to include more relevant terms through query expansion method. As the user query does not index properly all the relevant terms other than the query terms, it can lead to low precision results. Therefore, to retrieve relevant documents, the user's query has to be expanded by addition of more terms to original query. In this regard, an approach is proposed for expanding the query by finding the appropriate and relevant terms matching to query terms which can properly index query terms and then adding those related terms to the original query for efficient documents retrieval. Through this method, a query has been considered having some query terms, using which a WordNet graph 'G′ is constructed by exploring all the relations between the original query terms (i.e., hypernymy, hyponymy, meronymy, etc.,). Using this WordNet graph G having V vertices, a sub-graph G′ is extracted (which is empty initially) by considering the minimum distance L' between the query terms nodes so that the path between any two linked query terms is less than or equal to L'.

The value of L' in this approach is taken as:

$$\mathrm{L}' = (max(\min \textit{ distance between the respective query terms})) \tag{7}$$

If the path between any two linked query terms is less than or equal to L', then that path is added to the sub-graph G′. After constructing the entire sub-graph, the graph connectivity measures (discussed in Sect. 3) are calculated and the average score 'A' of all of the graph connectivity measures is found out. Then, select the nodes V′ from G′ such that node V′ is not equal to any query term and is having value of any three graph connectivity measures greater than the average score 'A' for the respective measure. Finally add the selected nodes to the original query for expanding the original user's query.

Consider a user query having 'i' terms where $1 \leq$ i $\leq$n. The below terms will be used further in the paper:

G	is the graph constructed around query terms from WordNet, i.e., WordNet Graph G.
G′	Sub-graph extracted from G
V′	is the nodes present in sub-graph
T_i	is the total number of query terms in the query where $1 \leq i \leq$ n
T_j	is the query term considered at a time where $1 \leq T_j \leq$ n
T_k	is the linked query term to T_j where j + $1 \leq T_k \leq$ n
L'	is minimum distance between the query terms

A is the average score calculated for graph connectivity measures.

The proposed algorithm is given in Fig. 1.

The description of the algorithm is given as below:

1. Initially query is entered by user having terms T1, T2, T3…Tn.
2. From the WordNet graph G, having V vertices using those the query terms create sub-graph G′ from the graph G such that G′ is subset of G.
3. For term T_j= 1,2 … n
   (a) Perform depth first search (DFS) on every term T_j of the WordNet graph (G).
   (b) For each T_k where $k \neq j$ & k is between j+1 to n, if there is a path $T_j \ldots T_k$ of length <= L', where $L' =$ 5, add all the intermediate nodes and edges of path from T_j to T_k to the WordNet sub-graph G′.
4. After all the query terms are examined sub-graph, i.e., G′ is obtained. For each node of the Sub graph G′, calculate the Graph Connectivity Measures Degree Centrality, KPP, Betweenness, PageRank &, HITS.
5. Select the nodes V′ from G′ such that, node V′ should not be equal to any query term Tj and having value of

// considering an initial user query having 'i' terms and the value of minimum distance i.e. L' is considered as L'=5

Step 1: Query is entered by user having 'i' terms where $1 \leq i \leq n$

Step 2: To create Sub-graph G' from WordNet graph G

Initialize the Sub graph G' to NULL

For j= 1 to n (Looping through the Query Terms),
Repeat

i. For every T_j perform depth first search (DFS) of the WordNet graph (G)
ii. For each T_k where $k \neq j$ & k is between j+1 to n, if there is a path $T_j \ldots T_k$ of length <= L', add all the intermediate nodes and edges of path from T_j to T_k to the WordNet sub-graph G'.

Where L' is maximum of Minimum Distances between any two Query Terms
i.e. L' = max [min (distance between respective query terms chosen)]
5 in our example

Step 3: Calculate the **Graph Connectivity Measures**

For each node of the Sub graph G' Calculate the below Graph Connectivity Measures

- Degree Centrality
- KPP
- Betweenness
- PageRank
- HITS

Step 4: Calculate the average score of all of the above Graph Connectivity Measures

Step 5: Select the nodes V' from G' such that
(i) Node V' should not be equal to any Query Term T_k
(ii) Any 3 Graph Connectivity Measures for V' are greater than their respective average score 'A'.

Step 6: Add the above selected terms to the Original Query for query expansion.

Fig. 1 The algorithm of the proposed method

any three measures greater than their respective average scores.
6. Add these selected terms (nodes) to the original query.

This algorithm can be shown with the help of a flowchart given in Fig. 2.

5 Illustration through example

The above approach can be explained through an example where a query is taken. The sample query taken is: "Inventions in Science and Technology". Here the stop words such as 'in', 'and' are neglected and the query terms selected are only 'invention', 'science', 'technology'.

A graph is constructed around all query terms from WordNet and is used for finding the shortest distance between a pair of terms by node counting method. The graph is given in Fig. 3

Using the graph given in Fig. 3, a particular query term is selected initially and length of the shortest path will be found out between all other query terms linked to the initially chosen query term. If the length of the shortest path between the query terms is less than or equal to the L' calculated, then that path between the two respective terms will be added to the sub-graph, which was initially empty. In this way the entire sub-graph is extracted from the WordNet graph consisting of query terms and other relevant terms.

The sub-graph G′ is shown in Fig. 4.

After the construction of sub-graph, the graph connectivity measures (discussed in Sect. 3) are calculated, and results are obtained shown in Table 1. For each graph connectivity measure, the average score is found out and based on that all those terms that have values of any three connectivity measures greater than the average score are selected as final

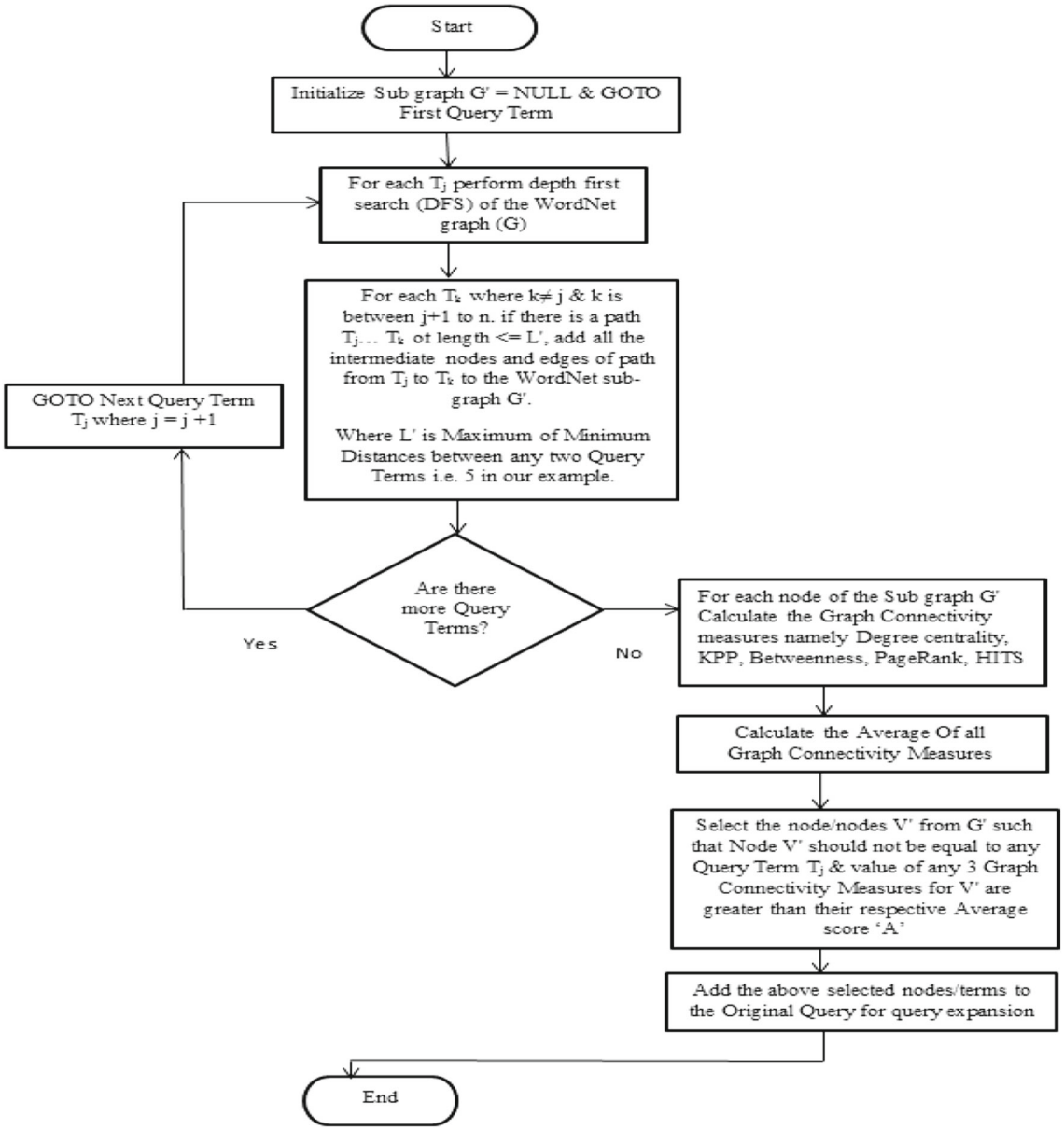


Fig. 2 Flowchart of the proposed method

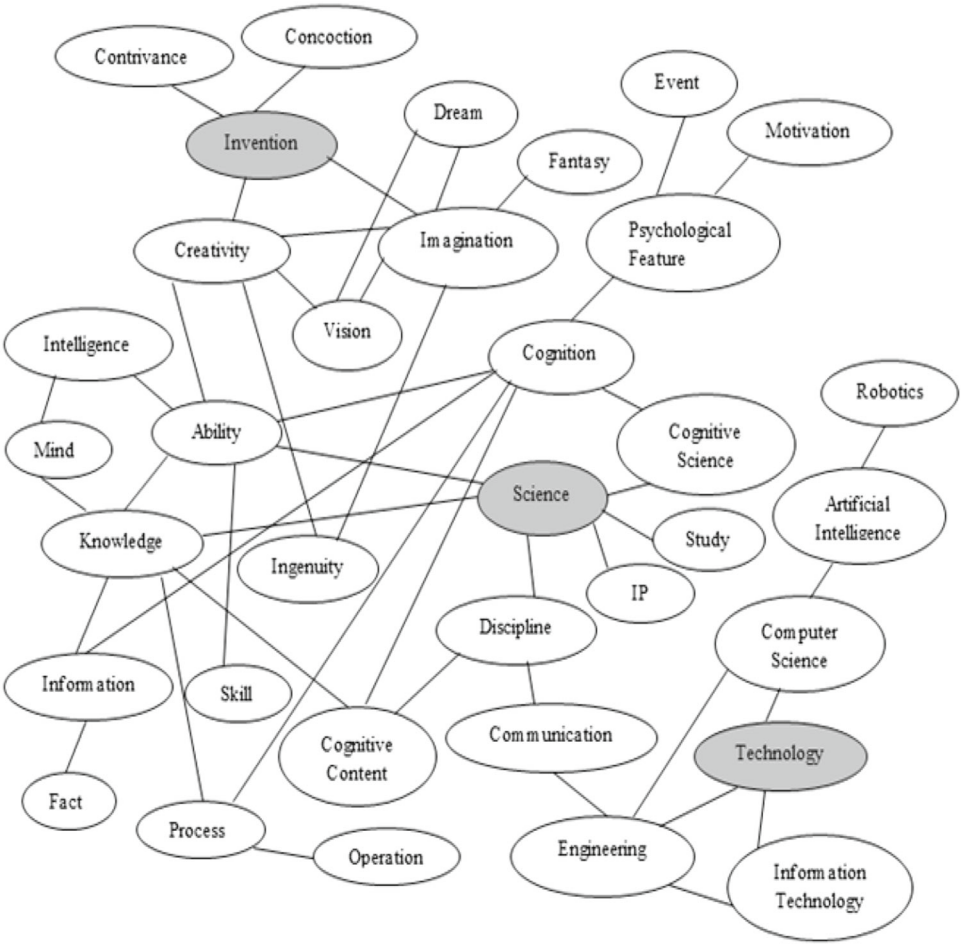
expansion terms out of all the candidate terms and will be added to the original query so as to retrieve more relevant documents and improve the efficiency of retrieval.

The table containing various evaluation parameters for each candidate term and the corresponding average score is given below:

Results: From the above values obtained, we can find out that four terms namely: creativity, ability, engineering and discipline (excluding the original query terms i.e., invention, science, technology) have higher average values than the other candidate terms. Therefore, query will be expanded by adding these terms to the original query.

6 Experiments, implementation and results

The given query expansion technique was tested on standard ADI data set. From this data set, we have considered queries having ambiguous terms, i.e., having polysemous words. We have used WordNet version 2.1 for determining context meaning of ambiguous terms. The TMG tool, which is a MATLAB toolbox for text to matrix generator, is used for generating term documents matrices, removing of stop words from query, frequent/infrequent terms removal, clustering of documents, retrieval of relevant documents, etc. TMG offers two alternatives for Text Mining.



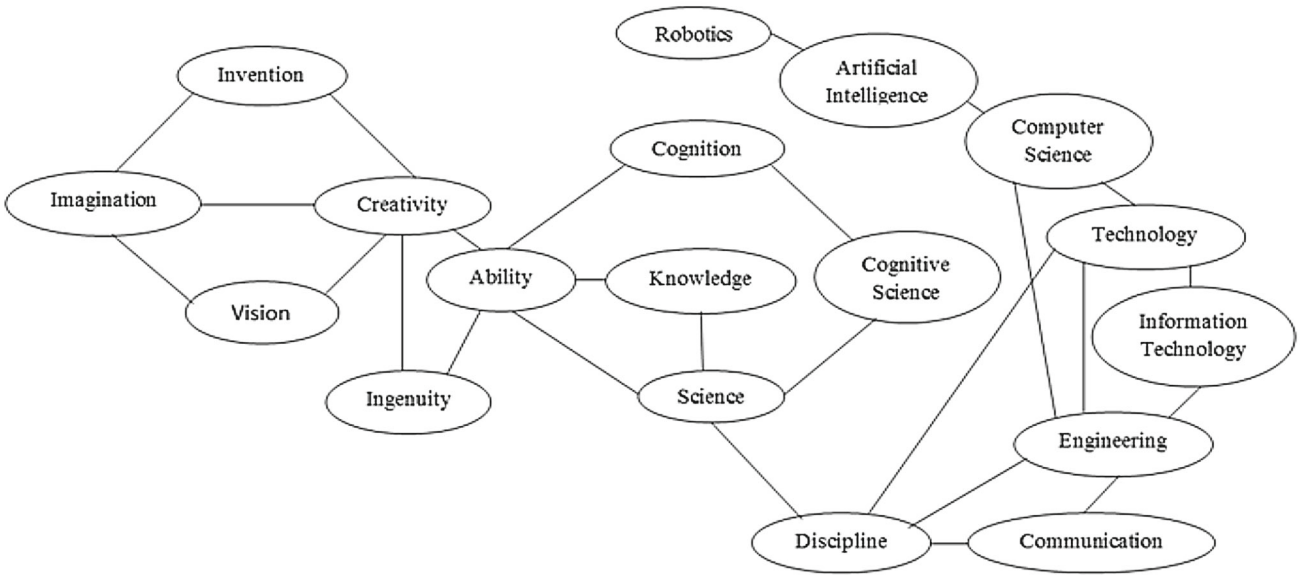


Fig. 3 An excerpt of WordNet around the query terms

Fig. 4 Sub-graph G' constructed from WordNet graph G

Table 1 Evaluation and results

Terms	Degree centrality	KPP (key player problem)	Betweenness centrality	PageRank	HITS
Invention	0.11	0.36	0.003	0.0334	0.30
Imagination	0.17	0.39	0.005	0.0554	0.33
Creativity	**0.29**	**0.50**	**0.025**	**0.0732**	**0.54**
Vision	0.11	0.36	0.001	0.0334	0.29
Ingenuity	0.11	0.41	0.001	0.0271	0.28
Ability	**0.29**	**0.55**	**0.025**	**0.0881**	**0.55**
Knowledge	0.11	0.43	0.003	0.0394	0.26
Cognition	0.11	0.39	0.001	0.0381	0.26
Cognitive science	0.11	0.39	0.001	0.0481	0.21
Science	0.23	0.51	0.018	0.0766	0.48
Discipline	**0.23**	**0.51**	**0.020**	**0.0591**	**0.57**
Communication	0.11	0.40	0.003	0.0294	0.25
Engineering	**0.29**	**0.50**	**0.025**	**0.0761**	**0.63**
Information technology	0.11	0.36	0.003	0.0294	0.29
Technology	0.23	0.48	0.011	0.0596	0.53
Computer science	0.17	0.40	0.007	0.0417	0.36
Artificial intelligence	0.11	0.33	0.003	0.0511	0.14
Robotics	0.05	0.25	0	0.0319	0.11
Average	**0.16**	**0.41**	**0.008**	**0.049**	**0.35**

The bold entries indicate the additional terms which are selected (on the basis of their centrality measures values)

- Vector Space Model (VSM)
- Latent Semantic Analysis (LSA)

Using the corresponding GUI, the user can apply a question to an ADI dataset using any of the aforementioned techniques and get HTML response. This retrieval GUI uses a set of parameters as insert query, stop-list, number of factors, similarity measure, etc. To test the efficiency of the graph connectivity approaches, we have primarily made use of the inset query parameter. We took the actual input query, used a set of Java API for WordNet Searching, for calculating and evaluating the measures (Degree Centrality, KPP, Betweenness Centrality, PageRank, HITS) to select the expansion terms.

As its name implies, the Java API for WordNet Searching (JAWS) is an API that provides Java applications with the ability to retrieve data from the WordNet database. It is a simple and fast API that is compatible with both the 2.1 and 3.0 versions of the WordNet database files and can be used with Java 1.4 and later.

Within the application, we can use JAWS by first obtaining an instance of WordNet Database with code like the following:

WordNetDatabase database = WordNetDatabase.getFileInstance();

After that we can begin to retrieve synsets from the database. We have used the same synsets to first construct the

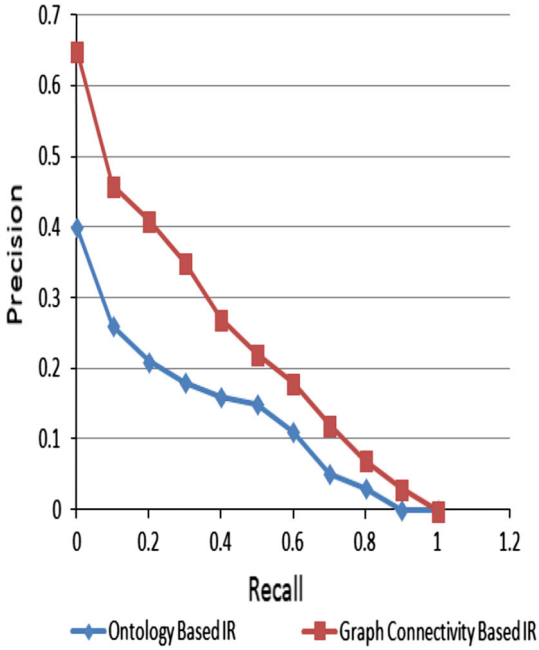


Fig. 5 Graph showing results for Ontology-based IR & graph connectivity-based IR

WordNet graph G and later the subgraph G′ (G and G′ being the graph used in the illustration earlier).

Once the input query is modified/expanded based on the measures evaluated above the same is fed to the insert query

parameter of the retrieval GUI of TMG. The resulting document set was then tested for precision and recall.. Precision is calculated for several values of N for only the top N documents. The method discussed in this paper is an improvement over the other query expansion methods, i.e., ontology-based query expansion [27], as this methodology takes into consideration the ambiguous terms and determines the correct sense of the same and results in a significant improvement in the precision by increasing the number of relevant documents. A graph is plotted between precision and recall in which, for particular values of recall, there is a significant increase in the precision values for graph connectivity-based expansion method. The result shows an improvement over the ontology-based query expansion method. The graph is given as below in Fig. 5.

7 Conclusion

The query expansion method proposed in this paper has shown improved precision and recall for the query having polysemy words. The terms were identified by determining the importance of a node/word sense in the graph created for the query and these terms served as the relevant expansion terms. The computational lexicon, WordNet is being used due to various relations between words are present and it was found that these relations play a significant role in representing the concepts/word senses in a semantically enriched way. This representation further helped in incorporating context meaning automatically while query is being expanded. Various graph connectivity measures used were able to successfully find out the importance of nodes. Our method provided better results as compared to similar methods deployed in the literature in past. In this way, the user's query is enriched with more relevant terms for efficient retrieval of relevant web pages. In future, the work can be expanded for all open class words such as verbs, adverbs, adjectives along with nouns.

References

1. Manning, C.D., Raghavan, P., Schutze, H.: An introduction to information retrieval, Cambridge University Press, (2009)
2. Stuckenschmidt, H.: "Data Semantics on the Web", J Data Semantics: JoDS (1), Springer, Berlin [u.a.] (2012)
3. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Process Manag **24**(5), 513–523 (1988)
4. Wang, C., Yajun, D.U., Zhang, P., Han, B.: A term-reweighting method for query expansion. J Comput Information Syst **6**(11), 3779–3785 (2010)
5. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill, New York (1988)
6. Buckley, C., Salton, G., Allan, J., Singhal: A, Automatic query expansion using SMART: TREC 3. In: D. Harman, ed., Overview of the Third Text Retrieval Conference (TREC-3), NIST Special Publication 500–225, pp. 69–80 (1994)
7. Kang, J.W., Kang, H.-K.: A term cluster query expansion model based on classification. Information in Natural Language Information Retrieval, International Conference on Artificial Intelligence and Computational Intelligence (2010)
8. Leroy, G. et al: Customizable and ontology-enhanced medical information retrieval interfaces, Methods of Info in Medicine, (2000)
9. Lioma, C., Ounis, I.: A syntactically-based query reformulation technique for information retrieval. Information Process Manag **44**(1), 143–162 (2008)
10. Cao, G., Nie-Y, J., Bai, J. : Integrating word relationships into language models. In proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pp. 298–305 (2005)
11. Collins-Thompson K., Callan, J.: Query expansion using random walk models. In: Proceedings of the 14th ACM Intl conference on information and knowledge management, pp. 704–711 (2005)
12. Sinha, R., Mihalcea, R.: Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: Proceedings of ICSC (2007)
13. George A. Miller: WordNet: a lexical database for English: International J. Lexicogr. (1995)
14. Navigli, R., Lapata, M.: An experimental study of graph connectivity for unsupervised word sense disambiguation, IEEE transaction on pattern analysis and machine learning, Vol. 32 No. 4, April (2010)
15. Kim, B.M., Kim, J.Y., Kim, J.: Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference. In: Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, Vol. 2, pp. 715–720 (2001)
16. Grootjen, F.A., van der Th Weide, P.: Conceptual query expansion. Data Knowledge. Eng. **56**(2), 174–193 (2006)
17. Singh, S.R., Murthy, H.A., Gonsalves, T.A.: Inference based Query Expansion Using User's Real Time Implicit Feedback. Knowledge Engineering and Knowledge Management, Communications in Computer and Information Science Vol. 272, 2013, pp. 158–172, Springer (2013)
18. Voorhees, E.M.: Query Expansion using Lexical-Semantic Relations. In: SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 61–69, New York (1994)
19. Gong, Z., Cheang, C., Leong Hou, U.: Web query expansion by WordNet. In: Andersen, K., Debenham, J., and Wagner, R., eds. Database and Expert Systems Applications, volume 3588 of Lecture Notes in Computer Science, pp. 166–175. Springer (2005)
20. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge Univ, Press (1994)
21. Brin, S., Page, M.: Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings Seventh Conference World Wide Web, pp. 107–117 (1998)
22. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings Ninth Symposium Discrete Algorithms, pp. 668–677 (1998)
23. Bonacich, B.P.: Factoring and weighing approaches to status scores and clique identification. J. Math. Sociol. **2**, 113–120 (1972)
24. Freeman, L.C.: Centrality in social networks: conceptual clarification. Social Netw. **1**, 215–239 (1979)
25. Borgatti, S.P.: Identifying Sets of Key Players in a Network. In: Proceedings Conference Integration of Knowledge Intensive Multi-Agent Systems, pp. 127–131 (2003)
26. Litvak, N., Scheinhardt, W., Volkovich, Y.: In-degree and PageRank of web pages: why do they follow similar power laws?"

Memorandum 1807, Department of Applied Math., University of Twente (2006)
27. Barathi, M., Valli, S. : Ontology Based Query Expansion Based on Word Sense Disambiguation", International Journal of Computer Science and Information, Security, Vol. 7, No. 2, February (2010)
28. Di Marco, A., Navigli, R.: Clustering and diversifying web search results with Graph Based Word Sense Induction", Computational Linguistics (2012)