CHAPTER 5

INFORMATION RETRIEVAL AND RANKING WITH SEMANTIC RELATIONSHIPS (RISeR & RaSeR)

Most of the information retrieval approaches are primarily focused on the number of occurrences of the given concept or keyword in the set of documents. The approaches are ineffective since the retrieved results are not completely satisfactory (Chen, 2009). It also becomes more and more complex as the information grows. The evolution of semantic based information retrieval approach enhances the accuracy of the retrieved documents. However, the information retrieval approaches does not learn the semantic relationships between the given concepts or keywords. The inclusion of semantic relationships makes the system more effective and robust than the system which is based on the number of co-occurrences of the given concept(s). The user query can be expanded with the identification of semantic relationships between the given keywords.

In this work, Retrieval of Information with Semantic Relationships (RISeR) has been introduced for obtaining more relevant information from the set of documents. Usually, the documents are ranked based on the maximum number of hits of a particular web page. However, the top n documents which are ranked using some ranking algorithm are unusable in most of the scenarios. In this work, ranking with semantic relationships (RaSeR) has been introduced to rank the documents on the basis of the relevance of the given keyword(s). This chapter starts with the need of the proposed RISeR and RaSeR algorithms. The RISeR and RaSeR algorithms are elaborated in the subsequent sections.

5.1 NEED FOR RISeR AND RaSeR ALGORITHMS

The keyword based search is most popular for its simplicity and ease of accessibility, but it has failed to retrieve the documents based on the semantics. As the information on the web increases dramatically, it becomes more and more complex to the user to find more relevant information. In semantic search, the given query is expanded with synonym of the given keyword in order to improve the accuracy of the information retrieval. Sometimes, it may also lead to retrieval of irrelevant documents. The
integration of semantic relationships in semantic search will help to reduce the number of irrelevant documents. Hence, the given query should be expanded with semantic relationships for effective information retrieval. The proposed system RISeR will help to accomplish this task.

The ordering of the results is another important challenge to exhibit the results to the end user. In general, the pages are ranked on the basis of number of hits and number of links. As the expectation differs among variety of users, it is futile. Hence, there is a need of some relevance measures for ranking the documents. RaSeR discovers the relevance using semantic relationships that exists between the given concepts and exposes the most relevant documents. The following section describes the proposed systems such as RISeR and RaSeR in detail.

5.2 RETRIEVAL OF INFORMATION WITH SEMANTIC RELATIONSHIPS (RISeR)

Due to enormous increase in information on the web, the pertinent information search from the web becomes more and more complex for the naïve users. The search engine is the most useful tool for obtaining relevant information from the web. The search engine focuses more on the presentation of the information rather than the semantics of the information. The evolution of the semantic web brings the system to understand the semantics of the information. It also helps to analyze the information on the web. It could recognize the concepts, semantic relationships and instances from the represented information.

The semantic relationship that exists between the concepts will facilitate the search engine to analyze more about the represented information and provides more appropriate information to the naïve user. The concepts are related to one another through some relationships and from the literature, it has been identified that the information is lost without the semantic relationships. The importance has been given to semantic relationships in the current semantic web to enhance the understanding of the concepts. The semantic web languages such as RDF (resource Description Framework) and OWL (Ontology Web Language) also have the provision of representing the semantic relationships. When the user has given the query, it must be expanded with the corresponding semantic relationships in order to retrieve the
documents. Hence, the number of irrelevant documents can be reduced with the semantic relationships.

For example, the user has given the query which consists of “Mac” and “Apple”. Generally, it retrieves the documents which are related to Mac computer. But Mac is related to Apple in many ways such as “Mac is developed by Apple”, “Mac eats Apple”, “Mac harvests Apple”, etc. If the user is interested in retrieving “Mac harvests Apple”, it could not be retrieved and displayed in top 10 results. Thus, the RISeR system is developed to expand the query with semantic relationships and retrieve the most relevant documents. The architecture of the RISeR is shown in Figure 5.1.

Here, the set of documents are collected from web using crawler and stored in the database. The documents are preprocessed and semantically annotated and then preprocessed documents are stored in the database for further processing. The set of ontologies are also stored in the knowledge base for achieving effective information retrieval. When the user has given the query, the query is expanded with SemPER algorithm. The identification of semantic relationships between the given concepts is essential for the enhanced understanding of the semantics between the given concepts.
or keywords. The query is expanded with the semantic relationships and then similarity measure is calculated using lexical resource(s) by the RISeR. The documents are retrieved from the knowledge base, based on the similarity measure and ontology and it has been presented to the user(s). Incorporation of semantic relationships between the given concepts helps to reduce the number of irrelevant documents. The workflow of the RISeR is shown in Figure 5.2.

![Workflow of RISeR](image)

**Fig 5.2. Workflow of RISeR**

The algorithm for RISeR is shown in Figure 5.3. In the algorithm, the similarity measure is calculated for the retrieved relationships. The similarity is measured in the range -1 to 1. The similarity is more if the calculated value is high. There is no similarity between the specified relationships if the value is 0. The value -1 represents that the specified relationships are completely opposite. In the RISeR algorithm, if the given concepts are not related, the documents are retrieved on the basis of term-document matrix. The term-document (TD) matrix describes the frequency of concepts that occur in the collection of documents. The matrix is formulated with the specified concepts and the set of processed documents and ontologies. The RISeR retrieves and displays the documents based on the number of occurrences of the given concepts among the set of documents. The implementation and the experimental
results of the retrieval of information with semantic relationship (RISeR) system have been dealt in Chapter 7.

```
Algorithm for RISeR

Input: User query Q
Output: Set of Documents D
RISeR (Q)

Begin
  Parse the Query
  Identify the key concepts
  For each concept in ontology and semantic annotated documents
     If (there exists a direct relationship between the concepts)
        Calculate similarity between each concept using WordNet
        Retrieve the documents which matches query
     Else if (there exists an indirect relationship between the concepts)
        Call SemPER // Reasoning algorithm
        Calculate similarity between two relations using WordNet
        Retrieve the documents which matches query
     Else
        // If the semantic relationship doesn’t exist between the concepts
        Retrieve documents using Term Document Matrix
  Display the documents
End
```

Fig 5.3. Algorithm for RISeR

RISeR system retrieves and displays the most relevant documents without any ordering of the documents. Next section describes ranking using semantic relationships (RaSeR) system which is used for ranking the documents.

5.3 RANKING USING SEMANTIC RELATIONSHIPS (RaSeR)

The RISeR displays the most relevant documents from the set of documents by identifying the semantic relationships that exists between the concepts and the similarity measures between the concepts. The concepts are related through multiple dissimilar relationships and thus various semantic paths have been formulated with
the relationships. The concepts are semantically related if the concepts are semantically connected (there exists the path between the given concepts) or semantically similar (properties of the two semantic paths must be same). The expanded query may result in many paths which are semantically related. As the paths between the given concepts are huge in number, the paths need to be graded according to the relevance of the user query. The user is interested only in top $n$ results. The RaSeR system should provide top $n$ results according to the relevance of the query. The relevance of the results can be measured based on the semantic relationships that exists between the extracted concepts and given query. To determine the relevance of the result, the following criteria have been considered:

- **Semantic Relation Strength (SRS):** The results have multiple semantic paths i.e. the given concepts are related through many dissimilar semantic relationships. The strength of the semantic relationships should also be considered for ranking.
  
  For example, consider Engine *is a* Machine and Engine *is a part of* Car. The Engine is a *part of car* is more related as compared to Engine *is a* Machine.

- **Relatedness degree (RD):** The given concepts are related through several concepts. The number of concepts associated in the extracted semantic paths need to be considered.

- **Semantic Relation Length (SRL):** The popularity of the semantic relationships over any type of semantic relationships derived between given concepts should be measured.

The relevance of the documents is measured through SRS, RD and SRL. The relevance measures have been explained below in detail.

The various categories of semantic relationships such as hypernym, hyponym, meronym, cause effect, spatial, and entailment exist between the given concepts. The weight has been assigned to each category to measure the importance of the semantic path. The semantic relationships can occur many times in the semantic paths. It can be considered to identify the number of meaningful semantic paths.
The semantic relation strength (SRS) is calculated using number of occurrences and weight of the semantic relationships.

\[
SRS = \frac{\sum_{i=0}^{n} c(R_i) \times w(R_i)}{\max(R_i) \times \max(w(R_i))}
\]

(5.1)

where \( n \) = no. of relations

The number of concepts associated with the given query provides more semantics. It helps to enhance the understanding of the user. The relatedness degree (RD) is calculated using

\[
RD = \frac{\text{freq}(C)}{\text{All } C \text{ between given concepts}}
\]

(5.2)

where \( C \) = no. of concepts

The semantic relation length is necessary to measure the popularity of the path which is calculated using

\[
SRL = \frac{\text{No. of semantic relation } R \text{ of each category}}{\text{All possible number of relations } R}
\]

(5.3)

**Fig 5.4. Architecture of RaSeR**
The architecture of RaSeR is shown in Figure 5.4. The RISeR retrieves set of documents and it has been ranked according to the relevance using the RaSeR algorithm which is shown in Figure 5.5.

**Algorithm for RaSeR**

**Input:** Set of retrieved Documents $D$

**Output:** Ranked Documents $D$

$RaSeR (Q)$

**Begin**

For each retrieved documents
- Calculate semantic relation strength, semantic relation strength and relatedness degree
- Find the sum and display in order

**End**

![Fig 5.5.Algorithm for RaSeR](image)

The detailed implementation and the experimental results of the ranking using semantic relationship (RaSeR) system have been dealt in Chapter 7.

5.4 SUMMARY

This chapter discussed the necessity of semantic relationship in information retrieval and ranking system. The working model of retrieval of information with semantic relationships (RISeR) system and ranking using semantic relationships (RaSeR) system are discussed elaborately in this chapter. Next chapter explains the way in which the semantic relationships are exploited in question generation system.