CHAPTER 6

CONCLUSIONS AND FUTURE ENHANCEMENTS

6.1 CONCLUSIONS

This thesis has focused on and presented the different approaches for computing the semantic similarity between the concepts or words in any domain. Semantic similarity is a computational method which quantifies the similarity among concepts belonging to single ontology or multiple ontologies. This quantification may be based on the properties, the relations a concept share with other concepts, their position in the taxonomy and corpus statistics. The proposed semantic similarity measures are tested with world knowledge represented in WordNet and Biomedical ontologies MeSH and SNOMED-CT. The measures proposed in this thesis are mainly based on the information content based approaches.

Based on the literature of the semantic similarity measures, a classification of those measures has been presented. The analysis revealed that the information content based approaches were corpus dependent and they suffered from sparse data problem. Hence a solution to address the sparse data problem is provided in this thesis by devising a new information content measure which was used in the well known information content based approaches.

Further as the similarity measures play an important role in information retrieval systems the similarity measures proposed in this thesis were tested for information retrieval. In this context a content based index was built and used in the information retrieval systems. The index terms of this content based index are the dominant concepts identified through the proposed semantic similarity measures. This content based index could be considered as an alternative for traditional keyword based index. The gain in the retrieval effectiveness in terms
of precision was appreciable when compared to the precision of the keyword based index.

Ontology has been widely used in many domains including database integration, bioinformatics, and the Semantic Web to facilitate the sharing of heterogeneous information. Semantic similarity techniques are becoming important components in the knowledge-based systems and semantic information retrieval (SIR) systems

Further from the investigation made it was evident that the information content based approach for measuring semantic similarity among ontologies is hardly available. Hence an attempt is made in this thesis to devise new information content based approaches for measuring semantic similarity among concepts belonging to different ontologies. Apart from this, the key contribution in this thesis is that various information based approaches for measuring semantic similarity among concepts belonging to different ontologies have been proposed.

The COSS, RDCOSS and RPCOSS approaches proposed in this thesis were based on the Tversky psychological model. The proposed methods were tested for the biomedical domain. The psycholinguistic approach of evaluation showed promising results as the correlations achieved by COSS and RDCOSS methods were appreciable. However, the RPCOSS method did not show significant results.

The approaches used for measuring similarity among concepts in a single ontology (Resnik, Lin and Jiang and Cornath methods) were extended to measure the similarity of concepts across ontologies. The results of these methods were comparably better than those of the existing path based method to measure similarity among ontologies. These similarity methods proposed for multiple ontologies could be used for applications like information retrieval, ontology mapping systems and ontology integration systems. Hence in this thesis, the proposed approaches were tested for ontology mapping and information retrieval.
Since the proposed approaches are corpus independent, they correlate well with the human judgments. Therefore, the correctness of the proposed similarity metrics is proved by comparing the result against the human judgments by computing the correlation coefficient. The proposed COSS measure results in higher correlation compared to the existing path length measure. The correlation achieved through COSS measure is 0.920. The refined Information Content based measure RRLIN achieve the next higher correlation of 0.911. The correlation of Refined Dice index (RDCOSS) measure for cross ontological concepts is quantified as 0.810. The similarity values computed through RJCOSS measure for measuring cross ontology concept similarity was not appreciable. But RRCOSS measure achieved correlation of 0.859 which is almost same as the RDCOSS measure.

The extrinsic evaluation of the proposed similarity measure has been carried out for information retrieval. The proposed semantic similarity measure has been used for conceptual query expansion using SNOMED-CT and MeSH ontologies. The information retrieval system’s performance is evaluated using the proposed semantic similarity measure by means of precision, recall and F-Measure. This has been compared against the traditional Vector Space information retrieval Model (VSM). The precision of the information retrieval application improved when the proposed COSS measure was used in the query document similarity matching. In comparison to the traditional vector space model and it is proved through experiments that there was 20% improvement in precision and recall.

Further, the proposed COSS measure was used for identifying corresponding concepts of diseases categories of SNOMED-CT and MeSH ontologies. The results obtained in the tests conducted could support ontology mapping applications.
6.2 FUTURE ENHANCEMENTS

Although all the semantic similarity measures developed are based on modeling the human similarity judgment, they differ fundamentally. This is, because there does not exist one single way of human similarity perception. To develop better semantic similarity measures, research must further investigate human cognitive processes and identify the influencing factors on similarity judgment. The major influencing factors have already been mentioned in the thesis: the application task and the context. This thesis has focused on a semantic similarity measure for a single specific application task such as semantic information retrieval to assess the relevance of information retrieved for a query. To develop a more general approach for semantic similarity measurement, future research must investigate similarity measures for various application tasks and should analyze the influence of the application tasks on similarity judgment.

There is ample room for further improvements and research beyond the current results. In this work, an indexing scheme had been projected by taking a generalized ontology, WordNet. But the domain specific ontology can give better results than a generalized ontology. So this work can be further be improved by merging two or more domain specific ontologies into a single generalized ontology. Furthermore, it can be extended by identifying dominant concepts in PDF documents.

For instance, semantic information integration of diverse data sources is an important task, subsequent to the identification of relevant data across various data sources. The similarity measure in information retrieval system is generally direction specific which means that according to query terms the documents are to be searched. But in certain applications like information integration systems the similarity computation need not be direction specific Hence this work could be extended to compute similarity in both directions.

Another unsolved question in semantic similarity measurement is the role of context dependence and symmetry of similarity. Context has undoubtedly a
great influence on human similarity judgment. For example, people generally consider water and sand as dissimilar, because they do not have many properties or relations in common. In case of fire, sand and water abruptly become very similar, because both share the ability to put out fire. The similarity model proposed in this thesis is based on the Tversky feature model. In this, the context of the concept is given inadequate importance. If the context is considered the performance of the information retrieval systems would further improve. Hence the psychological based models like Feature based model, geometric model, transformational may be investigated to include context in computing similarity.

The symmetrical behavior of similarity is a controversially discussed topic in the semantic similarity measurement literature. Various researchers proved empirically the existence of symmetric (Rada et al. 1989) as well as asymmetric similarity (Tversky 1977). Future research must analyze the various causes that lead to asymmetric similarity judgment. Hence it is essential to model the symmetry behavior of similarity adequately in a semantic similarity measure.

In general, the cognitive way of representing semantics is essential to assist humans to work in cooperation with machines. However, machines are better at processing semantics specified in the logic-based way. Future work must investigate whether cognitively adequate knowledge representations could be combined with logic based models to improve similarity computation.

The ontology mapping systems are gaining importance in the current trend of Semantic Web. The similarity aggregation determines the quality of the ontology mapping systems in terms of precision. The similarity aggregation is an amalgamation of word or string similarity combined with instance similarity. In future the study of effectively combining the similarity methods to improve the quality of similarity computation can be investigated.

This thesis has proposed information content based approaches for measuring similarity among concepts belonging to two different ontologies. This similarity measure could be extended to more ontologies and the scalability
issues could be studied. Further this can also be tested for more application domains like document clustering, document searching and semantic information retrieval in Peer to Peer (P2P) systems.