CHAPTER 9

CONCLUSION AND FUTURE WORK

The content of the web grows rapidly every fraction of a second. Blogs and social networking sites contribute to a very large extent. Understanding the content of the web becomes essential for efficient searching and ranking of these millions of pages. In this thesis, we have identified and grouped specific events, and indexed them based on Time, Place and Person; the important event based attributes were learned with the help of the semi-supervised bootstrapping approach, and the vague and complex temporal expressions were identified and normalized into a standard format and the specific events were clustered based on Time, Place and Person. All the document processing tasks utilized the semantic based features to provide multi-field semantic search results. These semantic based approaches search events semantically by considering both query and document context. The ranking algorithm described in this thesis ranks the documents based on the conceptual relations of the query, conceptual links between pages, page rank value and the learned conceptual weight based on the user preference. The ranking mechanism described in this thesis modified the existing concept relation based index (CoReX) with the Lucene Multifield search facility. The Lucene Multifield indexing mechanism helps to search concepts with multiple fields and facilitate users to search news events from different event-based perspectives. This search engine has been developed for Tamil news corpus, and we can extend it to other languages without much effort. The language independent semantic representation helps the user to work across languages. The limitation of this work is to automatically build the UNL dictionary for the different domains and languages. Once we obtain
the UNL dictionary and the grammatical structure of the language, we can convert the documents into an UNL based semantic graph. The other document text processing tasks remain same. In semantic based event indexer, we need to additionally include a language identifier as a new field to retrieve the documents multilingual. Therefore, if we extend the system for other languages, the people of different language community can be benefited. The same system can be extended to other useful domains such as agricultural, health, etc. The scalability issues were also tested by increasing the number of documents up to 5 lakhs, and we do not find any performance related issues.

In temporal based information processing, we have used 6 temporal relations to identify and normalize the temporal expressions, which do not follow any standard notations to tag the temporal expressions. Thus, we have planned to use a standard event based representation like TimeML, for representing the temporal expressions of the news documents. Therefore, the semantic graph based representation needs to use the annotations provided by the TimeML. This helps us to link the time and event in a standard way, which can further improve the precision of the search result.

Event aggregation is another important research area in which the user can find the relationship of one event with another. For an event which may consist of several sub events, people are often interested to know the cause and effect of an event, the whole picture of an event along the time limit, the relationship between one event to the other based on person, place and time based similarity. Though, this search system provides a convenient way to search events with different perspectives, it does not provide a graph like interface to show the related events. In the future, we plan to extend our work on aggregating events, and to provide a semantic graph with related events for a given user query. This interface can provide a convenient way to
navigate from one event to the other, to know the specific relationship of an event.

In summary, this thesis presents a variety of useful techniques for event-based document processing tasks for news documents, irrespective of the domain. The impact of this research is, the same approach can be adopted for the other social media content analysis, like trend analysis, blogs, financial domain and weather forecasting to extract useful information for their purpose. We have developed semantic based methods to know the valuable information of the news event for any event types. Therefore, this event search engine helps the user to search events with multiple fields, and ranks the required documents based on user preference and its conceptual meaning of the document with respect to the query and the document. This can further be extended with event aggregation based interface that shows the link between events to explore more information about an event, so that the user can study the whole picture of an event easily.