論 文 の 内 容 の 要 旨

論文題目　Research on Efficient Similar Sentence Extraction
（効率的な類似文抽出の研究）

氏　名　顧　彦慧　Yanhui Gu

Measuring the semantic similarity between sentences is an essential issue because it is the basis for many applications, such as text summarization, Web page retrieval, question-answer model, image extraction, and so forth. A few studies have explored this issue by introducing several kinds of techniques, e.g., knowledge-based strategies, corpus-based strategies, and hybrid strategies. However, most of these works focus on how to improve the effectiveness of the problem, i.e., precision. It is well known that the era of big data has arrived and the huge volume of data makes the efficiency become a critical important issue for most real applications. To cope with the challenge, the main purpose of this thesis is to improve the efficiency performance of similar sentence retrieval, by introducing effective index structures and efficient querying algorithms. In this thesis, I conduct research on four main issues which are related to the problem of similar sentences retrieval: efficient similar sentences extraction; trade-off between effectiveness and efficiency on similar sentences retrieval; similarity measurement by incorporating additional semantic resources; and similar spatial textual objects retrieval.

Searching similar sentences from large datasets is an important issue because it is the basis for many applications. From a given sentence collection, this kind of queries asks for those sentences which are most semantically similar to a given one. To tackle the efficiency issue while keeping high effectiveness, I introduce a general framework based on the previous work (i.e., TKDD'08) which has the highest precision. The basic idea is that it integrates several representative similarity strategies together, i.e., the syntax based strategy and the semantic based technique. For the syntax based strategy, it is composed of three different kinds of similarity metrics, i.e., Normalized Longest Common Subsequence (NLCS), Normalized Maximum Consecutive Longest Common Subsequence starting at character 1 (NMCLCS1) and Normalized Maximum Consecutive Longest Common Subsequence starting at any character *n* (NMCLCS*n*). These effective strategies measure the string similarity between words, and two essential factors are the longest common subsequence and the length of the words. The difference among these three similarity techniques is the concrete format of the longest common subsequence. To evaluate the semantic based similarity between words, the Second Order

Co-occurrence Pointwise Mutual Information (SOC-PMI) has been introduced. To cope with the mentioned four similarity metrics, I propose four corresponding different indexing techniques. For NLCS, the similarity score is related to the longest subsequence and the length of the two measured words. Based on the property, I propose an effective index structure and present theoretical analysis to deduce a boundary that can be used to terminate the search process while testing the candidates, i.e., the process can be stop if the length of candidate is smaller than the boundary. The efficiency performance is improved through this mechanism, because only a small part of the data collection is necessary to be evaluated. In a similar way, I propose the corresponding effective index structures for NMCLCS1 and NMCLCSn, respectively. For the semantic based strategy, i.e., SOC-PMI, I build an index structure by sorting all the candidates in ascending order of their frequencies in the preprocessing and measure the similarity of candidates while testing them one by one. A boundary is dynamically estimated according to the so far tested candidates. When the current frequency of candidate is larger than this boundary, the process can be terminated and thus avoid to test the remaining words. To integrate the four similarity metrics, I introduce an efficient assembling approach, i.e., Threshold Algorithm (TA), to hasten the process. Moreover, the same technique can be applied when integrating words to sentences. Thorough experiments have been conducted to evaluate the proposed strategies. Specifically, I conduct experiments on two real dataset with regard to different aspects, i.e., the size of data collection, the value of k, and the size of query. It demonstrates that the proposed approach performs 10 times faster than the baseline strategy. The reason is that the baseline strategy needs to evaluate all the pairs between the query and sentence in the data collection while the proposed strategy only tests a small part of the candidates. Furthermore, the top-k results can be output progressively by the introduced approach. I also evaluate the effectiveness of both methods on the benchmark Miller-Charles dataset and find that the proposed strategy can keep the same high precision as that of the baseline strategy.

Effectiveness and efficiency are two important issues to evaluate the performance of a similar sentence retrieve system. In this thesis, I present theoretical analysis on the time and space cost of the introduced approaches. Specifically, the space complexity is related to the size of the data collection and the average length of sentences. The time complexity needs to consider both the offline preprocessing and the online query processing. In the offline preprocessing, because all the words of the data collection are needed to be indexed based on the order of each strategy (four corresponding strategies in the general framework), the offline time complexity is related to the size of data collection and the average length of sentences. In the online query processing, based on the property of the threshold algorithm that is employed, the time complexity is related to the sorted access and random access. Specifically, cost of the

sorted access is estimated by counting the total number of items that are sorted accessed. The cost of random access is measured by the total number of items that are randomly accessed. Consider on the precision, an intuitive idea is that more similarity metrics should be applied if higher precision is preferred. However, more features may cause larger cost on execution. To evaluate the trade-off between effectiveness and efficiency, I conduct a set of experiments to explore the intrinsic relation between them. I first compare the efficiency performance of single feature in the baseline strategy with the proposed whole framework. Then I evaluate the efficiency performance of single feature in the baseline framework with the single feature in my framework. To evaluate the effectiveness, I employ a similar method, i.e., the single (or combination) strategy in the baseline and the single (or combination) strategy in my proposed framework are thoroughly compared. Furthermore, I also evaluate the execution time and index construct time by using both strategies. The experimental results illustrate that the single strategy in the baseline performs better than the proposed whole framework on execution time while it is worse on the effectiveness. Moreover, the proposed single index strategy achieves higher efficiency performance compared with its rival in the baseline. In a brief summary, there exists a trade-off between the effectiveness and efficiency, i.e., more similarity features improve the precision yet consume more execution time. How to design a general effective and efficient framework is still a challenge issue in this area. Nevertheless, the introduced strategies can achieve the best efficiency performance, which is the main research purpose of the thesis.

There are many existing studies that aim at obtaining high precision for similar sentence retrieval. A common general approach is to apply additional representative semantic resources, i.e., Wiki and Wordnet. The basic idea is that the similarity between sentences can be leant from these large labeled (or semi-labeled) data. For the difference between these semantic resources, i.e, Wiki and Wordnet, the former has large volume data yet many noises are included, while the latter is well labeled and can provide precise information yet it is small. In this thesis, I incorporate the two semantic resources by utilizing the advantages of them. Moreover, when assembling different similarity metrics together, I introduce the cross validation strategy to improve the effectiveness. The reason is that for different kinds of data, the weights of the similarity metrics should be varying and they can be learned from the training data. Furthermore, I have observed the scenario that many words are not included in Wordnet because of its small size. A corresponding dynamic weight tuning technique is proposed for the situation that many words are missing in the Wordnet. The experimental evaluation on four labeled datasets (benchmark, MSRpar, MSRvid, SMT) demonstrates that the proposed strategy can increase the precision by up to 10%. Nevertheless, additional semantic resources may introduce more cost on execution. To clarify the scenario, I further evaluate the effectiveness and efficiency by employing different combination of the semantic resources. The experimental results show that

there is a trade-off between effectiveness and efficiency.

Effective and efficient retrieval of similar spatial textual objects plays an important role for many location based applications, such as Foursquare, Yelp, and so forth. Although there are many studies exploring on this issue, most of them focus on naive similarity measurement (i.e., string similarity), yet few of them address the semantic similarity issue. The semantic information is important because the textual description is commonly short in spatial textual system. Traditional word co-occurrence based strategy can hardly evaluate the similarity between these texts. In this thesis, I propose a semantic aware strategy which can effectively and efficiently retrieve the top-*k* similar spatial textual objects based on a general framework. The proposed semantic similarity aware strategy builds a comprehensive index, which seamlessly integrates the spatial and textual information together. Two different kinds of strategies have been introduced, one is IRS and the other is DIRS. IRS incorporates semantic information into an information R-tree. It is essentially an R-tree and incorporates textual information into each node of the R-tree. When a query is issued to the tree index structure, it applies best-first search strategy to retrieve the top-*k* spatial and textual results. However, IRS tree is constructed by only taking into account the spatial information. Textual information is also important in the tree construction. DIRS tree takes both spatial and textual information into consideration when constructing the index structure. Through this approach, DIRS retrieves the top-k spatial textual results by traversing rather small numbers of nodes. I conduct thorough experiments and evaluate the performance of the proposed strategies. The results show that the introduced techniques outperform the state-of-the-art approach, i.e., the precision is improved by up to 40% and the efficiency is about 2 times faster. The proposed strategy is more effective than the baseline strategy because it incorporates semantic information on measuring the similarity between short texts. As far as I know, this thesis is the first work that explores how to effectively and efficiently integrate the spatial proximity and semantic similarity together.