The proposal of WWW Data Science course finial project

214806 Changmin Li
214821 Zhengyu Wang
214822 Yehao Mao
214824 Yanting Cao
214857 Zhibiao Mei

Summary of the Proposal

The purpose of this proposal is to determine the topic selection of the final course report, which includes the study and analysis of each topic selection by our group members, and finally determines the reason for the selection of topic 1: Intelligent applications based on the knowledge graph. Then the proposal briefly introduces the relevant background knowledge of topic one and evaluates the possible problems in the implementation process. The third part is the effect we expect to achieve for the project, and then it is preliminarily determined based on the current research knowledge. Implementation method, teamwork mode, and division of labor among members. Finally, a timetable drawn up based on our own situation is used to urge us to complete the final course project and defense on time and quality.

Discussion

This part mainly explains our analysis of the difficulties in selecting the three topics after a period of investigation and gives the reason for our choice of topic one. In the first class of WDS, the professor provided us with three directions for the final course report. In less than three months, we investigated all three topics, consulted related documents, and simply tested some codes, and became familiar with some processes. Our group's analysis of the three selected topics is as follows:

- Intelligent applications based on the knowledge graph: The content of topic one is to construct a knowledge graph and apply some applications to it. The difficulty of this topic selection lies in the acquisition of data sets, the construction of the knowledge graph schema, and the invocation of deep learning-related models in the intelligent application. The advantage is that the project contains most of the content of the knowledge graph, which allows us to have an overall systematic understanding of the knowledge graph. In addition, this is also a very practical topic, which can exercise our ability from theory to application in the future. At the same time, the professor also gave very detailed step-by-step instructions, which greatly saved our time.

- Reproducing the SOTA models or methods: The second topic is the most difficult topic we all agree on. The topic selection requires a certain prior knowledge reserve and strong innovation ability. After discussion, we felt that the ability to change the topic was beyond our current level, so we did not choose the topic.

- Evaluating the SOTA models or methods: We struggled for a while between the third topic and the first topic. Because the difficulty of choosing topic three is relatively low, only need to reproduce SOTA models or methods and analyze the results. Although the analysis in the field of science is also a complex task, combined with previous experimental experience, the analysis we can do at this stage is usually relatively simple. The reason for our entanglement is that choosing topic three can exercise scientific research ability, which will be helpful for subsequent papers. But in the end, we chose the former in practical ability and scientific research ability.

Goal and Methods

This part briefly explains the main phases of the project and the general implementation process. The main process of topic one is as follows:

1. We need to determine the field of the application first, and the field given by the professor is limited to military-related, so this step does not actually require us to do anything.

2. The real first step for us is to obtain topic-related data. In this process, you can use crawlers to crawl information from related websites or search for related data sets on the Internet. However, considering that data on military topics may not be so easy to obtain, we may need to collect data in this step to improve the design of the schema, and manually annotate the data according to the schema.

3. Ontology building[5] is the focus of the preliminary work because this step determines the structure of our knowledge graph, and in the later period will reflect whether the question and answer process meet our expectations. The schema used in the construction of the ontology is based on the data we have collected. After we define a more comprehensive schema, we can build a database through knowledge extraction[3], knowledge fusion (optional), and knowledge storage[1], and realize the visualization of the knowledge graph[6]. The storage tool plan uses neo4j (graph database) and MongoDB (store unstructured data) recommended by the professor. The size of the final knowledge graph must meet the conditions of greater than 100,000 entities and greater than 1,000,000 triples.

4. After the knowledge graph is built, we need to implement some applications on it. At present, the main application forms of the knowledge graph we have learned are recommendation systems and intelligent question and answer. We intend to use the constructed knowledge graph to realize intelligent question answering. Intelligent question answering requires the program to recognize intent based on the user's input before it can give the most relevant answer by using the knowledge graph. The process of intent recognition has been investigated and learned that it can be implemented using BERT[2] and textCNN[4] in deep learning. It is temporarily planned to use this method, and the final implementation method should be based on the construction of the knowledge map.

5. In addition to completing the basic functions, we also need to analyze the knowledge map and the effect of the final application, investigate and understand other methods learned during the experiment, and optimize and improve the parts of the program that do not meet expectations.

6. The final result presentation method is the experiment report (PDF), the data set file (if too large, choose the form of picture display), the program source code, the system demonstration video (if the defense link can be demonstrated on-site), and the final defense PPT.

Responsibility

This part introduces our preliminary work model of responsibilities distribution. The work mode can be roughly divided into two types. One is that each person receives a piece of function in the project for realization, and finally connects the work content; the other is that all members implement all functions according to the research at the same time. The method we chose is the second one for the following reasons, which is, the team members work together to advance the progress:

1. The team is not as efficient as the first in tackling one problem at the same time, but it allows everyone to get exercise. What we are after should not be the result, but the ability to grow from the experiment process.

2. The method of implementing functions separately and then docking sounds very efficient, but the experience of past experiments tells us that there are often various problems in the docking process, and for this kind of highly dependent project at various stages, this model is not conducive to members. There is a high probability that each part will be too different from the other.

In fact, through the analysis of the previous part, our work model does not have to complete one before proceeding to the next. It may be more reasonable to use a pipeline similar to the process scheduling, and at the same time, the research direction of the team members can be taken into consideration. For example, we can construct the knowledge graph schema while collecting the data set, and the two can interact with each other. For example, in the process of constructing the knowledge graph, a small off-the-shelf database can be used to test the application of intelligent questions and answers, and so on. The specific division of labor is shown in the defense report and PPT.

Schedule

This part shows the tentative brief project timetable.
10-12 weeks Data set collection, ontology construction;
13-14 weeks Knowledge graph construction;
15-16 weeks Smart QA application;
17 weeks of analysis, report writing, preparation of defense materials, and final defense.

The above is the content of the second set of proposals for the final project of the World Wide Web Data Science course. If the professor has any questions or comments, you can discuss and exchange them through WeChat or email. Thank you for reading.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014.

[4] Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374, 2019.

[5] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[6] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.