UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft
Faculty for Computer Science, Electrical Engineering and Mathematics
Department of Computer Science
Research Group DICE

Project Plan

Submitted to the DICE Research Group

Knowledge Graph Summarisation

by

Muhammad Haseeb Javaid
Pavan Kumar Sheshanarayana
Shreyas Kottur Shivananda
Usman Ashraf

Supervised by:
Diego Moussallem

Paderborn, April 29, 2020

Contents

1 Project Plan

Knowledge graphs (KG) are graphs/structures consisting of facts in the form of triples. Triples are a combination of subjects, objects and predicates. In general, knowledge graphs are huge in size and may contain redundant information. Knowledge Graphs' humongous size makes it difficult to process and represent them. The goal of this project group is to obtain a pruned KG. This consists of selecting a sub knowledge graph based on endpoints and summarizing those subgraphs by considering facts based on ranks.

1.1 Stages

The project will be carried out in two stages over the course of two semesters.

1.1.1 Stage 1

In stage 1, tasks are classified between seminar phase and development phase as mentioned below. The Stage 1 would conclude at the end of first semester.

1.1.1.1 Seminar Phase

The seminar phase consists of Training and Planning.

1. Training: In the training phase, the project group will get themselves familiarized with a basic understanding of KG. The training phase is divided into five tasks:
   - Understanding the basic concepts of Semantic Web, RDF, OWL, etc.
   - Getting familiarized with the LD2NL framework.
   - Getting familiarized with the BENGAL framework, mainly concentrating on the part to obtain Sub KG from a given KG.
   - Understanding of few graph summarization algorithms.
   - Getting familiarized with the design of SPARQL queries.
2. Planning: In this phase, the team sets milestones for carrying out the project and respective roles for the individual team members. Goals, Project plan and individual responsibilities of team members are finalized.

1.1.1.2 Development Phase

In this phase, team worked for the realization of set goals in the previous phase. The team understood the need for having the hands-on experience on few of the concepts learnt in the previous phase and to get to know the basics of few ideas that were to be important going forward. To realize this, the team decided to develop a short POC (Proof of concept) where end-to-end development from the selection of entities belonging to a certain type (Person,Country or Organization) to the subsequent summarization of each of those type of entities using an algorithm were done. Here, PageRank was chosen as the summarization algorithm and the knowledge base was set to DBpedia. The summarized entities were also visualized and was shown as a part of intermediate presentation. In stage 2, a few other summarization algorithms would be considered with the option of extending the knowledge base to Wikidata.

1.1.1.2.1 PageRank Algorithm PageRank [TR16], a proven algorithm to provide objective relevance scores for hyperlinked documents is also one of the foremost successful algorithms in providing objective scores of entities.These scores find their importance in many areas such as Recommender systems and Question-Answering systems. It is a query independent algorithm which is used to analyse the link structures in directed graph. Since the DBpedia was based of from wikipedia data, it was naturally important to the team to be also able to understand and implement the algorithm which was widely used on the Wikipedia's documents.

1.1.2 Stage 2

In the second stage team will be summarizing knowledge graph by considering other algorithms such as SALSA and LinkSum and also will be extending the knowledge base to Wikidata. Further, evaluation of summarized graphs and User Interface (UI) enhancements will also be performed in this stage. Here tasks are classified into three phases namely:

- Seminar Phase
- Development Phase
- Document Phase

1.1.2.1 Seminar Phase

During this phase team will perform analysis and ground work required for algorithm implementations and validation of summarized graphs. Further, research for different summarization algorithms, feasiblity check and ground work required for the implementation of selected summarization algorithms will be carried out in this phase.

1.1.2.2 Development Phase

This phase will include the below mentioned tasks:

1.1.2.2.1 Implementation of algorithms:LinkSUM and SALSA After thorough research,the team found the algorithms LinkSUM and SALSA to be relevant.Hence the implementation of both of these algorithms would be the main aim of this task.

1. LinkSUM
   It is a lightweight link-based approach for the relevance-oriented summarization of knowledge graph entities which goes beyond the state of the art by addressing the following

observed limitations of previously developed methods: lack of general applicability (commercial approaches) and the inclusion of redundant information in a summary (commercial and research approaches).To address these challenges, LinkSUM [TLR16] combines and optimizes techniques for resource selection with approaches for predicate selection in order to provide a generic method for entity summarization. LinkSUM is focused on global relevance measures and does not rely on personal or contextual factors like individual interests or temporal trends.

2. SALSA (The Stochastic Approach for Link-Structure Analysis)
   SALSA presented in [LM01] is query dependent algorithm, which is used to perform link structure analysis in directed graphs. It works by considering a set of entities obtained by performing a query on knowledge base and entities obtained in this set are expanded to form a base set. Further entities in base set are classified has hubs and authorities. Authorities are the nodes pointed by many hubs and hubs are the nodes pointing to many authorities. Two scores namely authority score and hub score is assigned to nodes. A node is having a good authority score if it is pointed by hubs with good hub score and is suppose to have good hub score if it is pointing to authorities with good authority scores. Based on these scores entities are ranked. By considering these rankings knowledge base will be summarized.

1.1.2.2.2 Extension to WIKIdata In the previous stage, team used DBpedia as the sole knowledge base for summarization. Here,in this task, the team will also be utilising wikidata as the other knowledge base for summarization.

1.1.2.2.3 Evaluation of obtained summarised graphs After the implementation of the algorithms and essentially ,the summarization of knowledge graphs, the team will evaluate the obtained summarization using a valid evaluation method.

1.1.2.2.4 Improvements in UI and New Features Integration In this task,enhancements to the existing UI along with other additional features such as providing download capabilities of the summarized result will be realized by the team.

1.1.2.3 Documentation Phase

In this phase, the team will document all the activities carried throughout the project. The main goal of this phase to collaborate previously created documents and present them to the audience.

1.2 Organization

Project Group: Knowledge Graph summarization
Supervisor: Diego Moussallem
PG lead: Shreyas Kottur Shivananda
PG team: Muhammad Haseeb Javaid, Pavan Kumar Sheshanarayana, Usman Ashraf, Ammar Mustafa
Weekly Meeting: Monday 9:00 AM/Thursday 9:30 AM
Weekly Update: Status updates through a google form.
Code repository: Github
Communication Channel: Slack
Task Management: Trello

1.3 Milestones

In this section tentative milestone dates for the tasks included in phase 2 are provided in figure 1.1.

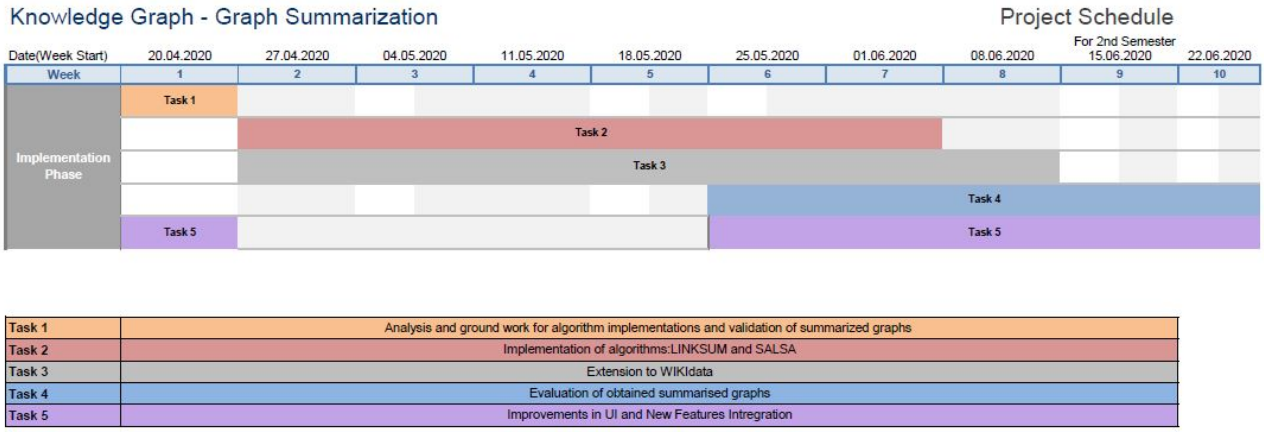

Figure 1.1: Milestone for Phase 2

Bibliography

[LM01] Ronny Lempel and Shlomo Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160, 2001.

[TLR16] Andreas Thalhammer, Nelia Lasierra, and Achim Rettinger. Linksum: using link analysis to summarize entity data. In *International Conference on Web Engineering*, pages 244–261. Springer, 2016.

[TR16] Andreas Thalhammer and Achim Rettinger. Pagerank on wikipedia: towards general importance scores for entities. In *European Semantic Web Conference*, pages 227–240. Springer, 2016.