Using AI and LLM techniques to find topics and visualize large amounts of research information

Rik D.T. Janssen, June 24, 2025

Graduation project description for University of Applied Sciences Utrecht, Education program HBO-ICT.

Description

In this graduation project, we use AI and LLM techniques to find topics and use these topics to visualize large amounts of research information. Research information is about anything related to research: research results, the persons in a research team, where they work, their collaborations, their skills, projects in which they have participated, as well as the relations between these entities. Examples of research results are publications, data sets, and software.

The tool to use in this graduation project is Ricgraph (www.ricgraph.eu), also known as Research in context graph. Ricgraph is software that is about relations between items. These items can be collected from various source systems and from multiple organizations. Insights can be obtained by combining information from various source systems, arising from new relations that are not present in each separate source system.

Usually, for Ricgraph, we use research information from multiple organizations and from multiple source systems. However, for this project, we only use OpenAlex (https://openalex.org) and research information from Utrecht University (https://openalex.org/works?page=1&filter=authorships.institutions.lineage:i193662353).

What problem or innovation is being addressed in this innovation project?

Authors assign keywords to their publications. These keywords are supposed to summarize the publication in a few words. However, they may be biased in some way. Also, for publications in the same research field that are about similar subjects, their author-assigned keywords may be very different.

In this graduation project, we will be using AI and LLM techniques to assign "topics" to publications. This means that there is a kind-of common "authority" assigning these topics, which probably gives a more consistent labeling. We can also use this approach to assign topics to research results that do not have keywords (sometimes data sets or software do not have them).

Having found these topics, we can use them to cluster research results in a meaningful way. If we select a number of topics, we will find all research results that correspond to all of these topics. Topics will also group the collaborators of research results. Usage scenarios can be researchers that use these topics to find colleagues who work on similar subjects, but who they do not know yet. This might be interesting for future collaborations, such as co-authoring future research. Other persons may use these topics to find experts in a certain field, since topics that group research results also group their authors (who are experts in the publications they wrote).

Research results can be from any field, even from research fields that do not exist yet. Therefore, we need to restrict the AI/LLM to use only a limited number of topics. We do this by using a thesaurus that has a hierarchy of broader and narrower terms (https://en.wikipedia.org/wiki/Thesaurus). We set a level in the thesaurus and limit the AI/LLM to use only terms from that level. For this project, we will use the thesaurus used for OpenAlex concepts (https://docs.openalex.org/api-entities/concepts).

Challenges are:

- How to get reproducible topics from AI/LLMs? How to prevent hallucinations?
- How many topics should be assigned to a research result?
- Which type of vector space should be used? What size of vectors? Why?
- Which open source and/or local LLM model should be used? What is the performance for topic finding when “small” and “large” models are compared? How about the speed?
- How can this process be done in an efficient way? How to store the (intermediate) results?
- How to visualize the topics found in Ricgraph Explorer? How to interact with the topics? What do users need? What do they want?
- What can the HU or their researchers learn from these topics? How can it be used to improve/expand research and education at HU?

What does the ideal end result look like?

In this project the open source software Ricgraph (https://github.com/UtrechtUniversity/ricgraph) is used. All results from this project will be open.

This project consists of two parts:

1. Finding topics using AI and LLM techniques.
2. Visualizing these topics in Ricgraph Explorer.

Every sprint review, a working topic finder and visualization user interface are to be demonstrated.

Ad 1
For finding the topics, the open source software Ricgraph-topics (https://github.com/UtrechtUniversity/ricgraph-topics) will be used. This software is written in Python.

The main challenge of this part is to find meaningful topics in a chosen level in the thesaurus (see above).

Basically, this part results in a “box” that takes a list of DOIs (https://en.wikipedia.org/wiki/Digital_object_identifier) as input, and produces DOIs and their topics as output. This can easily be converted to a format that can be imported in Ricgraph.

The challenge is to “optimize” this box (many of these will need to be configurable by using a config file). That means:

- Research on the LLM. Options are: what LLM models to use; what vector space to use; what size of the vectors in the vector space to use; what prompts to use; what temperatures to use; whether to use the title and full abstract as input for the LLM, or the separate sentences in the abstract; whether to combine several methods; etc.
- Research on the thesaurus. Options are: exploring the influence of the level chosen in the thesaurus; exploring alternative thesauri.
- Improve the ricgraph-topics software.

As a measure of quality we use reproducibility: for the same research result, every time the LLM is run, the same topics should be produced, even if the runs are weeks or months apart. Also, research results that are similar, should produce mostly the same topics. At some points in time, experts need to be involved to check the usefulness of the topics generated.

Ad 2
The visualizations need to be implemented in Ricgraph Explorer, part of Ricgraph. This software is written in Python, and has some JavaScript parts.

The main challenge of this part is to create and implement a number of meaningful visualizations and user interfaces. Preferably they have been created after consultation with researchers, and they have been tested on their usefulness.

There will be at least two types of visualization of topics and results:

A. For research results. A web page that allows the user to choose any number of topics on the left side of the page. After a user has chosen a topic on the left side, the right part of the web page updates the results that correspond to all of the selected topics. Then, the user can choose an additional topic, etc.

B. For persons. Similar to A, but now the right part shows persons that have coauthored research results that correspond to the topics chosen.

Besides these two user interfaces, we welcome new approaches to show and interact with the topics.

To which technologies and/or concepts will the students be exposed?

- AI, LLM
- topics
- thesauri; level chosen in a thesaurus
- graphs
- visualization
- reasoning techniques
- software development, software efficiency
- working with large data sets
- user interface design, user interfaces
- interaction with researchers
- documentation writing