Master Project:

Keywords Extraction and Visualization

University of Zurich[UZH]

Context

Topic detection and visualization techniques are commonly used to study and analyze the scientific literature to gain deeper understandings of, say, the long term research trends, or relations among different research topics [1]. Specifically, co-word analysis is of particular interest in machine learning and data visualization communities due to two main challenges: First, extracting meaningful and accurate keywords (keyphrase) from a document is a non-trivial task. Second, the complex relations of keywords with other literature-related information (citation, co-author, etc.) are challenging to analyze and visualize. In this project, we will focus on keywords extraction and visualization.

The most relevant work in this specific topic is by Isenberg et al. [1] where they analyzed all the IEEE VIS publications from 1990 to 2015. However, they only analyzed the keywords as provided by the paper authors, which may fail to faithfully reveal the topics and research methodologies used in all the papers. To address this limitation, in this project, we would like to apply keywords extraction techniques from machine learning community to help extract meaningful keywords from VIS journals and conferences, as well as a preliminary visualization prototype to analyze the extracted keywords.

Assignment

In this project, first, titles, abstracts and keywords information will be downloaded from all the VIS journals and conferences for the last 20 decades. Then, different keywords extraction algorithms (e.g., [2]) will be studied and applied to extract meaningful keywords. Third, follow the paper [1] and conduct a data analysis for those keywords, compare your results with the results from that paper.

In the end, the extracted keywords will be visualized for knowledge discovery. Two types of visualization techniques should be applied: word cloud and streamgraph [3].

Requirements

Preferably you have taken the BINF4234 Data Visualization Concepts course, with some machine learning background. The main part of this project will be conducted in Python, then either a Python or JavaScript visualization libraries will be used for the visualization.

Work Load

- 20% theory
- 60% implementation
- 20% visual analysis

Student Project Type

This topic can be done as a Master Project in a group of 2-3 persons. Goals are adjusted depending on the project type and the number of students.

Supervision

Prof. Dr. Renato Pajarola
Haiyan Yang (assistant)

Contact

Write an E-Mail to haiyan@ifi.uzh.ch

References

[1] Isenberg, Petra, et al. "Visualization as seen through its research paper keywords." *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2016): 771-780.

[2] Keyword Extraction: from TF-IDF to BERT: "https://towardsdatascience.com/keyword-extraction-python-tf-idf-textrank-topicrank-yake-bert-7405d51cd839"

[3] Byron, L., & Wattenberg, M. (2008). Stacked graphs–geometry & aesthetics. *IEEE transactions on visualization and computer graphics*, *14*(6), 1245-1252.

VISUALIZATION AND MULTIMEDIA LAB

Prof. Dr. Renato Pajarola
Visualization and MultiMedia Lab
Department of Informatics
University of Zürich