University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestrasse 14
CH-8050 Zürich
Phone +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch

UZH, Department of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Renato Pajarola
Visualization and MultiMedia Lab
Phone +41 44 635 43 70
Fax +41 44 635 68 09
pajarola@ifi.uzh.ch
vmml.ifi.uzh.ch

Master Project (18 ECTS)

Keywords Extraction and Visualization

Names: Chenglong Lei, Peng Yan

Student ID: 19763309, 19762780

From: 18 Mar 2021 To: 13 Sep 2021

Introduction

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 [1,2,3]. 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 (key-phrase) 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 our 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.

Description

The main goal of this project is to extract and analyze the keywords from VIS papers from 2000 to 2020. The papers prepared for this project are from Transactions of Visualization and Computer Graphics (TVCG), EuroVis and PacificVis. This project contains three main stages which are described below in detail:

1. Keywords extraction using available algorithms mentioned in [4] (12w)
   - 1.1. *Extract initial keywords from the title and abstract using existing algorithms.*
   - 1.2. *Construct the refined keywords set by combining the keywords extracted and the keywords directly given in the paper. Clean this dataset by excluding the influence of singulars/plurals, spelling mistakes, acronyms, etc.*
   - 1.3. *Compare the results from different algorithms and choose the one that has the best performance.*
   - 1.4. *Do multiple passes until the final keywords dataset is clean.*
   - 1.5. *Each student has to implement one algorithm from these four representative candidates: EmbedRank, TopicRank, YAKE, BERT [4].*
2. Keywords analysis as in paper [1] (6-8W)

4

2.1. Build keywords-document matrix and compute the correlation matrix.

2.2. Perform hierarchical clustering on the correlation matrix based on Ward's method.

2.3. Generate a keyword network where keywords are linked if their correlation is larger than 0.

2.4. From this network, compute the density of each cluster.

2.5. Build a co-occurrence matrix of all the keywords. Split all the keywords into two sets: one problem set and one visualization technique set, then build a co-occurrence matrix of these two sets. Reorder these two matrices by applying the algorithm in [6] and report your findings.

3. Keywords visualization (4-8W)

3.1. Implement a keywords word cloud for each year and visualize all the word clouds' temporal evolution.

3.2. Implement a sorted streamgraph [5] in which streams are vertically sorted based on the papers' influence (citation).

3.3. Each student should implement one visualization.

Remarks

Application development will be done using Python and JavaScript. The applicant is expected to provide a written report according to the rules of the IfI and give two oral presentations (including a live demo) of the implemented results. The first presentation will happen in the middle of the project as a midterm presentation. The code itself belongs to the deliverables of this work. In addition, the presentation slides, screenshots, and video of the application showing the features should be delivered. The project is supervised by Prof. Dr. Renato Pajarola and Haiyan Yang.

References

[1] Isenberg, Petra, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. "Visualization as seen through its research paper keywords." *IEEE Transactions on Visualization and Computer Graphics* 23, no. 1 (2016): 771-780.

[2] Matejka, Justin, Tovi Grossman, and George Fitzmaurice. "Citeology: visualizing paper genealogy." In *CHI'12 extended abstracts on human factors in computing systems*, pp. 181-190. 2012.

[3] Liu, Yong, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. "CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 3553-3562. 2014.

[4] Papagiannopoulou, Eirini, and Grigorios Tsoumakas. "A review of keyphrase extraction." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.2 (2020): e1339.

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

[6] Behrisch, Michael, et al. "Matrix reordering methods for table and network visualization." *Computer Graphics Forum*. Vol. 35. No. 3. 2016.