UMEÅ UNIVERSITET

Evaluating Document Clusters through Human Interpretation

Anton Eklund

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Lindellhallen 3, UB.A.230,
torsdagen den 3 april, kl. 13:15.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Maria Veronika Laippala, School of Languages and Translation Studies, University of Turku, Finland.

Department of Computing Science

Organization
Umeå University
Dept. of Computing Science

Document type
Doctoral thesis

Date of publication
13 March 2025

Author
Anton Eklund

Title
Evaluating Document Clusters through Human Interpretation

Abstract
Document clustering is a technique for organizing and discovering patterns in large collections of text, often used in applications such as news aggregation and contextual advertising. An example is the automatic grouping of news articles by theme, which is the focus of this thesis. For a clustering to be successful, typically the resulting clusters need to appear interpretable and coherent to a human. However, there is a lack of efficient methods to reliably assess the quality of a clustering in terms of human-perceived coherence, which is essential for ensuring its usefulness in real-world applications.

To address the lack of evaluation methods for document clustering focusing on human interpretation, we introduced Cluster Interpretation and Precision from Human Exploration (CIPHE). CIPHE tasks human evaluators to explore document samples from a cluster and collects their interpretation. The interpretation is collected through a standardized survey and then processed with the framework metrics to yield the cluster precision and characteristics. This thesis presents and discusses the development process of CIPHE. The feasibility of performing the exploratory tasks of CIPHE in a crowdsourcing environment was investigated, which resulted in insights on how to formulate instructions. Additionally, CIPHE was confirmed to identify characteristics other than the main theme such as the negative emotional response.

CIPHE was paired with a standard clustering pipeline to evaluate its capabilities and limitations. The pipeline is widely applied for its adaptability and conceptual simplicity, and also being part of the popular topic model BERTopic. The empirical results of applying CIPHE suggest that the pipeline, when integrated with a Transformer-based language model, generally yields coherent clusters.

Additionally, topic models have a similar aim as document clustering which is to automate the corpus processing and present the underlying themes to a human. Topic modeling has rich research on the human interpretation of topic coherence. In the thesis, the human interpretation collected with CIPHE was related to established research in topic coherence. Specifically, the human interpretation collected with CIPHE was used to highlight limitations with the keyword representations that topic coherence evaluation relies on.

Keywords
document clustering, topic modeling, information retrieval, human evaluation, human-in-the-loop, news clustering, natural language processing, topic coherence, human interpretation

Language
English

ISBN
print: 978-91-8070-646-9
PDF: 978-91-8070-647-6

ISSN
0348-0542

Number of pages
44 + 5 papers