Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics

Clustering of scientific papers for easy information retrieval

Team Member: Daniela Fichiu, 3552717, BSc Applied Computer Science, BSc Mathematics
daniela.fichiu@stud.uni-heidelberg.de
Team Member: Christian Homeyer, 3606476, PhD Computer Science
ox182@uni-heidelberg.de
Team Member: Jessica Kaechele, 3588787, MSc Applied Computer Science
uo251@stud.uni-heidelberg.de
Team Member: Jonas Reinwald, 3600238, MSc Applied Computer Science
am248@stud.uni-Heidelberg.de

GitHub Repository: `https://github.com/DonatJR/ita_ws20`

1 Motivation

While typing a query like "how do I open the terminal on mac?" into Google, one already has an exact idea of what he wants to find: a set of steps that will lead to a terminal being displayed on the screen.

When starting a new project or trying to improve existing bodies of work it is often vitally important to do extensive research. This both helps to familiarize oneself with the topic that is to be worked on and provides the necessary new information that is needed to solve the problem at hand.

A student who wants to know more about text analytics might search for "text analytics" on Google Scholar. The query yields back 1.650.000 results, the top three being "Text Analytics with Python", "Text analytics in social media" and "Semantic interaction for visual text analytics". A subsequent query like "text analytics topics" yields back results with the following titles: "Analyzing educational comments for topics and sentiments: A text analytics approach" or "A text analytics approach for online retailing service improvement: Evidence from Twitter".

We know how time-consuming it is to spend hours on search engines or websites like Research Gate or Google Scholar looking for research papers, hoping for the best, but never quite finding what we are looking for.

Our project emerged from the need of finding an easy way of exhaustively searching for scholarly literature, while bringing to light the relations between the subfields of the research field of interest and/or of other fields, thus providing a deeper understanding of the material to be searched for and easing up the process of finding information.

We want to propose a solution to this by taking scientific texts and clustering them into relevant subgroups, which can then be more easily presented to and explored by people looking for specific topics and terms.

While we intend to first focus on a rather narrow subset of papers from a topic like Deep Learning, more or less specific subgroups for clustering therein can be explored to find a good balance. The proposed pipeline should later be usable on a broader range of fields.

Furthermore, we think a solution to the stated problem could be used on a grander scale by building good group visualization tools and providing existing websites with this technology.

In addition, a meta search site incorporating data from these other sources with a common, easily digestible search and presentation interface could be developed.

2 Research Topic Summary

There exist various approaches to simplify getting an overview of a research field without extensive research and reading countless papers. Browsing through digital libraries and search engines can yield papers matching search strings, but to get an overview of an entire research area this is not good enough.

In [10] a tool is presented that uses the citation network to divide papers into clusters and identify trends, gaps and outliers. Another way to gain insights is to find key statements of a paper. As finding these statements is computationally expensive, the authors use 'Multi-Document Summarization' which is only applied to abstracts and citation contexts. The dataset used is ACL Anthology Network (AAN) [1] and contains the network of citations, as well as the full text of each article, its metadata, summary, references and citation sentences.

These techniques make it possible to gain insight into an entire research field more quickly, but for this project we will concentrate only on the clustering. Attempts to cluster papers have been made for many years. In 1973, for example, it had already been tried to cluster journals by comparing reference patterns and looking at mutual references [9].

In [8] the context of the citations is used in addition to the citations itself to cluster. Citations are identified and text around it is extracted to then use link-based clustering approaches, term-based clustering approaches and hierarchical document clustering, as well as a combination of all three, on this data. For comparison, this technique is also applied to the entire document and contrasted against the approach of using only citation context.

In [11] abstracts and titles of documents from Web of Science [5] are used. They perform two types of pre-processing, the first treats each word as a token, and stopwords are deleted. The second method uses term-clumping to find noun phrases with significant commonality. Then, several topic modeling algorithms are used: The Latent Dirichlet allocation (LDA), Correlated Topic Models (CTM), Hierarchical Latent Dirichlet Allocation (Hierarchical LDA) and Hierarchical Dirichlet Process (HDP).

Abstracts are also used in [7] to perform clustering. They use tokenization, remove stopwords and then apply a stemming algorithm. Keywords are then grouped and weighted and the closeness of two documents is calculated using cosine similarity measure. Additionally, clustering methods are applied to the whole abstract. Three algorithms from three different approaches are used: the k-medoid method from the example-based approach, the nearest neighbor method from the hierarchy-based approach and the MajorClust method from the density-based approach. The data source consists of 48

human classified abstracts from [2].

There are multiple approaches we could take in our own project. The final goal is of course to achieve a better clustering than previous attempts. In regards to [7], [8] and [11] we can leverage abstracts and citation contexts for clustering in favor of only either one of them. There is also an opportunity in taking different clustering techniques (e.g. k-means, hierarchical clustering or one of the ones mentioned above) and comparing their respective performance with each other. In addition we could also use a neural network architecture to either automate the process of generating word vectors from the texts or to automate the clustering from word vectors that are obtained in a more traditional fashion, but this is more of a bonus goal given the time constraints.

3 Project Description

3.1 Main project goals

Our main goal is to make it easier for users to search and explore scientific papers belonging to a specific topic or theme. For this we want to specifically arrive at a clustering (representation) that, for one, separates the different documents into correct clusters, but is also easy to work with in downstream tasks (e.g. the mentioned inclusion in some search site). To achieve this we basically interpret the steps mentioned in the subsection 3.2 as some coarse sub goals, which can then be worked on by different team members. Some of these sub goals can also be further divided, for example downloading and preprocessing data from different sources or implementing distinct clustering algorithms can be done by a single team member respectively.

3.2 Pipeline

Our pipeline will consist of the stages that can be seen in figure 1. First of all data from two different sources will be downloaded, temporarily stored and cleaned up to disregard records with missing data. All cleaned up records will be stored permanently so we don't have to download and clean up the data every time we change some downstream settings. From this data storage the text we are going to use in the end will be extracted and some clustering algorithms will be used to build the final result.

3.3 Data Set

We have two main and one backup source of scientific papers:

- Crossref [3] is an official Digital Object Identifier Registration Agency that provides access to the full text of its registered research papers through their own API. Sending a HTTP request with "deep learning" as a query returns a JSON object with all the URLs to the PDFs of the registered papers found in Crossref's data base containing the terms "deep learning" in their title. The PDFs will then have to be converted into plain text.
- All About NLP [1] is a website maintained by Yale University's Learning and Information Group that provides a corpus consisting of over 400 scientific papers on NLP in plain text. Closer examination of the text files has, however, shown that most of them contain spelling mistakes.
- Journal of Machine Learning Research [4] is an international forum for the electronic and paper publication of scholarly articles in all areas of machine learning. The papers are available in pdf format, while the abstracts are in html format.

3.4 Evaluation

It is non-trivial to characterize clustering metrics for evaluation. Since we will not use anything out of the ordinary, we refer to [6]. In general we can distinguish between *internal* and *external* measures. This directly corresponds to degrees of supervision.

Unsupervised Clustering. Data is divided into clusters without knowing ground truth clusters, such that data inside share max. similarity w.r.t. a specific data attribute. Classification requires knowing the label set $\mathcal{L}$, which we do not know in this case. Example algorithms for unsupervised clustering are: *K-Means*, *Mean shift*, *Expectation-Maximization* and many more. Efficient implementations exist in *sklearn* and similar Python libraries.

Because no ground truth exists, internal measures are highly dependent on the application. Typical internal measures depend on the number of clusters, intra- and inter-cluster distance distributions. Examples are the *Davies-Bouldin index*, *Dunn index* or the *Silhouette coefficient*. However, we usually do not know much about the underlying data, which makes it hard to compare algorithms this way as internal measures only indicate if one algorithms is better than another in some situations. In the end, we might however not know if an algorithms produces more valid results than another.

Supervised Clustering. Supervision for clustering research papers is non-trivial. There are several possible solutions for computing a supervisory signal. One of the general ideas of this project was to exploit the standardized submission process in academic journals/conferences. Authors need to submit a list of key words that best describe their work. On top, all documents follow structural guidelines for their specific journal/conference. This gives us the chance to use the standardized format to simplify processing. Furthermore, we can exploit the labels of platforms hosting our data. Articles are usually already grouped according to their topics. We propose three supervision modes that we are planning to try out:

1. Create hard labels, by using the preexisting group assignments of hosting websites

2. Create labels from provided key words upon submission. We could create a label set across the data and perform classification on our training set. This would allow to a) assign likely key words to papers b) cluster papers according to assigned key words. A down-stream task that benefits from our improvements would be key word generation.

3. Generate hard assignments based on key words. By assigning cluster labels to key words, we could compute a hard cluster from a combination of key words for a given paper

Supervised clustering allows validating an algorithm based on a dataset split and the ground truth labels. Typical evaluation metrics include: *Precision*, *Recall*, *F1-score*, *Jaccard index* or a *confusion matrix*. The supervision would be needed to achieve our bonus goal with a learning algorithm. We will compare our algorithms with baselines from the related works w.r.t to the mentioned metrics.

References

[1] Acl anthology network (aan). http://aan.how/.

[2] Cicling: International conference on computational linguistics and intelligent text processing. https://www.cicling.org/.

[3] Crossref. https://www.crossref.org.

[4] The journal of machine learning research (jmlr). https://jmlr.csail.mit.edu.

[5] Web of science. www.webofknowledge.com.

[6] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[7] Mikhail Alexandrov, Alexander Gelbukh, and Paolo Rosso. An approach to clustering abstracts. In Andrés Montoyo, Rafael Muńoz, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pages 275–285, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[8] Bader Aljaber, Nicola Stokes, James Bailey, and Jian Pei. Document clustering of scientific texts using citation contexts. *Inf. Retr.*, 13:101–131, 04 2010.

[9] Mark P. Carpenter and Francis Narin. Clustering of scientific journals. *Journal of the American Society for Information Science*, 24(6):425–436, 1973.

[10] Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.

[11] c-k Yau, Alan Porter, Nils Newman, and Arho Suominen. Clustering scientific documents with topic modeling. *Scientometrics*, 100:767–786, 09 2014.

Appendix A Figures

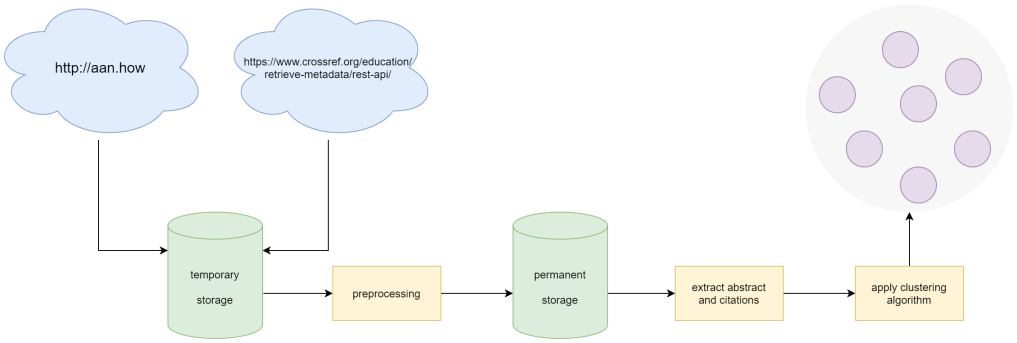

Figure 1: Coarse overview over the intended pipeline.