Nam, Jinseok; Kirschner, Christian; Ma, Zheng; ...

Knowledge discovery in scientific literature

Ruppenhofer, Josef [Hrsg.]; Faaß, Gertrud [Hrsg.]: Proceedings of the 12th edition of the KONVENS Conference. Hildesheim : Universitätsverlag Hildesheim 2014, S. 66-76

Quellenangabe/ Reference:
Nam, Jinseok; Kirschner, Christian; Ma, Zheng; Erbs, Nicolai; Neumann, Susanne; Oelke, Daniela; Remus, Steffen; Biemann, Chris; Eckle-Kohler, Judith; Fürnkranz, Johannes; Gurevych, Iryna; Rittberger, Marc; Weihe, Karsten: Knowledge discovery in scientific literature - In: Ruppenhofer, Josef [Hrsg.]; Faaß, Gertrud [Hrsg.]: Proceedings of the 12th edition of the KONVENS Conference. Hildesheim : Universitätsverlag Hildesheim 2014, S. 66-76 - URN: urn:nbn:de:0111-dipfdocs-182323 - DOI: 10.25657/02:18232

https://nbn-resolving.org/urn:nbn:de:0111-dipfdocs-182323
https://doi.org/10.25657/02:18232

Knowledge Discovery in Scientific Literature *

Jinseok Nam [1,2], Christian Kirschner [1,2], Zheng Ma [1,2], Nicolai Erbs [1], Susanne Neumann [1,2]
Daniela Oelke [1,2], Steffen Remus [1,2], Chris Biemann [1], Judith Eckle-Kohler [1,2]
Johannes Fürnkranz [1], Iryna Gurevych [1,2], Marc Rittberger [2], Karsten Weihe [1]
[1] Department of Computer Science, Technische Universität Darmstadt, Germany
[2] German Institute for Educational Research, Germany
http://www.kdsl.tu-darmstadt.de

Abstract

Digital libraries allow us to organize a vast amount of publications in a structured way and to extract information of user's interest. In order to support customized use of digital libraries, we develop novel methods and techniques in the Knowledge Discovery in Scientific Literature (KDSL) research program of our graduate school. It comprises several sub-projects to handle specific problems in their own fields. The sub-projects are tightly connected by sharing expertise to arrive at an integrated system. To make consistent progress towards enriching digital libraries to aid users by automatic search and analysis engines, all methods developed in the program are applied to the same set of freely available scientific articles.

1 Introduction

Digital libraries in educational research play a role in providing scientific articles available in digital formats. This allows us to organize a vast amount of publications, and the information contained therein, in a structured way and to extract interesting information from them. Thus, they support a community of practices of researchers, practitioners, and policy-makers. In order to support diverse activities, digital libraries are required to provide effective search, analysis, and exploration systems with respect to specific subjects as well as additional information in the form of metadata.

Our analysis is mainly focused on the educational research domain. The intrinsic challenge of knowledge discovery in educational literature is determined by the nature of social science, where the information is mainly conveyed in textual, i.e., unstructured form. The heterogeneity of data and lack of metadata in a database make building digital libraries even harder in practice. Moreover, the type of knowledge to be discovered that is valuable as well as obtainable is also hard to define. As this type of work requires considerable human effort, we aim to support human by building automated processing systems that can provide different aspects of information, which are extracted from unstructured texts .

The rest of this paper is organized as follows. In Section 2, we introduce the Knowledge Discovery in Scientific Literature (KDSL) program which emphasizes developing methods to support customized use of digital libraries in educational research contexts. Section 3 describes the sub-projects and their first results in the KDSL program. Together, the sub-projects constitute an integrated system that opens up new perspectives for digital libraries. Section 4 finally concludes this paper.

2 Knowledge Discovery in Scientific Literature

In the age of information overload, even research professionals have difficulties in efficiently acquiring information, not to mention the public.

An accessible, understandable information supply of educational research will benefit not only the academic community but also the teachers, policy makers and general public.

There are several related research projects. The CORE (Knoth and Zdrahal, 2012) project aims to develop a system capable of seamless linking of existing repositories of open access scientific papers. The CODE project developed a platform which facilitates exploration and analysis in research areas using open linked data.[1]

In contrast to general-purpose systems for managing scientific literature, we aim at building a system in specific domains including, but not limited to, the educational research where, for instance, users are allowed to navigate visually a map of research trends or are provided with related works which use the same datasets.

2.1 Structure of KDSL

The KDSL program is conducted under close collaboration of the Information Center for Education (IZB) of the German Institute for International Educational Research (DIPF) and the Computer Science Department of TU Darmstadt. IZB provides modern information infrastructures for educational research. It coordinates the German Education Server and the German Education Index (*FIS Bildung Literaturdatenbank*).[2]

Consisting of several related sub-projects, the KDSL program focuses on text mining, semantic analysis, and research monitoring, using methods from statistical semantics, data mining, information retrieval, and information extraction.

2.2 Data

All of our projects build up on the same type of data which consists of scientific publications from the educational domain. However, the publications differ from each other in their research approach (e.g., empirical/theoretical and qualitative/quantitative), in their topics and in their target audience / format (e.g., dissertations, short/long papers, journal articles, reviews). This leads to a vast heterogeneity of content which also follows from the broad range of disciplines involved in the educational research (for example psychology, sociology and philosophy).

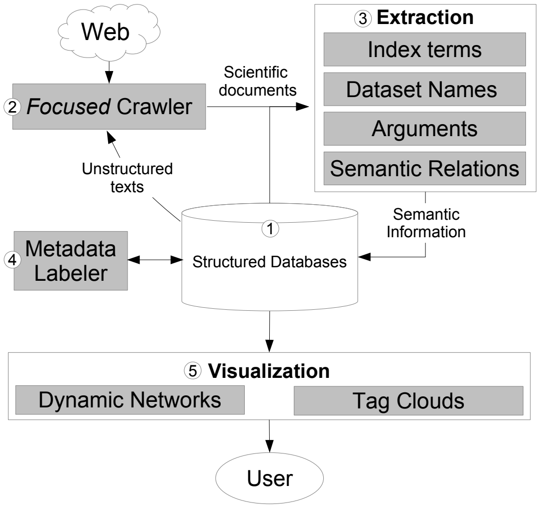

Figure 1: Links between sub-projects in KDSL for educational research

At DIPF, there are mainly two databases containing relevant publications for our projects: *pedocs* and *FIS Bildung*. *FIS Bildung* (Carstens et al., 2011) provides references to scientific articles collected from more than 30 institutions in all areas of education. Specifically, the database consists of over 800,000 entries and more than a half of them are journal articles in German. One-third of the references to articles published recently has full-text in a pdf format.[3] *pedocs* (Bambey and Gebert, 2010), a subset of *FIS Bildung*, maintains a collection of open-access publications and makes them freely accessible to the public as a long-term storage of documents. As of today, the total number of documents in *pedocs* is about 6,000.[4] Each entry in both databases is described by metadata such as title, author(s), keywords and abstract.

2.3 Vision and Challenges

The overall target of KDSL is to structure publications automatically by assigning metadata (e.g., index terms), extracting dataset names, identifying argumentative structures and so on. Therefore, our program works towards providing new

[1] `http://code-research.eu`

[2] `http://www.fachportal-paedagogik.de`

[3] Detailed statistics can be found at `http://dipf.de/de/forschung/abteilungen/pdf/diagramme-zur-fis-bildung-literaturdatenbank`

[4] April 2014

methods to identify and present the information searched by a user with reduced effort, and to structure the information regarding the specific needs of the users in searching the mentioned databases.

Figure 1 shows how the sub-projects interact with each other to achieve our goal. Each sub-project in KDSL acts as a building block of the targeted system, i.e., an automated processing system to help educational researchers. Getting more data, even unlabeled (or unannotated), is one of the key factors which lead to more accurate machine learning models. The focused crawler collects documents from websites in educational contexts (block ② in Fig. 1). Other sub-projects can benefit from a large corpus of the crawled documents that might provide more stable statistics in making predictions on unseen data. By using structured databases and the crawled documents, we perform several extraction tasks (block ③), such as identifying index terms (Sec. 3.2, 3.5), dataset names (Sec. 3.3), argumentative structures (Sec. 3.4), and semantic relations between entities (Sec. 3.1). Towards the enrichment of databases, we investigate methods to assign the extracted information in structured formats, i.e., metadata (block ④). In turn, we also aim at providing novel ways to visualize the search results and thus to improve the users' search experience (block ⑤), for instance through displaying dynamics of index terms over time (Sec. 3.6) and tag clouds (Sec. 3.7).

3 Projects

In the following sections, we describe sub-projects in KDSL with regards to their problems, approaches, and the first results.

3.1 Crawling and Semantic Structuring

A vital component of the semantic structuring part of this project is the process of reliably identifying relations between arbitrary nouns and noun phrases in text. In order to achieve high-quality results, a large in-domain corpus is required.

Task The corpus necessary for unsupervised relation extraction is created by enlarging the existing *pedocs* corpus (cf. Sec. 2.2) with documents from the web that are of the same kind. The project's contribution is thus twofold: *a*) focused crawling, and *b*) unsupervised relation extraction.

Dataset Plain texts extracted from *pedocs* pdfs define the domain of the initial language model for a focused crawler (Remus, 2014).

Approaches The *Distributional Hypothesis* (Harris, 1954), which states that similar words tend to occur in similar contexts, is the foundation of many tasks including relation extraction (Lin and Pantel, 2001). Davidov et al. (2007) performed unsupervised relation extraction by mining the web and showed major improvements in the detection of new facts from only few initial seeds. They used a popular web search engine as a major component of their system. Our focused crawling strategy builds upon the idea of utilizing a *language model* to discriminate between relevant and irrelevant web documents. The key idea of this methodology is that web pages coming from a certain domain — which implies the use of a particular vocabulary (Biber, 1995) — link to other documents of the same domain. The assumption is that the crawler will most likely stay in the same topical domain as the initial language model was generated from.

Using the enlarged corpus, we compute distributional similarities for entity pairs and dependency paths, and investigate both directions: a) grouping entity pairs, and b) grouping dependency paths in order to find generalized relations. Initial results and further details of this work can be found in (Remus, 2014).

Next Steps Remus (2014) indicates promising directions, but a full evaluation is still missing and still has to be carried out. Further, we plan to apply methods for supervised relation classification using unsupervised features by applying similar ideas and methodologies as explained above.

3.2 Index Term Identification

In this section, we present our analysis of approaches for index term identification on the *pedocs* document collection. Index terms support users by facilitating search (Song et al., 2006) and providing a short summary of the topic (Tucker and Whittaker, 2009). We evaluate two approaches to solve this task: (1) index term extraction and (ii) index term assignment. The first one extracts index terms directly from the text based on lexical characteristics, and the latter one assigns index terms from a list of frequently used index terms.

Task Approaches for index term identification in documents from a given document collection find important terms that reflect the content of a document. Document collection knowledge is important because a good index term highlights a specific subtopic of a coarse collection-wide topic. Document knowledge is important because a good index term is a summary of the document's text. Thesauri which are available for English are not available in every language and less training data may be available if index terms are to be extracted for languages other than English.

Dataset We use manually assigned index terms, which were assigned by trained annotators, as a gold standard for evaluation. We evaluate our approaches with a subset of 3,424 documents.[5] Annotators for index terms in *pedocs* were asked to add as many index terms as possible, thus leading to a high average number of index terms of 11.6 per document. The average token length of an index term is 1.2. Hence, most index terms in *pedocs* consist of only one token but they are rather long with on average more than 13 characters. This is due to many domain-specific compounds.

Approaches We apply index term extraction approaches based on tf-idf (Salton and Buckley, 1988) using the *Keyphrases* module (Erbs et al., 2014) of *DKPro*, a framework for text processing,[6] and an index term assignment approach using the *Text Classification* module, abbreviated as *DKPro TC* (Daxenberger et al., 2014). The index term extraction approach weights all nouns and adjectives in the document with their frequency normalized with their inverse document frequency. With this approach, only index terms mentioned in the text can be identified. The index term assignment approach uses decision trees (J48) with BRkNN (Spyromitros et al., 2008) as a meta algorithm for multi-label classification (Quinlan, 1992). Additionally, we evaluate a hybrid approach, which combines the extraction and assignment approach by taking the highest ranked index terms of both approaches.

Type	Precision	Recall	R-prec.
Extraction	11.6%	15.5%	10.2%
Assignment	**33.0%**	6.1%	6.6%
Hybrid	20.0%	**17.9%**	**14.4%**

Table 1: Results for index term indentification approaches

Table 1 shows results for all three approaches in terms of precision, recall, and R-precision. The extraction approach yields good results for recall and R-precision, while the assignment approach yields a high precision but a lower recall and R-precision. Assignment determines few index terms with high confidence that increases precision but lowers recall and R-precision, while extraction allows for identifying many index terms with lower confidence. The hybrid approach (Erbs et al., 2013), in which index term extraction and assignment are combined, results in better performance in terms of recall and R-precision.

Next Steps We believe that using semantic resources will further improve index term identification by grouping similar index terms. Additionally, we plan to conduct a user study to verify our conclusion that automatic index term identification helps the users in finding documents.

3.3 Identification and Exploration of Dataset Names in Scientific Literature

Datasets are the foundation of any kind of empirical research. For a researcher, it is of utmost importance to know about relevant datasets and their state of publications, including a dataset's characteristics, discussions, and research questions addressed.

Task The project consists of two parts. First, references to datasets, e.g. "PISA 2012" or "National Educational Panel Study (NEPS)", must be extracted from scientific literature. This step can be defined as a Named Entity Recognition (NER) task with specialized named entities.[7]

Secondly, we want to investigate functional contexts, which can be seen as the purpose of mentioning a certain dataset, i.e., introducing,

[5] We divided the entire dataset in a development, training, and test set.

[6] `https://code.google.com/p/dkpro-core-asl/`

[7] We extract the NEs from more than 300k German abstracts of the *FIS Bildung* dataset.

discussing, side-mentioning, criticizing, or using a dataset for secondary analysis.

Approaches First of all, the term *dataset* must be defined for our purposes. Although there is a common sense about what a dataset is, no formal definition exists. As a starting point, we use a list of basic descriptive features from Renear et al. (2010), which are *grouping, content, relatedness,* and *purpose*. As those features are not precise enough for our case, we need to further refine unclear aspects, like how to treat nested datasets,[8] or general names like PISA, which are not datasets in the strict sense, as they denote projects comprised of multiple datasets. Another question being discussed with domain experts is, if only primary datasets or also aggregated datasets, e.g., statistical data from the Zensus (German censuses), are relevant or if they should be treated differently.

There is a large number of approaches for NER (Nadeau and Sekine, 2007). Due to the lack of labeled training data and the high annotation costs, we have to resort to three un- and semi-supervised methods; *a*) an information engineering approach, where we manually crafted rules, *b*) a baseline classifier using active learning (Settles, 2011), and *c*) a bootstrapping approach for iterative pattern induction (Riloff and Jones, 1999), which has been used successfully by Boland et al. (2012) on a similar task.[9]

Challenges Apart from general NER challenges like ambiguity, variants, multi-word names or boundary determination (Cohen and Hersh, 2005), extracting dataset names comes with additional challenges. First, not even a partially complete list of names is available, and second, there is no labelled training data. A user study showed, that manual labelling is very costly. Furthermore, dataset names are sparse in our dataset and most names only occur once.

Next Steps After evaluating the different approaches, named entity resolution must be conducted on the results to map each name variant to a specific project or dataset entity. To finally explore the functional contexts, we will use clustering methods to determine clusters of contexts. After verifying and refining them with domain experts, multi-label classification can be applied to assign functional contexts to dataset mentions.

3.4 Identification of Argumentation Structures in Scientific Publications

One of the main goals of any scientific publication is to present new research results to an expert audience. In order to emphasize the novelty and importance of the research findings, scientists usually build up an argumentation structure that provides numerous arguments in favor of their results.

Task The goal of this project is to automatically identify argumentation structures on a fine-grained level in scientific publications in the educational domain and thereby to improve both reading comprehension and information access. A potential use case could be a user interface which allows to search for arguments in multiple documents and then to combine them (for example arguments in favor or against private schools). See Stab et al. (2014) for an overview of the topic Argumentation Mining and a more detailed description of this project as well as some challenges.

Dataset As described in section 2.2, the *pedocs* and *FIS Bildung* datasets are very heterogeneous. In addition, it is difficult to extract the structural information from the PDF files (e.g. headings or footnotes). For this reason, we decided to create a new dataset consisting of publications taken from PsyCONTENT which all have a similar structure (about 10 pages of A4, empirical studies, same section types) and are available as HTML files.[10]

Approaches Previous works have considered the automatic identification of arguments in specific domains, for example in legal documents (Mochales and Moens, 2011) or in online debates (Cabrio et al., 2013). For scientific publications, more coarse-grained approaches have been developed, also known as Argumentative Zoning (Teufel et al., 2009; Liakata et al., 2012; Yepes et al., 2013). To the best of our knowledge, there is

[8] E. g. the PISA project contains several datasets from multiple studies, like PISA 2000, PISA 2003, PISA-International-Plus, or even research specific sub-datasets could be considered.

[9] However, their dataset was completely different, so that it is unclear at this point if bootstrapping performs well on our task.

[10] http://www.psycontent.com/

no prior work on identifying argumentation structures on a fine-grained level in scientific fulltexts yet.

We define an argument as consisting of several argument components which are related: an argument component can either support or attack another argument component; the argument component being supported or attacked is also called claim. We set the span of an argument component to be a sentence. In the following (fictitious) example, each sentence (A, B, C, D) can be seen as an argument component connected by support and attack relations as visualized in figure 2.

A *Girls are better in school.* **B** *In the XY study, girls performed better on average.* **C** *One reason for this is that girls invest more time in their homework.* **D** *However, there are also other studies where no differences between girls and boys could be found.*

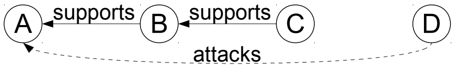

Figure 2: Visualization of an argumentation structure: The nodes represent the four sentences (A, B, C, D), continuous lines represent support relations, dotted lines represent attack relations

Next Steps Due to the lack of evaluation datasets, we are performing an annotation study with two domain experts and two annotators who developed the annotation guidelines. Next, we plan to develop weakly supervised machine learning methods to automatically annotate scientific publications with argument components and the relations between them. The first step will be to distinguish non-argumentative parts from argumentative parts. The second step will be to identify support and attack relations between the argument components. In particular, we will explore lexical features, such as discourse markers (words which indicate a discourse relation, for example “hence”, “so”, “however”) and semantic features, such as text similarity.

3.5 Scalable Multi-label Classification for Educational Research

This project aims at developing and applying novel machine learning algorithms which can be useful for providing methods to automate the processing of scientific literature. Scientific publications often need to be organized in a way of providing high-level and structured information, i.e., metadata. A typical example of a metadata management system is assigning index terms to a document.

Task The problem of assigning multiple terms to a document can be addressed by multi-label classification algorithms. More precisely, our task is to assign multiple index terms in *FIS Bildung*, to a given instance if we have a predefined list of the terms. There are two problems for multi-label classification in the text domain; 1) What kinds of features or which document representations are useful for our task of interest? 2) How do we exploit the underlying structure in the label space?

Dataset and Challenges In *FIS Bildung* database, tens of thousands of index terms are defined, because it is a collection of links to documents coming from diverse institutions each of which deals with different subjects, thereby requiring expertise of index terms maintenance. The difficulty of predicting index terms for a given document is divided largely into two parts. First, only abstracts are available which contain a small number of words compared to fulltexts. Secondly, given a large number of distinct labels, it is prohibitively expensive to use sophisticated multi-label learning algorithms. To be more specific, we have about 50,000 index terms in *FIS Bildung* which most of current multi-label algorithms cannot handle efficiently without a systematic hierarchy of labels. Hence, as a simplified approach, we have focused on 1,000 most frequent index terms as target labels that we want to predict because the rest of them occur less than 20 times out of 300K documents.

Approaches Multi-label classifiers often try to make use of intrinsic structures in a label space by generating subproblems (Fürnkranz et al., 2008) or exploiting predictions of successive binary classifiers for the subsequent classifiers (Read et al., 2011).

Neural networks are a good way for capturing the label structure of multi-label problems, as has been shown in BP-MLL (Zhang and Zhou, 2006). Recent work (Dembczyński et al., 2012; Gao and Zhou, 2013) find inconsistency of natural (convex) rank loss functions in multi-label learning.

Based on these results, Nam et al. (2014) showed that the classification performance can be further increased with methods that have been recently developed in this area, such as Dropout (Srivastava et al., 2014), Adagrad (Duchi et al., 2011), and ReLUs (Nair and Hinton, 2010), on the *FIS Bildung* dataset as well as several text benchmark datasets. Specifically, for multi-label text classification task, the cross-entropy loss function, widely used for classification tasks, has shown to be superior to a loss function used for BP-MLL which try to minimize errors resulting from incorrect ranked labels. Even though the former does not consider label ranking explicitly, it converges faster and perform better in terms of ranking measures. More details can be found in (Nam et al., 2014).

Next Steps Even though our proposed approach has shown interesting results, the original problem remains unsolved. How do we assign multiple labels to an instance where tens or even hundreds of thousands of labels are in our list? To answer this, we are going to transform both instances and labels into lower dimensional spaces while preserving original information or deriving even more useful information (Socher et al., 2013; Frome et al., 2013) which enables us to make predictions for unseen target labels at the time of training.

3.6 Temporally Dynamic Networks of Topics and Authors in Scientific Publications

In this part of the KDSL program, we build a probabilistic network for various aspects of scientific publication. The important entities are authors, ideas and papers. From authors, writing style and communities can be modelled. From papers, index terms, citations and arguments can be extracted. In reality, all these factors affect each other and when they are considered in one probabilistic model, the precision of each model should be improved, as a result of enhanced context.

Task and Data At first, we took the *pedocs* dataset and performed temporal analysis as the first dimension of the probabilistic network. By tracking the occurrence of index terms in the last 33 years, we monitor the development of topics in the corpus. The first assumption is that trendy topics lever-up the frequency of their representing keyword in the corpus at each period of time. The second assumption is that the significant co-occurrence of keywords indicates the emergence of new research topics.

Approach Co-occurrence has been used in trend detection (Lent et al., 1997). To capture more interesting dynamic behaviors of the index terms, we experimented with different measures to find index term pairs of interest. Covariance, co-occurrence, Deviation-from-Random, Deviation-from-Lower-Envelop are some of the measures we used to detect the co-developing terms. The covariance, co-occurrence are the standard statistical measures in temporal relation analysis (Kontostathis et al., 2004). The other measures are developed in our work, which exhibit the capability to gain more insights from the data.

Interestingly, some of the measures can reveal strong semantic relatedness between the index terms, e.g., Internationalisierung - Globalisierung (Internationalization - Globalization). This phenomenon indicates a potential unsupervised semantic-relatedness measure. And generally, our methodology can find interesting pairs of index terms that help the domain researcher to gain more insight into the data, please see (Ma and Weihe, 2014) for detailed examples of the findings.

For the manually selected index terms (about 300), we collaborated with domain experts from DIPF to assign categories (Field, Topic, Method, etc.) to them. With the category, we can look for the term pairs of our interest. For example, we can focus on the method change of topics, by limiting the categories of a term pair to Topic and Method.

Next Steps One critical problem to these analyses is data sparsity. Some experiments can only output less than 10 instances, which may be insufficient for statistically significant results. We adapt the methods to larger datasets like *FIS Bildung*. Besides optimization, we will work on other new measures and evaluate the results with the help of domain experts.

3.7 Structured Tag Clouds

Tag clouds are popular visualizations on web pages. They visually depict a set of words in a spatial arrangement with font size being mapped to an approximation of term importance such as term frequency. It is supposed that by organizing the words according to some (semantic) term relation, the usefulness of tag clouds can be further improved (see e.g., (Hearst and Rosner, 2008; Rivadeneira et al., 2007)). The goal of this project is to investigate if this assumption holds true and to research the optimal design and automatic generation of such structured tag clouds (Figure 3).

Task To approach our research goal, three main tasks can be distinguished: First, we examine how humans structure tags when being told that the resulting tag cloud should provide a quick overview of a document collection. Second, based on the determined criteria that the participants of our study aimed at, when layouting the clouds, we develop methods for automatically generating structured tag clouds. Finally, the performance of users employing structured tag clouds is compared to unstructured ones for specific tasks.

Dataset As the name suggests, tag clouds are often employed to visualize a set of (user-generated) tags. In our research, we use user-generated tags from social bookmarking systems such as BibSonomy[11] or Edutags[12]. We expect that the results can be generalized to similar data such as index terms assigned to scientific publications or these extracted from a document (collection).

Challenges There are many ways to (semantically) structure tags (e.g., based on co-occurrences or lexical-semantic relations). However, our goal must be not to generate an arbitrary tag structure but to organize tags in a way that is conclusive for human users and thus easy to read. A key challenge here is that no ground-truth exists saying how a specific tag set is arranged best.

Approaches We conducted a user study in which the participants were asked to manually arrange user-generated tags of webpages that were retrieved by a tag search in the social bookmarking system BibSonomy. Being aware that no single ground-truth exists, we investigated the criteria underlying the layout in detailed post-task interviews. Those criteria are now the basis for researching automatic algorithms and visual representations that can best approximate the user-generated layouts. Finally, unstructured and structured tag clouds will be compared in a study in which the performance of users in specific tasks is measured.

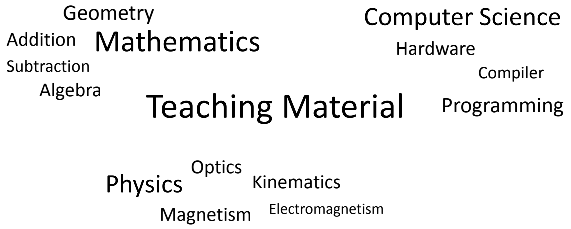

Figure 3: Example for a structured tag cloud.

Results & Next Steps In (Oelke and Gurevych, 2014) we presented the results of our user study. While previous work mainly relies on co-occurrence relations when building structured tag clouds, our study revealed that semantic associations are the main criterion for human layouters to build their overall structure on. Co-occurrence relations (i.e., two tags that are at least once assigned to the same bookmarked webpage) were only rarely taken into account, although we provided access to this information.

While some participants included all tags in their final layout, others consequently sorted out terms that they deemed redundant. Lexical-semantic relations (e.g., synonyms or hypernyms) turned out to be the basis for determining such redundant terms. Furthermore, small clusters were preferred over large ones and large clusters were further structured internally (e.g., arranged according to semantic closeness, as a hierarchy, or split into subclusters).

Next, we will work on the algorithmic design and finally evaluate the performance of structured tag clouds.

4 Conclusion

This paper describes 'Knowledge Discovery in Scientific Literature', a unique graduate program with the goal to make the knowledge concealed in various kinds of educational research literature more easily accessible. Educational researchers will benefit from automatically processed information on both local and global scopes. Local information consists of index terms (Sec. 3.2, 3.7, 3.5), relations (Sec. 3.1), dataset mentions and

[11] http://www.bibsonomy.org/

[12] http://www.edutags.de/

functional contexts (Sec. 3.3), and argumentation structures (Sec. 3.4). On the level of the entire corpus, temporal evolution of index terms and authors can be provided (Sec. 3.6).

Each sub-project aims at new innovations in the particular field. The close connection between computer science researchers and educational researchers helps us with immediate evaluation by end users.

Acknowledgments

This work has been supported by the German Institute for Educational Research (DIPF) under the Knowledge Discovery in Scientific Literature (KDSL) program. We thank Alexander Botte, Marius Gerecht, Almut Kiersch, Julia Kreusch, Renate Martini, Alexander Schuster from the Information Center for Education at DIPF for providing valuable feedback.

References

Doris Bambey and Agathe Gebert. 2010. Open-Access-Kooperationen mit Verlagen – Zwischenbilanz eines Experiments im Bereich der Erziehungswissenschaft. *BIT Online*, 13(4):386.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. 2012. Identifying References to Datasets in Publications. In *Theory and Practice of Digital Libraries*, pages 150–161, Paphos, Cyprus.

Elena Cabrio, Serena Villata, and Fabien Gandon. 2013. A Support Framework for Argumentative Discussions Management in the Web. In *The Semantic Web: Semantics and Big Data*, pages 412–426. Montpellier, France.

Carola Carstens, Marc Rittberger, and Verena Wissel. 2011. Information search behaviour in the German Education Index. *World Digital Libraries*, 4(1):69–80.

Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Dmitry Davidov, Ari Rappoport, and Moshe Koppel. 2007. Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 232–239, Prague, Czech Republic.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 187–192, Baltimore, MD, USA.

Krzysztof Dembczyński, Wojciech Kotłowski, and Eyke Hüllermeier. 2012. Consistent Multilabel Ranking through Univariate Losses. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1319–1326, Edinburgh, UK.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Nicolai Erbs, Iryna Gurevych, and Marc Rittberger. 2013. Bringing Order to Digital Libraries: From Keyphrase Extraction to Index Term Assignment. *D-Lib Magazine*, 19(9/10):1–16.

Nicolai Erbs, Pedro Bispo Santos, Iryna Gurevych, and Torsten Zesch. 2014. DKPro Keyphrases: Flexible and Reusable Keyphrase Extraction Experiments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 31–36, Baltimore, MD, USA.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26*.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. 2008. Multilabel Classification via Calibrated Label Ranking. *Machine Learning*, 73(2):133–153.

Wei Gao and Zhi-Hua Zhou. 2013. On the Consistency of Multi-Label Learning. *Artificial Intelligence*, 199–200:22–44.

Zellig S Harris. 1954. Distributional structure. *Word*, 10:146–162.

Marti A. Hearst and Daniela Rosner. 2008. Tag Clouds: Data Analysis Tool or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pages 160–160.

Petr Knoth and Zdenek Zdrahal. 2012. CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine*, 18(11/12).

April Kontostathis, Leon M Galitsky, William M Pottenger, Soma Roy, and Daniel J Phelps. 2004. A Survey of Emerging Trend Detection in Textual Data Mining. In *Survey of Text Mining*, pages 185–224. Springer.

Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. 1997. Discovering Trends in Text Databases. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, volume 97, pages 227–230, Newport Beach, CA, USA.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Dekang Lin and Patrick Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, pages 323–328, San Francisco, CA, USA.

Zheng Ma and Karsten Weihe. 2014. Temporal analysis on pairs of classified index terms of literature databases. In *Proceedings of the 10th International Conference on Webometrics, Informetrics, and Scientometrics (WIS) and the 15th COLLNET Meeting 2014*, Ilmenau, Germany.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, Haifa, Israel.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale Multi-label Text Classification – Revisiting Neural Networks. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–452, Nancy, France.

Daniela Oelke and Iryna Gurevych. 2014. A Study on the Layout of Human-Generated Structured Tag Clouds. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, Como, Italy.

John Ross Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier Chains for Multi-label Classification. *Machine Learning*, 85(3):333–359.

Steffen Remus. 2014. Unsupervised Relation Extraction of In-Domain Data from Focused Crawls. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–20, Gothenburg, Sweden.

Allen H. Renear, Simone Sacchi, and Karen M. Wickett. 2010. Definitions of Dataset in the Scientific and Technical Literature. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Ellen Riloff and Rosie Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the National Conference on Artificial Intelligence and the Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, pages 474–479, Orlando, FL, USA.

Anna W. Rivadeneira, Daniel M. Gruen, Michael J. Muller, and David R. Millen. 2007. Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 995–998, San Jose, CA, USA.

Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Burr Settles. 2011. Closing the Loop: Fast, Interactive Semi-supervised Annotation with Queries on Features and Instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, UK.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26*.

Min Song, Il Yeol Song, Robert B. Allen, and Zoran Obradovic. 2006. Keyphrase Extraction-based Query Expansion in Digital Libraries. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 202–209, Chapel Hill, NC, USA.

Eleftherios Spyromitros, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2008. An Empirical Study of Lazy Multilabel Classification Algorithms. In *Artificial Intelligence: Theories, Models and Applications*, pages 401–406.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Frontiers and Connections between Argumentation The-*

ory and Natural Language Processing, Bertinoro, Italy.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Edinburgh, UK.

Simon Tucker and Steve Whittaker. 2009. Have A Say Over What You See: Evaluating Interactive Compression Techniques. In *Proceedings of the 2009 International Conference on Intelligent User Interfaces*, pages 37–46, Sanibel Island, FL, USA.

Antonio Jimeno Yepes, James G. Mork, and Alan R. Aronson. 2013. Using the Argumentative Structure of Scientific Literature to Improve Information Access. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 102–110, Sofia,Bulgaria.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351.