Facet Embeddings for Explorative Analytics in Digital Libraries

Sepideh Mesbah(✉), Kyriakos Fragkeskos, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
{s.mesbah,k.fragkeskos,c.lofi,a.bozzon,g.j.p.m.houben}@tudelft.nl

Abstract. With the increasing amount of scientific publications in digital libraries, it is crucial to capture *"deep meta-data"* to facilitate more effective search and discovery, like search by topics, research methods, or data sets used in a publication. Such meta-data can also help to better understand and visualize the evolution of research topics or research venues over time. The automatic generation of meaningful deep meta-data from natural-language documents is challenged by the unstructured and often ambiguous nature of publications' content.

In this paper, we propose a domain-aware topic modeling technique called *Facet Embedding* which can generate such deep meta-data in an efficient way. We automatically extract a set of terms according to the *key facets* relevant to a specific domain (i.e. scientific objective, used data sets, methods, or software, obtained results), relying only on limited manual training. We then cluster and subsume similar facet terms according to their semantic similarity into facet topics. To showcase the effectiveness and performance of our approach, we present the results of a quantitative and qualitative analysis performed on ten different conference series in a Digital Library setting, focusing on the effectiveness for document search, but also for visualizing scientific trends.

1 Introduction

In light of the increasing amount of scientific publications, there is a growing need for methods that facilitate the exploration and analysis of a given research field in a digital library collection [1]. Existing approaches rely on word-frequency analysis [2], co-citation analysis [3], co-occurrence word analysis [4], and probabilistic methods like Latent Dirichlet Allocation (LDA) [5]. While popular, these approaches suffer from one major shortcoming: by offering a generic solution, they fail to capture the intrinsic semantics of text related to a specific domain of knowledge. For instance, probabilistic methods like LDA are designed to be generic and widely applicable; however, they often miss out on topics that are relevant from a user's point of view.

To support richer retrieval experience, we advocate extracting *"deep meta-data"* from scientific publication, i.e. meta-data able to represent domain-specific properties and aspects (*facets*) in which a document can be considered and understood within its (research) domain.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 86–99, 2017.
DOI: 10.1007/978-3-319-67008-9_8

Let us consider, for instance, the domain of *data processing and data science*, which is gaining popularity due to the availability of great amount of digital data, and progress in machine learning. In this domain, researchers and practitioners need to develop an understanding of the properties of available *datasets*; of existing data processing *methods* for the collection, enrichment and analysis of data; and of their respective implementations as *software* packages. The availability of deep meta-data about the facets (*datasets*, *methods*, and *software*) would enable rich queries like: *Which methods are commonly applied to a given dataset?*; *Discover state of the art methods for point of interest recommendation that have been applied to geo-located social media data with high accuracy results*. To the best of our knowledge, no state-of-the-art system is currently able to provide answers to the previous queries.

This paper presents an approach for generating domain-aware *"deep meta-data"* from collections of scientific publications. We focus on the data processing domain, and address the main facets described in the `DMS` ontology [6], namely *datasets*, *methods*, *software*, *objectives*, and *results*. We build upon a basic distant supervision approach for sentence classification and named entity extraction [7], and extend it with *facet embeddings* to automate the creation of *Facet Topics*, i.e. clusters of semantically similar facet terms which allow for easier querying and visualization. Our contributions are as follows:

- We introduce and formalize the concept of *facet topics*, which subsume a set of facet terms into higher level topics more suitable for exploration, visualization, and topic centered queries.
- We describe a novel approach for facet topic identification through *facet embeddings*. The approach combines distant supervision learning on rhetorical mentions for facet-specific sentence classification; semantic annotation and linking for facet terms extraction; and semantic clustering.
- We quantitatively and qualitatively assess the performance of our approach, and compare to established techniques like LDA topic modeling.
- We showcase our approach with a study exploring and visualizing trends and changes within the domain of data processing research, based on deep meta-data extracted from 11,589 research publications.

2 Related Work

The information overload in digital libraries is a crucial problem for researchers. Online digital libraries like the ACM Digital Library (DL), IEEE Xplore, CiteSeer etc., provide search options for finding relevant publications by using *"shallow"* meta-data such as the title, the authors, keywords or other simple statistical measures like the number of citations and download counts. However they are not designed to support the analysis of *"deep"* meta-data such as the topic, or methods and algorithms used in scientific publications.

There has been a large body of research focused on *deep* analysis of publications in scientific domains such as Software Engineering [1], Bio-informatics [8],

Digital Library evaluation [9], or Computers science [10]; for different purposes, such as finding topic trends in a domain [1,10] and evolution of scientific communities popularity [11]. Common approaches rely on methods such as word-frequency analysis [2], co-citation analysis [3,10], co-word analysis [4], and probabilistic methods like latent Dirichlet allocation [5]. In contrast to existing literature which is either specially tailored to a domain or fully generic, our work combines the strength of both approaches by being partially domain-aware: after defining domain-aware facets using (limited) expert feedback, our approach automatically extracts topics by analyzing the co-occurrence of named entities related to the facets, thus is scalable within a domain while still taking advantage of domain-specific knowledge and peculiarities.

While most current research [1,2,11] limits the analysis of a publication's content to its title, abstract, references, and authors, we extract facet terms from the full text of scientific publications, in order to obtain more descriptive and accurate topics. In addition, our method is not only based on selecting the most frequent keywords (e.g. nouns, verbs set and proper nouns) [2], and, differently from probabilistic methods like Latent Dirichlet Allocation [5], it considers the semantics of terms for topic identification.

Some existing methods for domain-specific concept extraction and categorization are based on noun phrase chunking [11,12] and use a bootstrapping approach to identify scientific concepts in publications. More recent research [13] used both corpus-level statistics and local syntactic patterns of scientific publications to identify and cluster similar concepts. Our method follows a distant supervision approach, a simple feature model (bags-of-words), and does not require prior knowledge about grammatical [12] and part-of-speech characteristics of facet terms. However, we do require a brief training phase for adapting our approach to a new domain.

3 Problem Description and Modeling

The goal of our work is to annotate n documents $D = \{d_1, ..., d_n\}$ of a domain-specific (scientific) corpus with faceted semantic meta-data. This meta-data goes alongside already available structured meta-data like for example author names, publication year, or citations. In particular, we aim at annotating documents with both *facet terms* and *facet topics*, as discussed in the following:

Facets and Facet Sets: The central elements of our approach are *facets*. Facets represent a perceived aspect relevant to user's understanding of documents in corpus D. When adapting our method to a given corpus, a *facet set* has to be defined which is used for describing documents in D, denoted as $F = \{f_1, f_2, \ldots, f_n\}$. Defining a good facet set requires some domain expertise. In the study presented in this work, we used specific facet set designed based on [6], namely the F_{DMS} facet set covering facets for a document corpus focused on data processing research. This facet set covers the five facets dataset, methods, software, objective, and result. We denote this as $F_{DMS} = \{DST, MET, SFT, OBJ, RES\}$.

Facets Terms: For each document $d \in D$ and facet $f \in F$, we extract a set of *facet terms* FT_f^d. A facet term $ft \in FT_f^d$ represents a term (usually a named entity, but also short phrases are possible) found in the full text of document d, and which can be clearly associated with facet f. We denote the set of all facet terms related to a given facet f found in any document of D as FT_f. Typical examples of facet terms for the method facet $MET \in F_{DMS}$ in our document collection are "Latent Dirichlet Allocation", "Support Vector Machine", or "Description Logic".

Facets Topics: Facet Terms are directly extracted from the full text of documents, and describe a document at a rather low level. In order to also allow for high-level analytics and queries, we introduce the concept of *facet topics*. Facet topics group multiple semantically related facet terms into a larger subsuming topic. In our use case scenario, when focusing on the methods facet, facet topics intuitively relate to research topics. For example, the terms "Support Vector Machine" and "Random Forest" can be subsumed by the facet topic "Machine Learning". The set of all facet topics for a given facet f is denoted as $FTP_f = \{t_1, t_2, \ldots t_k\}$, and each facet topic t is a subset of all facet terms, i.e. $t \in FTP_f : t \subseteq FT_f$. Furthermore, each term can be attributed to a topic, i.e. $FT_f = \bigcup_{t \in FTP_f} t$, and topics of a given facet are disjoint, i.e. $t_i, t_j \in FTP_f, t_i \neq t_j : t_i \cap t_j = \emptyset$ (however, there might be an overlap between topics of different facets, see next section). Terms in a facet topic show strong semantic cohesion.

4 Facet Term Extraction and Facet Topic Identification

In this section, we present our approach for *facet terms* and *facet topics* extraction from a collection of scientific publications, extending our previous work [7] by introducing additional steps for facet topic identification. An overview of our approach is shown in Fig. 1. Our approach is domain-aware in the sense that it requires some limited efforts to adjust it to a new domain (like deciding on facet sets), but is not inherently limited to a specific domain. In the following, we focus on the *data processing* domain, and address the five main *facets* (i.e. datasets, methods, software, results, and objectives) identified in the DMS ontology [6].

The process can base summarized as: First, we identify rhetorical mentions of a *facet* in the full text of documents. In this work, for the sake of simplicity, rhetorical mentions are identified at sentence level (i.e., each sentence is classified whether it contains a rhetorical mention of a given facet or not). Future works will introduce dynamic boundaries, to capture the exact extent of a mention.

After a rhetorical mention was found, we extract potential *facet terms* from it. These terms are filtered and, when applicable, linked to pre-existing knowledge bases. Finally, all filtered facet term candidates finally form the document-specific facet term sets FT_f^d.

The identification of rhetorical mentions is obtained through a workflow inspired by distant supervision, a training methodology for machine learning algorithms that relies on very large, but noisy, training sets. The training sets are

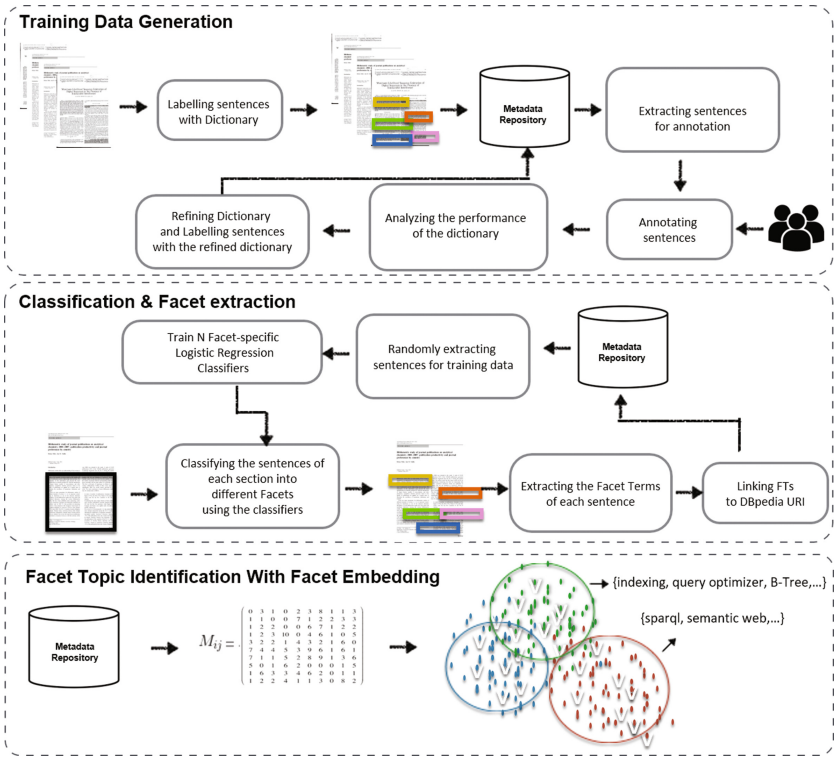

Fig. 1. Domain-aware Facet Modeling Workflow [7]

generated by means of a simpler classifier, for instance a mix of expert-provided dictionaries and rules, refined with manual annotations. Intuitively, the training noisiness can be canceled out by the huge size of the semi-manually generated training data. The method requires significantly less manual effort, while at the same time retaining the performance of supervised classifiers. Furthermore, this approach is more easily adapted to different application domains and changing language norms and conventions (more details in [7]).

Data Preparation: Scientific publications, typically available in PDF, are processed using state-of-art extraction engines, e.g. GeneRation Of BIbliographic Data (`GROBID`) [14]. `GROBID` extracts a structured full-text representation as Text Encoding Initiative(TEI)-encoded documents, thus providing easy and reliable access paragraphs and sentences.

Test and Training Data Generation: We created training and benchmarking datasets for evaluating our rhetorical mention classifier by relying on a phrase dictionary for each facet (as described in [7]), automatically labeling all sentences in the document corpus if they contain a mention of relevant for a facet or not. Then, we randomly select a balanced set of 100 mentions of each facet.

As the dictionary-based classifier is not fully reliable, we manually inspect and reclassify the selected sentences using feedback from two expert annotators, and rebalance the sentence set as needed. The inter-annotator agreement using the Cohen's kappa measure averaged over all classes was .58. Using this approach, we can create a reliable manually annotated and balanced test dataset quicker and cheaper compared to annotating whole publications or random sentences, as the pattern-classifier usually delivers good candidate sentences.

Machine-Learning-based Rhetorical Detection: As a next step in our distant supervision workflow, we train a simple binary Logistic regression classifier for each of the (facet) classes using simple `TF-IDF` features for each sentence. This simple implementation serves as a proof of concept of our overall approach, and can of course be replaced by more sophisticated features and classifiers.

As a test set, we use the aforementioned test set of 100 sentences for each facet. The *method* classifier showed the best performance with respect to its F-measure(0.71). From this, we conclude that our approach is indeed suitable for extracting *DMS* facet terms in a meaningful and descriptive fashion. However, there are still some false positives which cannot easily be recognized using simple statistic means, thus inviting further deeper semantic filtering in future works.

Facet Extraction, Linking, and Filtering: We extract *facet terms* from the labeled rhetorical mentions identified in the previous section, filtering out those terms which are most likely not referring to one of the *facet*, and retaining the others as an extracted term of the class matching the sentence label.

Facet extraction has been performed using the TextRazor API. TextRazor returns the detected *facet terms*, possibly decorated with links to the DBpedia or Freebase knowledge bases. As we get all *facet terms* of a sentence, the result list contains many *facet terms* which are not specifically related to any of the five *facets* (e.g. terms like "software", "database"). To filter the *facet terms*, we decided on a simple filtering heuristic assuming *facet terms* to be not relevant if they come from "common" English language (like software, database), while relevant terms are from domain-specific language or specific acronyms (e.g. SVM, GROBID). In our current prototype, we implement this heuristics by looking-up each term in *Wordnet*. Terms that can be looked-up are removed as we consider them common language. While this simple approach works for the "data science" domain, when extending our approach to a wider range of domains, this implementation should be replaced by more sophisticated heuristics, e.g., based on corpus statistics.

Facet Topic Identification With Facet Embedding After extracting all facet terms, we now strive to discover meaningful *facet topics*. Here, a central goal is to subsume facet terms based on their semantic similarity. We implement a measurement for semantic similarity of terms by *Facet Embeddings*, which exploit co-occurrence of facet terms. For each $t_i, t_j \in FT_f$, we count how often these terms co-occur within the same document: $co_{t_i,t_j} = |\{d \in D : t_i \in FT_f^d \wedge t_j \in FT_f^d\}|$.

This results in a large (sparse) co-occurrence matrix. We reduce the dimensionality of the matrix using truncated Singular Value Decomposition. This step ensures the removal of less informative terms, while increasing the performance and usability of our approach (a smaller matrix is computationally cheaper to process). Using the reduced matrix, we now obtained an embedding of each facet term of a given facet (i.e., each term is represented as row vector in the reduced co-occurrence matrix).

Finally, we now cluster all facet terms of a given facet in order to discover facet topics using K-means clustering, using Euclidean distance between the row vectors of two given terms as a distance measure. In order to find the optimal number k of clusters, we rely on Silhouette analysis, measuring the closeness of each point in a cluster to the points in its neighboring clusters. In addition to the Silhouette analysis, we also manually inspected the resulting clusters, but found that also from an qualitative point of view, the number of clusters determined by the Silhouette analysis is indeed the most satisfying one.

As a last processing step, we have two expert annotators label each facet topic in an iterative process until full agreement between the annotators was reached (see Sect. 5 for more details).

We also implemented an alternative version of facet embeddings, relying on neuronal word embeddings (in our case word2vec [15]) instead of co-occurrence in rhetorical mentions. However, initial qualitative inspection of the results indicate that the distance measure between the term embeddings is an inferior representation of perceived similarity of facet terms from our experts' point of view. A deeper analysis of these results will be the subject of a later study.

5 Evaluation and Experimentation

In this section, we analyze the performance of our facet topic modeling workflow. We analyze and discuss the quality of facet terms extracted from the classified sentences. Next we qualitatively evaluate the quality of the topics extracted using Facet Embeddings. Finally we present some examples of information needs of researcher that can be fulfilled using our approach.

Corpus Analysis: We focused on 11,589 papers from ten conference series: The Joint conference on Digital Libraries (JCDL); the International Conference on Theory and Practice of Digital Libraries (TPDL); the International Conference on Research and Development in Information Retrieval (SIGIR); the Text Retrieval Conference (TREC); the European Conference on Research and Advanced Technology on Digital Libraries (ECDL); the International Conference on Software Engineering (ICSE); the Extended Semantic Web Conference (ESWC); the International Conference On Web and Social Media (ICWSM); the International Conference on Very Large Databases (VLDB); and the International World Wide Web Conference (WWW).

Table 1. Quantitative analysis of the rhetorical sentences and facet terms extracted from ten conference series. Legend: PUB (publications), SNT (sentences), OBJ (objective), DST (dataset), MET (method), SFT (software), RES (results)

Conf.	Size		Rhetorical sentences						Unique facet terms				
	Years	#PUB	#SNT	#OBJ	#DST	#MET	#SFT	#RES	#OBJ	#DST	#MET	#SFT	#RES
ESWC	2005–2016	626	84439	12725	13528	26337	9614	22245	4197	4910	6987	4557	6416
ICWSM	2007–2016	810	34987	6096	4277	8936	1830	13848	2830	2241	3658	1538	4499
VLDB	1975–2007	1884	272380	30360	56647	77123	13317	94933	8008	13207	15319	6262	17532
WWW	2001–2016	2067	322801	47134	40449	97760	21347	116111	10902	10917	17783	8863	19822
ECDL	1997–2010	820	65470	12008	8079	18638	8130	18615	4634	3650	5894	4125	5376
ICSE	1976–2016	1834	182029	29850	16284	57494	26042	52359	8169	5841	12503	8776	11728
JCDL	2001–2016	1416	99747	19290	13002	27786	9692	29977	6524	5240	7754	5037	7979
SIGIR	1971–2016	412	39688	5080	4813	13214	2050	14531	2144	2377	4126	1588	4068
TPDL	2011–2016	276	23176	4660	3342	6032	2489	6653	2168	1871	2625	1719	2503
TREC	1999–2015	1444	122456	11828	14760	39121	8825	47922	6616	3085	4095	3286	7668

Due to changes in publication platforms and PDF format, the corpus does not contain all publications of each conference.[1] We believe the absence of few publications not to have an impact on the significance of our findings, but might still be reflected in the shown diagrams and results. Table 1 provides basic statistics for the analyzed corpus, including: the range of years, the number of publications, the number of extracted rhetorical sentences and mentions, and the distinct facet terms extracted from rhetorical sentences. *Method* and *results* facets are the most frequent, followed by *objectives*.

Quality of extracted topics: We investigated or domain-aware facet embedding compared to the domain-independent technique Latent Dirichlet Allocation (LDA) by asking two domain experts to label the topics derived by each method, while assessing which are more meaningful. For the sake of presentation, we set the maximum number of topics to $T = 30$, and performed the Silhouette analysis to find the number of optimal topics, which resulted in 27 topics.

In order to allow for a more informative comparison, we applied both approaches to the full text of publications, and also to only pre-classified sentences (because LDA is usually applied to full texts. Thus, in one case we use our facet embedding without restricting to classified facet sentences, but we also consider a case where LDA is applied to the set of all sentences which belong to a given facet). For the sake of brevity, we consider only the *method* facet when performing a facet pre-classification of sentences. The *method* classifier has shown the best performance with respect to its F-measure. Our analysis shows comparable results with the other facets.

Full Text without Facet Classification: For full text experiments, the corpus has been pre-processed by removing stop words, and representing each document as a bag-of-words. We use the LDA implementation provided by the `scikit-learn`

[1] For instance, around 100 JCDL papers for 2014 are not included in the analysis, as the proceedings were, only for that year, published by ieee.org.

library. For compatibility, we trained for 27 topics. For evaluating facet embeddings without any domain specific pre-classification on full texts, we are assuming that there is only a single facet, and each sentence of a document is classified as such (note: this is not how we usually intend our method to work).

Consider only Sentences classified as Method facet: In this experiment, we perform the *facet topic* extraction as described in Sect. 4, including facet-based sentence classification, facet term extraction, and facet embedding, but limited to only the *Method* facet. As a comparison, we also perform LDA on only those sentences classified as methods (therefore also giving LDA the chance to take advantage of the domain-aware training).

Results: A manual inspection on the resulting topics show that those identified by LDA are very hard to label and are perceived as semantically less meaningful by our experts, while the topics based on Facet Embeddings produced coherent and interpretable topics which were perceived as understandable and useful. In Table 2, we provide an example of 3 randomly selected topics for each aforementioned experimental setup. It can be observed that topics generated from sentences pre-classified as *method* show better semantic cohesion than those generated from full texts. Furthermore, we provide the full result of labeling all 27 topics for the method facet in Table 3. The top-40 term can be found in the companion website[2]

Table 2. Example top terms extracted using the generic (LDA) and domain-aware (FE) topics, using either full texts or only those sentences related to the *method* facet

Full text	*LDA*	reference, abstracts, linking, sofm, similarity annotations, backup, linkservice, annotation, digital query, data, user, web, information
	FE	sparql, semanticweb, linkeddata, rdf, dbpedia, sql, relationaldatabase, tuple, queryoptimization, datawarehouse, socialnetwork, facebook, randomwalk, pagerank, powerlaw
Facet	*LDA*	documents, used, classification, libraries, digital measure, performance, given, recommendation, used, social, twitter, media, popular, past
	FE	searchalgorithm, timecomplexity, datastructure, dynamicprogramming, sparql, semanticweb, linkeddata, dbpedia, rdfs, socialmedia, lda, latentdirichletallocation, topicmodel, socialnetwork

Application Example: Scientific Publication Retrieval: In this section we show scenarios of how computer science researchers could use our approach for their work. Furthermore, all faceted deep meta-data used in those scenarios has been published as an RDF knowledge base according to the DMS ontology, accessible from a SPARQL endpoint on the companion website.

[2] http://www.wis.ewi.tudelft.nl/tpdl2017.

Table 3. Top five *method* terms for each facet topic. Topic labels have been assigned manually by two xperts.

Topic name	Top five terms
Social Media Analytics: Text-based	social media, lda, latent dirichlet allocation, topic model, social network
Semantic Web: Knowledge Engineering & Representation	sparql, semantic web, linked data, dbpedia, rdfs
Semantic Web: Logic	description logic, dl, abox, tbox, semanticweb
Misc Topics: Web Information Systems	information retrieval, data structure, dataset, natural language, electronic media
Databases: Query Processing	tuple, hash join, sort, relational database, hash table
Databases: Modelling	data model, sql, query language, query optimization, tuple
Web Technologies	side, client, server, javascript, web application
Digital Libraries	digital library, information retrieval, xml, user interface, computer science
Machine Learning	machine learning, support vector machine, supervised learning, dataset, information retrieval
Web Engineering: P2P & Distributed Systems	peer, to, ip address, rdf, webservice
Social Graph Algorithms	greedy algorithm, approximation algorithm, optimization problem, social network, electronic media
Social Graph Analysis	pagerank, random walk, social network, webpage, adjacency matrix
XML Databases	xml, xpath, xquery, xmlschema, sql
Software Engineering: Testing & Formal Methods	source code, test case, control flow, test suite, program analysis
Software Engineering: Systems	software development, software engineering, software development process, software system, case study
Web Engineering: System Modelling	use case, web service, model checking, case study, semantic web
Web Engineering: Client-Side	web page, user interface, web browser, web content, javascript
Information Retrieval: QA, NLP, and Complex Queries	trec, question answering, document retrieval, information retrieval, query expansion
Information Retrieval: Evaluation	adhoc, trec, query expansion, information retrieval, relevance feedback
Information Retrieval: Ranking	query expansion, language model, relevance feedback, trec, information retrieval
Information Retrieval: Mining	score, f1, supervised learning, crf, bic
Microsoft Technology	microsoft, microsoft sqlserver, sql, xml, microsoft word
Databases: Indexing	tree, trees, data structure, access method, search algorithm
Databases: Transaction Mangement	concurrency control, lru, serializability, aries, tion
Databases: Algorithms	search algorithm, time complexity, data structure, dynamic programming, dataset
Recommendation	collaborative filtering, recommender system, gradient descent, singular value decomposition, social network
System Engineering: Architecture	operating system, programming language, file system, data structure, software engineering

Table 4. Examples of papers applying methods (MET) to given datasets(DTS)

Paper title	Dataset and method facet
Personalized Interactive Faceted Search [16]	IMDB(DST), Faceted search(MET)
refeREE: An Open Framework for Practical Testing of Recommender Systems using ResearchIndex [17]	IMDB(DST), Recommender system(MET)
The Party is Over Here: Structure and Content in the 2010 Election [18]	Facebook(DST), Sentiment analysis(MET)

Find publications that applied method X on a given dataset: Table 4 shows the result of an example query for methods which have been applied to movie dataset (i.e. IMDB and MovieLens) or Social media data (i.e. Facebook). For instance, [17] is a paper containing both the facet terms "Recommender system" labeled as *method*, and "IMDB" labeled as *dataset*.

Retrieve the most used methods of a given conference series: To answer this question, we use the number of papers for each *method* facet topic shown in Table 3 for a given conference. Results are shown in Fig. 2. The value in each cell denotes the values normalized by the number of publications in each conference overall. The figure also demonstrate the quality of our approach: topics like "Machine Learning" and "Information Systems" are popular for all considered conferences. "Database" topics are mostly popular in the VLDB conference series, while the topic "Digital Library" appears in JCDL and TPDL. Clearly, the extracted facet topics match the research focus of each conference. Also,

	ECDL	JCDL	TPDL	ICSE	VLDB	SIGIR	TREC	ICWSM	WWW	ESWC
Databases: Algorithms	0.0214	0.0249	0.0217	0.0176	0.1075	0.0362	0.0161	0.0288	0.0369	0.0264
Databases: Indexing	0.0045	0.0031	0.0007	0.0031	0.0758	0.0079	0.0015	0.0057	0.0069	0.0033
Databases: Modelling	0.0291	0.0116	0.0245	0.0292	0.1081	0.0225	0.0083	0.0076	0.0159	0.0279
Databases: Query Processing	0.0086	0.0056	0.0035	0.0117	0.1465	0.0150	0.0054	0.0057	0.0116	0.0160
Databases: Transaction Management	0.0024	0.0020	0.0035	0.0073	0.0728	0.0049	0.0012	0.0014	0.0043	0.0038
Digital Libraries	0.1909	0.1577	0.1058	0.0089	0.0063	0.0154	0.0076	0.0104	0.0116	0.0122
Information Retrieval: Evaluation	0.0122	0.0049	0.0014	0.0093	0.0046	0.0300	0.0797	0.0042	0.0061	0.0083
Information Retrieval: Mining	0.0003	0.0043	0.0014	0.0004	0.0016	0.0000	0.0017	0.0057	0.0052	0.0026
Information Retrieval: QA and NLP and and Complex Queries	0.0410	0.0282	0.0182	0.0085	0.0096	0.0780	0.2392	0.0212	0.0164	0.0191
Information Retrieval: Ranking	0.0395	0.0439	0.0357	0.0093	0.0098	0.1054	0.2548	0.0434	0.0299	0.0217
Machine Learning	0.0689	0.1415	0.1458	0.0542	0.0355	0.1592	0.1131	0.1619	0.1296	0.1070
Microsoft Technology	0.0217	0.0137	0.0070	0.0203	0.0160	0.0026	0.0073	0.0019	0.0143	0.0086
Misc Topics: Web Information Systems	0.2591	0.2549	0.2614	0.2067	0.1806	0.2513	0.1304	0.2011	0.1790	0.1215
Recommendation	0.0125	0.0280	0.0154	0.0080	0.0058	0.0450	0.0117	0.0477	0.0461	0.0191
Semantic Web: Knowledge Engineering & Representation	0.0178	0.0228	0.0596	0.0071	0.0114	0.0150	0.0100	0.0151	0.0449	0.2395
Semantic Web: Logic	0.0068	0.0011	0.0028	0.0026	0.0021	0.0009	0.0010	0.0024	0.0144	0.0753
Social Graph Algorithms	0.0042	0.0038	0.0077	0.0050	0.0152	0.0141	0.0041	0.0198	0.0417	0.0060
Social Graph Analysis	0.0255	0.0390	0.0294	0.0089	0.0154	0.0454	0.0327	0.0760	0.0664	0.0298
Social Media Analytics: Text-based	0.0065	0.0296	0.0308	0.0083	0.0052	0.0454	0.0151	0.2587	0.0561	0.0205
Software Engineering: Systems	0.0237	0.0181	0.0182	0.1705	0.0131	0.0123	0.0078	0.0109	0.0131	0.0176
Software Engineering: Testing & Formal Methods	0.0098	0.0060	0.0070	0.2095	0.0107	0.0093	0.0066	0.0071	0.0196	0.0136
System Engineering: Architecture	0.0175	0.0078	0.0112	0.0397	0.0216	0.0097	0.0034	0.0024	0.0097	0.0036
Web Engineering: Client-Side	0.0748	0.0658	0.0736	0.0283	0.0198	0.0498	0.0202	0.0387	0.0821	0.0350
Web Engineering: P2P & Distributed Systems	0.0116	0.0069	0.0021	0.0024	0.0099	0.0062	0.0034	0.0019	0.0134	0.0100
Web Engineering: System Modelling	0.0359	0.0307	0.0736	0.0892	0.0194	0.0093	0.0071	0.0109	0.0486	0.1063
Web Technologies	0.0062	0.0125	0.0112	0.0207	0.0048	0.0022	0.0027	0.0033	0.0389	0.0102
XML Databases	0.0478	0.0314	0.0266	0.0131	0.0707	0.0071	0.0080	0.0061	0.0371	0.0353

Fig. 2. Heatmap showing the relation between *research methods* and conferences. The values are normalized based on the numbers of papers in each conference.

other popular topics can be explored: conferences like JCDL or TPDL favor methods like Machine Learning, Digital Libraries, Web Information Systems, and Information Retrieval.

What are the trends for methods?: In order to answer this question, we visualize the number of publications covering a *method* facet topic (as listed in Table 3) over the course of the last 10 years. The results are shown in Fig. 3, giving an intuition about the quality of our approach: e.g., methods related to machine learning, software testing, or social media analytics gained great popularity in the last 10 years; while, as expected, topics related to core databases techniques or XML processing are becoming less popular.

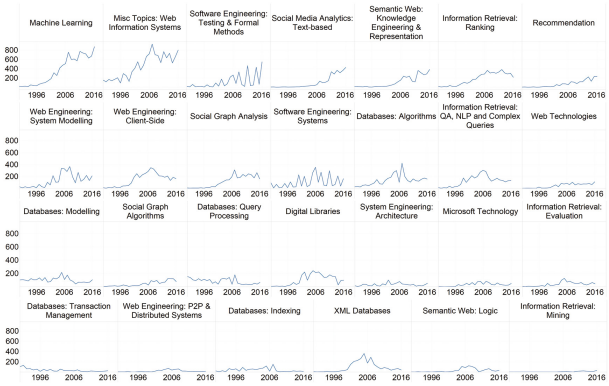

Fig. 3. The trends of *research methods* over years. The y axis shows the contribution of each topic in a certain year by means of the number of method-occurrence

6 Summary and Outlook

This paper presents the design and evaluation of a novel method for domain-aware topic identification for collections of scientific publications. Our method aims at improving the ability of digital libraries systems to support the retrieval, exploration, and visualization of documents based on topics of interest. In contrast to previous work, is taking advantage of some domain-specific insights which vastly improves the quality of the resulting topics, while still being adoptable to other domains by minimal efforts.

Our proposed method relies on a combination of sentence classification, semantic annotation, and semantic clustering to identify *Facet Topics*, i.e. clusters of semantically related *terms* that are tied to an *facet* relevant to an user's understanding of a document collection. The method specializes on the extraction of facet-specific information through the classification of rhetorical mentions in sentences. A lightweight distant supervision approach with low training costs (compared to traditional supervised learning) and acceptable performance,

allows for simple adaptation to different domains. Facet terms are extracted from candidate sentences using state-of-the-art semantic annotation tools, and are filtered according to their informativeness. *Facet Topics* are identified using a novel *Facet Embedding* technique.

We applied this novel method to a corpus of 11,589 publications on *data processing* from 10 conference series, and extracted metadata related to the 5 facets of the DMS [6] ontology for data processing pipelines. An extensive set of quantitative and qualitative analysis shows that, despite its simple design, our methods allows for topic identification performance superior to state-of-the-art topic modeling methods like LDA.

While promising, results leave ample space for future improvements. We are interested in investigating the performance of more complex machine learning classifiers (e.g. based on word-embeddings), possibly applied to group of related sentences. We also plan to investigate new techniques for facet terms extractions, and study the performance of our approach with larger amount of *Facet Topics*. Finally, we plan to expand our analysis to additional domains, and investigate new facets of interest.

References

1. Mathew, G., Agarwal, A., Menzies, T.: Trends in topics at SE conferences (1993–2013). arXiv preprint arXiv:1608.08100 (2016)
2. Shubankar, K., Singh, A., Pudi, V.: A frequent keyword-set based algorithm for topic modeling and clustering of research papers. In: 3rd Conference on Data Mining and Optimization (DMO), 2011, IEEE, pp. 96–102 (2011)
3. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inform. Sci. Technol. **57**(3), 359–377 (2006)
4. Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J., Möller, T.: Visualization as seen through its research paper keywords. IEEE Trans. Visual Comput. Graphics **23**(1), 771–780 (2017)
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
6. Mesbah, S., Bozzon, A., Lofi, C., Houben, G.J.: Describing data processing pipelines in scientific publications for big data injection. In: WSDM Workshop on Scholary Web Mining (SWM). Cambridge, UK (2017)
7. Mesbah, S., Fragkeskos, K., Lofi, C., Bozzon, A., Houben, G.-J.: Semantic Annotation of Data Processing Pipelines in Scientific Publications. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10249, pp. 321–336. Springer, Cham (2017). doi:10.1007/978-3-319-58068-5_20
8. Song, M., Heo, G.E., Kim, S.Y.: Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in DBLP. Scientometrics **101**(1), 397–428 (2014)
9. Afiontzi, E., Kazadeis, G., Papachristopoulos, L., Sfakakis, M., Tsakonas, G., Papatheodorou, C.: Charting the digital library evaluation domain with a semantically enhanced mining methodology. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference On Digital Libraries, pp. 125–134. ACM (2013)

10. Hoonlor, A., Szymanski, B.K., Zaki, M.J.: Trends in computer science research. Commun. ACM **56**(10), 74–83 (2013)
11. Gupta, S., Manning, C.D.: Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers
12. Tsai, C.T., Kundu, G., Roth, D.: Concept-based analysis of scientific literature. In: Proceedings of the 22nd ACM International Conference On Conference On Information & Knowledge Management - CIKM 2013, pp. 1733–1738 (2013)
13. Siddiqui, T., Ren, X., Parameswaran, A., Han, J.: FacetGist: Collective extraction of document facets in large technical corpora. In: Proceedings CIKM 2016 (2016)
14. Lopez, P.: GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 473–474. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04346-8_62
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **21**, 3111–3119 (2013)
16. Koren, J., Zhang, Y., Liu, X.: Personalized interactive faceted search. In: Proceeding of the 17th International Conference On World Wide Web - WWW 2008, pp. 477–485 (2008)
17. Cosley, D., Lawrence, S.: REFEREE: An open framework for practical testing of recommender systems using ResearchIndex. In: Proceedings of the 28th VLDB Conference, pp. 35–46 (2002)
18. Livne, A., Simmons, M.P., Adar, E., Adamic, L.a.: The Party is Over Here: Structure and Content in the 2010 Election. vol. 161(3), pp. 201–208 (2010)