SmartPub: A Platform for Long-Tail Entity Extraction from Scientific Publications

Sepideh Mesbah
Delft University of Technology
Delft, Netherlands
s.mesbah@tudelft.nl

Alessandro Bozzon
Delft University of Technology
Delft, Netherlands
a.bozzon@tudelft.nl

Christoph Lofi
Delft University of Technology
Delft, Netherlands
c.lofi@tudelft.nl

Geert-Jan Houben
Delft University of Technology
Delft, Netherlands
g.j.p.m.houben@tudelft.nl

ABSTRACT

This demo presents SmartPub, a novel web-based platform that supports the exploration and visualization of *shallow meta-data* (e.g., author list, keywords) and *deep meta-data* – long tail *named entities* which are rare, and often relevant only in specific knowledge domain – from scientific publications. The platform collects documents from different sources (e.g. DBLP and Arxiv), and extracts the domain-specific named entities from the text of the publications using Named Entity Recognizers (NERs) which we can train with minimal human supervision even for rare entity types. The platform further enables the interaction with the Crowd for filtering purposes or training data generation, and provides extended visualization and exploration capabilities. SmartPub will be demonstrated using sample collection of scientific publications focusing on the computer science domain and will address the entity types Dataset (i.e. dataset presented or used in a publication), and Methods (i.e. algorithms used to create/enrich/analyse a data set).

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; *Document structure*; *Content analysis and feature selection*;

KEYWORDS

Information Extraction. Document Metadata, Named Entity Recognition, Long-Tail Entity Types, Training Data Generation

ACM Reference Format:
Sepideh Mesbah, Alessandro Bozzon, Christoph Lofi, and Geert-Jan Houben. 2018. SmartPub: A Platform for Long-Tail Entity Extraction from Scientific Publications. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186976

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186976

1 INTRODUCTION

For years, online digital libraries like the ACM Digital Library, IEEE Explore, ArXiv, etc. provided search functionalities for exploring academic publications, and have thus become a fundamental part of modern research processes. However the retrieval functionality of current systems are often limited to searching on *shallow* meta-data such as the title, the authors, keywords. They are usually not designed to support the analysis of *deep* meta-data such as topics of domain-specific interests like used datasets or algorithms relevant for scientific computer science publications. While such systems exist for some domains like medicine or biology, the costs for obtaining deep meta-data are generally prohibitive for widespread application.

Discovering *deep* meta-data from scientific publications could enable complex entity-centric queries. For instance, a researcher in the field of machine learning could be interested in a query like: *discovering the state of the art image classification research methods that have been successfully applied to the Imagenet dataset.* For answering the query above, a system requires to have access to entities such as the dataset used (e.g Imagenet), the research methods that have been applied on the datasets (e.g. LSTM neural network), etc. The automatic recognition and typing of such named entities rely either on supervised machine learning models, trained on expensive type-labeled data produced by human annotators or the generation of labeled training data from knowledge bases which is not suitable for long-tail entity types that are not very representative in knowledge bases.

Contribution. In this demo we introduce SmartPub, a web-based platform that extracts long-tail entity types from scientific publication based on minimal human input, namely a small seed set of instances for the targeted entity type. Furthermore it supports the exploration and visualization of *deep* meta-data of scientific publications, i.e. meta-data able to represent domain-specific properties and aspects in which a document can be considered and understood within its (research) domain.

Users of the demo can interactively explore and visualize a collection of scientific computer science publications, by e.g. browsing for specific entities, tracking trends, discovering central concepts, or explore the usage of given entities over time. An example of the demonstration is available as a video screencast at the following address: https://youtu.be/zLLMwOT5sZc.

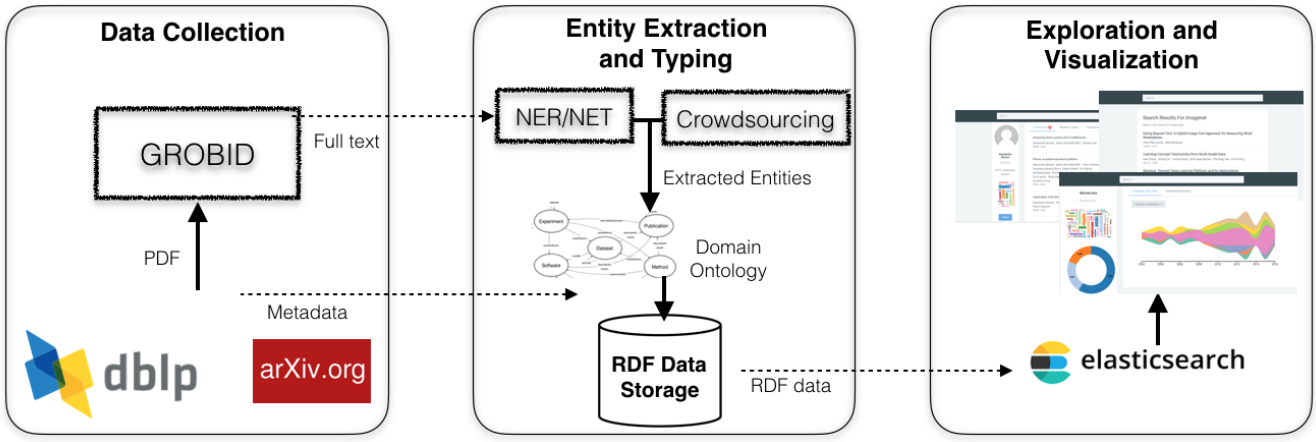


Figure 1: Architecture of the SmartPub platform.

Paper Organisation. The remainder of the paper is structured as follows. Section 2 describes the architecture of the SmartPub system, detailing its components and provided functionality. Finally, Section 3 describes the demonstration provided to conference attendees.

2 THE SMARTPUB SYSTEM

The architecture of the SmartPub system is depicted in Figure 1, where three major components are highlighted. The *Data Collection* is responsible for the retrieval of full texts and standard metadata (e.g. title, authors) of scientific publications. The *Entity Extraction* component focuses on the extraction of domain-specific entities from the publications' text, and builds a knowledge repository based on a pre-defined domain ontology. Finally, the *Exploration and Visualization* component offers user interfaces for exploration of the publications in the collection based on the extracted entities.

2.1 Data Collection

In the current implementation the data collection component retrieves scholarly data from DBLP[1] (a computer science digital library) and ArXiv[2]. For each paper, DBLP provides an XML entry that contains bibliographic meta-data (i.e. title, author names, year of publication) as well as the DOI url from which the publications' PDF can be retrieved. ArXiv offers open access to 1.4 million PDFs of scientific publications in different domains. In the next step, the retrieved PDFs are processed using GROBID (GeneRation Of BIbliographic Data) [3], a state-of-the-art extraction engine. GROBID extracts a structured full-text representation as Text Encoding Initiative(TEI)-encoded documents, thus providing easy and reliable access paragraphs and sentences.

2.2 Entity Extraction and Typing

The entity extraction component is designed to identify and type the *domain-specific entities* contained in the fulltext of a publication. All the metadata from a paper are then published in a RDF repository, encoded according to the DMS (Dataset, Method, Software) ontology [4]. In this demo we focus on the entity types *Dataset* (i.e. dataset presented or used in a publication), and *Methods* (i.e. algorithms used to create or analyse a data set).

The Entity Extraction and Typing component is organized into two sub-components namely *NER/NET* and *Crowdsourcing*. The extraction of entities relies on NER/NETs (Named Entity Recognition/Named Entity Typing) algorithms trained with minimal human supervision. SmartPub allows the interaction with crowds for training data creation or filtering purposes.

2.2.1 NER/NET. The goal of *NER/NET* sub-system is to address the problem of long-tail entity recognition with minimal human input. The training of *domain-specific* NER/NETs is a challenging task due to: 1) the *long-tail* nature of such entity types, both in existing knowledge bases *and* in the targeted document collections [8]; and 2) the high cost associated with the creation of hand-crafted rules or human-labeled training datasets for supervised machine learning techniques.

SmartPub integrates results from our previous work [5, 6], and extends them as depicted in Figure 2. Starting from a seed set of instances of the targeted entity type (e.g. method), (1) we obtain text snippets from the publication corpus to be used in a first training data extraction step; (2) the set of seed instances are then semantically expanded to include potential yet unknown instances. For this, the word2vec model (100 dimensions) is trained on the whole corpus, as described in [7], to learn all uni- and bi-gram word vectors for all terms in the corpus. Then, we use a pre-trained entity recognition library (e.g. the one provided by the NLTK package) to obtain a list of all entities contained in the training data. Entities are then clustered with respect to their embedding vectors using K-means clustering; silhouette analysis is used to find the optimal number k of clusters. Finally, clusters that contain at least one of the seed terms are assumed to (only) contain entities of the same type. In the third step (3) the set of training snippets are semantically expanded to include sentences which are unlikely to contain instances of the desired type, but are still very similar in semantics and vocabulary to serve as informative negative examples in order to boost the NER training accuracy. For this, we rely on *doc2vec*

[1] http://dblp.uni-trier.de/
[2] https://arxiv.org/

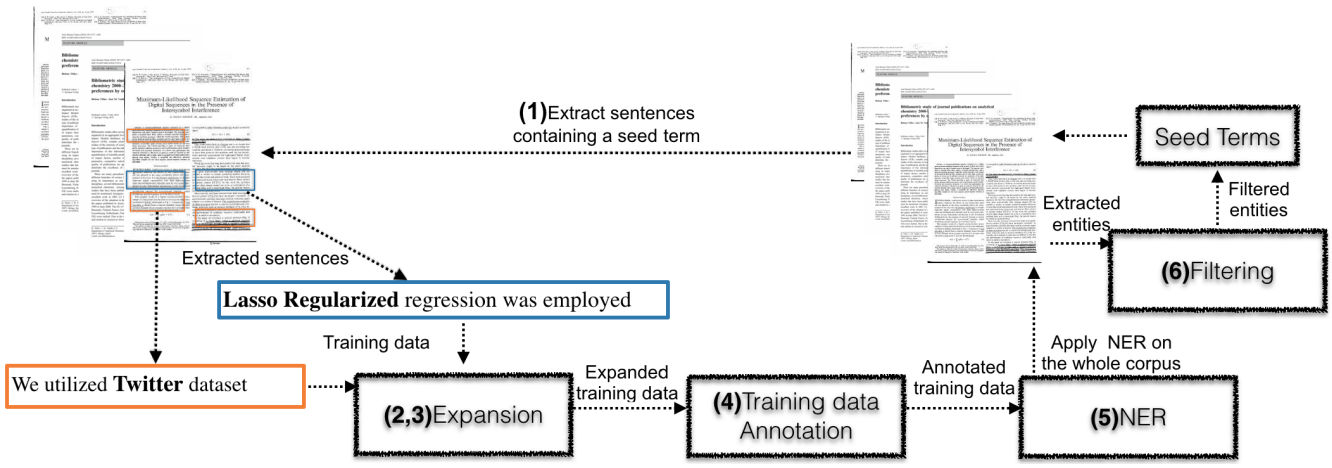


Figure 2: Overview of the domain-specific long-tail named entities recognition approach

document embeddings [2], a variant of *word2vec*, to learn vector representations of the sentences in the corpus. For each sentence in the development set, we use *doc2vec* (100 dimensions) to discover the most similar sentence which does not contain any known instance of the targeted type (i.e., expanded terms). Such sentences sometimes are likely to contain an unknown instance of the targeted entity type, which would now be misclassified in the training set. In the fourth step (4) the training data is annotated with the expanded instance list, so to (5) train a NER that is then applied on the document corpus to extract new named entities of the given types. In a final step (6) the extracted entities are processed trough a set of filters that heuristically exclude likely misclassified instances (e.g excluding general english words using wordnet[3]), thus yielding the final result set. For training a new *NER*, we used the Stanford NER tagger[4] to train a Conditional Random Field (CRF) model.

This automatic approach relies on minimal human input (the seed set of entities), and can operate in an iterative fashion by being repeated using the result set as a seed for the next iteration. We compared our method with the BootStrapping (BS) based concept extraction approach [9], a commonly used state-of-the-art technique in scientific literature.

Experiments [5, 6] shown that our approach can provide good quality results in terms of precision/recall/fscore for the `dataset` entity type (0.77/0.30/0.43) compared to *BS* (0.08/0.13/0.10) and for the `method` entity type (0.68/0.15/0.25) compared to *BS* (0.11/0.32/0.16), with a seed set of 100 entities. We infer that different expansion strategies augment the performance of our technique compared to the *BS* which just relies on features such as unigrams, bigrams, closest verb, etc.

2.2.2 Crowd-sourcing. The *Crowd-sourcing* [1] component is responsible to close the loop with the final users to help improving the performance of the NER/NET model. The crowd-sourcing component samples annotated sentences from the corpus and offers them the possibility to filter out irrelevant entities, so to reduce the number of false positives detected by the noisy NER. The current version of `SmartPub` uses the uncertainty sampling strategy[5] (e.g. least confidence, smallest margin), to rank unlabeled examples for annotation. To assess the quality of users' annotations, `SmartPub` currently implements a simple labeling aggregation scheme based on majority voting. Crowd-labeled sentences are then used to re-train the existing model, to achieve higher accuracy and/or identify new entity types. Moreover, the crowd-sourcing component also generates entity linking tasks. The task requires linking entities to an instance in the knowledge base, which entails annotating an ambiguous entity mention (e.g. SVM) with a link to the unique instance (e.g Support Vector Machine).

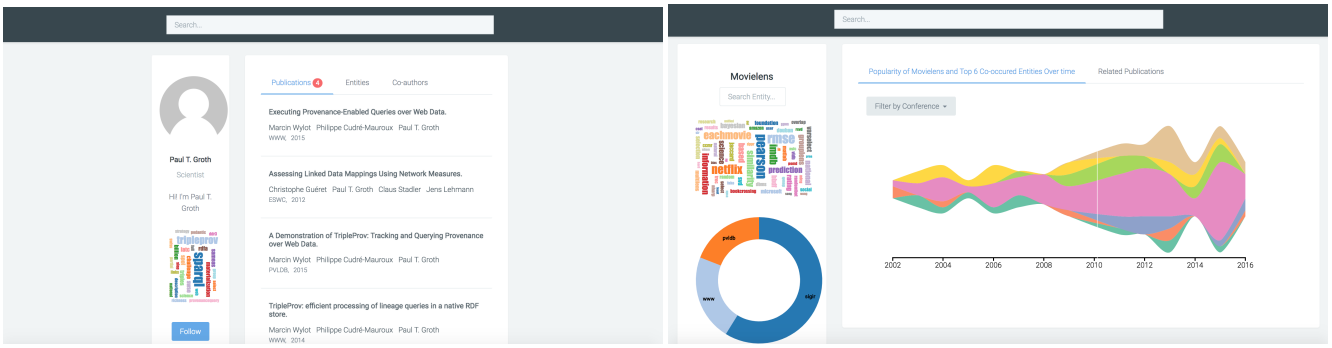
2.3 Exploration and Visualization

All the documents as well as the extracted entities in the corpus are indexed using Elasticsearch[6]. We designed an easy to use user interface to explore publications, authors and the domain specific entities (as in Figure 3).

The publications can be explored using the title, authors name or the fulltext. For each publication, `SmartPub` shows the entities extracted from the fulltext. Figure 3a shows an example of exploring authors. For each author we show, the list of publications in our corpus, list of co-authors as well as the extracted entities from the full text of the authors publications.

`SmartPub` currently offers the following set of visualizations for a given entity: *1) Popularity Over Time* in the shape of a stream graph. As depicted in Figure 3b on the right, the Stream Graph displays the contribution of a given entity and its top six co-occurred entities in a certain year by means of the number of entity-occurrence. The thickness of the graph shows the popularity of the entity in a year. Stream Graphs are ideal for discovering trends over time across a wide range of categories. Different colors in the graph are indicators of different entities and the name of the entities are displayed with hover interactivity. The Stream Graph can be further

[3]http://wordnet.princeton.edu/
[4]https://github.com/dat/stanford-ner
[5]https://github.com/ntucllab/libact
[6]https://www.elastic.co/



(a) Explore Authors **(b) Visualize entities**

Figure 3: Examples of data visualisations dashboard of the SmartPub platform

filtered according to the conference using a multi-select dropdown list. *2) Popularity Over Conferences* given conference in the shape of a pie chart. Figure 3b left shows the number of papers including a given *entity* in different conference series. 3) *Co-occured* entities in the shape of word cloud. Figure 3b left shows the word cloud, a graphical representation of the frequency of co-occured entities.

3 DEMO HIGHLIGHTS

We will present the demo using sample of scientific publications with a focus on data science and processing. In our corpus, we have 11,589 papers from ten conference series. The Joint conference on Digital Libraries (JCDL – 1,416 papers, 2001–2016); the International Conference on Theory and Practice of Digital Libraries (TPDL – 276 papers, 2011–2016); the International Conference on Research and Development in Information Retrieval (SIGIR – 412 papers, 1971–2016); the Text Retrieval Conference (TREC – 1,444 papers, 1999–2015); the European Conference on Research and Advanced Technology on Digital Libraries (ECDL – 820 papers, 1997–2010); the International Conference on Software Engineering (ICSE – 1834 papers, 1976–2016); the Extended Semantic Web Conference (ESWC – 626 papers, 2005–2016); the International Conference On Web and Social Media (ICWSM – 810 papers, 2007–2016); the International Conference on Very Large Databases (VLDB – 1884 papers, 1975–2007); and the International World Wide Web Conference (The Web Conference – 2067 papers, 2001–2016). The demonstration will focus on exploring scientific papers, authors as well as visualizing entities extracted from the full text of the publications by means of their popularity over time or conferences.

The demonstration starts by searching for publications containing an entity name (e.g. `clueweb`). A list of relevant publications is listed, showing meta data such as the title,authors name as well as the venue and publication year. By clicking on the author name, we can navigate to the *author* page. For each author, `SmartPub` shows publications in the corpus, the list co-authors, and the list of entities extracted from the author's publications, which are shown as a word cloud below the name of the author. Entities in the word cloud are clickable, leading to a separate tab called *Entities* which contains the list of entities with their corresponding entity types.

By clicking on each of the publications title we can navigate to the *publication* page which contains the abstract, the references as well as the entities extracted from the full text of the papers. By clicking on each of the entities listed in the entity tab we navigate to the *entity* page. For each entity, `SmartPub` offers a set of visualizations described in Section 2.3. As an example for the entity name `Clueweb`, in the stream graph we show the popularity of `Clueweb` and its top six co-occured entities (i.e. `wikipedia`, `urls`, `trec`, `nist`, `dbpedia`, `bm25`) in a certain year which can further be filtered based on a given conference. The Pie chart on the left shows that the `Clueweb` entity is mostly popular in information retrieval conferences such as TREC and SIGIR. The word cloud below the entity name depicts the co-occured entities with the given entity, which are all clickable. The users are able to search for any entity using the search box below the entity name on the left. Finally an example of a crowdsourcing task is shown, where the users are asked to select the appropriate label for the highlighted token.

Acknowledgments. This research has been supported in part by the Dutch national e-infrastructure with the support of SURF Cooperative (Grant Agreement No. e-infra170126).

REFERENCES

[1] A. Bozzon, P. Fraternali, L. Galli, and R. Karam. Modeling crowdsourcing scenarios in socially-enabled human computation applications. *Journal on Data Semantics*, 3(3):169–188, Sep 2014.
[2] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.
[3] P. Lopez. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *European Conference on Digital Library (ECDL)*, Corfu, Greece, 2009.
[4] S. Mesbah, A. Bozzon, C. Lofi, and G.-J. Houben. Describing data processing pipelines in scientific publications for big data injection. In *Proceedings of the 1st Workshop on Scholarly Web Mining*, pages 1–8. ACM, 2017.
[5] S. Mesbah, K. Fragkeskos, C. Lofi, A. Bozzon, and G.-J. Houben. Facet embeddings for explorative analytics in digital libraries. In *International Conference on Theory and Practice of Digital Libraries*, pages 86–99. Springer, 2017.
[6] S. Mesbah, K. Fragkeskos, C. Lofi, A. Bozzon, and G.-J. Houben. Semantic annotation of data processing pipelines in scientific publications. In *European Semantic Web Conference*, pages 321–336. Springer, 2017.
[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
[8] R. Reinanda, E. Meij, and M. de Rijke. Document filtering for long-tail entities. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 771–780. ACM, 2016.
[9] C.-T. Tsai, G. Kundu, and D. Roth. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1733–1738. ACM, 2013.