Increasing Datasets Discoverability in an Engineering Data Platform using Keyword Extraction

Parthasarathy Gopavarapu
Purdue University
112 E Wood St
West Lafayette, IN 47906
+1 765 771 9122
pgopavar@purdue.edu

Line C. Pouchard
Purdue University
504 W State Street
West Lafayette, IN 47907
+1 765 494 3875
pouchard@purdue.edu

Santiago Pujol
Purdue University
550 W Stadium Ave
West Lafayette, IN 47907
+1 765 496 8368
spujol@purdue.edu

ABSTRACT
In this paper we describe the use of keyword extraction in a data management platform for the storage, publication, and sharing of scientific and engineering datasets primarily related to the stress of concrete structures under earthquake conditions. To improve discoverability of datasets and assist scientists who upload data, we designed an automated keyword extraction system that will propose keywords for uploaded datasets.

CCS Concepts
• **Computing methodologies→Artificial Intelligence**

Keywords
Data discovery, data repository, keyword extraction.

1. INTRODUCTION
Datacenterhub (http://datacenterhub.org) is a platform for the storage, publication, and sharing of scientific and engineering datasets primarily related to the stress of concrete structures under earthquake conditions. Datacenterhub was created to provide a searchable and browsable platform to individual researchers who accumulate and analyze measurements and observation datasets of concrete structures that they have taken in the course of their research. Heterogeneous datasets include text descriptions, measurements, pictures, reports and drawings. Datasets are described with extensive metadata that includes experiment and case names, titles, contributor names, geo-coordinates, related publications, article citations, keywords, abstracts and other. Datacenterhub was developed as an extension to HUBzero®, a generic content management and collaboration platform developed and maintained at Purdue University.

One of the goals of Datacenterhub is to facilitate the discovery and exploration of datasets that have been uploaded in the platform, either by researchers who are interested in submitting datasets or by those who want to explore datasets underlying a publication. One obstacle to realize this goal is that researchers who submit their datasets often do not provide enough keywords for their datasets to enable powerful searches in Datacenterhub. This is compounded by the fact that datasets in the platform are self-published; thus the publication does not benefit from the assistance of a curator, for suggesting or adding new keywords.

JCDL2016, June 19-23, 2016, Neewark, NJ, USA.

DOI: http://dx.doi.org/10.1145/12345.67890

We have designed a strategy to mitigate this situation and we are implementing a series of steps to increase the number of keywords for searching for datasets. Keywords obtained by keyword extraction are offered to researchers for annotating their data when they upload their experiment. The unique contributions of this poster include a method for increasing the searchability and access to datasets in the platform and the application of text mining methods to keyword extraction.

2. PROJECT DESCRIPTION
Figure 1 shows the flow of processes required for performing keyword extraction.

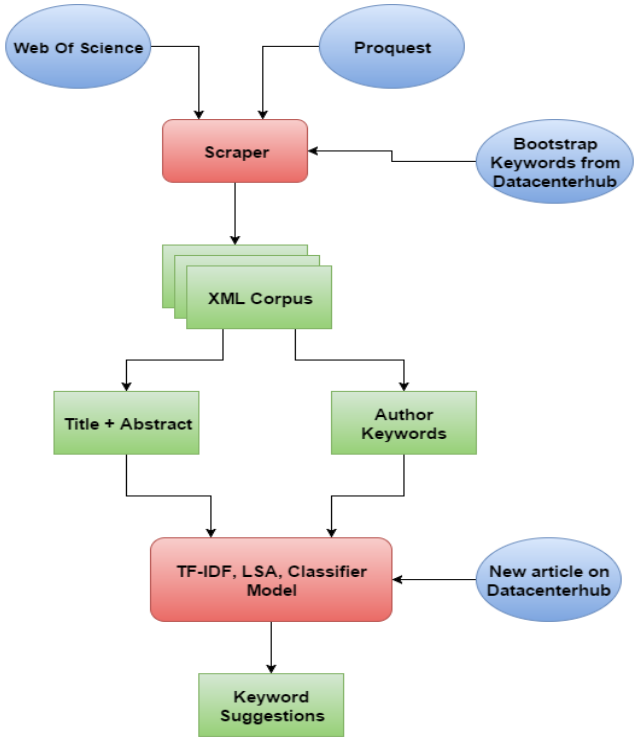

Figure 1: Workflow diagram for the Keyword Extraction Process. Ovals represent existing resources, rounded rectangles represent code/algorithms and rectangles represent created resources.

The keyword extraction process consists of **two major phases**. The **first phase** involves building a corpus of scientific articles in structural engineering, using title, abstract, and keywords from

web based resources such as Web of Science (http://webofknowledge.com) and Proquest (http://search.proquest.com). The **second phase** consists of utilizing the corpus created in phase one, to train machine learning models that can extract keywords, given new documents such as abstracts. These abstracts are uploaded by users in the platform to document their experiments and datasets.

2.1 Phase 1

We use a bootstrap set of keywords obtained from Datacenterhub to initialize the process of corpus creation. Using these keywords, we build a corpus of scientific articles in the structural engineering domain. We use the title and abstract fields of these articles.

Bootstrap Keyword Set: This set is the list of keywords obtained from publicly available datasets on Datacenterhub. Some of the non-relevant keywords are eliminated and keyword normalization is performed.

Scraper: Once the normalized keyword list is created, each of these keywords is used as a search phrase on both Web of Science and Proquest. We use python scripts that utilize modules such as Mechanize (wwwsearch.sourceforge.net/mechanize) and BeautifulSoup (www.crummy.com/software/BeautifulSoup) to perform the search programmatically and process the results. Mechanize is used to fill and submit the search forms on these websites and BeautifulSoup is used to process the results obtained on submission. Once the search results are obtained, each of the result articles are processed to extract specific fields to create an XML document.

XML Corpus: The XML corpus contains one xml document per search result made of title and abstract. Duplicates are eliminated during the scraping process. Our corpus currently includes 36132 XML documents representing scientific articles amounting to 174 MB.

2.2 Phase 2

In Phase 2 we use the XML corpus of documents corresponding to the search results of the bootstrap keywords. We extract Title, Abstract and Keywords. The string obtained from concatenating Title and Abstract (Text) is used to represent the document and Keywords are considered labels for processing algorithms. We train a Multi-Label Binary Classifier using the Text and Keywords dataset created above. [1]

Dataset Representation: For each document the Text (Title+Abstract) is replaced using a feature representation learnt from the corpus. We use Latent Semantic Analysis (LSA) [2] along with TF-IDF [3] features so that the conceptual information of the document can be represented in addition to the importance of the terms in the document.

TF-IDF: Term Frequency - Inverse Document Frequency (TF-IDF) is a statistic used to measure the relevance of terms present in a set of documents to the content of the documents. The documents are represented using a TF-IDF matrix where each row represents one document from the corpus and each entry in the row represents the TF-IDF value for the corresponding document and word pair.

Latent Semantic Analysis: LSA is a document modeling technique which uses the occurrence matrix (M) of a corpus to find the topic distributions of the documents in the corpus. This matrix M is split into three matrices, U, Σ, and V using Singular Value Decomposition (SVD). The matrix V_{mxk} represents the documents in a lower dimensional space and is considered as the topic distribution of the documents where k represents the number of topics and m the number of documents. The TF-IDF and LSA features obtained are concatenated to form a new feature representation for the documents.

Keyword Suggestion: A list of words containing all the distinct keywords from the corpus is created. For each document a new binary vector of the same size as the keyword list is created which contains 1 if the corresponding word in the list is also a keyword for the document. This vector acts as the label for the document represented using TF-IDF and LSA.

Once the dataset is prepared with the feature representation of documents and the labels, we can use it to train classification tools [1]. We use K-Nearest Neighbor classifier in our implementation due to its ease of use and speed of training. Once the classifier is trained, the keywords for any new document can be predicted using the following process:

- Use the TF-IDF and LSA model to project the document into the new feature representation.

- Provide the feature representation obtained above as input to the classifier model trained using the corpus.

- This will output a binary vector of the same size as the list of keywords. All the words corresponding to entries containing 1 in the vector can be reported as the **keyword suggestions** for the new document.

2.3 Phase 3

In this phase, we integrate the results of the keyword extraction into the Datacenterhub interface. This phase is not represented in Figure 1, and is only in the planning stage at the time of this writing. We plan for the integration of a drop-down menu of keywords based on the suggestions extracted in Phase 2. This menu will reside on the upload interface of Datacenterhub.

3. CONCLUSION

We collected an XML corpus of scientific abstracts covering the domain of the datasets collected in Datacenterhub. We use it to create keyword extraction models so that researchers who submit their datasets to Datacenterhub are provided with additional keywords for their datasets. We offer these keyword suggestions to users to assist them in increasing the discoverability of their datasets.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers CIF21-DIBBS 1443027.

5. REFERENCES

[1] Hasan, KS. And Vincent Ng.2014. Automatic Keyphrase Extraction : A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1262–1273, 2014.

[2] Dumais, S.T. 2005. Latent Semantic Analysis. *Annual Review of Information Science and Technology*. 38,1(Sep. 2005) 188-230.

[3] Salton, G. and Buckley, C. 1988. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management.* (1988), 513-524.