Delve: A Data set Retrieval and Document Analysis System

Uchenna Akujuobi (✉) and Xiangliang Zhang

King Abdullah University of Science and Technology
{uchenna.akujuobi,
xiangliang.zhang}@kaust.edu.sa

Abstract. Academic search engines (e.g., Google scholar or Microsoft academic) provide a medium for retrieving various information on scholarly documents. However, most of these popular scholarly search engines overlook the area of data set retrieval, which should provide information on relevant data sets used for academic research. Due to the increasing volume of publications, it has become a challenging task to locate suitable data sets on a particular research area for benchmarking or evaluations. We propose Delve, a web-based system for data set retrieval and document analysis. This system is different from other scholarly search engines as it provides a medium for both data set retrieval and real time visual exploration and analysis of data sets and documents.

1 Introduction

The area of scholarly search engines although sparsely studied, is not a new phenomenon. Search engines provide a new insight into scholarly information searchable on the web, incorporating functionalities to rank and measure academic activities [3]. However, due to the unprecedented rate in the number of scholarly papers published per year [4], researchers often go through an exhaustive step of re-searching and reading through many documents to locate usable data sets (i.e., relevant benchmark/evaluation data sets) that fits their research problem setting. It is therefore, desirable to have a platform where experts and non-experts are able to access not just topic or document information but also relevant data sets, together with the ability to analyze their interconnection. This task can be structured as an information retrieval task [5]. Current systems are designed either for data set search[1] or for scholarly search[2]. One system [1] incorporated the use of data set as a filter agent for their document search results. However, users are often interested in *locating data sets relevant to their search* rather than using the data sets to filter their search.

In Delve[3], we take a different approach by designing a system that allows users to locate both relevant documents and data sets, and also to visualize

[1] http://www.re3data.org/
[2] http://www.scholar.google.com/
[3] The system can be seen in action at https://youtu.be/bF6PUj8801U

and analyze their relationship network. Delve borrows ideas from label propagation [2] algorithm and adopts methods proposed in ParsCit [6] for text mining. Our system also provides a simple and easy-to-use interface built on the d3.js[4] framework which facilitates visualization and analysis of papers and data sets.

2 System Design

Our data set was constructed with an initial focus on academic documents published in 17 different conferences and journals between **2001** to **2015**, including ICDE, KDD, TKDE, VLDB, CIKM, NIPS, ICML, ICDM, PKDD, SDM, WSDM, AAAI, IJCAI, DMKD, WWW, KAIS and TKDD. Using the Microsoft graph data set[5], we then extended these documents, adding their references and the references of their references (up to 2 hops away). In total, we currently have **2,116,429** academic publications from more than 1000 different conferences and journals.

Data set and document analysis. Our system is built on the citation graph of these more than 2 million papers. Formally, in a directed citation graph $G = \{V, E\}$, two nodes v_i and v_j are linked by edge $E(v_i; v_j)$ if v_i cites v_j. Since the system is designed for data set relevant retrieval, an edge $E(v_i; v_j)$ between v_i and v_j can be labeled as: 1 - if v_i cites v_j because v_i uses the data set available/used in v_j ; and 0 - otherwise. Then based on the labels, we can extract the data set labeled citations. The initial labeling work was conducted by crowd-sourcing on papers and data sets cited by papers published in ICDE, KDD, ICDM, SDM and TKDE from 2001 to 2014. These labels (accounting for ***5% of the whole graph edges***) have been manually verified to be correct by three qualified participants. Due to the high cost, it is infeasible to label the remaining 95% of edges manually. Therefore, the main challenging task is to develop a correct and yet efficient algorithm to efficiently assign labels to the large amount of unlabeled edges using the limited amount of verified labels. To solve this problem, we developed a semi-supervised learning method ***"link label propagation algorithm"*** using ideas borrowed from label propagation algorithm [2].

Label Assignment. The original label propagation (LP) algorithm predicts labels for nodes, our task is to predict labels for edges. Therefore we restructure the original graph to $G' = \{V', E'\}$ where V' is the set of edges E in graph G, and E' is the set of generated edges. The edges E' are generated by linking each edge E_i in G (V'_i in G') to the top 10 similar edges E_j (V'_j in G') that have the same target node as E_i or where the target node of E_i is the source node of E_j. To define the similarity between citations, we extract the number of data set keywords[6] from each citation context (i.e. the sentences which encompass the citations). We then defined a Gaussian similarity score between pairs of edges (E_i, E_j) $Sim_{ij} = \exp(-\frac{\|d_i - d_j\|^2}{2\sigma^2})$, where $d_i = \frac{n_d}{n_c}$. n_d is the number of data set

[4] `https://d3js.org/`
[5] `https://academicgraph.blob.core.windows.net/graph-2015-11-06/index.html`
[6] Manually compiled list of data set related words

related words in the sentences which encompasses the citation depicted as E_i, and n_c is the number of such sentences in the source papers. For edges having the same target nodes, we assign a weight of $1 + Sim_{ij}$, and $0.5 + Sim_{ij}$ otherwise.

With the constructed graph $G' = \{V', E'\}$ where a small portion of V' have verified labels, label propagation algorithm is run to propagate the given labels to unlabeled V'. We conducted extensive experiments to evaluate our designed method. Our system achieves an average precision of 82%.

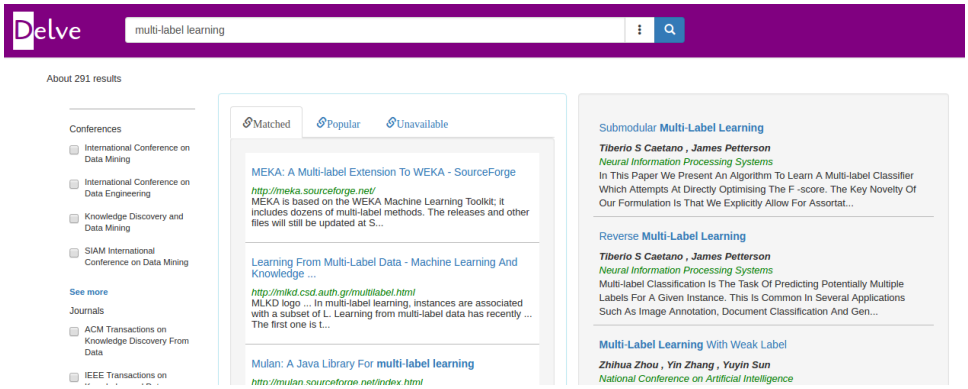

Fig. 1. Results from searching for "multi-label learning" in Delve

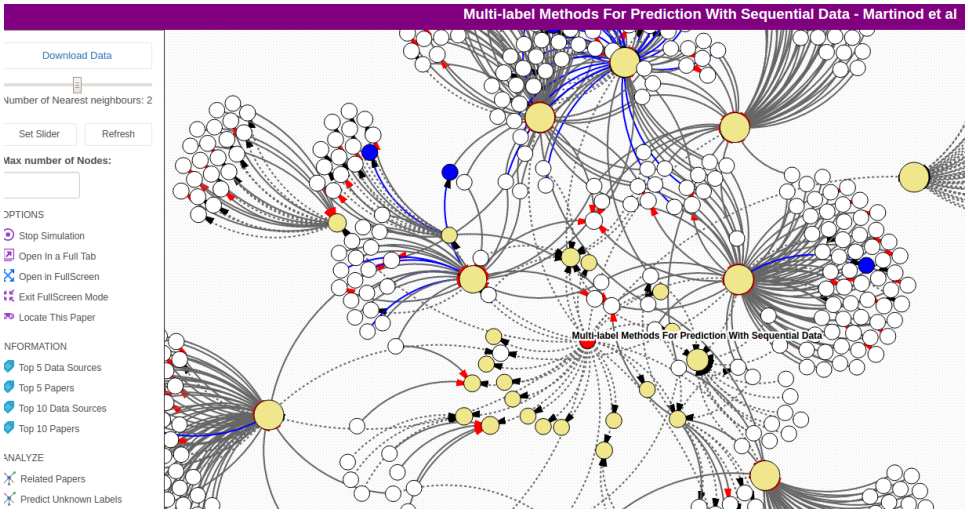

Fig. 2. Final output of uploaded file analysis in Delve

3 Use cases

Delve is based on two components: search and online document analysis.

Search: This enables users to search on a keyword, author or phrase for both documents and data sets. Delve analyzes this query and presents the user with results (outputs) ranked by relevance. Fig. 1 shows the result of the query "multi-label learning". The search result is split into two : data set results and scholarly document results. The data set result is futher split into three parts: 1. Matched data sets (data sets matching the search query). 2. Popular data set (data sets used by the papers matching the search query ordered by popularity). 3. Unavailable data sets (currently temporary or permanently inaccessible relevant data sets, e.g., invalid or closed links). Data sets can be either papers where the data sets are described or web links where the data sets are located.

On-line document analysis: This function enables a user to analyze a paper by understanding its relationship with other papers and data sets without having to go through the references; searching each of them manually. It can also be used by authors to discover which papers are advisable to cite in their work. A user can either analyze any document in our database or upload a scholarly document file for analysis, e.g., a PDF file. When a document is uploaded for analysis, Delve mines and analyzes the document text, translates the results as a query and displays the result as a visual citation graph, as shown in Fig. 2, which gives the result of analyzing *Multi-label methods for prediction with sequential data* [7]. We would like to point out that this paper is not in our system at the moment of writing this paper. However, based on its references and citations, our system can analyze its relevant papers and visualize the citation relations.

Note that in Fig.2, the blue edges indicate data set relevant relationships, and the size of the nodes show its importance in the network measured based its citations in the subgraph. Mouse hovering over a node displays the item title and clicking on a node displays more information about the item. In addition, the red edges show a non data set relevant relationship, and broken edges have unknown labels. The unknown labels can be inferred using label propagation.

References

1. "Semantic Scholar." Web. 21 Feb. 2017. `https://www.semanticscholar.org`.
2. Fujiwara, Yasuhiro, and Go Irie. "*Efficient label propagation*". (ICML-14). 2014.
3. Ortega, Jos Luis. "*Academic search engines: A quantitative outlook*". Elsevier, 2014.
4. National Science Board. "*Science and engineering indicators*". 2012. Arlington, VA, USA: National Science Foundation.
5. Sathiaseelan, J. G. R. "*A technical study on Information Retrieval using web mining techniques*". (ICIIECS) 2015.
6. Councill, Isaac G. et al. "ParsCit: an Open-source CRF Reference String Parsing Package" LREC (2008).
7. Read, Jesse, et al. "Multi-label methods for prediction with sequential data". Pattern Recognition, Volume 63, 2017: 45-55.