Automatic Citation Extraction and Analysis

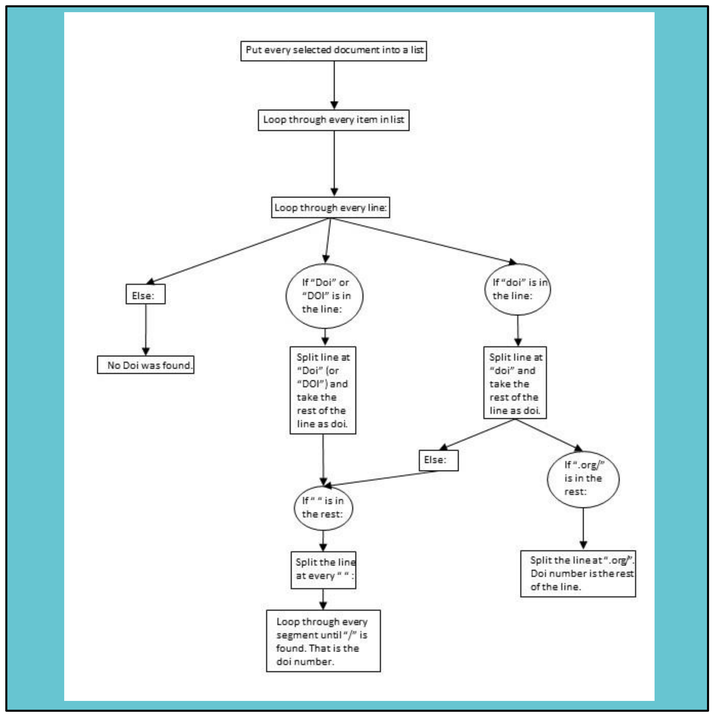
Tyler J. Conn, Samuel Hill, Ahmed Mustafa, Dr. Andrea Tartaro

Introduction

Large scale scientific research projects can result in breakdowns in communication, planning, and eventually goals. To promote and measure cohesion within a project, we propose analyzing bibliographic cross citing and textual similarities in research papers. To decrease effort required from the research scientists, this information must be automatically extracted from the research papers. This information can then be used to identify cohesion in the research project and possibly suggest new collaborations.

DOI Extraction

A DOI, or Digital Object Identifier, is a unique string of letters and numbers used to identify a single piece of published work online. By automatically extracting a DOI from a research paper, we can then use it to identify the article within the PubMed database and extract its information for later use.

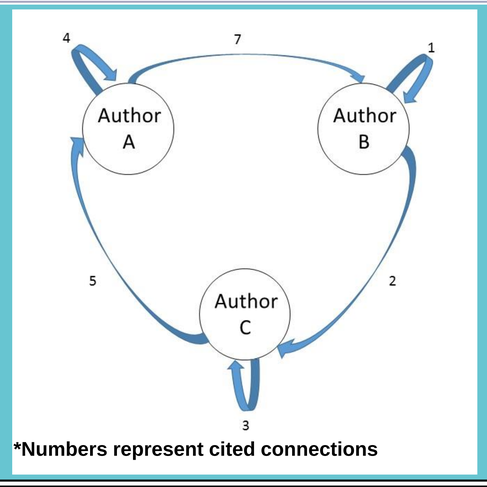


Basic Algorithm for DOI Extraction

Information Extraction

After obtaining the DOI, it is entered into the MetaPub module to extract information on both the original article and all of its citations [1]. The pulled information ranges from author names to dates or subjects in order to find the most possible connections. This data is then transferred to a CSV that is read by a python script that creates the network of information in a RDF(directed, labeled graph data format in the Web) database called SPARQL. The data is then analyzed through the Weka data mining software [2]

```
SELECT ?title
WHERE
{
<http://example.org/book/book1>
  <http://purl.org/dc/elements/1.1/title> ?title .
}
```



*Numbers represent cited connections

Database Sample Query (Right)
General Concept behind Connections(Left)

Results

10.1063/1.3529919	######	10.1088/1758-5082/3/3/0341(	10.1007/springerreference_1736	10.2351/1.3633221	10.2961/jlmn.2011.02.0013	10.1016/j.matlet.2011.02.057
10.1371/journal.pcbi.1002588	######	10.1016/j.ejor.2012.01.027	10.2316/p.2012.769-059	10.1137/110845732	10.1515/9783110288537.111	10.1371/journal.pone.0038448
10.4049/jimmunol.0900789	######	10.2174/187482870090201000	10.4049/jimmunol.0901517	10.1007/978-3-540-47648-1_1	10.1007/978-3-540-47648-1_53	10.1128/mcb.00001-09
10.3390/ma3094668	######	10.1109/icsens.2010.5690288	10.1007/978-4-431-99703-0_11	10.1111/j.1472-8206.2010.008	10.2330/joralbiosci.52.26	10.1016/j.bone.2010.01.175
10.1115/1.4023646	######	10.4028/www.scientific.net/ar	10.1109/smartgridcomm.2013.66	10.4028/www.scientific.net/an	10.4156/jdcta.vol7.issue3.55	10.1007/978-3-642-41175-5_7
10.1152/ajprenal.00183.2010	######	10.1152/ajprenal.00586.2009	10.1152/ajprenal.00670.2009	10.1042/cbi20100006	10.1152/ajpregu.00207.2009	10.1152/ajprenal.00735.2009
10.1007/s10439-010-0099-y	######	10.1007/s10439-010-0138-8	10.1109/icgec.2010.190	10.1093/med/9780195369779	10.1115/1.859599.paper51	10.1016/j.bpj.2010.08.033
10.1016/j.jmbbm.2010.01.005	######	10.1016/j.dental.2010.03.006	10.5005/jp/books/11019_16	10.1016/j.dental.2010.08.039	10.1016/j.denabs.2010.02.014	10.1016/j.dental.2010.01.010
10.1159/000334595	######	10.1177/1947603510381095	10.1016/s1063-4584(11)60296-3	10.1177/0363546510390476	10.4323/rjlm.2011.265	10.1016/s1063-4584(11)60228-8
10.1002/term.291	######	10.3724/sp.j.1206.2010.00052	10.1159/000272316	10.1007/s10561-010-9175-7	10.1016/j.expneurol.2010.01.00	10.15283/ijsc.2010.3.2.69
10.1177/0022034511415275	######	10.1016/j.joca.2011.03.002	10.1016/j.carbon.2010.12.062	10.1007/springerreference_76	10.1007/springerreference_773	10.1007/springerreference_21037
10.1186/1755-1536-4-15	######	10.1007/s10571-011-9694-1	10.3410/f.12819956.14121054	10.1172/jci44778	10.1007/978-1-4419-5774-0_34	10.1007/978-1-4419-5774-0_35
10.1074/jbc.M111.273474	######	10.1161/atvbaha.110.220988	10.1016/j.numecd.2011.06.002	10.1111/j.1742-4658.2011.082	10.1373/clinchem.2011.176800	10.1016/j.atherosclerosis.2011.02
10.1117/1.3662457	######	10.1002/sca.20219	10.5772/14949	10.1111/j.1365-2818.2011.035	10.1117/12.879792	10.1002/sca.20284
10.1016/j.joca.2011.03.002	######	10.1177/0022034511415275	10.1111/j.1469-8137.2011.03674	10.1007/978-0-85729-923-9_2	10.1111/j.1365-2052.2011.0230	10.1007/springerreference_30794
10.1016/j.ydbio.2011.06.041	######	10.1016/b978-1-4377-2262-8.(	10.5772/20052	10.5772/20242	10.3410/f.13365978.14736086	10.1111/j.1939-1676.2011.00846.
10.3791/3521	######	10.1164/ajrccm-conference.20	10.1164/ajrccm-conference.2011	10.1164/ajrccm-conference.20	10.1158/1541-7786.mcr-10-050	10.1164/ajrccm-conference.2011.

Table of results when comparing DOIs

After running the program on our corpus of documents, and extracting 5 citations per document, the system identified one connecton. In the particular example found, two articles were found to be citing each other. While one connection may seem insignificant, further analysis will likely result in many more connections.

Future Work

The current work is a proof of the concept. For the future, there are several important aspects to build upon is the size of the database,. With more data to work with, the algorithm could find countless more connections amongst the scientific community. Further implementation could compare author, subject, and even publisher connections. Eventually the use of learning algorithms such as Neural Nets and Support Vector Machines could help to identify groups of similar papers and classify text to find similarities on the textual level [3]. These tools could then be used to report on collaboration amongst groups and to suggest collaborations similar to Spotify but for research [4].

Sources

[1] metapub 0.3.17.1; https://pypi.python.org/pypi/metapub/0.3.17.1
[2] Weka 3; http://www.cs.waikato.ac.nz/ml/weka/
[3] Russell, Stuart, and Peter Norvig. Artificial Intelligence: A Modern Approach. Upper Saddle River, NJ: Pearson, 2010. Print.
[4] Spotify; https://www.spotify.com/us/

Acknowledgements

This research is based upon work supported by the National Science Foundation-EPSCoR program under Grant Number EPS-0903795. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.