SciTo trends: visualising scientific topic trends

Serafeim Chatzopoulos[1,2], Panagiotis Deligiannis[2], Thanasis Vergoulis[1], Ilias Kanellos[1,3], Christos Tryfonopoulos[2], and Theodore Dalamagas[1]

[1] IMSI - “Athena” Research & Innovation Center,Athens, 15125, Greece
{schatz,vergoulis,ilias.kanellos,dalamag}@imis.athena-innovation.gr
[2] Univ. of Peloponnese, Dep. of Informatics & Tel/tions, Tripoli, 22100, Greece
{cst11017,trifon}@uop.gr
[3] NTUA, School of Electrical & Computer Engineering, Athens, 15780, Greece

Abstract. Monitoring trends in scientific disciplines is a common task for researchers and other professionals in the broad research and academic community, like research and innovation policy makers and research fund managers. We demonstrate SciTo, a powerful tool that assists monitoring trends in scientific disciplines. SciTo supports keyword-based search for the identification of scientific topics of interest and comparison of interesting topics to each other in terms of their popularity inside the academic community.

Keywords: Information retrieval · Topic modeling · Scientific impact

1 Introduction

Monitoring trends in scientific disciplines, comparing different scientific topics in terms of their popularity in the academic community, or identifying new areas that will attract much attention in the near future, are useful tasks for researchers drafting their research plans for the next months or years. However, apart from scientists, such tasks are also of great interest for other professionals in the broad research and academic community, like research fund managers trying to develop new calls for research projects.

Although many tools to explore scientific literature have been introduced in recent years (e.g., Google Scholar, AMiner[4] [4], Semantic Scholar[5], CiteSeerX[6] [5]), most of them focus solely on the keyword-based publication retrieval and do not provide advanced topic trend monitoring options. Even those providing such features either give insights about the evolution of topics based solely on the number of relevant publications (e.g., AMiner’s Trend[7] and Scholar Plotr[8]), or provide trends for particular keywords or predefined topics (e.g., Scholar Plotr,

[4] https://aminer.org/

[5] https://www.semanticscholar.org

[6] https://citeseerx.ist.psu.edu

[7] https://trend.aminer.org/

[8] https://www.csullender.com/scholar/

Dimensions[9]). Providing trends for keywords is not very convenient since users often ignore important keywords that would facilitate or expand their searches (e.g., in the case of very recently created or alternative technical terms). Finally, relying on predefined topics is not sustainable since it requires curation of topic lists by domain experts since domain-specific research topics evolve.

We introduce "SciTo trends"[10] (Scientific Topics trends), a Web-based tool for topic trend monitoring and comparison, addressing the previous issues. It is built on top of a very large, interdisciplinary dataset containing information (abstracts, citations, etc.) for more than $12M$ scientific articles. Our main contributions are the following:

- SciTo automatically extracts topics from the abstracts of the articles it stores avoiding the need for manual curation.
- It provides a powerful keyword-based search on the stored scientific topics.
- Apart from trends based on the number of publications, it also provides trends based on citation counts and the average short-term impact of the topics (based on RAM [2] scores of its papers).

2 Functionality

Figure 1 illustrates SciTo's search interface where a user enter keywords that describe an interesting topic. Upon submitting a query on these keywords, SciTo returns the set of tag clouds, each representing a related topic. The user can select to review the popularity trends for any of the displayed topics (clicking on the "information button"), or to compare the trends for $2-4$ different topics (clicking on the "comparison button" for each topic to be part of the comparison).

The page that displays the popularity trends of a particular topic contains two different types of infographics (top-left corner of Figure 1): (a) the *pyramid infographic* and (b) the *trend infographics*. Both visualise the information related to the topic popularity according to three indicators (number of publications, number of citations, average short-term impact). The former informs if the topic is among the top (1% or 20%) according to each popularity indicator. The latter displays (a) yearly numbers of topic-related articles published, (b) yearly numbers of citations attracted by topic-related articles, and (c) the average short-term impact for topic-related articles published each year. The topic comparison page (top-right corner of Figure 1) contains only the trend infographics, however displaying the time series for all topics under comparison.

3 Data collection and processing

SciTo's database stores (a) paper citation data, (b) article impact scores, and (c) article abstracts. The citation data is collected from the latest version of

[9] https://app.dimensions.ai

[10] http://scito.imsi.athenarc.gr

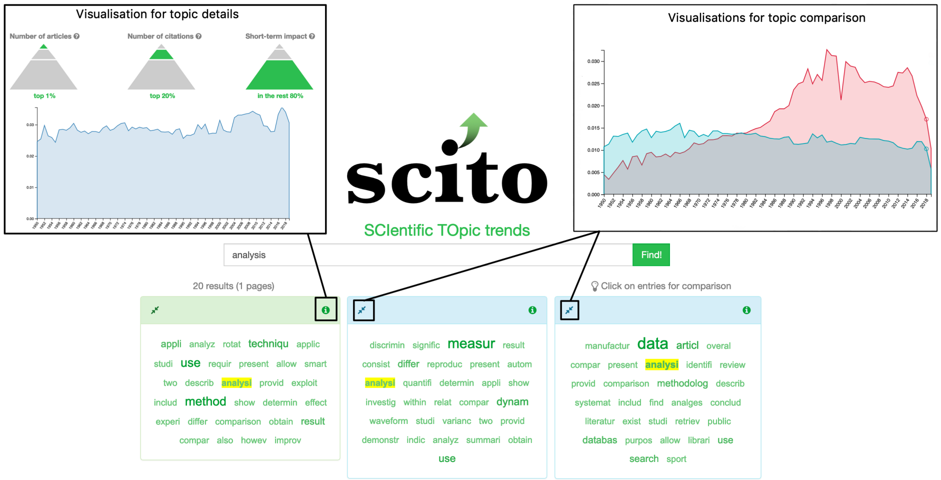
Visualisation for topic details
Number of articles
Number of citations
Short-term impact
top 1%
top 20%
in the rest 80%
0.00
0.01
0.02
0.03
1950 1952 1954 1956 1958 1960 1962 1964 1966 1968 1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018
Visualisations for topic comparison
0.000
0.005
0.010
0.015
0.020
0.025
0.030
1950 1952 1954 1956 1958 1960 1962 1964 1966 1968 1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018
scito
SCIentific TOpic trends
analysis
Find!
20 results (1 pages)
Click on entries for comparison
appli analyz rotat techniqu applic studi use requir present allow smart two describ analysi provid exploit includ method show determin effect experi differ comparison obtain result compar also howev improv
discrimin signific measur result consist differ reproduc present autom analysi quantifi determin appli show investig within relat compar dynam waveform studi varianc two provid demonstr indic analyz summari obtain use
manufactur data articl overal compar present analysi identifi review provid comparison methodolog describ systemat includ find analges conclud literatur exist studi retriev public databas purpos allow librari use search sport

Fig. 1. SciTo's search interface.

OpenCitations COCI dataset[11] ($\sim 450M$ citations for $> 45M$ articles). Based on this data, SciTo calculates and stores paper citation counts and RAM [2] scores, used to measure overall and short-term paper impact, respectively.

Since COCI does not provide article abstracts, SciTo collects abstracts for those papers also indexed by the Crossref API[12] and Open Academic Graph[13] [3, 4] (about $\sim 12M$ papers). Based on these abstracts it trains an LDA [1] model, using the gensim[14] topic modelling library. All extracted topics are indexed by a full-text search engine powered by Apache Solr[15] and running on a 3 VM cluster (8 cores & 16GB RAM/node) to facilitate keyword-based search.

Finally, SciTo's Web UI was implemented using PHP under the MVC architecture (Yii2 framework was used). Custom JS code, based on third-party libraries, was used for SciTo's visualisations (e.g., *D3.js*).

4 Demonstration

At the conference, the audience will have the opportunity to interact with SciTo to examine its full capabilities However, we will also demonstrate some interesting scenarios we have identified, two of which are described below:
Scenario 1: An audience member explores topic trends relevant to the keyword "gene". Since she is interested in the field that studies gene expression, she finds interesting a topic containing the terms "express", "mrna", "rna", and

[11] http://opencitations.net/download
[12] https://www.crossref.org/services/metadata-delivery/rest-api/
[13] https://www.openacademic.ai/oag/
[14] https://radimrehurek.com/gensim/
[15] http://lucene.apache.org/solr/

"transcript". She opens the details of this particular topic and discovers that it is rather popular, since the pyramid infographic displays it in the top 20% of all topics according to all indicators provided. Moreover, by reviewing the trend infographics she realises that this topic became very popular after the mid 80s. **Scenario 2:** The same audience member wants to compare the "gene"-related topic she discovered in scenario 1 with another topic from life sciences, in particular, the research field studying drug effects. Using SciTo she identifies a relevant topic containing the terms "drug", "treatment", and "effect". She selects both this topic and the "gene"-related topic for comparison. In the comparison page the trend infographics reveal that, although the "drug"-related topic was traditionally more popular than the "gene"-related one, after 1995 the latter started to become equally or, even, more popular (depending on the indicator used).

5 Conclusion and Future Work

We demonstrated SciTo, a Web-based tool that assists monitoring trends in scientific disciplines. It supports (a) keyword-based search for the identification of interesting scientific topics and (b) comparison of topics in terms of their popularity inside the academic community. These features make SciTo a powerful tool for professionals in the broad research and academic community. In the future, we plan to extend SciTo to include topics of different levels of granularity (more and less generic in comparison to the current ones) that will be organised in hierarchies. Moreover, we plan to capture and exploit topic history (evolution).

Acknowledgments

We acknowledge support of this work by the project "Moving from Big Data Management to Data Science" (MIS 5002437/3) which is implemented under the Action "Re-inforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR (2003)
2. Ghosh, R., Kuo, T.T., Hsu, C.N., Lin, S.D., Lerman, K.: Time-aware ranking in dynamic citation networks. In: IEEE ICDMW. pp. 373–380. IEEE (2011)
3. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.j.P., Wang, K.: An overview of microsoft academic service (mas) and applications. In: WWW. pp. 243–246 (2015)
4. Tang, J., et al.: ArnetMiner: Extraction and mining of academic social networks. In: ACM SIGKDD. pp. 990–998 (2008)
5. Wu, J., Williams, K.M., Chen, H., Khabsa, M., Caragea, C., Tuarob, S., Ororbia, A., Jordan, D., Mitra, P., Giles, C.L.: Citeseerx: AI in a digital library search engine. AI Magazine **36**(3), 35–48 (2015)