MASARYK UNIVERSITY
FACULTY OF INFORMATICS

Google Scholar

http://scholar.google.com/

Zuzana Zatrochová

3rd semester Mgr.

Authors: Alex Verstak, Anurag Acharya

Date: 26.11.2014 - 05.12.2014

1 Introduction

Google Scholar (GS) is a digital library that provides an access to enormous amount of scholary literature including journal and conference papers, dissertations, academic books, abstracts, technical reports, patents and more. It tries to cover the widest research area possible by collecting documents using search robots crawling the web sites and indexing documents according to the relevance and ranking of a search term. Robots return from web sites with valuable index information of documents and store it to the GS database. Even though, GS is the most widely used tool for retrieval of academic documents it is often criticized for its document ranking mechanism that prioritizes the citation count of document and suppresses the age attribute strengthening the Matthew Effect [2][3]. Moreover, there is no emphasis on frequency of term in the full-text. Nevertheless, it remains the powerful tool for identifying and retrieving the scholary literature accessible in the web environment.

2 Current State and Goals

Google Scholar was released in November 2004 and made a great progress ever since. It became mainly exploited search engine for scholarly literature. It was shown in recent research papers [6] that search engines are used by 89% of researchers as their first step towards the scholary literature. GS takes 65% share of the amount. The rest is covered by Yahoo and Bing search engines. GS is increasing on popularity also among authors of publications that realized the importance of making the public profile of their work. In August 2014, Google claims that almost 75% of search results are accompanied by links to the author profiles [8]. Even though, Google does not advertise the amount of scholary literature present in the Google Scholar's database, the research in May 2014 showed almost 160 milion documents [4].

The main goal of Google Scholar is to provide an access to as many research papers as are available on the Web. The users should be able to retrieve academic literature easily, link publications with related documents, citations and authors and locate complete documents freely available on the web. It is also desirable to provide access to the recent discoveries in any research field. Moreover, to the benefit of the authors, GS tries to make the tracking of publication citations easier. Authors should be able to know who is citing their articles and monitor the development of their research work.

3 The Work Behind

Google Scholar uses robots that crawl through repositories and publisher sites. Robots proactively gather metadata for document indexing, create inverted index for full-text searching and report information back to the search engine [5]. Data sent by robots are algorithmically analysed, considering many factors to decide including metadata in the index. Users requesting and searching literature with specific terms are provided by set of links to indexed documents. Google also provides access to network of citations discovered by link analysis algorithm. However, the direct access to the documents penalizes repository websites that may not be even visited, hence, not covered in statistical information. Another great problem of these sites is to be found by Google crawlers. In 2010 only 38% of digital objects searched by title were found in Google index [6]. Reasons that often inflict the problem include: slow server responses, deadlinks failures, labyrinths created by repository software or poor application of metadata.

The mechanism of GS to retrieve database information is criticized due to the lack of formal control over what is supposed to be stored in it [10]. Therefore, content of the database may become inconsistent, claiming the need of cleaning mechanisms to make the content more useful. The problem mainly includes the ability to obtain database evaluations and analysis with correct information. As a result Google Scholar is not a reliable source as bibliometric tool. However, it is unarguable that even though, GS contains unreliable sources of material the great coverage of freely accessible documents successfully prevails in contrast to databases such as Web of Knowledge and Scopus, both commercial and invaluable sources for scientific activity analysis.

3.1 Google Metrics

In 2012, Google has released a tool named Google Scholar Metrics as a part of evasive Google Scholar project. Scholar Metrics list the journals with major impact on scientific field within last five years. Journals are classified according to their scientific impact measured according to their h-indexes. The h-index is easily computable parameter that provides reliable information on the journal or author productivity in relation to citation counts of published articles. The h-index is the largest number that represents at least h articles of the journal or the author that have at least h citations. According to this index, Scholar Metrics provide the table of first hundred journals with biggest scientific impact categorized by publication language. In [7] Scholar Metrics is criticized for its mixture of products with different nature (journals, repositories, databases, conference proceedings and working papers) in the same table. The variability of products and their coverage should prevent their bibliometric comparison. Moreover, standardization and result browsing is criticized as well. Fortunately, since the release of 2014 version of Scholar Metrics, users can browse journals categorized into the field of their study. The field service is available only in the English language for the moment. The general Scholar Metrics cover journals also in several worldwide languages.

3.2 Google Scholar Citations

Another feature of GS is a release of Google Scholar Citations in 2012. Scholar Citations is a tool that measures the impact of researches to the scientific community and their productivity. It also provides the way for authors to track citations of their articles. Each author can view the citation increase since the beginning of his publishing career or the increase over last five years. The aim is to provide citation statistics, visualization of research networks and track the most important documents in selected research field. The tool uses three kinds of metrics: h-index, i10-index and total number of citations. i10-index measures the number of articles of an author that were cited at least ten times. The main drawback of citation counts is the chance of some citations not being discovered or possibly being modified from the Web.

3.3 Google Scholar Library

In 2013, a new feature Google Scholar Library provides the personal collection of literature for anyone who has registered an Google account. Users may save their results into the library right from a discovered page and label them for easier future use. Currently, GS provides an organisation of documents according to the date of their release.

3.4 Searching

Searching in GS can be done using an arbitrary word or sentence that a user evaluates as the appropriate term for the content he is looking for. However, if we want to make the results more accurate, we can hold on to several recommended searching rules. The default search is not case sensitive, combines words using AND logical operator and searches full-text, citations and abstract of the document. Another options for search are:

- **Author**: author:"Name"
  For example, using the search term author:"Leslie Lamport", GS hands 411 documents organised according to the relevance. Searching term without author tag delivers 9260 results including the documents that cite or mention author in their body. In cases where authors name resembles some important term from an area of research it is highly probable that results will be more irrelevant without use of recommended search language (author:"Handle" - 626 results, Handle - 4 850 000 results)

- **Phrase**: "The Phrase"
  The search for the exact phrase in the document is used when we search for a words that have a specific meaning when they are used together ("game theory") but separately have a different meaning ("game" and "theory")

- **Title**: intitle:memory
  Even though, first search results of GS almost always contain the search term in the title, if a search term relating to the object of our study is very important, it is safe to include the name in the title to avoid unrelated content in results. For example, if we study some specific tool such as Google Scholar it is easier to search with the term intitle:Google Scholar as there is very high probability that all documents about GS will contain the name in the title. On the other hand the term as the "computer" would be very general and provide great area of unrelated document results

- **Institution**: site:institutional-domain or site:top-level-domain
  For example site:edu or site:cam.ac.uk will provide the results on the topic that are stored only on websites within the domain. The search term site:muni.cz provides 198 000 results. Currently, this parameter shows results consisting only from primary document versions

- **Synonyms**: ∼synonym
  Using synonyms for the most searches provides only small increase of the amount of documents in results and takes three times longer in average. For example, a search term computer hands 5,720,000 results in 0.05 seconds while search term ∼computer provides 5,820,000 results in 0.57 seconds. We can infer that GS still suffers from unreliable evaluation of synonyms. It is recommended for researchers to learn the basic terms of their work and then browse the internet for higher education with more relevant words.

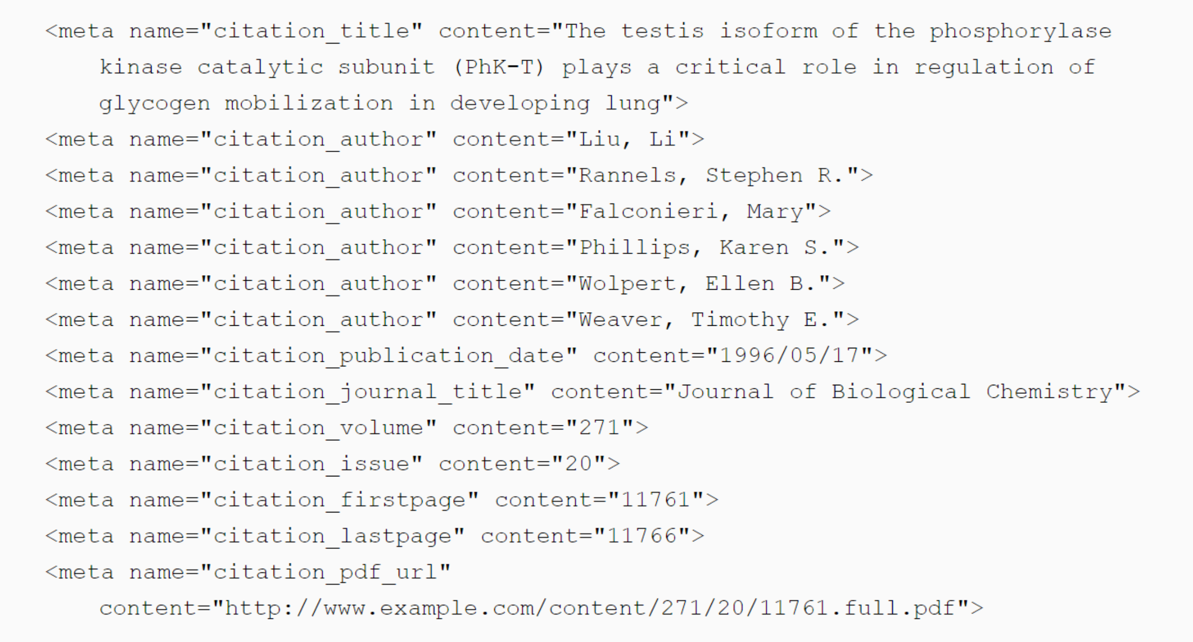
- **Filetype:** filetype:type
  Usually PDF file formats are preferred in the search results. However if a person wants to search for a presentation on a particular topic a term filetype:ppt is relevant to use.

- **Include, Exclude:** +;-
  Using these tags, we can include automatically excluded search terms such as "the" or exclude words we do not want to have in the results

GS also provides standard advanced searching interface supporting the search for articles with exact phrases, excluding words, at least some of the words and interval of years when the article was published. In addition, the articles published since 2014, can be organised either according to the relevance and the release date.

3.5 Metadata, Indexing

Documents that are indexed by Google Scholar consist of academic literature in HTML or PDF format with searchable text not exceeding 5MB in size. Larger files should be uploaded to the Google Books format also included in GS searching results. GS uses parsers (automatic software) that identifies bibliographic data references of documents. Meta-tags consist of:



```
<meta name="citation_title" content="The testis isoform of the phosphorylase
    kinase catalytic subunit (PhK-T) plays a critical role in regulation of
    glycogen mobilization in developing lung">
<meta name="citation_author" content="Liu, Li">
<meta name="citation_author" content="Rannels, Stephen R.">
<meta name="citation_author" content="Falconieri, Mary">
<meta name="citation_author" content="Phillips, Karen S.">
<meta name="citation_author" content="Wolpert, Ellen B.">
<meta name="citation_author" content="Weaver, Timothy E.">
<meta name="citation_publication_date" content="1996/05/17">
<meta name="citation_journal_title" content="Journal of Biological Chemistry">
<meta name="citation_volume" content="271">
<meta name="citation_issue" content="20">
<meta name="citation_firstpage" content="11761">
<meta name="citation_lastpage" content="11766">
<meta name="citation_pdf_url"
    content="http://www.example.com/content/271/20/11761.full.pdf">
```

Figure 1: Examples of Meta-tags. Taken from [1]

- **Title tag** - citation_title or DC.title contains title of the paper not journal, repository or the book
- **Author tag** - citation_author or DC.creator contain only authors of the paper not contributors or advisers. Each author name is provided in separate tag. At least one is required
- **Publication date** - citation_publication_date or DC.issued contains the full date of the publication
- **Journal and Conference Papers** - bibliographic citation data include: citation_journal_title (conference), citation_issn, citation_isbn, citation_volume, citation_issue, citation_firstpage, citation_lastpage or DC: DC.relation.ispartof, DC.citation.volume, DC.citation.issue, DC.citation.spage, DC.citation.epage
- **Theses, Dissertations and Technical Reports** - bibliographic citation data include: citation_dissertation_institution, citation_technical_report_institution or DC.publisher and citation_technical_report_number
- **Other Documents** - should follow tags as how they would be cited in the References of other papers.
- **Item Location** - citation_pdf_url or DC.identifier tags

Google does not recommend to use Dublin Core based metadata because they work poorly for journal papers. Dublin Core does not have unambiguous fields for journal title, volume, issue, and page numbers [1].

4 Conclusion

Generally, GS is one of the greatest sources of electronic scholary literature freely available on the Internet. Even though, it has its pros and cons there is not and will not be a major competitor at least for another several years. The main advantage of GS is that it disposes with the largest database of variously formatted research works. Even though, sources of information are not always reliable, the major share of results covers related material and with right choice of words anybody can achieve successful browsing. GS have been main source of information for my research work for last three years and I was always able to identify the most important research documents present in my field of study. Therefore, I would recommend it for everyone in spite of the amount of critique among researchers. I would like to conclude with some interesting facts I have learned during my research about Google Scholar. As of year 2011, there were 3,223 universities in Europe with 2,657,514 index items in GS. Moreover, Czech universities have produced 61,667 indexed academic items. The Masaryk University belongs to the list of largest universities in the world according to the number of records indexed in GS in August 2010. It earned the 20th place with 30,800 records. The first is the Harvard university with total of 1,170,000 records in GS [10].

References

[1] http://scholar.google.com/

[2] Beel, Jöran, and Bela Gipp. ”Google Scholar's ranking algorithm: an introductory overview.” Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09). Vol. 1. 2009.

[3] Beel, Jöran, and Bela Gipp. ”Google scholar's ranking algorithm: The impact of articles' age (an empirical study).” Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on. IEEE, 2009.

[4] Orduńa-Malea, Enrique, et al. ”About the size of Google Scholar: playing the numbers.” arXiv preprint arXiv:1407.6239 (2014).

[5] Pomerantz, Jeffrey. ”Google Scholar and 100 percent availability of information.” Information Technology and Libraries 25.2 (2013): 52-56.

[6] Arlitsch, Kenning, and Patrick S. O'Brien. ”Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar.” Library Hi Tech 30.1 (2012): 60-81.

[7] Cabezas Clavijo, Á, Delgado López-Cózar, E (2012). Scholar Metrics: the impact of journals according to Google, just an amusement or a valid scientific tool? EC3 Working Papers, 1.

[8] Alex Verstak, ”Fresh Look of Scholar Profiles.” Google Scholar Blog, August 21, 2014

[9] Arlitsch, Kenning, and Patrick S. O'Brien. ”Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar.” Library Hi Tech 30.1 (2012): 60-81.

[10] Aguillo, Isidro F. ”Is Google Scholar useful for bibliometrics? A webometric analysis.” Scientometrics 91.2 (2012): 343-351.