Description

As the volume of published academic research material continues to grow, the prospects for new researchers looking to navigate through all of that data in order to compile a literature review or research proposal becomes evermore daunting. Whether or not data aggregation will provide the solution remains to be seen.

What is Data Aggregation?

Data aggregation is the process of collecting data and presenting it in a summarized manner. Accuracy of data analysis and result delivery depends on the amount and quality of collected data. Therefore, data aggregation is a crucial step in literature review.

Data aggregation is used to provide statistical analysis for collated research data and to create a summarized data. The collation, curation, and presentation of data in summary format has been a recognized practice in the commercial world for decades now.

How is Data Aggregation Done?

Aggregation is often done using data aggregation tools such Google Looker, Zoho Analytics, Cloudera Distribution, etc. These are called data aggregators. They typically include features for collecting, processing, and presenting aggregate data.

Experian, one of the providers of your credit score, is one of the examples of data aggregation that has managed to amass data points on individual consumers that run into the thousands. Beyond the provision of a simple three-digit score, that data can now be mined for some very targeted marketing campaigns.

Data Aggregation Techniques

Data aggregation can be done using 4 techniques following an efficient path.

1. In-network Aggregation:

This is a general process of gathering and routing information through a multi-hop network.

2. Tree-based Approach:

The tree based approach defines aggregation from constructing an aggregation tree. Tree structure is minimum spanning tree, sink node observe as a root and source node consider as a leaves. Information developed of data start from leaves node up to root means sink node(base station).

3. Cluster-based Approach:

Cluster-based approach is used to collate larger data on the entire network. It is divided in to a few clusters. Each cluster has a cluster-head which is selected between cluster members.

4. Multi-path Approach:

In multi-path approach, partially aggregated data is sent to single parent node that is “root node” in aggregation tree, a node could send data above various paths. In which each and every node can transmitted data packets to its possibly many inputs.

Data Mining

Aggregation of data only serves one-half of a prospective user’s needs. Having so much information available in one large database has the potential to save a considerable amount of time from having to work with multiple individual databases. However, that time can only be saved if the collated data can be searched or mined to find the information you are looking for quickly and accurately.

Garbage in, Garbage out (GIGO)

The old maxim about the information you get out of a computer being only as good as the information put into it can be seen to apply to these big data repositories. Restricted-access databases such as *Academic Search Premier* can be expensive if your library doesn’t have access, but they represent your best hope for the latest research. *JSTOR* (short for Journal Storage) offers limited free access with other subscription options for current journal publications.

Open Access Databases

Open Access Publishing is committed to the free availability of research data as a stand against the alleged elitism of research journals with expensive subscription rates. Databases such as the Public Library of Science (PLoS) and Stanford's HighWire have grown dramatically as a result of this campaign. Highwire, for example, lists over 2.5 million free full-text articles and a total article count of almost eight million, but these exemplars set a standard that many other open access databases donot achieve:

- *The Social Science Research Network* (SSRN) offers total article counts in the hundreds of thousands, but many of them are pre-publication "working papers" that may require further follow-up with the original authors.
- *PubMed* from the National Institutes of Health (NIH) contains over 24 million articles, but many of them have restricted access.
- *Google Scholar* leverages the power of the 'Mother Google' search algorithm, but many of the search results will include restricted articles and journals that have been flagged as being questionable for their lack of a rigorous peer review process.

A Mixed Blessing

Data aggregation performs a valuable service in casting a wide net to compile relevant data into one big database. However, in these days of open access journals that charge article processing fees instead of journal subscription fees, the quality of the research material that gets captured in that net has become increasingly unpredictable. Sophisticated search algorithms may help you identify relevant material by topic and date, but some of those results may be restricted rather than full access, and some of them may be of questionable authorship. Proceed cautiously!

Category

1. Career Corner
2. PhDs & Postdocs

Date Created
2015/12/17
Author
editor