Scholarly Big Data Mining: Opportunities, Methods and Challenges

Teshome M.

28 August 2015

Personal Profile

Education

- Bachelor of Science in Computer Science → Hawassa Univesrity, 2006
- Master of Engineering in Software Engineering → Chongqing University, China, 2011
- PhD student at Dalian University of Technology, China since Sept. 2014

Work Experience

Teaching Experience

- April 2007- December 2007, Softnet college, Addis Ababa, Ethiopia
- Feb. 2008-Sept. 2012, Arba Minch University, Arba Minch, Ethiopia
- Sept. 2012- till date, Adama Science and Technology University, Adama, Ethiopia

Company Experience

- ASTU Enterprise, Software Development Business Unit, Feb. 2014-Agust 2014, Adama, Ethiopia
- July 2013-August 2014, DHIT Solutions, Addis Ababa, Ethiopia
- LOCI Software Development PLC., Jul 2010-Sept 2010 Donghai, China

Contents

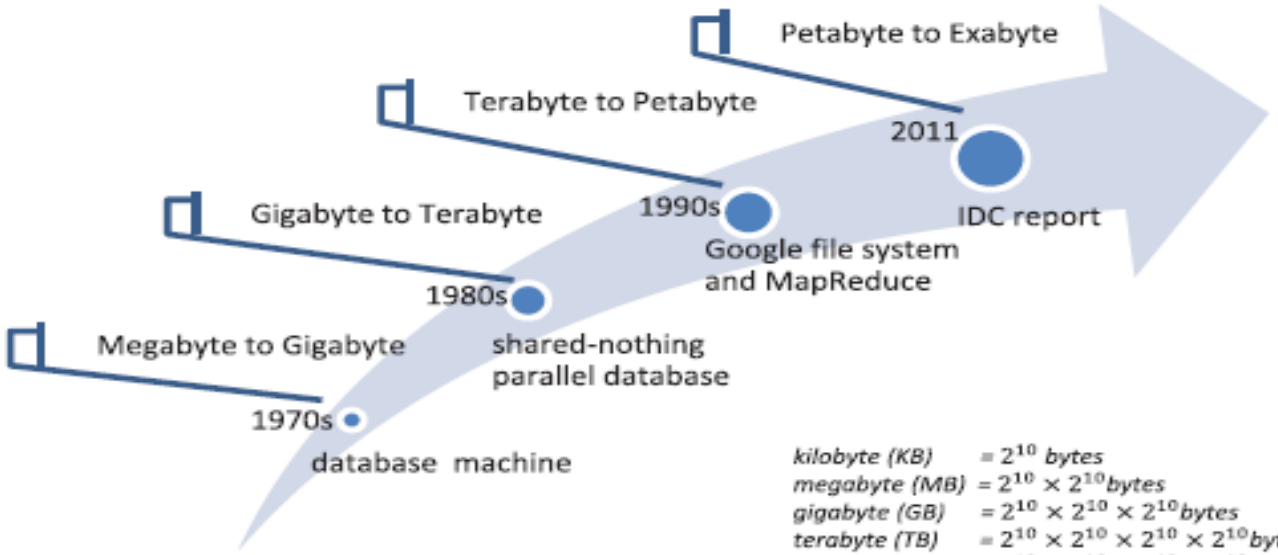
Big Data Analytics-Overview

- In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range.
- According to (McKinsey, 2013), over 2 billion people worldwide are connected to the Internet, and over 5 billion individuals own mobile phones.
- By 2020, 50 billion devices are expected to be connected to the Internet→ predicted data production will be 44 times greater than that in 2009, (McKinsey ,2013).
- As information is transferred and shared at light speed on optic fiber and wireless networks, the volume of data and the speed of market growth increase.

Big Data Analytics-Overview (cont…)

- Domains that contribute for massive data generations
  - Health care
  - scientific sensors
  - Internet
  - financial companies
  - User-generated data (e.g. social networks)
  - Scholarly Communication

Big Data Analytics-Overview(Cont…)



kilobyte (KB) $= 2^{10}$ *bytes*
megabyte (MB) $= 2^{10} \times 2^{10}$ *bytes*
gigabyte (GB) $= 2^{10} \times 2^{10} \times 2^{10}$ *bytes*
terabyte (TB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ *bytes*
petabyte (PB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ *bytes*
exabyte (EB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ *bytes*

A briefly history of big data with milestones→ adopted from H.Hu et al.: Toward scalable systems for big data analytics

Big Data Analytics-Overview(cont...)

- The fast growth rate of such large data generates numerous challenges such as the rapid growth of data, transfer speed, diverse data, and security.
- Current data volumes are driven by both unstructured and semi-structured data in contrast to traditionally data is stored in a highly structured format to maximize its informational contents.

Big Data Analytics-Overview (cont…)

- End-to-end processing can be impeded by the translation between structured data in relational systems of database management and unstructured data for analytics.
- Big data management and Analysis requires new technologies and architectures to extract value from it by capturing and analysis process.
- Addressing various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security are key inputs to determine opportunities and several open issues in Big Data domination.

Big Data: Definitions and Attributes

- possesses rapidly growing nature, vary in nature, with large volume and challenging to manage and analyze using existing traditional techniques, due to:-
  - Variety- raw, structured, semi-structured or unstructured data
  - Volume-petabytes, increase to zettabytes (10^21)
  - Velocity-speed of incoming data and at which the data flows.
  - Variability- the inconsistencies of data flows

Big Data: Definitions and Attributes (cont...)

- Complexity- Cleaning data, Transforming across the system and connect and correlate the relationships etc.
- Value- Discovery knowledge patterns and trends e.g. Exploring Co-citation in bibliographic networks to investigate the possible future collaboration among researchers.

Big Data: Definitions and Attributes (cont...)

Table 1. Comparison between big Data and traditional Data

	Traditional Data	Big Data
Volume	GB	Constantly update(TB or PB currently)
Generated rate	Per hour, day,...	More rapid
Structure	structured	Semi-structured, unstructured
Data Source	centralized	Fully distributed
Data Integration	Easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Access	interactive	Batch or near real-time

Big Data Management

- Need to synchronize big data architecture with the support infrastructure of the organization
- Information from machines or sensors and large sources of public and private data are disorganized and messy.
- Most companies were unable to either capture or store these data, and available tools could not manage the data in a reasonable amount of time.
- New Big Data technology should improve performance, facilitates innovation in the products and services of business models, and provides decision making support.

Big Data Management (cont...)

- Big Data technology aims to minimize hardware and processing costs and to verify the value of Big Data before committing significant company resources.
- Properly managed Big Data are accessible, reliable, secure, and manageable.
- Big Data Analytics encompasses Data Acquisition, Data Storage, Visualization and Data Analysis

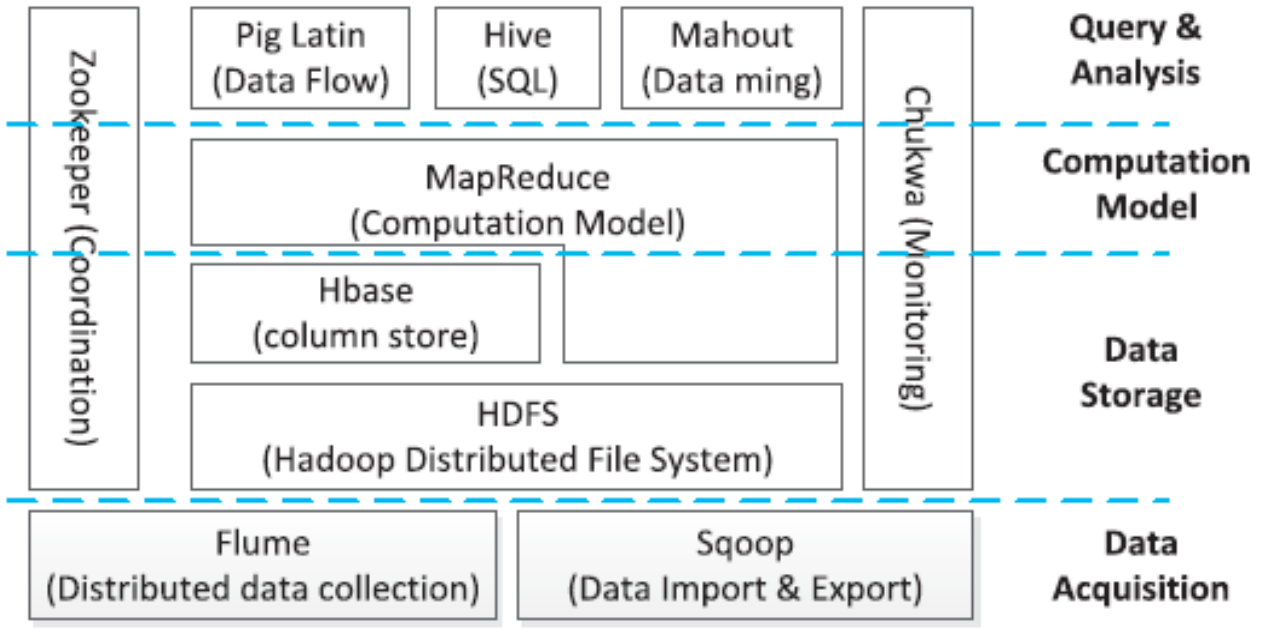
Data Management Tools

- Data management tools and techniques includes Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort
- Most commonly used tools and techniques for big data are Hadoop, MapReduce, and Big Table.
- Redefined data management as they effectively process large amounts of data efficiently, cost effectively, and in a timely manner.

Hadoop Framework

- **Hadoop Framework- clones Google's distributed file system and the MapReduce computation model in handling massive data processing.**
  - Hadoop enables distributed processing of large amounts of data on large clusters of commodity servers.
  - Hadoop Features, suitable for big data management and analysis:
  - **Scalability:** Hadoop allows hardware infrastructure to be scaled up and down with no need to change data formats.
  - **Cost Efficiency:** Hadoop brings massively parallel computation to commodity servers, leading to a sizeable decrease in cost per terabyte of storage, which makes massively parallel computation affordable for the ever growing volume of big data.
  - **Flexibility:** Hadoop is free of schema and able to absorb any type of data from any number of sources.
  - **Fault tolerance:** Missing data and computation failures are common in big data analytics. Hadoop can recover the data and computation failures caused by node breakdown or network congestion.

Hierarchical Architecture of Hadoop



Big Data Analysis

Common Methods

- **Data visualization-** to communicate information clearly and effectively through graphical means.
- **Statistical analysis - can serve two purposes for large data sets: description and inference.**
  - Descriptive statistical analysis can summarize or describe a collection of data
  - inferential statistical analysis can be used to draw inferences about the process.
  - Complex multivariate statistical analysis uses analytical techniques such as aggression , factor analysis, clustering, and discriminant analysis.

Big Data Analysis (Cont…)

Common Methods

- **Data mining- is the computational process of discovering patterns in large data sets.**
  - Data mining algorithms (machine learning, pattern recognition, statistics, and database communities.
  - Includes C4.5, k-means, SVM (Support Vector Machine), a priori, EM (Expectation Maximization),PageRank, AdaBoost, kNN, Naive Bayes, and CART.
  - Covers classification, clustering, regression, statistical learning, association analysis and link mining, neural network and genetic algorithms, are useful for data mining in different applications.

Scholarly Big Data

- Advancement in science→ scientists steadily produce a large volume of research articles
- Scholarly big data refers to the vast quantity of data that is related to scholarly undertaking, such as journal articles, conference proceedings, theses, books, patents, presentation slides and experimental data.
- Microsoft Academic is reported to have over 50 million academic document records and in 2010 it was estimated that the annual growth rates of several popular databases from 1997-2006 ranged from 2.7 to 13.5%. (P.Larsen, 2010)
- An average of 43% of articles published between 2008 and 2011 were freely available online .

Scholarly Big Data (cont...)

- Big scholarly data varies (e.g. articles, lecture slides etc)
- Scholarly big data is of significant interest to groups involved in decision making in funding, education and government, as well as scientists, businesses and the general public.
- Scholarly services → Google Scholar, PubMed, the ArXiv and CiteSeer were built to collect, analyze and provide access to this data.

Is Scholarly Data Big enough?

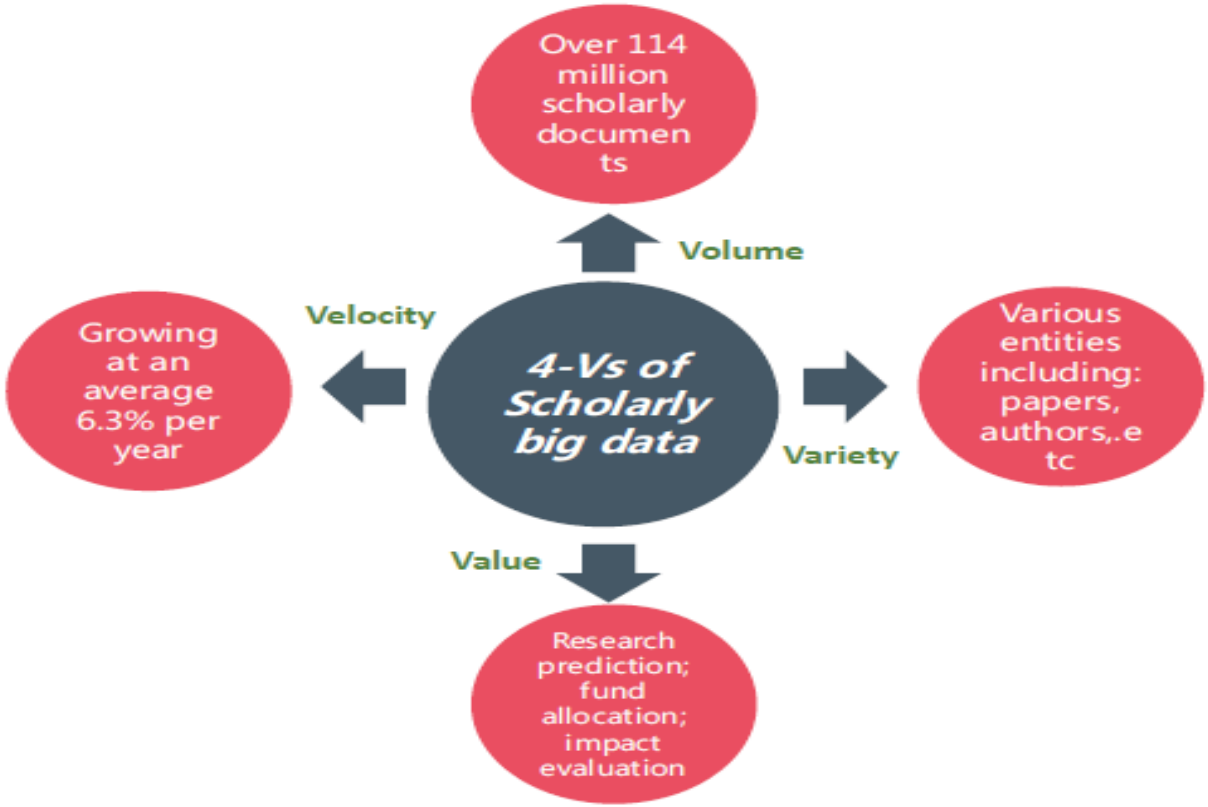


Fig. 1. 4Vs features of Big Scholarly Data

Scholarly Big Data Analysis Challenges

- Data collection, integrating information from multiple sources and extracting meaningful information from the data.
- Designing algorithms and processes that are able to deal with heterogeneity
- Scholarly data is also highly relational: citations among papers result in a rich citation network; co-authorship results in a co-authorship network; research projects are related to specific grants; and authors are related to specific institutions and publications.
- Entity linking and name disambiguation in scholarly big data
- Sharing data- issues related to intellectual property and copyright may limit the copying and sharing of data among different groups.
- The size of the data may also be a prohibiting factor.
- Research Data Management

Why Scholarly Big Data Analytics?

- To design better scholarly data management and sharing approaches
- To design efficient algorithms and tools to foster research collaborations
- To help researcher find relevant and high quality papers
- To detect research groups/communities with similar interest
- To assess research area and researcher impacts

Why Scholarly Big Data Analytics? (cont…)

- To identify researcher career and collaborations trends patterns
- To predict future research trends and possible evolving interdisciplinary researches and help funding organizations
- To exploit knowledge from scholarly big data and provide better services for scholars and understand the rules and laws of science itself.
- E.g. Analyzing citation relationships extracted from large collections of data may evaluate the impact of a given paper or scholars→ help to allocate reputations and funds to scientists.
- Coauthor behavior analysis among scholars→ help to find the distribution of academic communities→ to built map of science

Scholarly Big Data Analytics approaches

- Scholar Data Acquisition
- Visualization techniques
- Analysis techniques
- Scholarly Data Analysis

Scholarly Big Data Management

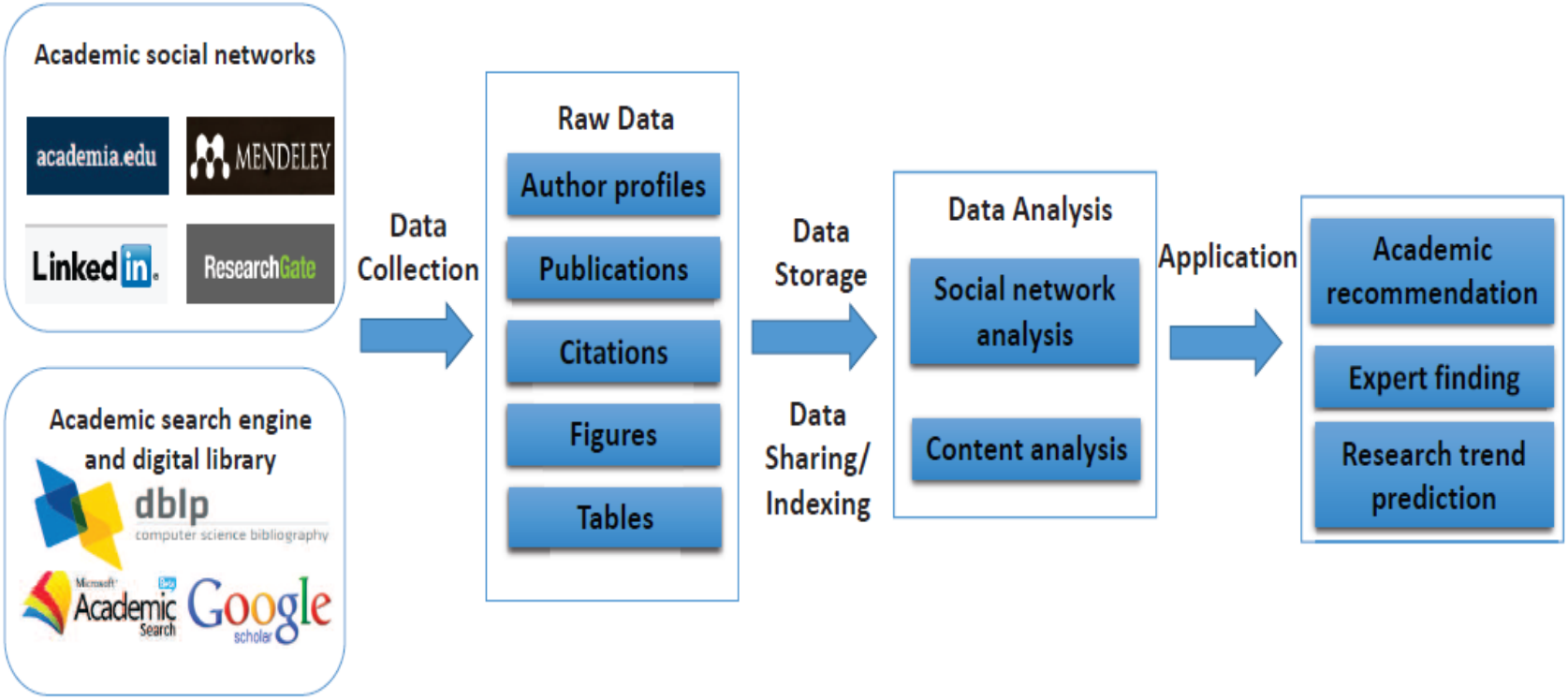


Fig. 2. Framework of Scholarly big data analysis

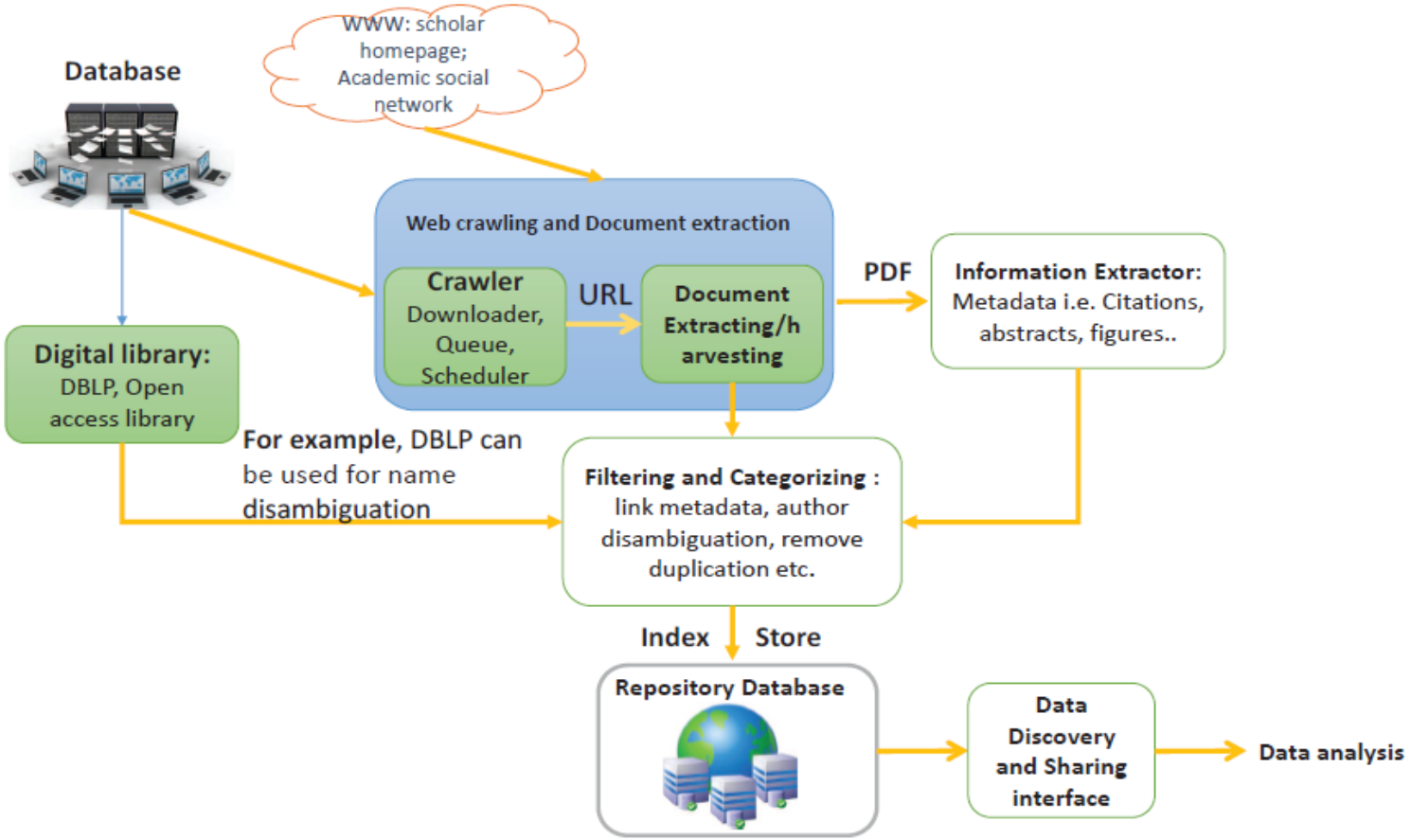
Scholarly Big Data Management(cont…)

- Digital Libraries→ serve community in publishing, accessing and securing information, promote visibility of research output
- Academic Search Engines→ find relevant research documents.
- Academic Social Networks→ enhance sharing and disseminating scientific knowledge and discoveries, provide for scientific collaborations, promote institutions impact in education and research, and scholars to share their research works and expertise
- Data Sharing and Indexing-DOI

Scholarly Data Acquisition techniques

- The major sources of scholarly data are:
  - Google Scholar, PubMed,DBLP, the ArXiv and CiteSee etc..
  - Citation Data
  - Website(university website, research home pages etc.) through web structure mining or web data extraction
- Scholarly information encompasses:- metadata, citations, algorithms, figures, abstracts, tables etc.
- Citations→ scientific work impacts, understand scientific knowledge diffusion patterns between fields and to identify emerging hot research topics
- Author Details Extraction and Profiling

High Level View of Scholarly Data Collection



Scholarly Big Data Visualization

- Academic Social Network Analysis(Mining)
  - The scholarly communication or collaboration can depicted through Scholarly/bibliographic networks.
  - nodes the social networks are usually denoted as an academic entity , such as paper, a venue(journal/conference),an author etc.
  - A link usually denotes relationships such as citation, co-authorship, co-citation, bibliographic coupling or co-word.
    - In citation networks, each node is a piece of knowledge and a link denotes the knowledge flow.

Scholarly Big Data Visualization (Cont…)

- The interaction of research aggregates can be explored from different types of scholarly networks to study a range of perspectives of research interactions and scholarly communications.
- For Example
  - co-authorship networks focus on finding patterns of contacts or interactions between social actors.
  - Similarity-based networks such as co-citation networks, bibliographic-coupling networks.
  - Co-word networks focus on identifying research topics or disciplines.

Scholarly Big Data Analysis Methods

- Scholarly Network Analysis→ methodical analysis of social networks, which aims at viewing social relationships in terms of network theory(average path length, Clustering coefficient, Degree centrality)
- Scholarly Text Mining→ focuses on analysis of the content (document clustering and classification)

Scholarly Networks

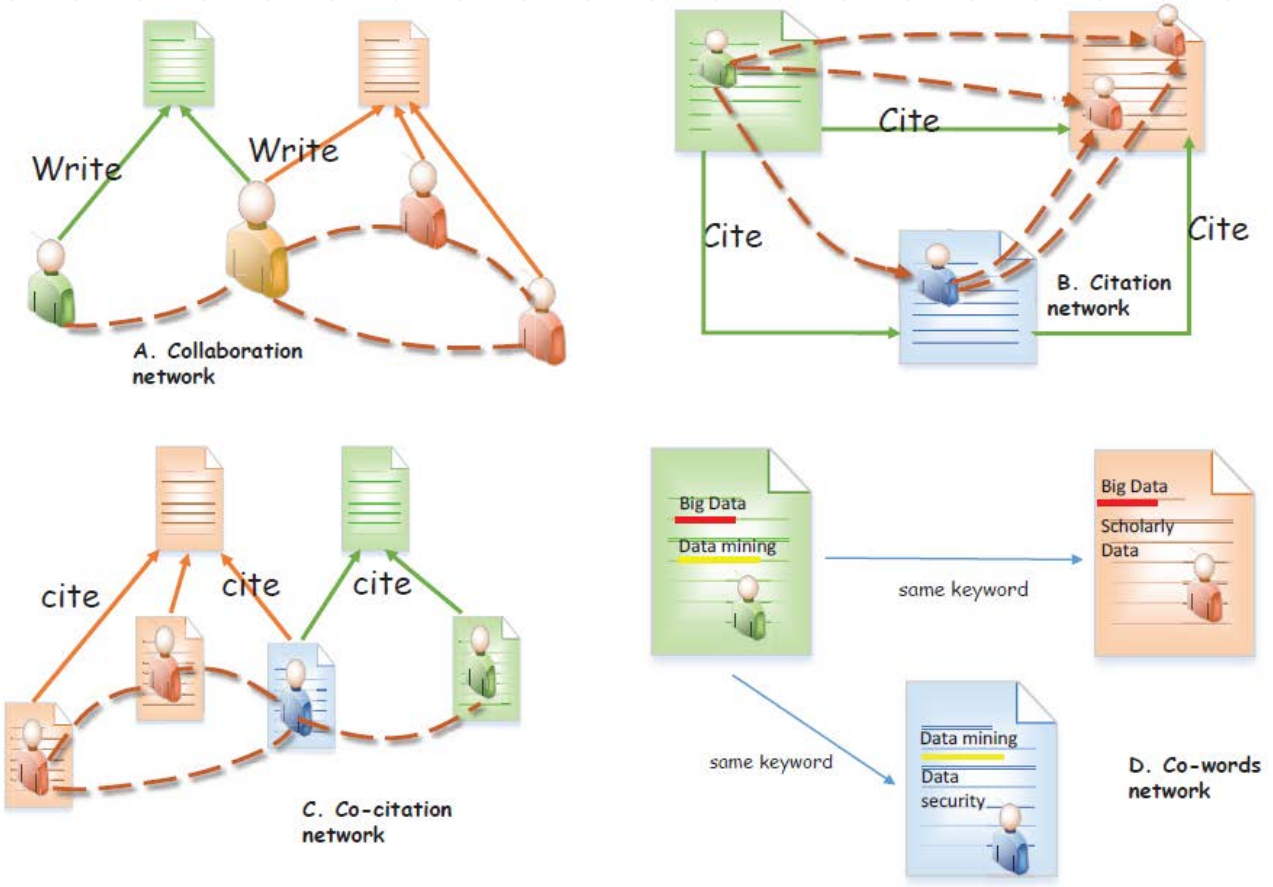


Fig. 4. Four most popular types of Scholarly networks

Scholarly Big Data Analysis Methods(cont…)

- Scholarly Text Mining→ focuses on analysis of the content (document clustering and classification)
- Issues of scholarly text mining
  - Co-word Analysis
  - Citation Mining
  - Patent Analysis
  - Full Text Analysis→ to study knowledge flow, identify the most significant given a specific domain

Scholarly Big Data Application

- Academic recommendation
  - Literature recommendation
  - Collaboration Recommendation
  - Venue Recommendation
- Expert Finding→ prominent researchers in a given fields are a cornerstone to understand the extent to which their field of research is progressed and future directions
- Research Trend Prediction
- Scientific Community Detection

Open Issues and Challenges

- Scholarly Data Acquisition
- Altmetrics→ with the easily access to more scholar information, we can evaluate the impact of a publication more efficiently and quickly.
- Heterogeneous Networks
- Data Sharing and Indexing etc...

Conclusion

- Big Data is characterized by four factors:- volume, variety, velocity and value
- Big Data Analytics from Data generation to Analysis were discussed.
- The availability of unprecedented amounts of scholarly big data on scientists' collaboration, documents sharing and publications open the possibility of investigating science itself as well as scientists ourselves.
- The data can greatly promote the development of science by promoting scientific collaboration, scholar data sharing, and more fair fund allocation methods.
- Scholarly Big Data Analytics is a promising research area.

References *(Some of referred papers)*

1. Hu et al., Toward Scalable Systems for Big Data Analytics, IEEE Access, vol. 2, pp. 652-687, 2014
2. Rong Hu et al., ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application, IEEE Transactions on Emerging Topics in Computing, vol. 2, NO. 3, pp. 302-313, 2014.
3. Rui Xu and Donald Wunsch , Surveying of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005
4. Quanquan Gu and Jiawei Han, Towards Feature Selection in Network, Proceedings of Information and Knowledge Management(CIKM,2011), Glasgow, Scotland, 2011
5. Charu Aggarwal and Karthik Subbian. 2014. Evolving network analysis: A survey. ACM Computer Survey. 47,1, Article 10 (April 2014), pp.36.
6. Manish Gupta, Charu Aggarwal, Jiawei Han and Yizhou Sun, "Evolutionary Clustering and Analysis of Bibliographic Networks", Proc. of 2011 Int. Conf. on Advances in Social Network Analysis and Mining (ASONAM'11), Kaohsiung, Taiwan, July 2011
7. Xiaozhong Liu,Yingying Yu, Chun Guo, Yizhou Sun, and Liangcai Gao, Full-Text based Context-Rich Heterogeneous Network Mining Approach for Citation Recommendation, ACM/IEEE Joint Conference on Digital Libraries (JCDL'14), London, 2014.
8. Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, Jiawei Han: Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. ASONAM 2011: 121-128

References *(Some of referred papers)*

9. Nawsher Khan et al., Big Data: Survey, Technologies, Opportunities and challenges, Scientific World Journal Volume 2014.
10. Fangbo et al., Research-Insight: Providing Insight on Research by Publication Network Analysis, SIGMOD'13, June 22–27, 2013, New York, New York, USA.
11. Zhen Chen et al., AVERec: A Random Walk Based Academic Venues Recommendation, Mobile and Social Computing Lab, School of Software, Dalian University of Technology, 2014.

Thank You!