Indian Journal of Science and Technology, *Vol 10(21), DOI: 10.17485/ijst/2017/v10i21/112828, June 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Graph Data: The Next Frontier in Big Data Modeling for Various Domains

Angira Amit Patel[1] and Jyotindra Dharwa[2]

[1]Department of MCA, Shri Chimanbhai Patel Post Graduate Institute of Computer Applications, S. G. Highway, Ahmedabad – 380015, Gujarat, India; angira.it@gmail.com
[2]Department of MCA, Acharya Motibhai Patel Institute of Computer Studies, Ganpat University, Ganpat Vidyanagar, Mehsana – 384012, Gujarat, India; Jyotindra.dharwa@ganpatuniversity.ac.in

Abstract

Graph is considered as next frontier in the era of Big data due to its flexibility and self-explaining property. The prime objective of this research is to reveal graph database as an alternative of traditional relation database in the field of database. This research illustrates potential of graph structure for diversified modeling along with its convenience for various domains. Extensive literature review demonstrates use of diversified graph structures as means of data storage and analysis as it can cope up any kind of complex structures ranging from multi linked web data, complex chemical structure, gene data, network structure, social network, e-commerce to text data. A formal conclusion of this review revealed use of various graph models according to state-of-affairs of various domains as well as data modeling challenges and complexity of Big data. This investigation anticipates use of an appropriate graph structure and provides guidelines for solving data modeling challenges for structure, semi structure and unstructured data. The diversified graph structure along with its characteristics has been suggested for real world problems of various domains. This research could lend a helping hand to anyone who wants to implement graph data model for their data management challenges and computational problems.

Keywords: Big Data, Graph Data, NoSQL, Property Graph

1. Introduction

There are various real world problems associated with Big data which require solving multifaceted data processing challenges. Mainly it leads towards data representation challenge with sparse, uncertain and incomplete data. A lot of research has been already done in the field of data modeling and many models have been already proposed under the umbrella of NoSQL technology. The proper selection of suitable data structure facilitates to cope up with processing complexity and improves efficiency and effectiveness of the analysis technique. The major selection criteria comprise of volume of the data, growth rate for volume of the data, need of the flexibility into schema, time-evolving factor, further processing requirements and many more. The versatility of graph structure for various domains is also feasible due to diversification of the structure. It shows adaptability of graph data modeling in wide range of domains like social network, web, image, bi-informatics, GIS, e-commerce, digital library and citation network representations. It

**Author for correspondence*

has been observing that graph structure is most suitable for multi-linked and massive data set due to its effectiveness for associative operations.

The efforts of this research work demonstrate use of diverse graph data models for wide-range real world problems. This comprehensive survey attempts to encourage software engineers in the direction of graph based data representation.

2. Road Map

This paper organized in three main sections as follow: The section 3 introduce booming technologies such as Big data, NoSQL and graph database. The section 4 throw lights on benefits of implementing graph data modeling.It also explores possibility of implementing graph modeling for various domains. As graph is flexible and versatile structure having many characteristics which makes it suitable to represent any complex structure of real world. The section 5 covers few trade-offs and implementation issues. The sections 5 also illustrate literature review of graph modeling for wide range of computational problems. The last section concludes possibility of graph modeling for various domains if suitable graph models designed for representation.

3. Big Data and NoSQL

Nowadays, billions of users are connected to the World Wide Web and spending significant amount of time via mobiles, computers and other devices. Consecutively, there are collections of petabyte scale unstructured data and it is increasing with constant growth rate every day. Hence, it emerges into necessity of a database technology that could support wide range of data storage, scalable processing and analysis of unstructured data. In this scenario, big data technologies evolve as revolutionary solution. The big data solution also revolved around some architectural artifacts, scalable storage, cloud computing, parallel processing and new data managing paradigms. Big data defines with 3V's characteristics[1]. The first 'V' is symbolization of extra large scale of the data **volume**. The second 'V' is symbolization of **variety** of data that emphasis on heterogeneous data (structure, unstructured and semi structure). The third 'V' is symbolization of **velocity** of data that highlight on data-analytics. Figure 1 shows detail description about 3V characteristics of big data.

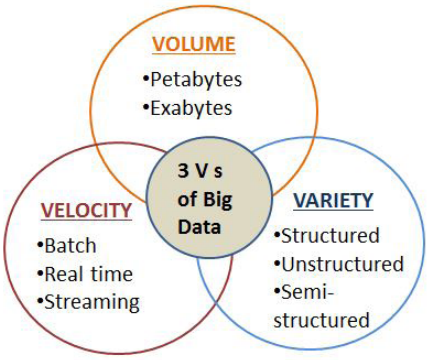

Figure 1. 3V's of Big data.

The main challenges when working with Big Data are storage management and optimized processing. An alternative of the traditional RDBMS must be invented to manage scalable database. The imminent data management approach referred as NoSQL (Not only SQL) technologies. This technology was expanded and implemented with Key-value store, BigTable, Document-store and Graph data store. Figure 2 shows various implementation methodology of NoSQL.

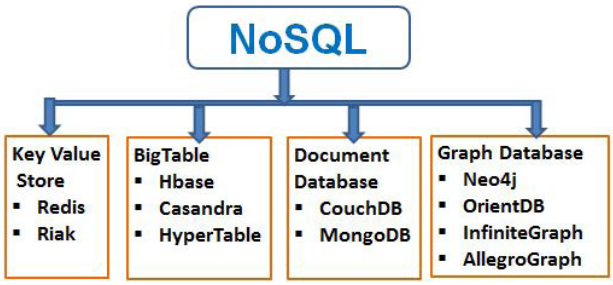

Figure 2. NoSQL Implementation.

4. Graph Data

Under the umbrella of NoSQL database movement many data models were proposed, mainly Key-Value pair, BigTable implementation, Document-store and Graph Database. Recently, graph becomes very well-known structure for data management in research area for its versatile applicability to represent any kind of real world problems. A graph database is uses graph structures to store information, which is basically collection of nodes, edges, and properties[2]. Node is primitive building block of graph which represents entities of real worlds. The relations between nodes are represented as edges. Property

describes entities and edges[3,4]. Figure 3 shows characteristics of graph data structures.

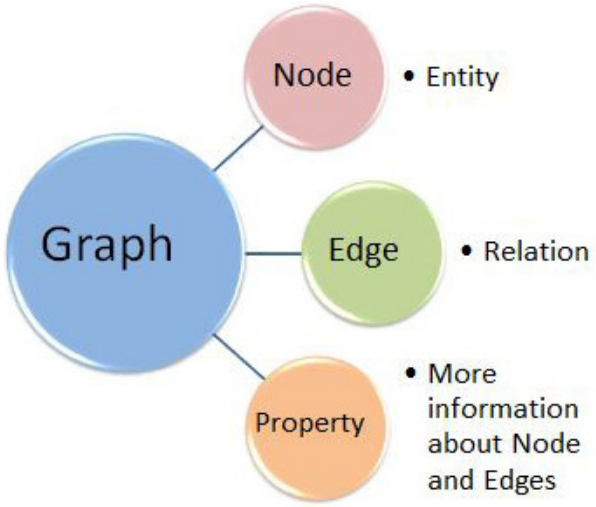

Figure 3. Graph structure.

4.1 Benefits of Graph Data Modeling

The gaining popularity of graph database is due to many advantages offer by it. The graph databases has been proven beneficial with schema free structure, the flexibility of structure for representing real world data, efficiency for associative operations and diversification of the structure. The detail comparison between relational data and graph data is presented in literature[5].

4.1.1 Gain for Multiple Entity Representation

Graph consists of three main building blocks: node, properties and edges. Node is primitive building block which represents single or many types of entities. Nodes can use to represent entity of the domain such as people, businesses, accounts, protein, book, research paper, web page or any other thing you might want to maintain information of it. Properties of nodes are relevant information about node entity, which are essential for expressive function as well as data preservation. For instance, if "Book" is entity and "Programming with NoSQL" is one of the nodes, it may tie with properties such as "Author", "Price", "Publisher" and many others. It depends on which aspects of entities are relevant and useful to particular database. Edges represent the relation between nodes. Most of the important information really stored in the form of edges. A graph can naturally model many connected and complex structures like web link, program control flow, traffic flow, genome database, medical records, workflow and social network.

Graph comprised of same type of nodes is known as single entity graph or single relation graph, which can helpful for dataset having single entity like person, city and research paper for social network, city map and graph of friends network. Figure 4 shows relation among various persons using single entity graph.

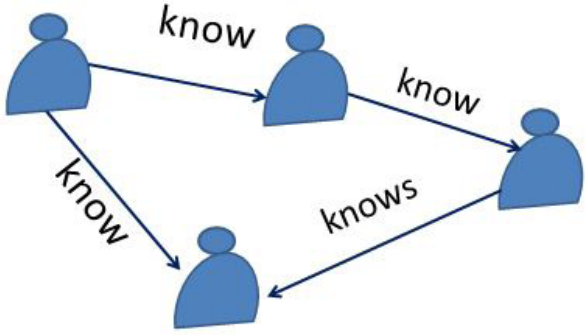

Figure 4. Single entity graph structure.

Graph structure also allows establishing relations among multiple types of nodes which represent distinct entities and their attributes as a single structure. Such graph which comprises links between various different types of nodes is known as multiple entity graphs or multi relation graph. For example, web graph can include nodes about web pages, author, title, language, image and metadata of pages. In short, graph structure provides added gain for the representation of multiple distinct entities along with their associated entities. The link can be established between different neighboring entities affiliated with each other, even dissimilarity of the nature. The close physical association of multiple entities and their relations positively affect on performance of data retrieval. Figure 5 shows relation among various persons and books using multiple entity graphs.

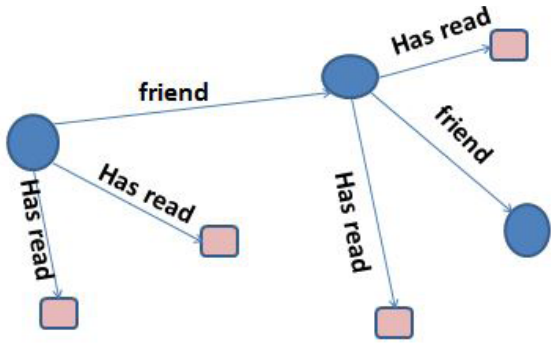

Figure 5. Multiple entity graph.

4.1.2 Flexibility due to Diversity of Structure

The graph structure mostly comprised of three building blocks: nodes, edges and property which represents entity, relationship and attributes respectively. Many meaningful models emerge when ones examine the all probable combinations of nodes, properties, and edges to represent connections and interconnections of entities. In literature, many models are illustrated such as directed graph, undirected graph, labeled graph, unlabeled graph, simple graph, hyper graph, property graph, multi-link graph, tree like structure, weighted graph, un-weighted graph etc[6-8]. Graph is a general model where as trees, lattices, sequences, path and items are degenerated graphs[9].

In mathematics, various kind of graph structures are defined, which offer in-hand guidelines for data modeler[6,7,10]. Several sophisticated graph structures recommended in graph theory, out of that most significant and common types are discussed here. A finite graph is a graph with a finite number of edges and a finite number of vertices. Note that in a simple graph, a finite number of edges follows directly from a finite number of vertices. In the case of a multi-graph this may not apply, as there may be an infinite number of edges between two given vertices. A graph which has either an infinite number of edges or vertices is an infinite graph. Single graph data base is represented as a single graph, which is generally used to represent single drug compound, chemical compound, and biological structure. Multiple graph database is consisting of many disjoin graph, which is generally used to represent many drug compounds, work flow or sentences in text. A bipartite graph is a graph having two distinct sets A and B of nodes, such that each edge pair to a vertex in A and a vertex in B. For example, internet movie database consists of nodes for actors and movies; each node of actor category is linked with node of movie category. A multi-graph is a graph with multiple edges between the same pair of vertices. A simple graph is a graph which is not a multi graph, i.e. there is no more than one edge could be between each pair of vertices. It is generally used to represent data of social network where relationship between two persons shown by edge. An undirected graph is a graph with the edges which are defined as doubleton sets of vertices but, neither ordered pairs nor numbered. A directed graph is a graph with directed edges. It can be used to represent web structure, where web page has links towards many other web pages. Property graph is a very expressive graph[11] which stores attributes related to entities and/or relations. It can be used for social network, email network and many other applications. Weighted graph is a special type of graph where weight (numeric property) is assign to the edges for special purpose. It can be used in geographical system, citation network etc. This graph model has been implemented by neo4j framework which is gaining popularity nowadays.

5. Challenges with Graph Data Modeling

Graph can be used for many domains which are extremely diverse this may require a range of different graph models and diverse way of processing. Many techniques have been proposed were for the case of specific graphs under a requirement of that specific domain, complexity of data and need of processing methodology. Out of that, few are discuss in detail in literature review section. Different applications required graphs of different sizes, density, diameter and complexities. For example, the graph model which are designed for the web or social networks need to be constructed for scalable data with extra-large size and having distinct node labels. Here, to cope up with large number of edges, it is suggested that the edges can be derived as data stream, so special processing techniques are required for the same. On the other hand, the graph models which are designed for chemical data includes many small sized graphs and processing need to take into account repetitions in node labels with different isomorphism. Whereas, the graph model for biological data are large in size and bipartite in nature. Such variation makes graph modeling much more challenging and needs modifications in generalized graph models for various applicant domains.

5.1 Graph Data Models for Various Domains

In this section, we illustrate literature review of different graph models used in various domains like biology, computer networking, social networking, text analysis and e-commerce. Another research has also shown use of weighed graph data for image analysis[12]. The main idea behind this is demonstrating possibility of graph data modeling for various domains. This explorative study of diverse graph based data representations that could motivate researchers in the future.

5.2 Graph Data Model for Computer Network and Web Data

The internet and computer networks are similar to graph in nature. Graph based representation of massive internet data is known as Web graph. The Web of Data (RDF) is characterized with massive number of nodes, very large number of directed edges and multi-relational graphs. It is scale-free graph and used for personalized Page Rank computation[13]. Investigators are interested to search of structural features of graph, common or uncommon sub-network patterns and to understand the propagation of influence, information, diseases and computer viruses over the network using graph based representation[14]. It is challenged by massive data management, domain issue related to networked data, condensed representations of the graph data and graph summarization techniques[15]. As a part of large graph management challenges, the edges in the underlying graph may arrive in the form of a data stream so, different techniques are essential to be influence for future processing of the underlying graphs[16].

5.3 Graph Data Model for Social Network Data

Social networking site has open up many avenues for research and investors have started thinking about social network analysis many years ago[17]. The main challenge of social network analysis includes local and global pattern analysis, finding local influential entities, exploring network dynamics and many more[18]. Earlier research work has shown study and analysis of web of group affiliations. The literature introduces a time-varying social network data model[18] using property graph and Neo4j graph database management framework. The literature introduces unique methodology of finding communities using dynamic weighted directed graphs[19]. The literature shows new approach of mining frequent and high weighted cliques from bird's migration networks[20].

5.4 Graph Data Model for Bio-Chemical Data

Graph data modeling and mining started primarily in practice for biological data like genome, protein structure and drug compounds. Bio-chemical data can easily represent as graphs in which the nodes correspond to atoms or amino acids and the links correspond to bonds between the atoms. Graph proved as powerful tool to represent complex structure of biological data as it naturally in the form of graph. In the case of gene data, the individual graph is massive and having significant repetitions among the different nodes. In this field, the main computation challenges are sub graph discovery, graph matching, frequent sub structure pattern mining and sequence extraction with respect to isomorphism of the structure[21]. The isomorphism challenge is that the nodes in a given pair of graphs may match in a variety of ways. Many algorithms has been proposed and proven efficient such as SUBDUE, gSPAN, GASTON, FFSM, GREW etc[22].

5.5 Graph Data Model for Text Data

Recently, graph representation of unstructured text data was proposed by many researchers[23,24] which increase expressive power as well as content and structural representation of data. The graph representation of text data is emerged as an alternative to the bag of words/phrases representation[25]. There presentation serves to capture a range of aspects: 1. word stem 2. word Part Of Speech (POS) 3. word order 4. word hyponyms 5. sentence structure 6. sentence division and 7. sentence order[24]. The document classification technique which uses bag-of-words/ phase representation with use of vector space model, that fail to cover semantic and structural details[26]. In recent times, distance graph representations of text data has been proposed which preserve information about the relative ordering and distance between the words in the graphs in contrast to traditional vector based model[25]. The wide spectrum of text mining algorithms is available which can directly leveraged with graph based text representation. It shows many advantages and richness of the graph structure[25].

5.6 Graph Data Model for GIS Data

The graph based representation of GIS data is proposed for the reason that graph is powerful and flexible knowledge representation structure. It can easily combine spatial and non-spatial data at the same time along with three types of spatial relations: topological, orientation and distance[26]. Specifically, geometric network is composed of junction points (nodes), flows between junction points (edges) and weights assign to the junction points. These types of geometric networks are often used to model road network and public utility networks such as electrical, gas and water networks. It can be extensively utilized for transportation, infrastructure and other public utility network planning, design and maintenance.

Table 1. Various graph models.

Domain	Data Set	Graph Property	Node Property	Edge Property
Web	Web pages	Massive Dynamic	Single Relation /Multi Relational Labeled Nodes	Directed
Social Networking	Friend Network	Massive Dynamic	Single Relation / Multi Relation Labeled Node	Directed/ Undirected Labeled / Unlabeled
Biochemistry	Drug data	Small Multi-graph Static	Multi Relation Labeled Node	Undirected
Text Mining	Words	Multi-graph	Multi Relation Labeled node	Directed
Bioinformatics	Gene data	Bipartile	Labeled node	Undirected
Network	Routers data	Massive Dynamic	Labeled	Undirected
E-commerce	Knowledge	Dynamic	Labeled	Directed Weighted

5.7 Graph Data Model for E-Commerce Data

Since the late 20th century, the number of Internet users has increased dramatically and many e-commerce giants started performing well. A very wide product catalogue, a huge amount of customer data, extensive range of product features and real time processing with this Big data is really challenging. On the other side personalization, information filtering and recommendation system become integrated part of e-commerce website. In general, the nature of e-commerce website data is scalable, multi-dimensional, semi-structured to unstructured and volatile. In this scenario, few initiatives have taken to move towards use of graph database. For instance, the research literature has proposed use of graph for recommendation system called knowledge graph and proven more effective[27]. The research proposed use of graph for digital library recommendation system[28-30]. The core of the research work consequence various graph models proposed for solving real world problems by exploring underneath data property and suggesting supportive graph structure to improve efficiency for the further processing as shown into Table 1.

6. Conclusion

This research effort is emphasis on potential use of graph structure specifically for large, complex, unstructured, time variant and dynamic data set. The major characteristics of graph structure include flexibility due to diversity of structure, efficiency for associative operations, schema-free representation of data, gain for multiple entity representation and adaptability for diverse domains. In contrast to relational databases where typically requires expensive join operations, graph databases are often proven faster for associative data sets as they do not typically require expensive join operations. Graph based data modeling - attracts many people working with big data as they can perform efficient with large data sets. The schema free nature is more suitable to manage ad-hoc and changing data with evolving schemas such as ad-hoc network, social network and web.

The outcome of this study lend helping hand in narrow down the complexity of data modeling process and provide guidelines for the selection of the suitable graph model for complex data. Still more research is expected towards survey of various frame works implemented of the various graph models.

7. Reference

1. Sagiroglu S, Sinanc D. Big Data: A review. In: Collaboration Technologies and Systems (CTS), International Conference IEEE; 2013 May. p. 42–47. Crossref
2. Aggarwal CC, Wang H. Managing and Mining Graph Data, New York: Springer. 2010. 40, p. 181–90.
3. Angles R, Gutierrez C. Survey of Graph Database Models, ACM Computing Surveys (CSUR). 2008; 40(1):1. Crossref
4. Ghrab A, Romero O, Skhiri S, Vaisman A, Zimányi E. Grad: On Graph Database Modeling. arXiv preprint arXi. 2016. p. 28.
5. Vicknair C, Macias M, Zhao Z, Nan X, Chen Y, Wilkins D. A Comparison of a Graph Database and a Relational

Database: A Data Provenance Perspective. In: Proceedings of the 48th Annual Southeast Regional Conference ACM; 2010 Apr 15, p. 42. Crossref
6. Rodriguez MA, Neubauer P. The Graph Traversal Pattern. arXiv preprint arXiv; 2010. p. 1–18.
7. Essam JW, Fisher ME. Some Basic Definitions in Graph Theory, Reviews of Modern Physics. 1970; 42(2):271. Crossref, Crossref.
8. West DB. Introduction to Graph Theory Upper Saddle River: Prentice Hall, 2ne Edition, 2001.
9. Pokorný J. Graph Databases: Their Power and Limitations. In: IFIP International Conference on Computer Information Systems and Industrial Management. Springer International Publishing; 2015 Sep, p. 58–69. Crossref.
10. Aggarwal CC, Wang H. Graph Data Management and Mining: A Survey of Algorithms and Applications. In: Managing and Mining Graph Data. Springer US; 2010. p. 13–68. Crossref
11. Glossary of Graph Theory Terms. Date accessed: 01/01/2015. Crossref.
12. Miller JJ. Graph Database Applications and Concepts with Neo4j. In: Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA: USA; 2013. 2324, p. 36.
13. Cheng B, Yang J, Yan S, Fu Y, Huang TS. Learning with ℓ^1 -Graph for Image Analysis, IEEE Transactions on Image Processing. 2010; 19(4):858–66. Crossref.
14. Jeh G, Widom J. Scaling Personalized Web Search. In: Proceedings of the 12th International Conference on World Wide Web; 2003. p. 271–79. Crossref.
15. Chakrabarti D, Faloutsos C. Graph Mining: Laws, Generators, and Algorithms, ACM Computing Surveys (CSUR). 2006; 38(1):2.
16. Phillips C, Swiler LP. A Graph-Based System for Network-Vulnerability Analysis. In: Proceedings of the 1998 Workshop on New Security Paradigms; 1998. p. 71–79. Crossref.
17. Babcock B, Babu S, Datar M, Motwani R, Widom J. Models and Issues in Data Stream Systems. In: Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM; 2002. p. 1–16. Crossref. PMid:12776715.
18. Wasserman S, Faust K. Social Network Analysis: Methods and Applications, Cambridge University Press. 1994; 8:1–6.
19. Cattuto C, Quaggiotto M, Panisson A, Averbuch A. Time-Varying Social Networks in a Graph Database: A Neo4j use Case. In: First International Workshop on Graph Data Management Experiences and Systems ACM; 2013. p. 11. Crossref.
20. Dongsheng D, Li Y, Jin Y, Lu Z. Community Mining on Dynamic Weighted Directed Graphs. In: Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information and Knowledge Management; 2009. p. 11–18. PMid:19237778 PMCid:PMC2861560.
21. Tang M, Wang W, Jiang Y, Zhou Y, Li J, Cui P, Liu Y, Yan B. Birds Bring Flues? Mining Frequent and High Weighted Cliques from Birds Migration Networks. In: Database Systems for Advanced Applications. Springer Berlin/ Heidelberg; 2010. p. 359–69. Crossref.
22. Butler G, Wang G, Wang Y, Zou L. A Graph Database with Visual Queries for Genomics. In: APBC; 2005. p. 31–40. Crossref.
23. Faloutsos C, McCurley KS, Tomkins A. Fast Discovery of Connection Sub-graphs. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM; 2004. p. 118–27. Crossref.
24. Wörlein M, Meinl T, Fischer I, Philippsen M. A Quantitative Comparison of the Subgraph minersMoFa, gSpan, FFSM, and Gaston. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg; 2005. p. 392–403.
25. Jin W, Srihari RK. Graph-Based Text Representation and Knowledge Discovery. In: Proceedings of the 2007 ACM Symposium on Applied Computing, ACM; 2007. p.807–11. Crossref. PMCid:PMC1891340.
26. Jiang C, Coenen F, Sanderson R, Zito M. Text Classification using Graph Mining-Based Feature Extraction, Knowledge-Based Systems. 2010; 23(4):302–08. Crossref.
27. Aggarwal CC, Zhao P. Towards Graphical Models for Text Processing, Knowledge and Information Systems. 2013; 36(1):1–21. Crossref.
28. Palacio MP, Sol D, González J. Graph-Based Knowledge Representation for GIS DATa in Computer Science. In: Proceedings of the Fourth Mexican International Conference; 2003 Sep. p. 117–24. Crossref.
29. Catherine R, Cohen W. Personalized Recommendations using Knowledge Graphs: A Probabilistic Logic Programming Approach. In: Proceedings of the 10th ACM Conference on Recommender Systems. ACM; 2016 Sep. p. 325–32. Crossref.
30. Huang Z, Chung W, Ong TH, Chen H. A Graph-Based Recommender System for Digital Library. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries, ACM; 2002 July. p. 65–73 Crossref.