A Multi-Source Data Aggregation and Multidimensional Analysis Model for Big Data

Pan LIU, Lin CHEN

School of Computer, National University of Defense Techonology ChangSha, China
1185307014@qq.com,agnes_nudt@qq.com

Abstract:With the rise of Internet applications such as search engines, social networks, and e-commerce, the amount of data in the Internet is rapidly expanding. There are a lot of data generated every moment, and the global information is also increasing. Therefore, the big data is driven by the Internet industry, and it is also a subversive technological innovation compared to the cloud computing and Internet of things. How to carry on the fast retrieval in the massive and different types of data? How to discover potential associations between different data? How to mining the potential value of the data? And how to create a multidimensional view of the data? These urgent problems need to be solved. In this paper, a multi-source data aggregation and multidimensional analysis model for big data (DAM_AM) is proposed. The model adopts the hierarchical structure and introduces data aggregation mechanism, multi-source processing mechanism, object and association mapping mechanism, and "walk" mechanism. Using these mechanisms, multi-source data is normalized to a coherent and consistent representation pattern. And then the fields that represent a class of entity in the representation pattern are aggregated into a set of fields. By mapping different field sets into different objects and associations and combining with the time dimension and space dimension, we can build a multifaceted visual model. Through the concrete case analysis and verification, it indicates that the DAM_AM model can analyze the data from multidimensional and multi-level, and shows the potential correlation between different data. The model not only has high computational efficiency and has high scalability, but also shows the analysis results clearly and intuitively.

1 Introduction

With the rise of Internet applications such as search engines, social networks, and e-commerce, the amount of data in the Internet is rapidly expanding. According to incomplete statistics, Google's daily processing of data amounted to 20PB [1]; Facebook 's user data has reached 15PB and grows at a rate of 60TB per day; Taobao's data center stores at least 14PB of data[2]. International Data Corporation(IDC) predicts that the global information will reach 35ZB [3] in 2020. How to deal with the large-scale data and extract the value from the large-scale data has become an urgent problem. Therefore, The big data is driven by the Internet industry, and it is also a subversive technological innovation compared to the cloud computing and internet of things.

Compared with the traditional data, the characteristics of big data can be summarized as [4]: huge data volume, variety , data update fast, high uncertainty . these features make big data processing face enormous challenges: data volume increases ceaselessly; data in various forms; the authenticity and quality of data is uncertain.. In response to these difficulties and challenges, different platforms of big data processing have been introduced. At present, platforms of big data processing are divided into four categories: batch computing platform, flow computing platform, interactive computing platform and graph computing platform.

The representatives of batch computing platform are the Apache Hadoop [5] and Microsoft Dryad. The Hadoop is the open source implementation based on the MapReduce programming model, and it is established as a open source project in 2006.

The representatives of stream computing platform are Twitter's Storm and Yahoo's S4. In Storm, the processing logic of a Storm application is encapsulated into an object named Topology which is composed of several message source components named Spout and message handler named Bolt, and the processing logic is presented in the form of a directed graph. Storm uses the master / slave architecture.

Interactive computing is the feature of database system (DBMS, Database Management Systems) . Users enter a SQL query command at the database system terminal, the database system will return the search results [6]. The first interactive computing platform is Google's Dremel [7], which is an extensible interactive data query system for read-only nested data. Dremel is able to perform aggregated queries on the table including trillions rows in

seconds by using . service tree and column storage technology of data.

At present, the major graph computing platforms are Apache's Giraph [8] , GraphX [9] developed by the University of California at Berkeley, and the GraphLab[10] developed by Carnegie Mellon University .

Through studying on the interactive computing and the graph computing, we find that the interactive computing can get query results efficiently from large-scale data, but it is difficult to deal with different types of data, and the results are not intuitive enough because the results are always presented in the form of table. Thus it is not easy to find data association in deep level. . The graph computing displays the results of calculation and query in a graphical way. Although the results are very clear and intuitive, the computational efficiency is not as good as the interactive computing. In this paper, we propose and design a multi-source data aggregation and multidimensional analysis model for big data (DAM_AM). Not only can the the model deal with different types of data by normalizing the data in several formats into the same representation pattern, but also makes the final results display more intuitive and clear by creating different views to display the results of querying and matching. The DAM_AM model not only has higher efficiency of calculation and query, and compared with the graph computing, the analysis architecture has a strong scalability. Instead of rebuilding a new model , the new data can be directly updated to the existing models based on new data and previous data,

2 A Multi-Source Data Aggregation and Multidimensional Analysis Model for Big Data

2.1 Model Concepts

Definition 1 Object: A specific instance of a type. An object can be a specific type, node, or entity and it is uniquely defined by the key and described by attributes.

Definition 2 Association: A relationship between two objects. Associations are directional, and it is uniquely defined by the key and described by attributes.

Definition 3 Key: A unique data that identifies an object or association and can vary based on context.

Definition 4 Attribute: Data that describes an object or association and shows features of an object or association. Attribute can be used for searching, detailing, grouping the objects and describing associations and objects.

In the DAM_AM model, objects and associations have types to organize and classify the data. the creation of an object or association is based on a combination of key and type. Different object types can be created to support different types of analyses. Any attributes that are used for searching or querying should have an index put on the corresponding database fields to speed up the results and they do not necessarily need to be "real" values queried from the underlying database.

2.2 The DAM_AM Model Description

The framework of the DAM_AM model is divided into three layers: data resource layer, model layer and data presentation layer. The overall framework is shown in Fig. 1.

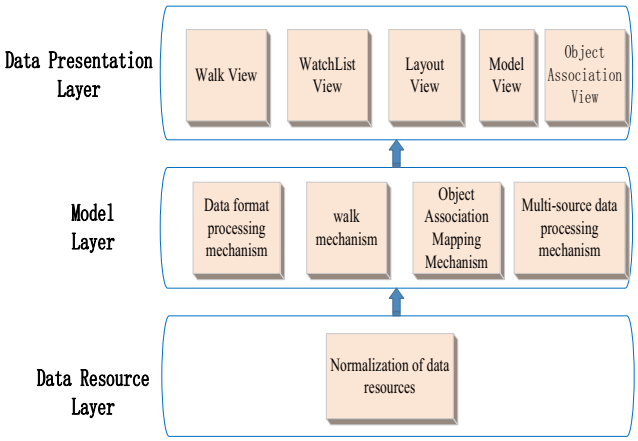

Figure 1. The hierarchical frame diagram of the DAM_AM model

The data resource layer normalizes the different types data which can be accessed and read from different data sources. Only after normalization can you map objects and associations correctly from The set of fields to be aggregated.

The model layer is used to deal with problems when we create the analytical model and map objects and associations from tables. This layer contains multi-source data processing mechanism, walk processing mechanism, object and association mapping mechanism and data format processing mechanism.

The data presentation layer is used to visualize the DAM_AM model and the results of searching and querying. It can display the associations between different objects in a more intuitive way. It contains object and association view, walk view,"WatchList" view, layout view, model view and so on.

2.3 Key Technologies of the DAM_AM Model

1) Data Resource Aggregation Mechanism

Data resource layer first to access and read data from different data sources, These data formats are different, such as the data formats may be csv, xls, xml or database tables and so on. These data must be brought together into a common representation format so that they have a consistent and coherent presentation pattern. And then the fields that represent a class of entity in the representation pattern are aggregated into a set of fields. There may be intersections between different sets of fields, but the key used to identify the entity in sets of fields can not be crossed, and the other fields that used as the attributes of the entity can be crossed. The DAM_AM model maps different sets of fields into different objects and associations according to the needs of the analysis. The data resource aggregation is shown in Fig. 2.

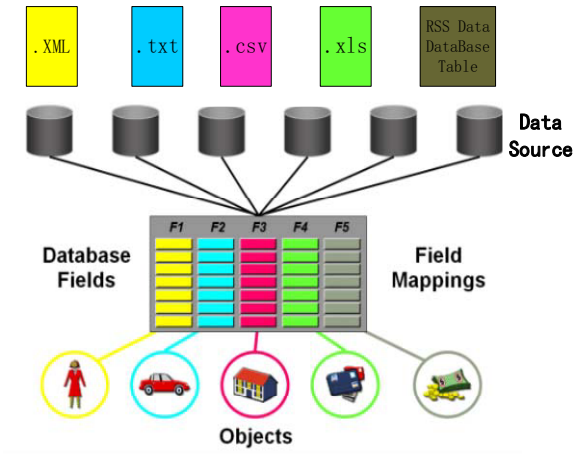

Figure 2. Data aggregation display

2) *Format Processing Mechanism*

The reliability and quality of the data directly affects the creation of the analysis model. So it is necessary to deal with the data before creating the analysis model. We found that the problems of data can be divided into four areas: value errors, data missing, unreasonable data structure, incomplete formats. show in Fig. 3.

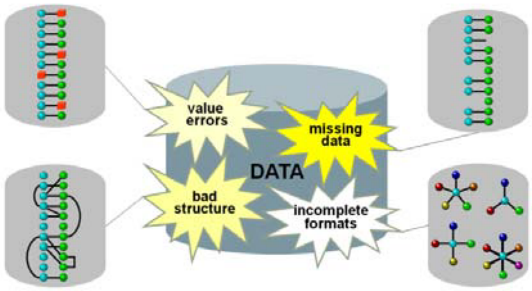

Figure. 3. Classification of data problems

In order to avoid data format errors and data inconsistencies, the data format processing mechanism will deal with the data properly, such as concatenating field values , representing names as SMITH/JOHN/Q and do not break the name into separate fields, uppercasing all values, removing extra spaces as well as representing dates differently.

3) *Multi-Source Data Processing Mechanism*

When you create a DAM_AM model, you must be aware of the correspondence between the model you are creating and the existing data sources. When there are multiple data sources, the correspondence between the model and the different data sources can be roughly divided into three categories (show in Fig. 4). On the basis of that, we analyze the conditions for generating these three modes in detail.

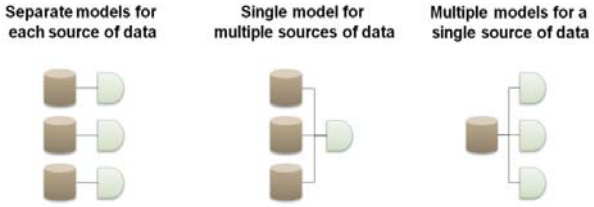

Figure 4. Corresponding modes between multi-sources and models

Separate models for each source of data. the conditions for generating this mode includes: When there aren't any common keys between any of the data sources; When there aren't any common values between any of the data sources; When the sources are very dissimilar to each other – even if there are common values; When each data source can provide its own unique set of patterns and trends.

Single model for multiple sources of data. the conditions for generating this mode includes: When only attribute values can be derived from a source; When there are limited numbers of associations between objects in the data; When there aren't any patterns independently contained within any single source of data.

Multiple models for a single source of data. the conditions for generating this mode includes: When different viewpoints on the data are needed for exposing different analytical situations; When there is more than one way to model the data set; When the data source is very complex - containing multiple tables/fields for the primary objects

4) *Object and Association Mapping Mechanism*

The object and association mapping mechanism is the basis for building a DAM_AM model. Because the process of establishing a DAM_AM model is to map the fields and tables in the database into the objects and associations. For this reason, we must understand the concepts of fields and tables in the database schema before building the model. In addition, we should know the SQL syntax such as Select, Join and Where clauses, because all the objects in the DAM_AM model are created through Select statements , and Select statements is a very fast database operations.

Before you building any model, you should first get familiar with the database and need to understand the structure of the database (tables/fields) as well as review the data to determine which fields are populated. An example of creating objects and associations is shown in Fig. 5.

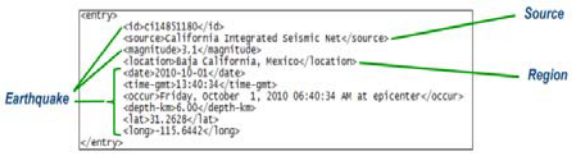

Figure 5. Example of creating objects and associations

Through the analysis of the original XML data, it can be seen that an event entity should consist of three objects: Earthquake , Region and Source. These three objects can represent the event of earthquake clearly. Therefore, we need to create these three objects which mapped from the original data. In the process of creating the objects, we also need to set the names, keys, attributes, icons, and related descriptive labels for different objects. Three objects are created, and the next step is to create associations between the objects. Since the original file is in XML format, this raw data need to be normalized to the table representation via the data resource layer. Association is mapped based on the tables. so as long as the two keys of objects exist in the same table, the association between the two objects can be established. Association are also directional. Before you establishing the association, you need to determine the "FromObject" and "ToObject" to ensure there is a directional association between two different objects. You also need to set key, attributes, and the color of the wires for the association. As shown in the figure above, we can establish the association between Earthquake and Source, and we can also establish association between Region and Earthquake.

5) Walk Processing Mechanism

Walk processing mechanism is to "walk" the objects in the same model or different models. In the association processing mechanism, we propose the concepts of "walk" and level. "walk" refers to an in-depth search on the specified object which based on associations between different objects. A level which equals to 1 means that only "walk" the objects which are directly associated with the specified object. We can "walk" from one object to other objects, and so on, until find the target object. The whole process of the "walk" completely shows the relevancy between objects at different layers, which helps to analyze and find the potential association between different data. In addition, the walk processing mechanism can also cross-search between different models, as long as the keys of the objects in different models are same.

3 Case Analysis and Validation

In this section, we will use the DAM_AM model to analyze and model the example, and show the results of the analysis. Firstly we read four tables named as tblTelephone, tblSuspect, tblOrganization and tblSuspect_tblOrganization from a database. The tblOrganization table includes fields such as name, organization_id and classfiction and so on. The tblSuspect table includes fields such as suspect_id, first name, last name and sex and so on. The tblTelephone table includes fields such as start time ,end time, duration,from number and to number and so on. The tblSuspect_tblOrganization table includes fields such as unique_id, suspect_id and organization_id and so on.

These four tables will be normalized into a unified representation pattern by the data resource layer. The data is formatted to exclude some wrong format, such as the data with the date type, which does not need to display a specific time, will be dropped time. Since these four tables are all read from the same data source, we select the separate models for each source of data. The next step is to use object and association mapping mechanism to create the corresponding objects and associations. Analysis of the different fields of the table, we can see the "FromNumber" field and the "ToNumber" field in the tblTelephone table can be used as different keys to create two objects. And they represent the caller phone and callee phone respectively. On the basis of the analysis, we create two objects. These two objects are both named Phone .The keys of these two objects are not same but types of these two objects are the same. According to the definition of the object for the third quarter, they should be the same object. Thus, these two objects in the model are the same icon and have the same name. In the DAM_AM model, such objects are referred to as merged object.

And then we set the key, attributes, and icon for these two objects whose object types are both Phone. On the basis of that, we establish the association of these two objects based on the tblTelephone table. And in order to set the direction of the association, we set "FromObject" to the Phone object whose key is "FromNumber" and set the "ToObject" to the Phone object whose key is ."ToNumber". Finally, we set the key of the association and the color of the connection line. Similarly, analysis of the fields for the remaining three table, it is noticeable that tblSuspect table can create Suspect object and tblOrganization table can create Organization object. And then the association between the Suspect object and the Phone object can be established based on the tblTelephone table. Then again, the association between the Organization object and the Suspect object can also be established based on tblSuspect_tblOrganization table. The steps about creating the new objects and establishing the new associations are the same as for creating a Phone object, so we will not repeat them again.

Each time a new object and association is added to the same model, the model will update. We can see the model update process (show in Fig. 6) after these three objects and associations are created.

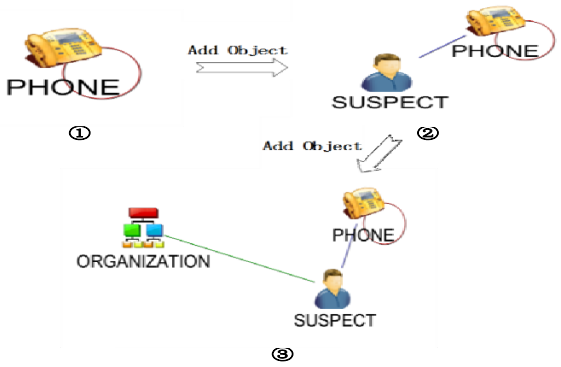

Figure 6. The update process of the DAM_AM model

From the view of the model update, we can see that ① denotes a self-linked object, ② denotes the transition

model, ③ denotes the final DAM_AM model. Open the object and association view of this model to see the Fig. 7.

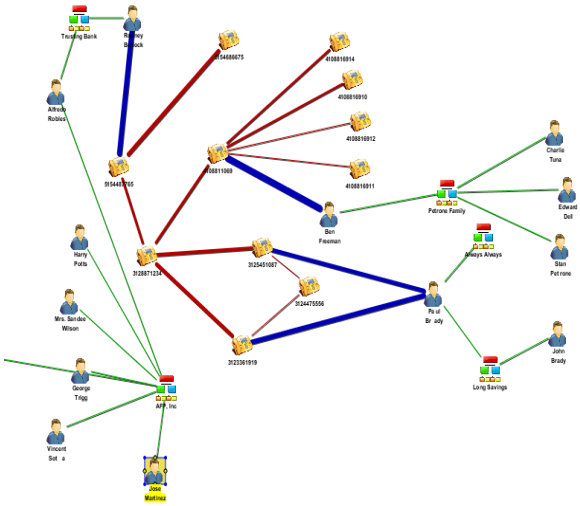

Figure 7. Object and association view

Through the analysis of the object and association view, we can see clearly which two Phone objects have made a phone call, and the thickness of connecting lines represents the number of calls. the thicker the connecting lines, the higher the frequency of calls. Similarly, we can see which Suspects are included in Organizations.

After creating the DAM_AM model, we can create a "WatchList" file. To match the contents of the "WatchList" file in the DAM_AM model ,and then the matched objects will be marked in a special icon in the view. In order to obtain the relationship between the two special Suspect objects and other objects in the model, You can select these Suspect objects and "walk" them. Finally we can see the relationship between these two objects and other objects in the view. Show in Fig. 8~10.

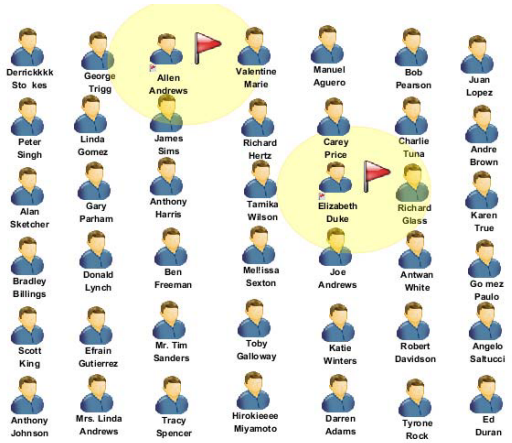

Figure 8. "WatchList" view

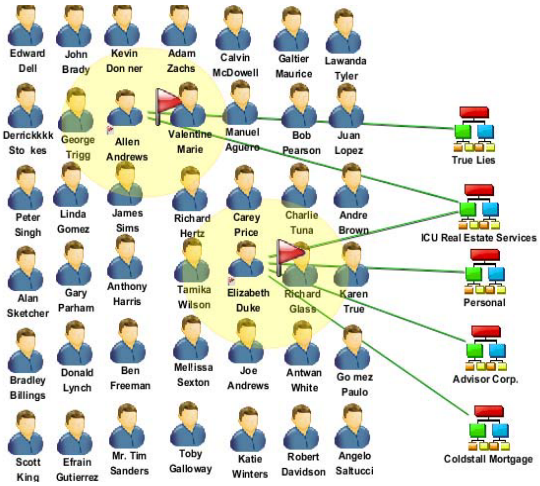

Figure 9. "walk" view

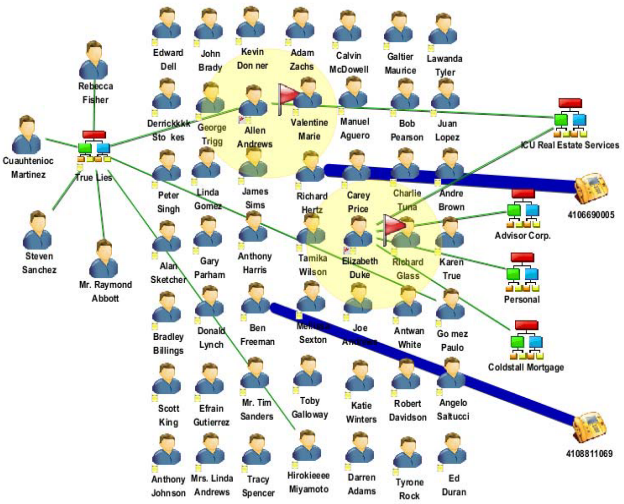

Figure 10. "walk" view

4 Conclusion

In this paper, we design a multi-source data aggregation and multidimensional analysis model for big data (DAM_AM). The model adopts the hierarchical structure and introduces data aggregation mechanism, multi-source processing mechanism, object and association mapping mechanism, and walk mechanism. Using these mechanisms, multi-source data is normalized to a coherent and consistent representation pattern. And then the fields that represent a class of entity in the representation pattern are aggregated into a set of fields. By mapping different field sets into different objects and associations we can build a multifaceted visual model. The DAM_AM model is further improved and enriched the model by putting forward some new concepts such as merged object, "WatchList", and "walk". the DAM_AM model can analyze the data from multidimensional and multi-level and shows the potential correlation between different data. The model not only has high computational efficiency and has high scalability, but also shows the analysis results clearly and intuitively.

References

[1] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters.Communications of the ACM. 2008, 51 (1): 107–113.

[2] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows,and biggest growth in the far east. IDC iView: IDC Analyze the Future. 2012, 2007: 1–16.

[3] China Computer Society Big Data Expert Committee China Big Data Technology and Industry Development White Paper, 2013.

[4] Barwick H. The“ four Vs” of Big Data. Implementing Information InfrastructureSymposium, 2012.

[5] Apache Hadoop MapReduce. https://hadoop.apache.org.

[6] Pavlo A, Paulson E, Rasin A, et al. A comparison of approaches to large-scale dataanalysis. In Proceedings of the 2009 ACM SIGMOD International Conferenceon Management of data. 2009: 165–178.

[7] Melnik S, Gubarev A, Long J J, et al. Dremel: Interactive Analysis of Web-ScaleDatasets. In Proc. of the 36th Int’l Conf on Very Large Data Bases. 2010:330–339.

[8] Apahce Giraph. http://giraph.apache.org/.

[9] Xin R S, Gonzalez J E, Franklin M J, et al. Graphx: A resilient distributed graphsystem on spark. In First International Workshop on Graph Data ManagementExperiences and Systems. 2013: 2.

[10] GraphLab: A New Parallel Framework for Machine Learning. http://www.select.cs.cmu.edu/code/graphlab/.