US 20140046977A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2014/0046977 A1

(43) Pub. Date: Feb. 13, 2014

GOPALAKRISHNAN et al.

(54) **SYSTEM AND METHOD FOR MINING PATTERNS FROM RELATIONSHIP SEQUENCES EXTRACTED FROM BIG DATA**

(71) Applicant: **XURMO TECHNOLOGIES PVT. LTD.**, Bangalore (IN)

(72) Inventors: **SRIDHAR GOPALAKRISHNAN**, Bangalore (IN); **SUJATHA RAVIPRASAD UPADHYAYA**, BANGALORE (IN)

(73) Assignee: **XURMO TECHNOLOGIES PVT. LTD.**, BANGALORE (IN)

(21) Appl. No.: **13/755,047**

(22) Filed: **Jan. 31, 2013**

(30) **Foreign Application Priority Data**

Aug. 10, 2012 (IN) 3286/CHE/2012

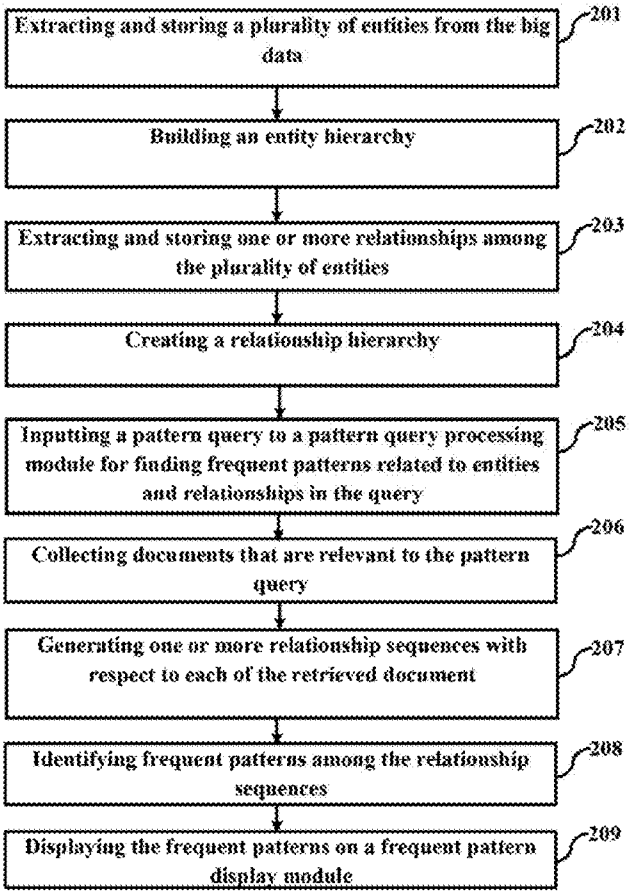
Publication Classification

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 17/30539*** (2013.01); ***G06F 17/30604*** (2013.01)
USPC **707/776**

(57) **ABSTRACT**

The various embodiments herein provide a system and method for mining frequent patterns in relationship space from a plurality of relationship sequences extracted from a big data The system comprises a data repository for collecting and storing the big data. An Entity Store for collecting and storing a plurality of entities from the big data, an Entity Hierarchy for representing a hierarchical structure of entities, a Relationship Store for collecting and storing relationship instances between the pluralities of entities, a Relationship Hierarchy for representing a hierarchical structure of relationship, a language/domain model for organizing entities and relationships in a hierarchical manner, a pattern query Processing Module (PQPM) for processing, a pattern query related to finding patterns in relationships and entities, and a Pattern Generation Module (PGM) to generate frequent patterns and a Frequent Pattern Display Module (FPDM) to provide a visual presentation of the mined patterns.



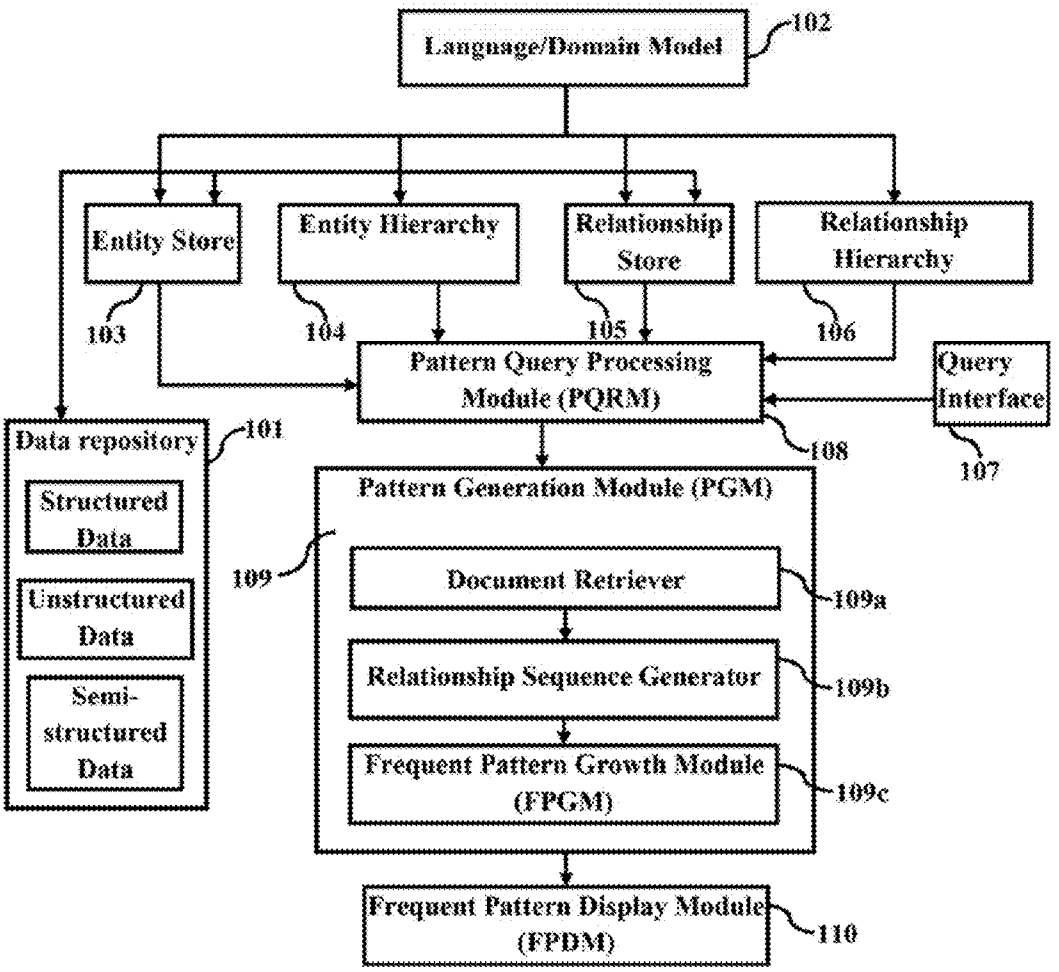

Language/Domain Model
102
Entity Store
Entity Hierarchy
Relationship Store
Relationship Hierarchy
103
104
105
106
Pattern Query Processing Module (PQRM)
Query Interface
107
108
Data repository
101
Structured Data
Unstructured Data
Semi-structured Data
Pattern Generation Module (PGM)
109
Document Retriever
109a
Relationship Sequence Generator
109b
Frequent Pattern Growth Module (FPGM)
109c
Frequent Pattern Display Module (FPDM)
110

FIG. 1

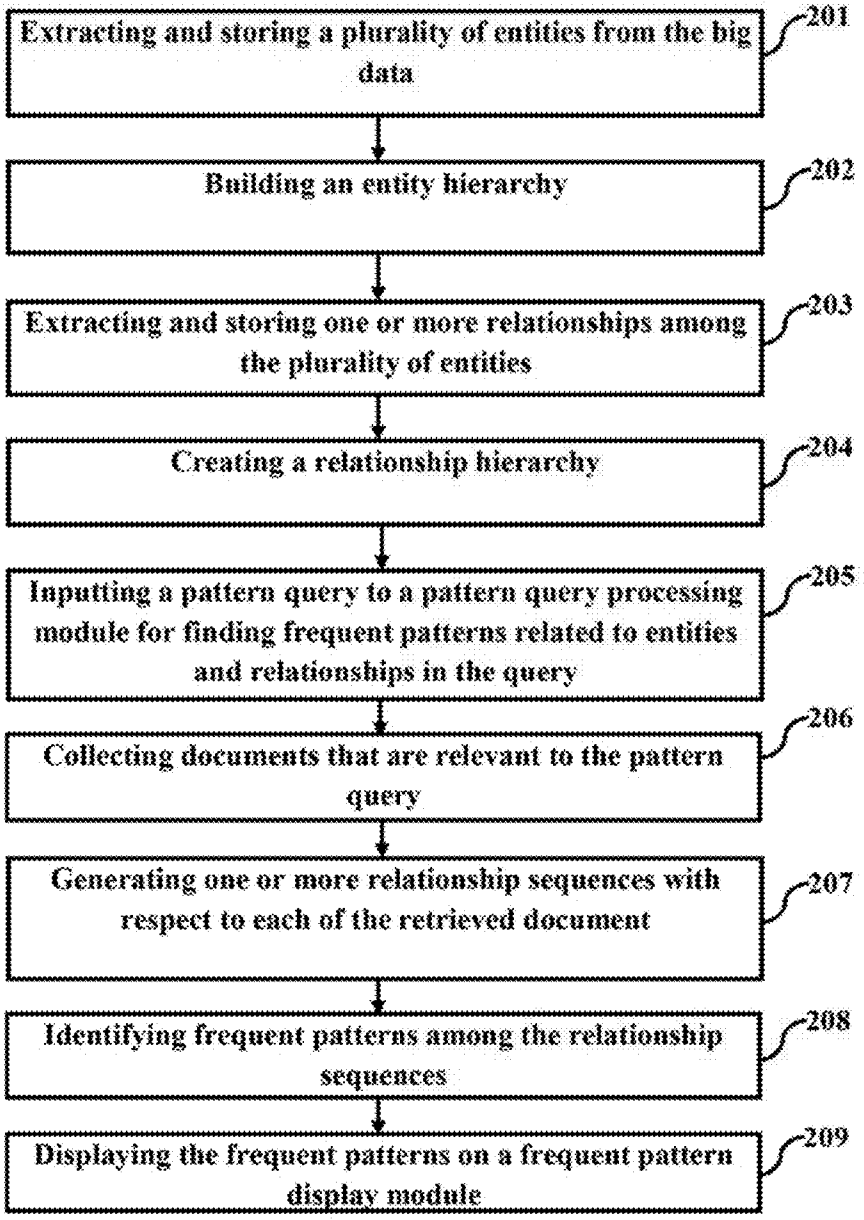

Extracting and storing a plurality of entities from the big data
201
Building an entity hierarchy
202
Extracting and storing one or more relationships among the plurality of entities
203
Creating a relationship hierarchy
204
Inputting a pattern query to a pattern query processing module for finding frequent patterns related to entities and relationships in the query
205
Collecting documents that are relevant to the pattern query
206
Generating one or more relationship sequences with respect to each of the retrieved document
207
Identifying frequent patterns among the relationship sequences
208
Displaying the frequent patterns on a frequent pattern display module
209

FIG. 2

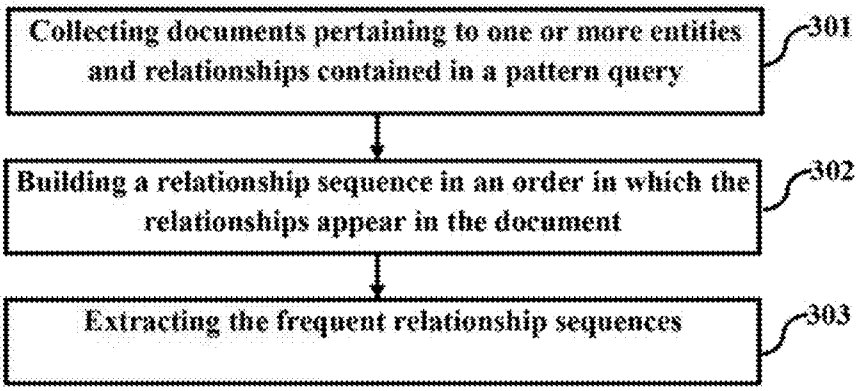

Collecting documents pertaining to one or more entities and relationships contained in a pattern query
301
Building a relationship sequence in an order in which the relationships appear in the document
302
Extracting the frequent relationship sequences
303

FIG. 3

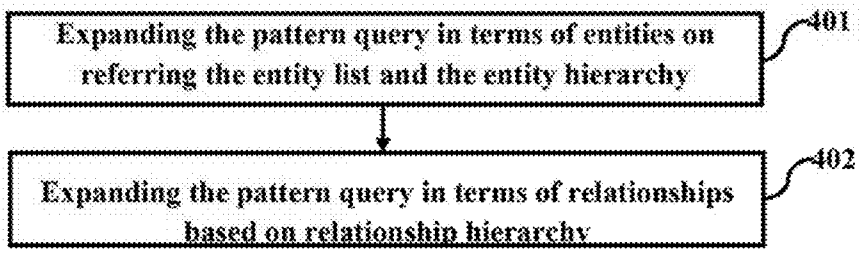

Expanding the pattern query in terms of entities on referring the entity list and the entity hierarchy
401
Expanding the pattern query in terms of relationships based on relationship hierarchy
402

FIG.4

SYSTEM AND METHOD FOR MINING PATTERNS FROM RELATIONSHIP SEQUENCES EXTRACTED FROM BIG DATA

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present application claims priority of Indian provisional application serial number 3286/C14E12012 filed on Aug. 10, 2012, and that application is incorporated in its entirety at least by reference.

BACKGROUND

[0002] 1. Technical Field

[0003] The embodiments herein generally relate to data mining and particularly relates to a mining patterns from structured, unstructured and semi-structured data from heterogeneous sources. The embodiments herein more particularly relates to a system and method for mining patterns in relationship sequences extracted from big data.

[0004] 2. Description of the Related Art

[0005] Information explosion within and outside the organization has led to exponential increase in unstructured data, while the systems currently used are especially meant for processing structured data. With the advent of big data systems such as columnar databases, map reduce frameworks such as Hadoop, it is now possible to store heterogeneous data at one point. A big data is any one or a combination of an unstructured data source, a semi-structured data source and a structured data source. However, making information available for analytics or deriving new perspectives from big data to enable analytics is something that is not understood clearly yet.

[0006] Pattern mining in structured data is a fairly well understood problem; however, pattern mining on unstructured data is much less understood. The approaches for pattern mining in structured and unstructured data are completely different. In both structured and unstructured pattern mining, the co-occurrence of entities decides the pattern, given that the entities can share multiple relationships among themselves. However, the co-occurrence of entities alone does not ensure that patterns are bound to the correct context.

[0007] One of the existing prior art provides a system and method for the automatic mining of new relationships which employs the use of "association rule mining" in discovering new relationships. The "association rule mining" technique basically uses the co-occurrence of words that are used to describe a relationship to find new relationships. The objective of this prior art is to discover new relationships between entities given that a statistical module asserts the significance of the relationship, and a relationship does not match existing relationships between the pair of entities already in the relation database. The prior art adds new relationships after trying to resolve it with the existing relationships.

[0008] Another existing prior art provides a state of art method for effective pattern discovery for text mining which follow a term based approach for closed sequence pattern mining. This effort too examines sequences that are formed by term occurrences. The prior art considers only the text data and the method of extracting patterns cannot be extended to other forms of data. Also, the prior art limits itself to mining patterns in entity space, where every term is considered as entity.

[0009] There exist many limitations in existing prior arts which explain pattern mining in relationship space. The existing systems attempt to perform pattern-mining on either structured or on unstructured data and not on amalgamation of both. Also, the approach of existing pattern-mining is based on co-occurrence of two or more entities, and mines patterns in entity space only. These methods do not ensure contextual resolution of entities, as same entities can co-occur in different contexts he existing pattern-mining approaches do not mine patterns upon resolution of both entities and relationships, although certain aspects of entity resolution have been addressed. Further, many forms of representation of relationships that occur between entities are rather complex and require expensive logical inference mechanism for realizing a hierarchy of the relationships. In unstructured data context, it is important to arrive at a suitable representation of relationship that facilitates easy resolution of relationships.

[0010] In view of the foregoing, there is a need for a system and method for mining patterns in relationship sequences extracted from big data. There is also a need for system and method for finding patterns based on co-occurring relationships. Further there exists a need for a system and method which can extract frequent patterns in relationship space from relationship sequences.

[0011] The abovementioned shortcomings, disadvantages and problems are addressed herein and which will be understood by reading and studying the following specification.

SUMMARY

[0012] The primary object of the embodiments herein is to provide a system and method for mining patterns in a relationship space from a collection of structured, unstructured and semi-structured, data.

[0013] Another object of the embodiments herein is to provide, a system and method for enabling pattern extraction in relationship space by storing entities and relationships, and maintaining entity hierarchy and relationship hierarchy respectively.

[0014] Yet another object of the embodiments herein is to provide a system and method for building relationship sequences from heterogeneous data sources to represent the order in which the relationships occur to facilitate pattern mining.

[0015] Yet another object of the embodiments herein is to provide a system and method for extracting relevant relationship sequences from stored relationships using entity and relationship hierarchies for pattern-mining.

[0016] Yet another object of the embodiments herein is to provide a system and method for generating most frequent patterns in relationship space from relationship sequences.

[0017] Yet another object of the embodiments herein is to provide a system and method for deriving new perspectives from big data to enable analytics of the derived data.

[0018] These and other objects and advantages of the present invention will become readily apparent from the following detailed description taken in conjunction with the accompanying drawings.

[0019] The various embodiments herein provide a system for mining frequent patterns in relationship space from a plurality of relationship sequences extracted from a big data. The system comprising a data repository for collecting and storing the big data, an entity store for collecting and storing a plurality of entities from the big data, an entity hierarchy for representing a hierarchical structure re of entities, a relation-

ship store for collecting and storing relationship instances between the plurality of entities from the big data, a relationship hierarchy for representing a hierarchical structure of relationships, a language/domain model for organizing entities and relationships in a hierarchical manner, a Pattern Query Processing Module (PQPM) for processing a pattern query related, to finding patterns in relationships and entities, a Pattern Generation Module (PGM) to generate frequent patterns from one or more relationship sequences from the data sources collected based on the pattern query and a Frequent Pattern Display Module (FPDM) to provide a visual presentation of the mined patterns. The pattern generation module performs frequent pattern mining by gathering relevant data sources using the entity hierarchy and the relationship hierarchy. It generates relationship sequences with respect to each of the data source and extracts the most frequent patterns in the collection of relationship sequences.

[0020] According to an embodiment herein, the big data comprises structured, unstructured and semi-structured data from heterogeneous data sources.

[0021] According to an embodiment herein, the entity store is a collection of entities extracted from the big data. The entity store stores specific information with respect to each entity.

[0022] According to an embodiment herein, the entity hierarchy represents a hierarchical structure of entities resolved using Natural Language Processing (NLP) techniques with a support of the Language and Knowledge Models.

[0023] According to an embodiment herein, the relationship store is adapted to store information related to each relationship instance.

[0024] According to an embodiment herein, the Relationship Hierarchy represents a hierarchical arrangement of relationships by resolving, the relationships through at least one of a word-sense disambiguation technique, syntactic resolution and semantic resolution in conjunction with the language/domain model for context resolution.

[0025] According to an embodiment herein, the Pattern Query Processing Module (PQPM) processes the pattern query by expanding the pattern query in terms of entities after consulting with the entity store and the hierarchy of entities. The pattern query is a list comprising entities and relationships of the entities.

[0026] According to an embodiment herein, the Pattern Query Processing Module (PQPM) performs a query expansion of the pattern query to provide a relevant result by disambiguation and resolution of the entities in the pattern query. The disambiguation of the entities in the pattern query is conducted by identifying explicit and implicit similar entities and ignoring the dissimilar entities.

[0027] According to an embodiment herein, the Pattern Generation Module (PGM) comprises a document retriever to collect documents pertaining to the entities and relationships suggested by the query expansion, a Relationship Sequence Generator to create a relationship sequence with respect to each of the retrieved documents, and a Frequent Pattern Growth Module (FPGM) for extracting relevant relationship sequences.

[0028] According to an embodiment herein, the Relationship Sequence Generator builds the relationship sequences by treating each relationship as an item. Each relationship sequence comprises the relationships in the order of appearance in the data source.

[0029] According to an embodiment herein, the Frequent Pattern Growth Pattern-Mining Module (FPGM) adapts a Frequent Pattern Growth (FPG) algorithm for extracting relevant relationship sequences which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

[0030] The embodiments herein further provide a method for mining frequent patterns from a plurality of relationship sequences extracted from a big data. The method comprising, extracting a plurality of entities from the big data, storing the extracted plurality of entities in an entity store, extracting and storing one or more relationships among the plurality of entities, building an entity hierarchy by arranging the plurality of entities in a hierarchical manner, creating a relationship hierarchy by arranging the relationships in a hierarchical manner, inputting a pattern query; where the pattern query is a list of entities and the relationship of entities, processing the pattern query to find patterns in relationships and entities, retrieving relevant data sources from data using the entity hierarchy and the relationship hierarchy based on the pattern query, building relationship sequences with respect to one or more retrieved data sources and extracting frequent patterns from the relationship sequences and displaying the frequent patterns on a frequent pattern display module.

[0031] According to an embodiment herein, the big data comprises structured, unstructured and semi-structured data from heterogeneous data sources for enabling data analysis on a single view.

[0032] According to an embodiment herein, generating frequent patterns among the relationship sequences is performed using a Frequent Pattern Growth Algorithm which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

[0033] According to an embodiment herein, the method of extracting frequent patterns comprises collecting data sources pertaining to one or more entities and relationships contained in a pattern query, building a relationship sequence pertaining to each of the data source by handling each relationship as an item in an item-set that represents a relationship sequence, building a relationship sequence in an order the relationships appear in the document and identifying the frequent relationship sequences,

[0034] According, to an embodiment herein, the method of processing the pattern query comprises extracting the hierarchy of the plurality of entities, expanding the pattern query in terms of entities based on the entity hierarchy and expanding the pattern query in terms of relationships based on the relationship hierarchy.

[0035] According to an embodiment herein, expanding the pattern query in terms of entity comprises disambiguating the entities the pattern query, including synonyms and implied entities in the query expansion and perforating context resolution by including similar entities and discarding dissimilar entities.

[0036] According to an embodiment of the present invention, expanding the pattern query in terms of relationships comprises resolving relationships according to the context, including the relationship which implies context similarity, including the relationships that are implied within the syntactic and semantic similarity and discarding the semantically and syntactically dissimilar relationships.

[0037] These and the other aspects of the embodiments herein will be better appreciated and understood when considered in conjunction with the following description and the

accompanying drawings. It should be understood, however, that the following descriptions, while indicating preferred embodiments and numerous specific details thereof, are given by way of illustration and not of limitation. Many changes and modifications may be made within the scope of the embodiments herein without departing from the spirit thereof, and the embodiments herein include all such modifications.

BRIEF DESCRIPTION OF THE DRAWINGS

[0038] The other objects, features and advantages will occur to those skilled in the art from the following description of the preferred embodiment: and the accompanying drawings in which:

[0039] FIG. **1** is a block diagram illustrating a system for frequent pattern mining in relationship space, according to an embodiment of the present disclosure.

[0040] FIG. **2** illustrates a flow chart of a method for performing frequent pattern mining in relationship space, according to an embodiment of the present disclosure.

[0041] FIG. **3** is a flow diagram illustrating a method for extracting frequent patterns, according to an embodiment of the present disclosure.

[0042] FIG. **4** is a flow chart illustrating a method for processing the pattern query, according to an embodiment of the present disclosure.

[0043] Although the specific features of the present invention are shown in some drawings and not in others. This is done for convenience only as each feature may be combined with any or all of the other features in accordance with the present invention.

DETAILED DESCRIPTION OF DRAWINGS

[0044] In the following detailed description, a reference is made to the accompanying drawings that form a part hereof, and in which the specific embodiments that may be practiced is shown by way of illustration. These embodiments are described in sufficient detail to enable those skilled in the art to practice the embodiments and it is to be understood that the logical, mechanical and other changes may be made without departing from the scope of the embodiments. The following detailed description is therefore not to be taken in a limiting sense.

[0045] The various embodiments herein provide a system for mining frequent patterns in relationship space from a plurality of relationship sequences extracted from a big data. The system comprising a data repository for collecting and storing the big data an entity store for collecting and storing a plurality of entities from the big data an entity hierarchy that represents a hierarchical structure of entities, a relationship store for collecting and storing relationship instances between the plurality of entities from the big data, a relationship hierarchy that represents a hierarchical structure of relationships and a language/domain model for organizing entities and relationships in a hierarchical manner. The system further comprises a Pattern Query Processing Module (PQPM) far processing, a pattern query related to finding patterns in relationships and entities, a Pattern Generation Module (PGM) to generate frequent patterns from one or more relationship sequences from the data sources collected based on the pattern query and a Frequent Pattern Display Module (FPDM) to provide a visual presentation of the mined patterns. The pattern generation nodule performs frequent pattern mining by extracting relevant relationship sequences from the relationship store using the entity hierarchy and the relationship hierarchy.

[0046] The big data comprises structured, unstructured and semi-structured data from heterogeneous data sources for enabling data analysis on a single view.

[0047] The entity store is a collection of entities extracted from the big data. The entity store stores specific information that enables in distinguishing with one or more entities to retrieve one or more documents containing relevant entities corresponding to the pattern query. The entity hierarchy represents a hierarchical structure of entities resolved using Natural Language Processing (NLP) techniques with the support of the language/domain models.

[0048] The relationship store is adapted to store information related to each relationship instance for distinguishing with one or more relationship instances. The Relationship Hierarchy represents a hierarchical arrangement of relationships by resolving the relationships through at least one of a word-sense disambiguation technique and context resolution technique in conjunction with the language/domain model.

[0049] The Pattern Query Processing Module (PQPM) processes the pattern query by expanding the pattern query in terms of entities alter consulting with the entity store and the hierarchy of the entity. The pattern query is a list comprising entities and relationships of the entities.

[0050] The Pattern Query Processing Module (PQPM) performs a context resolution of the pattern query to provide a relevant result by disambiguation of the entities in the pattern query. The disambiguation of the entities in the pattern query is conducted by considering synonyms and implied entities obtained during expansion of pattern query where similar entities are included and dissimilar entities are excluded.

[0051] The Pattern Generation Module (PGM) comprises a document retriever to collect documents pertaining to the entities and relationships contained in the pattern query. A Relationship Sequence Generator to create a relationship sequence with respect to each of the retrieved documents. A Frequent Pattern Growth Module (FPGM) for extracting relevant relationship sequences.

[0052] The Relationship Sequence Generator builds the relationship sequences by treating each relationship as an item. Each relationship sequence comprises the relationships in the order of appearance in the document.

[0053] The Frequent Pattern Growth Module (FPGM) adapts a Frequent Pattern Growth (FPG) algorithm for extracting relevant relationship sequences which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

[0054] The method for mining frequent patterns from a plurality of relationship sequences extracted from a big data. The method comprising, extracting a plurality of entities from the big data. An entity refers to concepts comprising language unit having an independent meaning. The plurality of entities extracted from the big data is stored in an entity store and the extracted entities are arranged in a hierarchical manner. Similarly one or more relationships among the plurality of entities are extracted and stored and a relationship hierarchy is created by arranging the relationships in a hierarchical manner. Further a pattern query is inputted to a pattern query recognition module which processes the pattern query to find patterns in relationships and entities, retrieve relevant data sources from data using the entity hierarchy and the relationship hierarchy based on the pattern query, build relationship

sequences with respect to one or more retrieved data sources, extract frequent patterns from the relationship sequences and display the frequent patterns on a frequent pattern display module. The pattern query is a list of entities and the relationship of entities. Here generating frequent patterns among the relationship sequences is performed using a Frequent Pattern Growth Algorithm which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

[0055] The method of extracting frequent patterns comprises collecting data sources pertaining to one or more entities and relationships contained in a pattern query. Then relationship sequence pertaining to each of the data source is built by handling each relationship as an item in an item-set that represents a relationship sequence. Further relationship sequence is built in an order the relationships appear in the document and finally the frequent relationship sequences are identified.

[0056] Similarly the method of processing the pattern query comprises extracting the hierarchy of the plurality of entities expanding the pattern query in terms of entities based on the entity hierarchy and expanding the pattern query in terms of relationships based on the relationship hierarchy.

[0057] Here the pattern query in terms of entity comprises disambiguating the entities the pattern query, including synonyms and implied entities in the query expansion and performing context resolution by including similar entities and discarding dissimilar entities. Similarly, expanding the pattern query in terms of relationships comprises resolving relationships according to the context, including the relationship which implies context similarity, including the relationships that are implied within the syntactic similarity and discarding the contextually and syntactically dissimilar relationships.

[0058] FIG. **1** is a block diagram illustrating a system for frequent pattern mining in relationship space, according to an embodiment of the present disclosure. The system comprises a data repository **101**, a Language/Domain Models **102**, an entity store **103**, an entity hierarchy **104**, a relationship store **105**, a relationship hierarchy **106**, aquery interface **107**, a Pattern Query Processing Module (PQRM) **108**, a Pattern Generation Module (PGM) **109** and a Frequent Pattern Display Module (FPDM) **110**.

[0059] The data repository **101** is adopted for collecting and storing big data. The big data is a collection of all forms of data comprising structured, semi-structured and unstructured data from heterogeneous sources and a language/domain model **102** to resolve and organize entities and relationships in a hierarchy. The language/domain model **102** is used to disambiguate sense in an unstructured data. The language/domain model **102** also disambiguates sense in the structured and semi-structured data contexts from data repository **101**.

[0060] The entity store **103** is a collection of entities extracted from the data repository **101**. The entity store **103** also stores certain specific information relating to entities that helps in distinguishing other entities. The entity store **103** is used only to retrieve the documents containing the relevant entities corresponding to a pattern query **108**. The entity hierarchy **104** is built using the Language/Domain Model **102**. The entity hierarchy is a hierarchical structure of entities that is built using Natural Language Processing (NLP) techniques with the support of the Language/Domain Model **102**. The the Language/Domain Model **102** is used to resolve and organize entities and relationships in a hierarchy. The Language/Domain Model **102** is especially used to disambiguate sense in an unstructured, it is useful to disambiguate sense in the structured and semi-structured data contexts also. After generation of the entity hierarchy, the entity hierarchy is made available to a pattern query Processing Module (PQRM) **108**.

[0061] The relationship store **105** includes a collection of relationship instances that also stores certain information specific to relationship instances. The relationship hierarchy **106** is a hierarchical arrangement of relationships that are contextually resolved by word-sense disambiguation with the help of the Language/Domain Model **102**. The relationship store **105** and the relationship hierarchy **106** functions in conjunction with the Pattern Query Processing Module (PQRM) **108**.

[0062] The Pattern Query Processing Module (PQPM) **108** receives a pattern query inputted through a query interface **107** and performs processing as per the required information. The pattern query comprises a list of entities and relationships. The PQPM **108** consults the entity store **103** and the entity hierarchy and expands the pattern query in terms of entities. This entity expansion process involves disambiguating the entities in the pattern query, including the synonyms and implied entities in query expansion, making a context resolution to include the similar and exclude the dissimilar entities.

[0063] The Pattern Generation Module (PGM) **109** comprises a Document Retriever **109***a*, a Relationship Sequence Generator **109***b* and a Frequent Pattern Growth Pattern Mining Module (FPGMM) **109***c*. The document retriever **109***a* collects all documents pertaining to the entities/relationships contained in the pattern query. The Relationship Sequence Generator **109***b* generates a relationship sequence with respect to each of document or data by treating each relationship as an item. The Relationship Sequence Generator **109***b* builds a relationship sequence in the order of appearance in the document. The Frequent Pattern Growth Pattern-mining module (FPGMM) module uses a Frequent Pattern Growth algorithm (FPG) for processing the pattern query. The FPG algorithm treats the relationship sequences like item-sets and extracts the most frequent item-sets/relationship sequences. The Frequent Pattern Display Module (FPDM) **110** provides for in visualizing the most frequent patterns extracted from relationship sequences in conjunction with the entity.

[0064] FIG. **2** illustrates a flow chart of a method for performing frequent pattern mining in relationship space, according to an embodiment of the present disclosure. The method comprises frequent pattern mining in relationship space. In particular, the method comprises processing of big data for recognizing plurality of entities. The plurality of entities are then extracted and stored in an entity store. The entity store, stores meaningful entities extracted out of big data irrespective of the form from which the entity originates (**201**). Entities are objects that make independent sense. Entities are a named and unnamed object which includes names of living and non living things, concepts, theories or simply the language units that make independent sense. Entities is any one of named entities such as names of places, people etc., or concepts that is represented by one or more terms (example, "Purchase power', 'Purchase' as noun and 'Purchase' as verb is three different concepts). In brief, the entity refers to named entities and concepts (language unit with independent meaning). An entity hierarchy is then built by arranging, the plurality of entities in a hierarchical manner (**202**). Further a set of relationships among a plurality of entities is extracted and

stored in a relationship store (**203**), and a relationship hierarchy is created by arranging the relationships in a hierarchical manner (**204**).

[0065] The method involves the use of the entity hierarchy arid the relationship hierarchy during response to the pattern query. In case of a pattern query, the pattern query is inputted to a Pattern Query Processing Module (PQPM) for finding frequent patterns related to entities and relationships in the query (**205**).

[0066] The document collector collects the documents that are relevant to the pattern query (**206**). Based on the contents of the pattern query, the Relationship Sequence Generator generates a relationship sequence for each of the retrieved document (**207**). The PGM adopts a Frequent Pattern Growth Module (FPGM) for identifying the frequent patterns among the relationship sequences (**208**). Finally, the identified patterns are displayed on a Frequent Pattern Display Module (FPDM) (**209**).

[0067] FIG. **3** is a flow diagram illustrating a method for extracting frequent patterns, according to an embodiment of the present disclosure. The method comprises receiving a pattern query in a pattern query Processing Module (PQPM). The PQPM processes the pattern query and communicates with a Pattern Generation Module (PGM). The PGM comprises three subunits as Document Retriever, a Relationship Sequence Generator and a Frequent Pattern Growth Module (FPGM). Once the PGM receives the command from the PQPM, the document retriever starts collecting, one or more documents (**301**). The one or more documents are related to the one or more entities and relationships contained in the pattern query. Once the related documents are collected, the Relationship Sequence Generator builds a relationship sequence in an order in which the relationships appear in the document (**302**). The relationship sequences that appear like "item-sets" enable frequent item set mining. The item-sets comprise relationship sequences in an orderly manner for easy processing. Once an ordered item-set is built, the Frequent Pattern Growth Module (FPGM) mines for the required pattern as desired by the pattern query (**303**). The result of the frequent relationships sequences are then displayed by Frequent Pattern Display Module (FPDM).

[0068] FIG. **4** is a flow chart illustrating a method for processing the pattern query, according to an embodiment of the present disclosure. The pattern query is raised by a user which is inputted to a Pattern Query Processing Module (PQPM). Depending on the content of the pattern query, the PQPM expands the pattern query in terms of entities on referring the entity list and the entity hierarchy (**401**). Expanding the pattern query in terms of entity includes steps of disambiguating the entities in the pattern query, including synonyms and implied entities in the query expansion and performing context resolution by including similar entities and discarding dissimilar entities. The PQPM then expands the pattern query in terms of relationships based on the relationship hierarchy (**402**). Here expanding the pattern query in terms of relationship includes resolving relationships according to the context, including the relationships which implies context similarity, including the relationships that are implied within the syntactic similarity and discarding the contextually and syntactically dissimilar relationships.

[0069] The embodiments of the present invention disclose an approach that looks for patterns in the relationship space. The embodiments of the present disclosure, provides a robust approach to find patterns and ensures context resolution effectively. The entities and relationships among the entities assist in understanding the big data. All the entities and relationships are derived and collected. This collection of entities and relationships serves as input to all intelligent processing of data. Data mining and data analysis applications, forecasting, predictive analytics applications and machine learning applications make use of the patterns to learn further insights. The embodiments herein enable an enterprise that intends to facilitate processing of big data and build applications on top. The embodiment herein also allows building of domain specific, niche applications that harness big data. The embodiments herein provides immense benefit to following sectors but is not limited to retail, health and pharmaceutical services, banking and insurance.

[0070] The foregoing description of the specific embodiments will so fully reveal the general nature of the embodiments herein that others can, by applying current knowledge, readily modify and/or adapt for various applications such specific embodiments without departing from the generic concept, and, therefore, such adaptations and modifications should and are intended to he comprehended within the meaning and range of equivalents of the disclosed embodiments. It is to be understood that the phraseology or terminology employed herein is for the purpose of description and not of limitation. Therefore, while the embodiments herein have been described in terms of preferred embodiments, those skilled in the art will recognize that the embodiments herein can be practiced with modification.

What is claimed is:

1. A system for mining frequent patterns in relationship space from a plurality of relationship sequences extracted from a big data, the system comprising:

a data repository for collecting and storing the big data;

an entity store for collecting and storing a plurality of entities from the big data;

an entity hierarchy for representing a hierarchical structure of entities;

a relationship store for collecting and storing relationship instances between the plurality of entities from the big data;

a relationship hierarchy for representing a hierarchical structure of relationships;

a language/domain model for organizing entities and relationships in a hierarchical manner;

a pattern query Processing Module (PQPM) for expanding a pattern query related to finding patterns in relationships and entities;

a Pattern Generation Module (PGM) to generate frequent patterns from one or more relationship sequences from the data sources collected based on the pattern query; and

a Frequent Pattern Display Module (FPDM) to provide a visual presentation of the mined patterns;

where the pattern generation module performs frequent pattern mining by extracting relevant relationship sequences from the relationship store using the entity hierarchy and the relationship hierarchy.

2. The system according to claim **1**, wherein the big data comprises structured, unstructured and semi-structured data from heterogeneous data sources.

3. The system according to claim **1**, wherein the entity store is a collection of entities extracted from the big data, wherein the entity store stores information specific to each entity.

4. The system according to claim **1**, wherein the entity hierarchy is a hierarchical structure of entities resolved using Natural Language Processing (NLP) techniques with a support of the language/domain model.

5. The system according to claim **1**, wherein the relationship store is adapted to store information related to each relationship instance.

6. The system according to claim **1**, wherein the Relationship Hierarchy provides a hierarchical arrangement of relationships by resolving the relationships through at least one of a word-sense disambiguation technique, syntactic and semantic similarity and context resolution technique in conjunction with the language/domain model.

7. The system according to claim **1**, wherein the pattern query Processing Module (PQPM) processes the pattern query by expanding, the pattern query in terms of entities after consulting the entity hierarchy, wherein the pattern query is a list comprising entities and relationships of the entities.

8. The system according to claim **1**, wherein the pattern query Processing Module (PQPM) performs a expansion of the pattern query to provide a relevant result by disambiguation of the entities in the pattern query, where the disambiguation of the entities in the pattern query is conducted by identifying explicit and implicit similar entities and ignoring the dissimilar entities

9. The system according to claim **1**, the Pattern Generation Module (PGM) comprises:

a document retriever to collect documents pertaining to the entities and relationships suggested by the query expansion;

a Relationship Sequence Generator to create a relationship sequence with respect to each of the retrieved documents;

a Frequent Pattern Growth Module (FPGM) for extracting relevant relationship sequences.

10. The system according to claim **9**, wherein the Relationship Sequence Generator builds the relationship sequences by treating each relationship as an item, where each relationship sequence comprises the relationships in the order of appearance in the document.

11. The system according to claim **9**, wherein the Frequent Pattern Growth Pattern-Mining Module (FPGM) adapts a Frequent Pattern Growth (FPG) algorithm for extracting relevant relationship sequences which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

12. A method for mining frequent patterns from a plurality of relationship sequences extracted from a big data, the method comprising:

extracting a plurality of entities from the big data, where an entity refers to concepts comprising language unit having an independent meaning;

storing the extracted plurality of entities in an entity store;

extracting and storing one or more relationships among the plurality of entities;

building an entity hierarchy by arranging the plurality of entities in a hierarchical manner;

creating a relationship hierarchy by arranging the relationships in a hierarchical manner;

inputting a pattern query, where the pattern query is a list of entities and the relationship of entities;

expanding the pattern query to include most relevant entities and relationships and ignore irrelevant patterns and relationships;

retrieving relevant data sources from data using the pattern query;

building relationship sequences with respect to one or more retrieved data sources;

extracting frequent patterns from the relationship sequences; and

displaying the frequent patterns on a frequent pattern display module.

13. The method according to claim **12**, wherein the big data comprises structured, unstructured and semi-structured data from heterogeneous data sources for enabling data analysis on a single view.

14. The method according to claim **12**, wherein generating frequent patterns among the relationship sequences is performed using a Frequent Pattern Growth Algorithm which considers the relationship sequences as item-sets and extracts the most frequent item-sets.

15. The method according to claim **12**, wherein the method of extracting frequent patterns comprises:

collecting data sources pertaining to one or more entities and relationships contained in a pattern query;

building a relationship sequence pertaining to each of the data source by handling each relationship as an item in an item-set that represents a relationship sequence;

building a relationship sequence in an order the relationships appear in the document; and

identifying the frequent relationship sequences.

16. The method according to claim **12**, wherein the method of processing the pattern query comprises:

extracting the hierarchy of the plurality of entities;

expanding the pattern query in terms of entities based on the entity hierarchy; and

expanding the pattern query m terms of relationships based on the relationship hierarchy.

17. The method according to claim **16**, expanding the pattern query in terms of entity comprises:

disambiguating the entities the pattern query;

including synonyms and implied entities in the query expansion; and

discarding dissimilar entities.

18. The method according to claim **16**, expanding the pattern query in terms of relationships comprises:

resolving relationships according, to the context;

including the relationships which implies context similarity;

including the relationships that are implied within the syntactic similarity; and

discarding the contextually and syntactically dissimilar relationships.

* * * * *