(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 April 2018 (19.04.2018)

W!P O | PCT

(10) International Publication Number
WO 2018/069836 Al

(51) International Patent Classification:
G06K 9/62 *(2006.01)*

(21) International Application Number:
PCT/IB20 17/056262

(22) International Filing Date:
10 October 2017 (10.10.2017)

(25) Filing Language: English

(26) Publication Langiuage: English

(30) Priority Data:
62/406,103 10 October 2016 (10.10.2016) US

(71) Applicant: **KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY** [SA/SA]; 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA).

(72) Inventor: **ALABDULMOHSIN, Ibrahim Mansour;** 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (SA).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.1 7(H))*
— *of inventorship (Rule 4.1 7(iv))*

Published:
— *with international search report (Art. 21(3))*

(54) Title: QUASI-CLIQUE PROTOTYPE-BASED HYBRID CLUSTERING

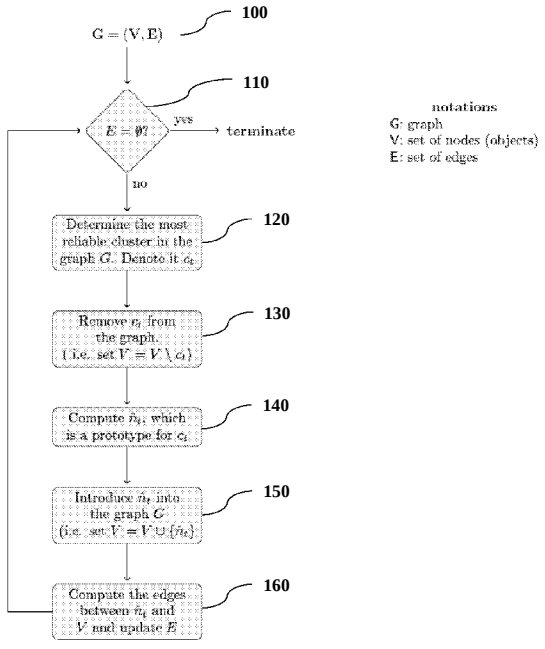


FIG. 1

(57) Abstract: Embodiments of the present disclosure describe a clustering scheme and system for partitioning a collection of objects, such as documents or images, using graph edges, identification of reliable cluster groups, and replacement of reliable cluster groups with prototypes to reconstruct a graph. The process is iterative and continues until the set of edges is reduced to a predetermined value.

WO 2018/069836 A1

QUASI-CLIQUE PROTOTYPE-BASED HYBRID CLUSTERING

BACKGROUND

[0001] Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a key task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

[0002] Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and error. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

[0003] Cluster analysis is commonly categorized under the broad field of unsupervised learning techniques because a "correct" solution is often undefined or unavailable. As a result, many clustering algorithms have been developed to achieve heuristic goals. In the popular k-means algorithm, for instance, clusters are assigned such that the sum of differences (or distances) between each object and its assigned cluster is small. This algorithm assumes that the number of clusters, k, is known beforehand.

[0004] Another closely related algorithm is the Gaussian mixture model, which is similar in spirit to the k-means algorithm and is often solved using the Expectation-Maximization (EM) procedure. In the Gaussian mixture model, the goal is to maximize the likelihood of the clustering assignments. It assumes a definite probability distribution on the data objects.

[0005] Moreover, a third family of clustering algorithms is connectivity based clustering, which are designed to optimize the "distance" between clusters. This latter family

of methods, such as "single-linkage" clustering and "complete linkage" clustering, differ from each other in how they measure the distance between clusters.

[0006] In some clustering applications, however, two distinctive characteristics are present. First, there exists a single correct clustering assignment. Second, the homogeneity and separation assumptions are satisfied, whereby each of a pair of objects in the same (correct) cluster tends to be similar to each other, whereas objects across different clusters tend to be dissimilar. Examples of applications that satisfy these characteristics include news aggregation and social network event detection.

[0007] In a news aggregation application, a large collection of articles are clustered into news stories, where each article in the same cluster describes the same news story. In this application, either assigning articles that describe the same story into two different clusters or assigning articles that describe different stories into the same cluster would be erroneous.

[0008] In a social network event detection application, a collection of short texts, such as tweets in Twitter or public statuses in Facebook, are clustered into groups according to their contents in such a way that short texts describing the same event are grouped together in the same cluster. The size and age of the cluster are, in turn, used to determine the importance and urgency of the respective event.

[0009] Similarly, in an image search and retrieval application, a large collection of images can be clustered according to content similarity. This allows search engines to avoid duplicate results and to improve user experience.

[0010] In a traditional clustering setting, there is no objectively "correct' clustering algorithm. Hence, most traditional clustering algorithms are designed to optimize some heuristic scores instead and the most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. In the popular k-means algorithm, for example, the sum of differences (or distances) between objects and their clusters is minimized. In probabilistic methods, such as the Gaussian mixture models, the probabilistic likelihood is maximized via the iterative Expectation-Maximization (EM) procedure. In connectivity-based clustering, which includes algorithms such as the single-linkage method and the complete-linkage method, the distance between clusters is minimized. Other heuristic scores have been proposed as well, including the Davies-Bouldin index and the Silhouette coefficient.

[0011] However, in some clustering settings, such as news aggregation, event detection in social networks, and image clustering, a different set of requirements is imposed. In these settings, there does exist a well-defined notion of "clustering accuracy". In news

aggregation, for example, a large collection of articles is clustered into news stories, where each article in the same cluster is supposed to describe the same news story. In this application, either assigning articles that describe the same story into two different clusters or assigning articles that describe different stories into the same cluster would be erroneous. Therefore, the goal here is not to optimize some artificial score. Rather, the primary objective is to improve the accuracy of the clustering results.

[0012] Therefore, the goal of these applications is to maximize the accuracy of the clustering method, which is a non-heuristic well-defined performance indicator. Generally, clustering algorithms that have been designed to optimize some alternative heuristic scores, such as distance or likelihood, do not perform well for such applications because they are not designed from the outset to maximize the clustering accuracy.

SUMMARY

[0013] In general, embodiments of the present disclosure describe methods for clustering. Accordingly, embodiments of the present disclosure describe clustering methods yielding correctness (there exists a single correct clustering assignment), homogeneity (any two objects in the same correct cluster tend, on average, to be similar to each other), and separation (objects in different clusters tend, on average, to be dissimilar to each other).

[0014] Embodiments of the present disclosure provide a highly accurate clustering assignment even when data is contaminated with noise or outliers. In one embodiment of the present disclosure, a new cluster is identified by partitioning the similarity graph into connected components and by identifying the largest clique in each connected component. Then, the largest cluster is collapsed into a single node and the graph is reconstructed afterward, with some similarities recomputed. The entire process is repeated until all nodes are isolated in the graph. The primary feature of this method is that it is clique-driven. It is a greedy algorithm, which has an important advantage over the popular k-means algorithm in that it does not require that the number of clusters be specified in advance. It can be used to cluster any objects, such as images, texts, and sequences, provided that a similarity score can be defined on those objects. It is also quite different from both "agglomerative clustering" and "hierarchical clustering". In an agglomerative clustering, every node forms its own cluster, and clusters are merged to form new clusters in a bottom-up fashion. In hierarchical clustering, a top-down approach is taken, where all nodes initially belong to the same cluster, which cluster is divided into smaller clusters afterward.

[0015] One embodiment of the present disclosure, by contrast, uses graph edges as evidence of reliability. It identifies the largest clique as the most reliable cluster. Then, this cluster is collapsed into a single object (prototype) and the graph is reconstructed. Iteratively, the method identifies the largest clique in the new graph to be the new cluster, which can include the previous cluster as well. Hence, it is a bottom-up approach, which is similar to agglomerative clustering, but it is clique-centric to provide more reliable results.

[0016] The details of one or more examples are set forth in the description below. Other features, objects, and advantages will be apparent from the description and from the claims.

BRIEF DESCRIPTION OF DRAWINGS

[0017] This written disclosure describes illustrative embodiments that are non-limiting and non-exhaustive. In the drawings, which are not necessarily drawn to scale, like numerals describe substantially similar components throughout the several views. Like numerals having different letter suffixes represent different instances of substantially similar components. The drawings illustrate generally, by way of example, but not by way of limitation, various embodiments discussed in the present document.

[0018] Reference is made to illustrative embodiments that are depicted in the figures, in which:

[0019] **FIG. 1** illustrates a process flow diagram for a clustering scheme in accordance with one or more embodiment of the present invention.

[0020] **FIGS. 2A-D** illustrates a conceptual graph diagram of a clustering process using the scheme of FIG. 1.

DETAILED DESCRIPTION

[0021] The invention of the present disclosure relates to cluster analysis and clustering methods. One embodiment of the present disclosure describes a hybrid method between the general family of "quasi-clique" clustering algorithms, such as the Highly Connected Subgraph (HCS) algorithm, and "prototype-based" clustering algorithms, of which the k-means algorithm is most well-known.

[0022] **FIG. 1** illustrates a high-level flow diagram of a clustering method in accordance with the present invention. Initially, object-to-object similarity is encoded in an undirected, unweighted graph G = (V, E), where V is the set of objects (or nodes) and E is the set of unweighted edges, as indicated by numeral 100. At step 110, the set of edges, E, is

evaluated and the process is terminated if E = 0. If E≠ 0, then the most reliable cluster in graph, G, is identified at step 120. Ideally, the most reliable cluster is the largest clique in the graph. At step 130, the identified cluster is removed from the graph. This signifies that all nodes in the group are permanently assigned to the same cluster, although the cluster itself is subject to change in future iterations. At step 140, a new node is computed and introduced as a prototype (i.e. representative) to the cluster (denoted $\hat{n}_t$ in FIG. 1) at step 150. This prototype can, for instance, be the mean of the cluster or its medoid. The edges between the new prototype node $\hat{n}_t$ and all remaining nodes in the graph are recomputed at step 160. The process is repeated until all nodes in the graph are completely isolated.

[0023] The method describes a process for partitioning objects into accurate clusters. It is assumed that the process can measure the similarity score between any two arbitrary objects. Given two objects $n_;$ and ***rij*****,** $5(\eta_i,$ ***rij)*** denotes their similarity score. It is also assumed that there exists a known threshold κ > 0, such that if for two objects $n_;$ and ***rij*****,** we have ***S(ni, rij) »*** κ**,** then the two objects are more likely to have come from the same cluster. Conversely $S(n_i,$ ***rij) «*** κ implies that the two objects n_i and ***rij*** are unlikely to have come from the same cluster. It is not assumed that the similarity score is perfect. The process is thus capable of handling noise and outliers. This threshold κ can be estimated from data. It is further assumed that for any collection of objects $\{n_1,$ ***n2,*** …, ¾}, there exists a procedure for finding a prototype to the collection of objects. If the objects are numeric, for instance, then the prototype can be the mean or median. Otherwise, the prototype can be a medoid. In a news aggregation example, news articles are represented using the bag-of-words representation, and the mean (average) of a collection is used as its prototype.

[0024] As described above, the first step of the clustering process is computing the pairwise similarity between any two objects. Let G = (V, E) be a graph, where V is the set of all objects of interest. If $S(n_i,$ ***rij)*** ≥ κ for ***ni, rij £ V,*** then an edge is added between nodes n_i and ***rij*** in the graph G. Otherwise, no edge connects the two nodes n_i and ***rij.***

[0025] In order to obtain accurate results, the process proceeds in an iterative greedy manner. At each round of the process, the most reliable cluster is identified. Ideally, the most reliable cluster is the largest clique in the graph. However, determining the largest clique is a computationally difficult problem. The most straightforward approach is to use an approximation algorithm, such as the algorithms of Brunato et al. (2007) and Feige et al. (2001). Another approach is to increase the threshold κ until the size of the largest connected

component in the graph is below a specified maximum number. The second approach is used in the examples herein.

[0026] Once a collection of objects $c_t = \{v_1, v_2, \ldots, v_m\}$ is identified as the next most reliable cluster, all of its nodes are removed from the graph G and a new prototype node is inserted to replace the entire cluster. More precisely, the new graph $G' = (V', E')$ is defined by:

$$V' = V \setminus c_t \cup \{n_t\}$$

$$E' = \{(u, v) \in E : u \in V' \wedge v \in V'\}$$

[0027] The rationale behind this approach is twofold. First, by identifying the most reliable cluster, the clustering accuracy is maximized. Second, by substituting a prototype for a cluster, the similarity scores between the prototype and the rest of the graph become a more reliable indicator of similarity since a prototype is, by construction, an aggregation of multiple nodes in the cluster. In our embodiment for news aggregation, for instance, a single news story can be conveyed in various ways. However, the differences between news articles that describe the same news story is generally attributed to superficial factors, such as writing styles. By forming a prototype, such superficial differences are removed and the resulting prototype becomes a more representative object to the original news story.

[0028] Once the cluster is removed and a single prototype is inserted in its place, the process proceeds by examining the similarity between the prototype and all of the remaining nodes (i.e. objects) in the graph. Let $\hat{n}_t$ denote the new prototype node and let n_i be some other object in the graph G'. If $S(n_t, n_i) \geq \kappa$, an edge is added between $\hat{n}_t$ and n_i into E'. Otherwise, no edge connects the two nodes directly in the graph. Once all of the edges are determined, the process is repeated in the new graph, where the next most reliable cluster is identified. Because the next most reliable cluster can contain prototype nodes, clusters that have been found in previous rounds may be merged together in future rounds.

[0029] A process for breaking ties is also defined. If two clusters are found to be of the same reliability (e.g. clique size), then the priority of a cluster is determined by the sum of priorities of its node. For example, a priority of a node may be equal to the size of the cluster it represents if it is a prototype, or is one otherwise.

[0030] **FIG. 2** depicts the clustering process as applied to a small similarity graph. In Fig. 2a, the original similarity graph is shown. The nodes of the graph are the objects to be clustered, where the shading of a node represents the cluster it belongs to. Due to imperfections in the similarity score, imperfections in the threshold κ, or due to noise, the

nodes are not perfectly separated in the graph according to their true clustering assignment. Hence, a reliable process for identifying the clusters is desired.

[0031] The first step of the process is to identify the most reliable cluster, which in this example is the largest clique in the graph. This cluster is contained within region 200 in **Fig. 2a.** Next, all of the four nodes are removed and the entire cluster 200 is replaced with a single prototype 210, which is marked with a larger size in **Fig. 2b. Fig. 2b** shows the new graph after all edges have been recomputed.

[0032] Next, the process identifies the most reliable cluster in the new graph. This cluster is identified as 220 **Fig. 2b.** Again, the cluster is removed and is replaced with a single prototype node 230 as shown in **Fig. 2c. Fig. 2c** shows the final graph when edges are recomputed. Here, it is noted that after aggregating the upper cluster with its prototype 230, its original apparent similarity to node 240 is no longer present because aggregation improves the reliability of similarity functions.

[0033] Next, the process looks for the most reliable cluster in the new graph. A tie exists between two clusters comprised of two nodes each. The first cluster connects node 240 with the top node 250. The second cluster connects node 250 with the prototype node 210. Because ties are broken according to the priority of the nodes, the cluster that contains the prototype node is selected. Once it is replaced with a single prototype 260 and edges are recomputed, all of the remaining nodes (230, 240, 260) in the graph are isolated as shown in **Fig. 2d.** Hence, the algorithm terminates.

[0034] One variant that can be used alleviates the computational burden of finding the maximum cliques in the graph. Instead of operating on the entire graph, the graph can be partitioned, first, into connected components. Then, the process is applied to each connected component separately. Once the process terminates, isolated nodes remain for each connected components. These nodes are, next, merged into a bigger graph, with edges being recomputed. The process is applied again to this new entire graph. The purpose of this approach is to speed up the algorithm and to reduce memory consumption. In one example embodiment, the process was implemented on a single quad-core workstation, and handled collections of over 50,000 objects in less than 15 minutes.

[0035] Another closely related variant is local clustering, which is most suitable for extremely large graphs that cannot be processed by a single workstation. Instead of partitioning a graph into connected components, the graph can be partitioned into multiple subgraphs. This can be embodied by, first, sampling at random from nodes according to their degrees, and, second, by traversing the neighborhood of the sampled nodes in a depth-first

(DFS) manner. The clustering process can, then, be implemented on those separate subgraphs by separate machines running in parallel, and the resulting prototypes can be merged together into a bigger graph. The process is repeated in the bigger graph afterward.

[0036] Other variants can be made for the choice of the prototype nodes as well. If objects are numeric, then the prototype node may correspond to the mean or average of the objects of the cluster. Otherwise, a medoid can be used instead. Two examples of medoids

1. Maximizing the minimum similarity to other nodes in the cluster:

$$prototype(\{n_1, n_2, \ldots, n_m\}) = \arg \max_{1 \le i \le m} \min_{1 \le j \le m} S(n_i, n_j)$$

2. Maximizing the average similarity to all other nodes in the cluster:

$$prototype(\{n_1, n_2, \ldots, n_m\}) = \arg \max_{1 \le i \le m} \sum_{1 \le j \le m} S(n_i, n_j)$$

[0037] Many algorithms have been proposed for object clustering, including quasi-qlick clustering methods. The most prominent example is the Highly Connected Subgraph (HCS) algorithm (Hartuv and Shamir, 2000). In HCS, clustering proceeds in a top-bottom fashion. At each round, a cluster is identified by the size of its cut from the rest of the graph. Variants of this algorithm have been proposed, such as by using the normalized cut or conductance (Schaeffer, 2007). The embodiments of the present invention, by contrast, operate in a bottom-up fashion, using cliques instead of minimum cuts, and using prototypes with graph transformations to update the clustering assignment at each round.

[0038] Another class of closely-related clustering algorithms are spectral methods (Schaeffer,2007). These methods form variants of the Laplacian matrix of the graph and infer clustering assignments via eigenvalue decomposition. In contrast, embodiments of the present provide a greedy iterative process that is designed to improve the clustering accuracy by combining the merits of both quasi-clique methods with prototype-based methods.

[0039] Furthermore, agglomerative algorithms, such as the pairwise nearest neighbor method (Franti et al., 2003) or the Wards method (Ward Jr, 1963), are similar to the present method in that they all operate in a bottom-up fashion. In traditional agglomerative methods, a distortion function, such as the sum of distances, is used to select the two clusters to merge at each iteration. In contrast, distortion functions are not used in some embodiments of the present invention. Instead, multiple nodes can be merged together at a single round by identifying the largest clique in the graph, or some approximation to it. In addition, agglomerative methods do not form prototypes to improve the clustering accuracy, which is a key process of embodiments of the present invention.

WHAT IS CLAIMED IS:

1. A method for clustering similar objects together, the method comprising:

evaluating pairwise similarity between a set of objects to define a set of edges between multiple nodes of a graph;

identifying a reliable cluster of nodes within the graph;

defining a first prototype node and replacing the reliable cluster of nodes with the prototype node;

evaluating pairwise similarity between the first prototype node and the remaining nodes of the graph to define a new set of edges between multiple nodes of the graph;

identifying the next reliable cluster of nodes within the graph;

defining a second prototype node and replacing the next reliable cluster of nodes with second prototype node;

evaluating pairwise similarity between the second node and the remaining nodes of the graph to define a new set of edges between multiple nodes of the graph; and

repeating said identifying the next reliable cluster of nodes, defining subsequent prototype nodes, and evaluating pairwise similarity between the nodes until no similarity exists between the nodes of the graph.

2. The method of clustering similar objects together of claim 1, wherein said identifying a reliable cluster of nodes includes identifying the most reliable cluster of nodes.

3. The method of clustering similar objects together of claim 2, wherein the most reliable cluster of nodes is the largest clique of the graph.

4. The method of clustering similar objects together of claim 1, wherein said identifying a reliable cluster of nodes includes an approximation algorithm.

5. The method of clustering similar objects together of claim 1, wherein said identifying a reliable cluster of nodes includes increasing a threshold κ until a size of a largest connected component in the graph is below a specified maximum number.

6. The method of clustering similar objects together of claim 1, wherein said defining a prototype node includes using a mean or median of the cluster or its medoid.

7. The method of clustering similar objects of claim 1, wherein said identifying the reliable cluster of nodes within the graph includes partitioning the graph to define isolated nodes for each partition segment and then merging the nodes.

8. The method of clustering similar objects of claim 1, wherein said identifying the reliable cluster of nodes includes partitioning the graph into connected subcomponents and applying a clustering process to each of the connected subcomponents.

9. A method of clustering objects comprising:

establishing pairwise similarity between two objects in a graph, G = (V, E), where V is the set of objects of interest, with a set of edges, E, defined between nodes n_i and n_j in graph G, according to $S(n_i, n_j) \geq \kappa$ for $n_i, n_j \in V$;

identifying a reliable cluster of nodes within the graph;

defining a first prototype node and replacing the reliable cluster with the prototype node;

updating the graph, G, to $G' = (V', E')$, with

$V' = V \setminus c_t \cup \{\hat{n}_t\}$ and $E' = \{(u, v) \in E : u \in V' \wedge v \in V'\}$;

evaluating similarities between the first prototype node and the remaining nodes, with $\hat{n}_t$ denoting the first prototype node and n_i another being another nodes in the graph G' and if $S(\hat{n}_t, n_i) \geq \kappa$, an edge is added between $\hat{n}_t$ and n_i into E'; and

repeating the steps of identifying a reliable cluster, defining next prototype nodes, replacing the next prototype node for the reliable cluster, updating the graph, and evaluating similarity between prototype nodes and remaining objects of the graph.

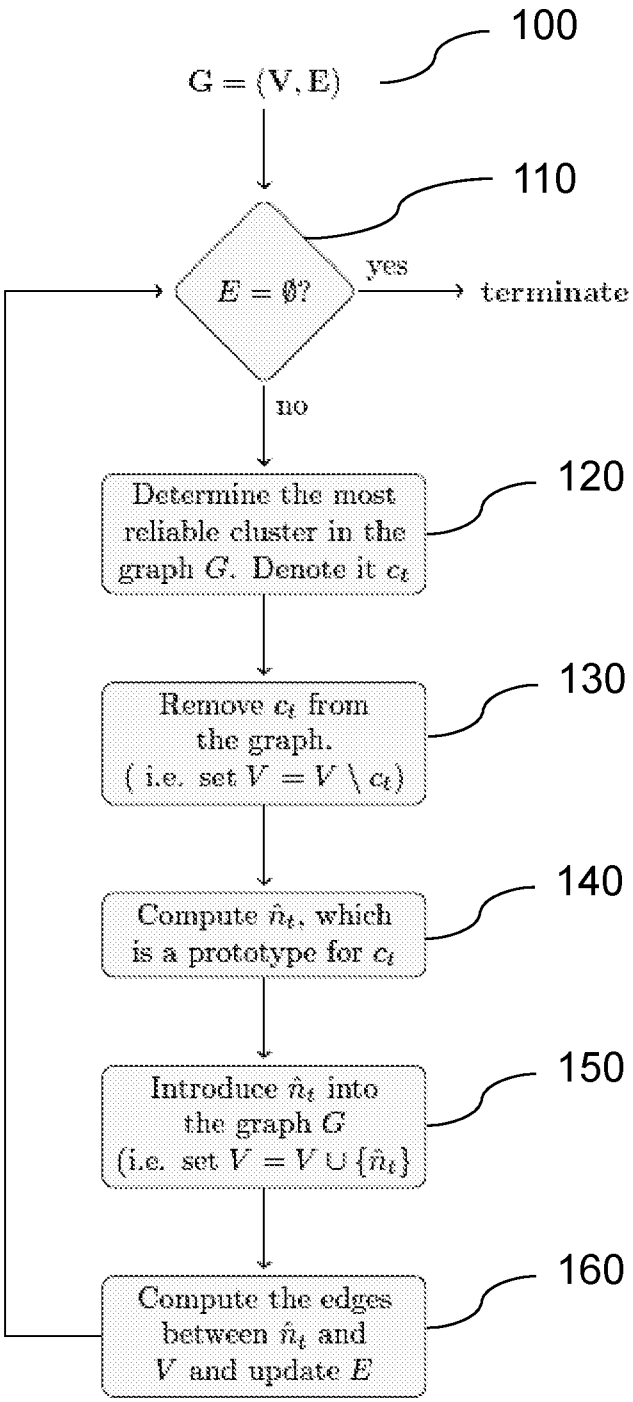
10. The method of clustering objects of claim 9, wherein said identifying a reliable cluster in the graph includes identifying the most reliable cluster of nodes.

11. The method of clustering objects of claim 9, wherein said identifying a reliable cluster of nodes includes an approximation algorithm.

12. The method of clustering objects of claim 9, wherein said identifying a reliable cluster of nodes includes increasing a threshold κ until a size of a largest connected component in the graph is below a specified maximum number.

13. The method of clustering objects of claim 9, wherein said defining a prototype node includes using a mean or median of the cluster or its medoid.

14. The method of clustering objects of claim 9, wherein said identifying the reliable cluster of nodes within the graph includes partitioning the graph to define isolated nodes for each partition segment and then merging the nodes.



notations

G: graph
V: set of nodes (objects)
E: set of edges

FIG. 1

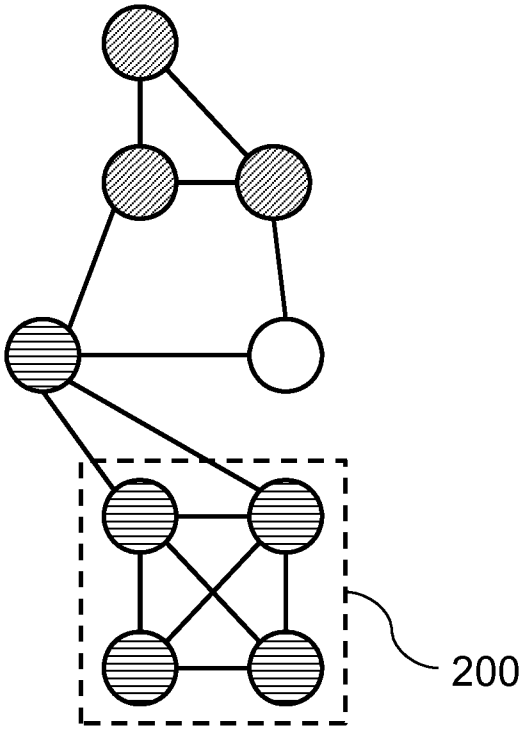


FIG. 2A

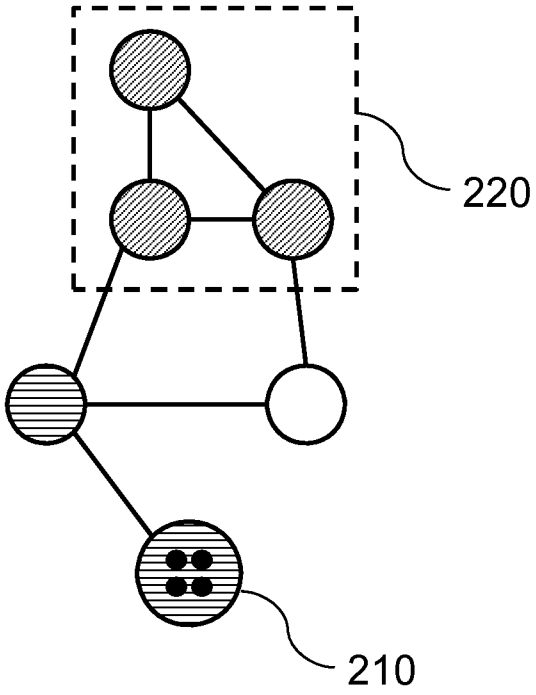


FIG. 2B

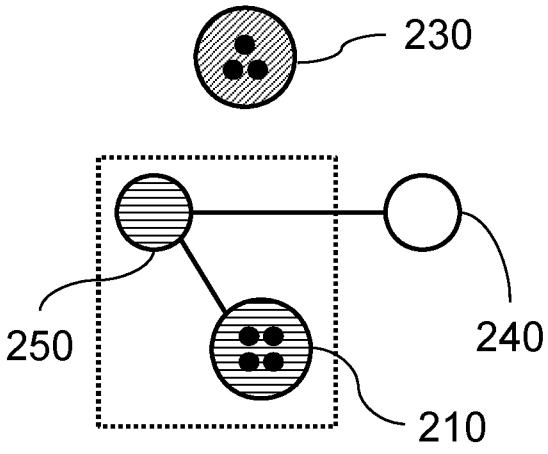


FIG. 2C

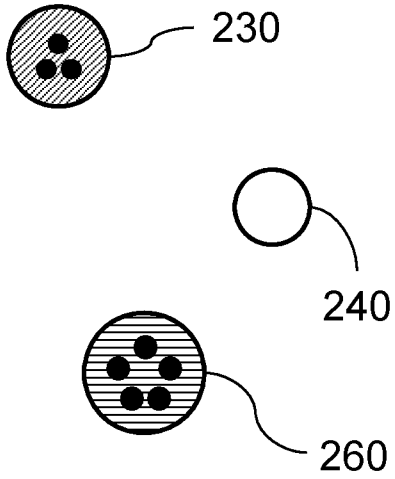


FIG. 2D

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2017/056262

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/62
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal , WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Jubin Edachery ET AL: "Graph Clusteri ng Using Distance-k Cliques" In: "Network and Parallel Computing" , 1 January 1999 (1999-01-01) , Springer International Publishing, Cham 032548, XP055439061 , ISSN: 0302-9743 ISBN: 978-3-642-01969-2 vol . 1731 , pages 98-106, D0I : 10. 1007/3-540-46648-7_10, the whole document	1-14
A	US 2010/004925 Al (AH-PINE JULIEN [FR] ET AL) 7 January 2010 (2010-01-07) the whole document	1-14
A	US 2013/294690 Al (URBACH SHL0M0 [IL] ET AL) 7 November 2013 (2013-11-07) the whole document	1-14

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" documentwhich may throw doubts on priority claim(s) orwhich is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 January 2018	19/01/2018

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Meier, Ueli

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2017/056262

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010004925 A1	07-01-2010	NONE	
US 2013294690 A1	07-11-2013	US 8509525 B1	13-08-2013
		US 2013294690 A1	07-11-2013