Mukherjee and Bandyopadhyay *J Big Data* (2017) 4:35
DOI 10.1186/s40537-017-0097-0

Journal of Big Data

RESEARCH Open Access

Learning topic description from clustering of trusted user roles and event models characterizing distributed provenance networks: a reinforcement learning approach

Sanjoy Kumar Mukherjee and Sivaji Bandyopadhyay*

*Correspondence:
sivaji.cse.ju@gmail.com
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

Abstract

This paper proposes a reinforcement learning based message transfer model for transferring news report messages through a selected path in a trusted provenance network with the objective of maximizing the reward values based on trust or importance based and network congestion or utility based cost measures. The reward values are calculated along a dynamically defined policy path connecting start topic or event node to a goal topic or event or issue nodes for incrementally defined time windows for a given network congestion situation. A hierarchy of agents of trusted roles is used to accomplish the sub-goals associated with sub-story or subtopic in the provenance structure where an agent role may assume the semantic role of the associated sub-topic. The twitted news story thread or plan of events is defined in this work from the starting topic or event node to the goal topic or event node for incrementally defined intervals of time. The graphs are clustered into subtopic and these sub-goals or sub topic nodes of a topic node at every level of granularity are associated with cluster of news reports which describe activities associated with sub-goal or sub-topic events. Such cluster of nodes may also represent drilled down sequence of sub-events describing a sub-topic or sub-goal node. The policy path in a topic or story graph model is defined by applying reinforcement learning principles on dynamically defined event models associated with evolution of topic definition observed from incrementally acquired samples of input training data spanning multiple time windows. We provide a methodology for unifying similar provenance graph models for adapting and averaging the policy path classifiers associated with individual models to produce a reduced set of unified models derived during training. A minimum set cover of classifiers is identified for the models and a clustering procedure of the models is suggested based on these classifiers. Other database clustering methods have also been suggested as alternatives for clustering these models. A collection of unified models are identified from the models identified within a cluster and the policy path classifiers associated with these models provide the story or topic descriptions destined to goal topic or event nodes characterizing these models within a cluster.

Keywords: Computational trust, Reinforcement learning, Q Learning, Policy path, Reward, Provenance, Bayesian model, Materialized views, Network congestion, Database roles

Introduction

This paper proposes a reinforcement learning based message transfer model for transferring news report messages through a selected path in a trusted provenance network with the objective of maximizing the reward values based on trust or importance based and network congestion or utility based cost measures. The important issue of determining the integrity (trustworthiness) of data provenance in data fusion and knowledge fusion activities has been described in [1]. The provenance sketch of a piece of data based on files or processes that act as input data flow to the other processes has been described in [2]. The granularity of provenance information influencing communication and processing overhead has been described in [2]. A networked provenance system with migration of information between network nodes has been described in [3, 4]. The use of concepts of provenance applied to news reports analysis has been described in [5–9]. The application of markov decision process to planning, event modelling has been described in [10, 11]. The model for a message passing network has been considered in [12]. A description of the rationale for storage and retrieval of trusted information using a probabilistic temporal database approach has been described in [12].

The application of clustering to news reports analysis comprising of events has been described in [13, 14]. The high level general topics have been discussed for the entire event, as compared to specific topics that have been discussed during specific segment of the event [15]. The theory of additive compositionality has been described in [16]. This has relevance to the computing frequency count of occurrences of intersect-subset of attributes words in the merged record as has been described in "Problem statement and solution" section. The detection of previously unspecified events using feature pivot techniques has been described in [17]. The paper on clustering of news reports [18] provide a description of aspect or topic based [19] tree model construction and merging of tree models for incrementally acquired reports for new time intervals. The SOAR system [20] which has a procedure of conflict resolution for choice of next sub-state of the current state and reinforcement learning procedure has been considered for accomplishing this resolution. Aspect based regression tree clustering [18] produces clusters of documents as leaf nodes, where every leaf node can be associated with a news topic or a news category [21] or a news issue [22, 23].

A cluster node expansion and a cluster edge expansion measure are considered in our paper for growing a cluster spanning a hierarchy of sub-topic sub-goal clusters using a derived cluster node expansion capability measure. Thus quality of identified grouping or clustering of nodes in a module group can be represented using this distance measure. The identification of sub goal states or nodes as cluster heads has been described in [24]. The calculation of weights associated with classifiers has been described in [25–30]. The computation of similarity and complement measures associated with classifiers has been described in [14, 31]. A k-modes clustering algorithm has been described in [32] for categorical data. A number of cluster analysis methods has been described in [32] for identifying the number of clusters, which have been defined based on gap statistics, and other measures based on information criteria such as AIC (Akaike Information Criteria) or BIC (Bayes Information Criteria) which can be relevant in intra model or across models clustering. The candidate models participating in the procedure for this merging have policy path classifiers which have been described as either only member of any set of

similar classifiers or are members in the minimum cover set [33] of such similar classifiers that are present in the cluster. Hierarchical topics provide an overview of topics from one corpora [34]. The method suggested in [34] provides a full picture of topics from multiple corpora which can represent time updated versions of earlier corpora, where the hierarchical topic models have been merged based on graph matching methods [34], such as graph edit distance and other such methods [34]. A survey of set covering problem solutions has been provided in [33]. A combination of partitioning and agglomerative clustering algorithm has been described in [35] where the data points have been partitioned and then clustered maintaining the total number of identified partitions. A method of merging models using agglomerative clustering approach where models have been merged based on a maximum classification likelihood measure has been described in [36].

Earlier works had only considered the provenance, computational trust, topic or subtopic user roles and data importance issues in isolation while this work integrates the considerations into a unified approach. The approach presented in our paper is novel to the news reports description and modelling problem for the above consideration. The Q Learning approach adopted is unique to our work in the news reports modelling application. Our Record Merging approach and calculation of net trust value at a database relation site hosted at a node is novel. This has been defined based on records arriving from other network node sites together with the records already present at the site or on the new event influencing the record at the site. The record merging approach and net trust calculation of output message of a node is unique to the news reports description and modelling application. The derivation of macro action (s) and the computation of reward(s) associated with these actions are unique to this work in producing a reduced clustered space in the news reports modelling application. The approach described in merging of event or topic models for identifying similar or discerning threads of news reports or messages over incremental time periods is unique to our work.

Related work

Provenance graph, computational trust and hierarchy of roles

The issues of trust and provenance have been described in [37, 38] where every state in the provenance graph has been associated with an activity. The security or access restrictions or storage constraints of data base access and the concept of associating cost with trust has been described in [3]. The use of information cascades for tracing the path associated with a piece of news has been described in [6]. A news article has been described as having been influenced or transformed with edit activities which define the path of provenance to the final version of news as described in [7]. The concepts of dependency provenance for describing event dependencies have been described in [39]. The modelling of a database view as a query tree evaluated in a bottom up fashion has been described in [40]. The identification of trusted groups of users with clustering has been described in [41]. A Hierarchical Solution of Markov Decision Process (MDP) has been described in [30, 42]. The comparison of same edge set in the query proposal graph and the provenance graph has been described with the SEC (same edge contribution) in [31]. A group operator has been described in [43] for defining groups based on actor and its invocation granularity. The situation trust values which have been averaged

over all situations to determine the trust in an agent of interest as has been described in [44]. A provenance system has been defined [4, 45] on collection of subjects *s*, objects *o*, and attributes *a*. A subject *s1* with attribute set *a1* performs an action. A migration policy decision has been taken based on input object *o1* and attribute set *a1* associated with subject *s1* and the state reached as a result of migration. The information transfer model developed in our work which includes network issues has relevance to the work described in [4]. The issue of network congestion and measuring its relevance has been described in [46–48]. The above description summarizes related work in the news reports provenance graph representation and modelling issues.

The application of markov decision process to decision theoretic planning has been described in [10]. The Q Learning based approach for construction of provenance based plan of records has strength of relevance with the reinforcement learning procedure as described in [10]. The concept of tokens received from event and or background topic or other originating topic [49] for producing a new theme describing events or topics [49] has been described in [50].

This method of reward calculation described in this paper has similarity with the precision measure of accurately classifying a target from a source as has been described in [23]. This mathematical model of cost or reward value calculation which has been described in this paper has similarity with a model of action as has been described in [42]. A networked model structure and distributed exchange of information have been considered in [5, 47, 51, 52]. The provenance trust scores of all data passing through a node "A" has been averaged to produce a measure of trust associated with the node "A" as has been described in [53]. A trust measure has been updated for every data or message passing through this link in its provenance path which provides a context for trust calculation. This trust is calculated as the average of the trust values computed from the context provided by every data or message passing through this link. The concept of averaging trust values over contexts has been described in [54]. A procedure of average trust calculation along a provenance or policy path learned with reinforcement learning Algorithm has been described in the context of calculating average reward along a policy path in [55, 56]. A trust value can be calculated [57] as the weighted average of the trust values computed along alternative provenance paths where the weights have been determined from the path length. The Q value has been expressed as a weighted sum of extracted information for each feature as has been described in [58]. The concept of trust computation from record weight and trust of agents has been described in [59]. The above description summarizes related work in the trust or reward value calculation issues.

The evolution of events over time stamps has been used to describe event threads [60] as has been described in [61]. The documents have been selected based on a pre-selected set of features based on time, and then partitioned into clusters of events which have been organized into a hierarchical structure as has been described with threads within [61]. A DAG defining a provenance graph has been utilized to model topic correlations as has been described in [62]. The correlations between few interpretable super topics and their sub topics have been examined in [62]. The above description summarizes the related work in representation of event or topic thread.

The distinctive structural features associated with a social role have been interpreted as "structural signatures" [63]. A concept of group context variable has been described in [64] for determining admissibility of users to a trusted group role with a representation of trust or risk variable corresponding to the role. A distance from ideal measures or the same-as chains property suggested in [65] can also be utilized for setting thresholds in performing clustering of nodes or actor roles. The concept of associating a semantic role for tracking [66] a topic or sub-topic has been described in [67]. The above description summarizes related work in the representation of role within our work.

Related work in story, topic, event

The detection of previously unspecified events using feature pivot techniques has been described in [17]. The cluster centres describe the initial news reports cluster [68]. The different structures or models [68] have been sourced from news reports that have higher degree of "contribution". Aspect level classification or clustering news reports has been described in [18, 23] which categorizes the news reports into classes or categories for exploring at multiple depths of detail. A likelihood of a particular aspect word in a snippet of information has been identified and the probabilities of words that describe the aspect have been described in [69]. The counts of a shared aspect [70] describing snippets in the corpus of documents have been collected for processing in [69]. The above description summarizes related work in the aspect based topic representation property.

The use of information cascades has been used for tracing the path a piece of news has traversed in the social media graph initiating from the source and traversing the information influencers [6] in the relevant provenance path. Multiple sources can produce stories about the same event which may be grouped for summarization and mining [71] purposes as has been described in [72]. The dependencies can be defined between aspects from which a relative weight-age can be calculated for every aspect describing its importance as has been described in [72]. The theory of additive compositionality for combining several words with an element wise sum of vectors for comparing similarities between whole tweets has been described in [16]. The above description summarizes related work in the dependencies between aspects and the relative weight-age calculation issues of aspects.

A topic has been represented with many sub-topics and as the granularity of an event is too small to describe a topic, and a less granular event has been used to describe only a sub-topic in [73]. The event indexing models has been used for characterizing a topic or story narrative with a plan of event executions as has been described in [74]. A similarity measure between news articles has been defined using weighted similarity of the persons and keywords describing the articles [75]. The topic splitting and merging research where an event causally influences another for establishing links between stories or topics or storylines have been described in [18, 76–80]. There can be multiple storylines related to higher level topics, people, location and time [71]. A storyline has been described [71] on a subset of the relevant topics, for instance a storyline has been described as to cover only the political or the economic aspect of Lehman Brother collapse. A cost has been associated with detection of false link between document pairs where the strength of a link has been described as representative of a measure of

cohesion between these documents [76] and this cost or reward value can be utilized for applying reinforcement learning approach. A topic has been described as a seminal event or activity alongside occurrences of other secondary events and activities [81]. The concept of evolutionary discovery of theme which has been interpreted as a semantically coherent topic or subtopic transition appears within [82]. Many themes can be active at an interval of time, and a theme evolution graphs has been represented with arcs connecting a theme to another across the time intervals [82]. The theme evolution arcs can represent threads of themes describing lessons from the event, aids, concerts for the event, personal experience from the event (s), or donation match [60], utilized for describing the topic of Asian Tsunami. Topic specific words can be identified by removing words which appear exceeding a threshold measure defined based on ratio of frequency count of documents in which this word appears and frequency count of documents describing the topic of interest [83]. The above description summarizes related work in the hierarchical topic sub topic and threads of themes or sub-topics representation property.

The weight calculated for an aspect has been described based on frequency of occurrences of terms associated with this aspect interpreted as importance measure as has been described in [12, 18, 72]. The incremental merging of acquired topic tree model for more recent times uses a procedure for detecting the attachment point of new model as has been described in [18]. A Gaussian random markov field approach has been adapted to model correlations between different corpora or document and markov topic model uses this approach to describe topic structure within and across corpora of documents [84]. The ART model [85] describes the per message topic distribution based on author and recipient pair. The high level general topics have been discussed for the entire event, unlike specific topics [86] that have been discussed during specific segment of the event [15]. Thus a general tweet has a weak topical influence from the event, unlike a specific tweet which has a strong topical influence from one segment of the event [15]. A hierarchical clustering from more abstract topics to more concrete topics based on time and conditions defined on aspect attributes has been described in [87]. An event has been described as a series of stories where these stories has been described as having been formed from core stories and their related secondary stories [88]. The development and evolution of these core stories can be described by a number of branches which have been detected from denseness of connected nodes in the neighbourhood [88]. The above description summarizes related work in the dynamics of story or topic representation.

Related work in clustering of nodes in the provenance graph

A clustering based on weight of link or edge connecting event pair as has been described in [89]. A MMHP (Marked Multivariate "Hawkes" Process) algorithm utilizes textual information cluster for activation of events into different clusters and the HTM ("Hawkes" Topic Model) algorithm models the evolution of textual information with Correlated Topic Model (CTM) through cascade of topics as has been described in [90]. A concept of role of macro action has been described in [91] where this macro action satisfies a sub-goal. The clustering of documents based on shared aspect or topic or event or sub-topic, or sub-event has been described in [18, 92–94].

The nodes of high centrality importance has been described as bottleneck states connecting the minimum cut arcs which are members of another cluster or component [95]. The "betweenness" centrality of a node or document can be important in detecting variation of the changes impacted by the cited sources as has been described in [96]. The clustering objective measures like MQ (module quotient) or the Quality Cut Measure has been examined for inclusion of newly identified nodes in a cluster of nodes in a graph. The cluster node expansion and cluster edge expansion measures have been considered in [24] for growing a cluster using a derived cluster node expansion capability measure. A concept of group context variable has been described in [64] for determining admissibility of users to a trusted group role with a representation of trust or risk variable corresponding to the role. A distance from ideal measures or the same-as chains property suggested in [65] can also be utilized for setting thresholds in performing clustering of nodes or actor roles. A concept of role of macro action has been described in [91] where this macro action satisfies a sub-goal with a reward function particular to the sub-goal or sub-topic [97]. The concept of associating a semantic role for tracking a topic or sub-topic has been described in [67]. The paper on clustering of news reports [18] provide a description of tree model construction and merging where the split or merge points have been described as providing a link from past story to more recent story description. Aspect based regression tree clustering described in [18] produces clusters of documents as leaf nodes, where every leaf node can be associated with a news topic or a news category [21] or a news issue [22]. The above description summarizes related work in the clustering of topic or sub topic or event nodes, both within a provenance graph model or across a pair of such models.

A recursive partitioning approach has been adopted instead of a mixture model for satisfying the conditions of interactions of variables, or for detecting these interactions, or for generating scarce interaction patterns, or for selecting these variables without requiring prior specification of these variables [98]. A recursive splitting of cluster nodes describing a regression tree as has been described in [18] reiterates these advantages. The documents have been analyzed for identifying a list of candidates for a target document based on titles similarity, content similarity, unique words and frequents words [99]. The importance measure of a term can be utilized for computing the similarity or distinctive relation between documents which signifies that these documents are describing a collection of similar or discerning topics. The "betweenness" centrality measures the importance of a node in the cluster. The degree centrality has been considered as not sufficient as a measure for the whole network and closeness centrality has not always proven as the ideal measure for measuring centrality. The importance of a node or story or event depends on the interpretation provided in describing the topic graph model and has been measured with the use of between-"ness" centrality in this paper. The key phrases which have been represented in these centrally important nodes have been considered as more relevant in constructing a policy path plan routed through these nodes. Similar articles or articles describing similar events and which have been clustered to produce an event class trigger have been used for describing events in the same class [100]. A Hierarchical Topic Detection algorithm has been described in [101] where a theme area comprising of the title and the initial paragraph of the document has been defined and the detail area has been used for additional details about the topic or

subtopic of the topic. The network metrices have been evaluated in [102]. The centrality has been described as highly characteristic of a hierarchical network, where patterns have been described as centralized on few individuals who attract attention from other network nodes [102]. The high density networks have been distinguished in terms of level of modularity which is the measure of interconnectedness of clusters [103]. However the clusters defined based on degree centrality measuring the interconnectedness between network nodes does not prove as sufficient for the entire cluster. The relevant algorithm examine large data sets and efficiently find subgroups and the algorithm uses edge "betweenness" as a metric for identifying boundaries of communities as has been described in [103]. The in-group social media network structure has users with unified interests, while networks with clustered communities [104] limit information flow to small silos of users which have been described as stable over time [103]. The distinctive structural features associated with a social role have been interpreted as "structural signatures" [63]. The above description summarizes related work in the "centrality" issues in clustering.

A cluster has been defined for news stories occurring at a snapshot interval of time which describe the same sub goal or sub topic of the goal topic [105] or whole story [106]. A clustering of events has been defined on the commonality of attributes describing events, where exact match of attributes causes events to be positioned in the same cluster, whereas a partial match causes a link to be described between events across their corresponding clusters as has been described in [106]. The story clusters have been defined for their corresponding snapshot time intervals and have been linked based on their similarity measures like Jaccard index, Sorensen-Dice coefficient or similarities defined based on measure of inclusion or exclusion. A k-modes clustering algorithm has been described in [32] for categorical data. A number of cluster analysis methods has been described in [32] for identifying the number of clusters, which have been defined based on gap statistics, and other measures based on information criteria such as AIC (Akaike Information Criteria) or BIC (Bayes Information Criteria). The above description summarizes related work in the other clustering methods applicable to both within and across provenance graph models.

Related work in classifier learning

A reinforcement learning procedure has been adopted for identifying the policy describing the narrative path for the whole story which can also represent the path from start topic or event to goal topic [107] or event. Machine learning algorithm such as reinforcement learning has been utilized in the task of Text Mining [108] applications for learning aspect or topic of document or story [109] as tasks in computing long term rewards and thus making these methods applicable to Big Data Analytics [71]. A reinforcement learning procedure has been applied to an aspect based representation of a data base of stories which have been considered as relevant to the news reports analysis or econometric problem domain as has been described in [18, 110]. Reinforcement learning method has been utilized for finding the stochastic shortest path in scheduling of tasks for satisfying the quality of service constraints [46] in the presence of drift of concept [111] associated with the incoming data in Big-Data [71] Streaming Applications as has been described in [112]. Q Learning procedure can be applied to produce a

plan [10] of event or topic executions [74] from the start topic or event [107] to the goal topic or event in [107]. Q Learning is a model free reinforcement learning method. The Q Learning procedure has been described as a topic in reinforcement learning in [43]. Q Learning procedure considers the weighted average of rewards present in a policy path as compared to reinforcement learning approach which considers a simple average of reward in a policy path [43]. The above description summarizes the related work in reinforcement learning, Q Learning and Big Data Applications.

The main difference between Bayesian and non Bayesian methods is the use of priors. The prior for a Bayesian network structure can be converted to the priors for an equivalent Bayesian structure by application of change of variables using a method of Jacobian Transformation as has been described in [113]. The concept of score for belief network has been presented in [114, 115]. The scores of two isomorphic belief networks must be considered as equal. A Bayesian metric with Dirichlet priors (BD) has been presented in [116] for calculating a score. A score equivalent BD metric has been presented (BDe) for identifying score equivalent Bayesian network structures. Such score equivalent metrics can be used for identifying equivalent task structure network for provenance path learning. A method based on scoring equivalent class operators has been described in [116]. A search algorithm has been described that moves along structures with the application of these equivalence class operators. A greedy algorithm has been applied over a local subspace which produces better results [117]. A unit task or activity node which has been added to or removed from a network produces the maximum improvement in this score [118]. However the greedy hill climbing approach can result in search getting trapped in local maxima thus requiring backtracking and restarts. An adaptive simulated annealing approach may also be applicable here [119]. The above description summarizes the relevance of Bayesian learning issues to our work.

The process of adapting classifier to the situation of change has been described in [120]. A concept drift can happen if prior probabilities of classes defined on target variables change, or if conditional probabilities associating the target class variable and input variable changes, with a corresponding impact of change on posterior probabilities [120]. The ensemble learning methods produce models that are either homogenous or heterogeneous. Homogenous algorithms which use the same algorithm with different parameter either by introducing randomness or by manipulating input attributes, model outputs or training instances with a process of "bagging" or "boosting" as has been described in [121]. The explore vs exploit strategies for selecting actions has been described in [120]. Reinforcement learning algorithm has been described as where all state action pairs have been observed and top updated Q values have been retained for future processing as has been described in [122]. The reward function values impact distance measures used by the clustering algorithm [122]. The three heuristics for credit assignments [27] have been described in [123]. Constraint Satisfaction and Emerging Algorithm based on Set Covering Problem Solutions have been described in [124]. A survey of set covering problem solutions has been provided in [33]. A combination of partitioning and agglomerative clustering algorithm has been described in [35] where the data points have been partitioned and then clustered maintaining the total number of identified partitions. A method of merging models using agglomerative clustering approach where models have been merged based on a maximum classification likelihood

measure has been described in [36]. The above description summarizes the related work in merging graph models.

The ideas described in this paper where the graph models are either first partitioned using a partitioning approach like PAM or CLARA and the representative models in each cluster are merged using either a Bayesian scoring approach or from using their property of markov equivalence thus producing an essential model have similarities with ideas discussed in [35] and in [36]. The distance between a pair of graph models has been described using a graph edit distance measure in [125].

A cluster has been defined for news stories occurring at a snapshot interval of time which describe the same sub goal or sub topic of the goal topic or the whole story [106]. The concept of Structured Stories has been described in [126] where a semantic zoom provides drill down into detail of a specific event in the story. An event within the story can be linked to another story containing a detailed narrative of the event [126]. A story has been described as a sequence of story fragments where a story fragment can act as a bridge based on term context dependent attractiveness between start topic and the end topic [107]. The concept of bridging topic has been described in the story of Lehman Brother Collapse which has been described using the Political Aspect or Topic and Economic Aspect or Topic and this can provide an example to our approach of describing splitting or merging using bridging topic in story description.

A window length has been defined for training the topic model which has been specified using a time range and a refresh rate of length less than the window length for considering if the training instances are old [127]. The model used for event similarity calculation lags an interval of refresh rate in minutes [127]. As the volume of tweets has been considered as not uniform the refresh rate compensates for non uniformity of the count of words. Alternatively, a sliding window has been updated every 15 min, and the model has been retrained with tweets of the past 24 h [127]. These are relevant in considering lengths of incremental time intervals for constructing model.

Clustering and topic models have been integrated into representing a storyline where there is a probability computed for assigning a new document to an existing storyline or a new storyline [128]. A storyline can represent a higher level topic where a link has been established to this topic for preserving this storyline [128]. A cluster has been identified for a newly acquired document based on either a hard decision where the Bayesian cluster models have been updated with assignment of a document to a cluster with the highest probability or a soft probabilistic updating has been performed. A cosine similarity value between topic pairs has been utilized to identify a bump in the cosine similarity graph for interpreting event topics [129]. The localized high cosine similarity bumps have been used for interpreting event topics, where a random change in the similarity value can indicate events that are not time specific [129]. An event topic present in another event topic's hot zone or with high similarity value have been linked together in the graphical representation and grouped together to represent list of sets of event topics [129]. In the genomics collection, divergence of topic stories from the general collection, can indicate that subsequent stories have been associated with a new topic within a new introduced time window [130]. The candidate models participating in the procedure for merging have policy path classifiers which have been described as either only member of any set of similar classifiers or are members in the minimum cover set [33] of

such similar classifiers that are present in the cluster. The above description summarizes related work in the Bayesian Model Merging and minimum cover set merging methods.

Description of our work

Problem statement and solution

News reports describing event(s) are associated with or are routed to an intermediate node in a network of nodes describing a provenance graph where each node in this graph is associated with an activity. Every news report is presented by a user whose trustworthiness is measured with a computational trust value. This activity describes a "match and extract" procedure for transforming these input reports to the destination activity or event or topic node thus providing a more detailed version of the narrative. This search and extract procedure which uses the importance measure associated with individual attributes or records describing the news reports and the computational trust measure associated with their originators have been described in "Development of relevance of computational trust" section. The relevance of the clustering to this work has been described in "Related work in clustering of nodes in the provenance graph" section and our approach to clustering of nodes has been described in "Our solution to clustering nodes in provenance graph" section. This procedure of learning the path in a provenance graph of activity or event nodes has been described in "Rationale for application of Q Learning algorithm in this trust based representation" section. The rationale for application of Q Learning algorithm to the Trust Provenance Graph Representation has been described in "Application of Q Learning algorithm for learning provenance path" section. The provenance graph models thus constructed from identified dependencies between events or activities is relevant for an interval of time window. These graph models are merged to produce consistent models spanning larger intervals of time windows which apply also to incrementally acquired models for more recent time interval windows. The narrative of splitting and merging for producing the goal topic or whole story description in these identified merged models have been described in "Our approach to provenance graph structure and classifier learning" section.

A block diagram or the design of the solution steps has been presented in Fig. 1. The examples in support of the design have been presented in this paper in Figs. 2 and 3. The Theme or Event or Topic or Subtopic evolution threads describe the Asian Tsumami depicted in Fig. 1. Here for example, Lessons from Event has been described with the Theme evolution Theme 1 → Theme 1 → Theme 4. Possible threads for Aids include Theme 4 → Theme 3 → Theme 5 and thread for Personal Experience include Theme 5 → Theme 6 → Theme 3 → Theme 3. The confluence or merging of Theme 4 and Theme 5 into Theme 6 and then evolving into Theme 3 described the Personal Experience thread as Theme 4 → Theme 6 → Theme 3 → Theme 3 or as Theme 5 → Theme 6 → Theme 3 → Theme 3. Theme 1 at time interval 4 splits into and evolves towards Theme 1 at time 5 and Theme 6 at time 5, corresponding to Lessons from the event thread and Criticism on Iraq thread. The learning or construction of provenance path to a goal path discovery has been described in Fig. 3 as exemplified in [131].

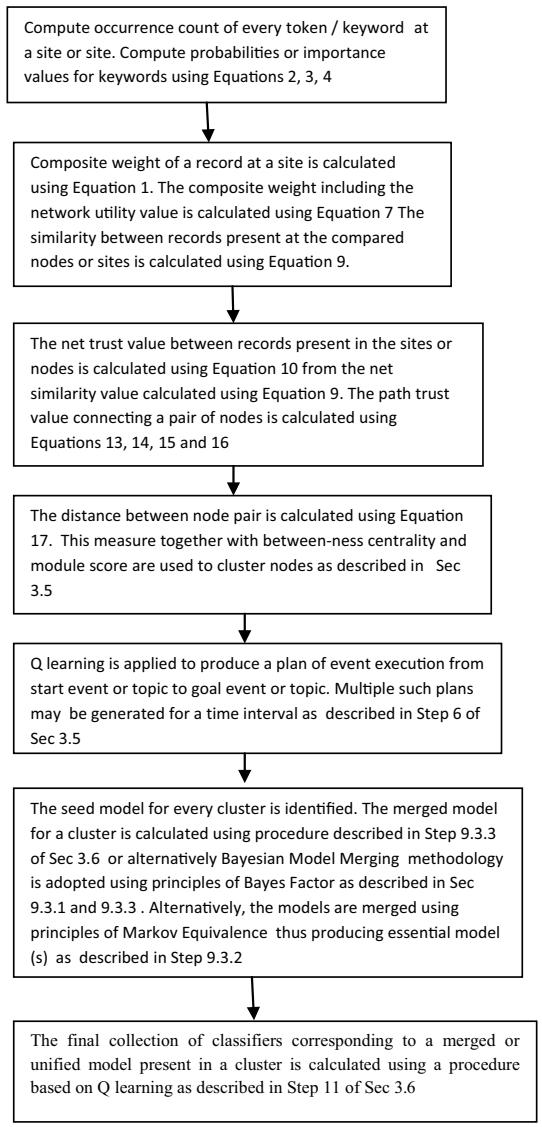

Fig. 1 Flow chart or block diagram summarizing design of solution steps

Development of relevance of computational trust

The model for a message passing network is considered in the present work. Trust needs to be associated with messages as well. Two trust values are defined, one for the communicating node and the other for the message sent. In order to conclude about the trustworthiness of a message, a composite trust needs to be computed. A summarized description of issues of storage and retrieval of trusted information using a probabilistic temporal database approach is available in [12]. A measure has been introduced as a composite of calculated weight of record or attributes importance value and the computational trust of the agent attached to the node. The utility measure associated with any node can be calculated based on probability of action of forwarding data from that node to the network node option for the current congestion situation. This probabilistic utility measure may indicate the fraction of the sent

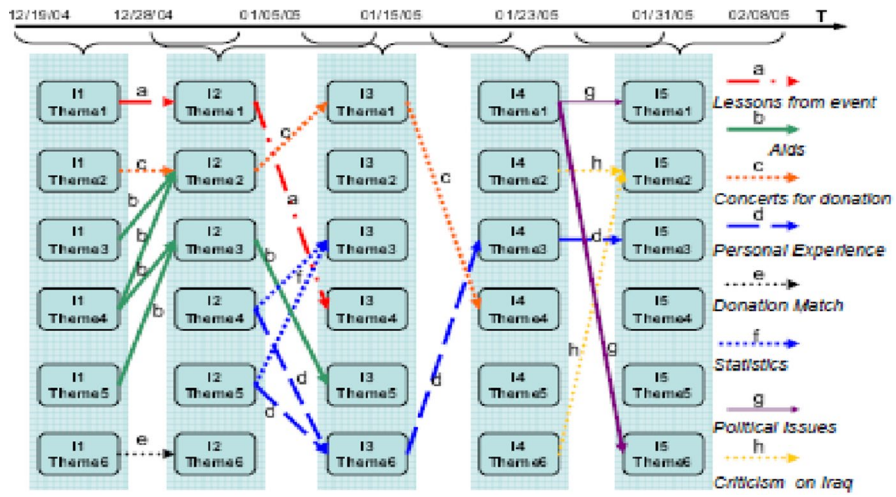

Fig. 2 Theme evolution graph describing Asia Tsunami [82]

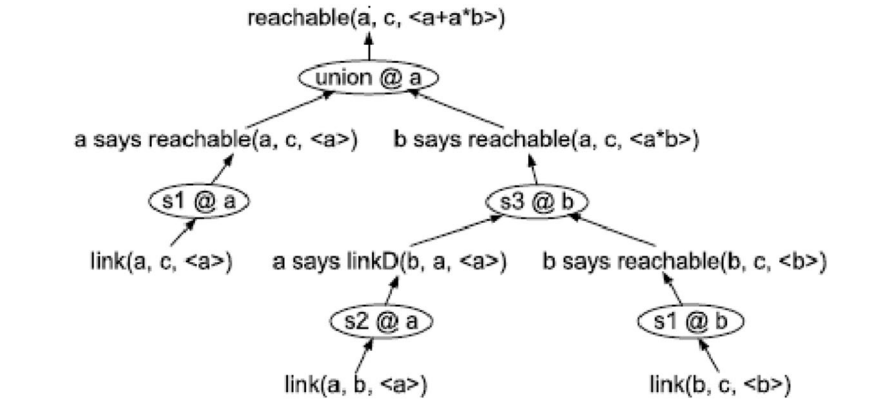

Fig. 3 Example of provenance path construction [131]

packets from the source that are successfully delivered at the destination thus avoiding dropped packets (Eq. 8). This net trust value which is the composite trust of attribute importance values and network utility values and the computational trust of the agent associated with the node is derived from application of principles described in Eq. 10. An event or topic modelling approach is considered where the dependency between news report event or subtopic nodes is represented from their correlated property derived from similarity or distance measure associating these news report events or subtopics. The reward value calculation for transferring message between these nodes is described and which is utilized for deriving a policy path of message transfers using a reinforcement learning approach.

$$\text{The weight associated with a source record} = \left(\sum\nolimits_i (p_i * mc_i)/(ascm)\right) \tag{1}$$

where i represents a valid or selected attribute in the merged record where merging records have identical value for this attribute, mc_i is occurrence count of token associated with attribute i in merged record which is calculated as the sum of occurrence counts of corresponding attribute value in the individual records, *ascm* is the cumulative occurrence count of tokens in the merged news report for all valid or selected attributes i present in the merging records and p_i determined from application of Eqs. 2, 3 and 4.

The value is normalized with a division by the summation of such probabilities identified for all attributes of the source record.

A measure can be associated with any field value based on the occurrence count of this value in the news report from where this record is represented. This can be calculated based on the iteratively developed conditional probability of a field token value given an event as a maximization step of EM algorithm solution as described in [132]. The EM algorithm maximizes the posterior probability of occurrence of the event.

The expectation step is $p(ej|xi)(t+1) = p(ej)(t) * p(xi|ej)(t)/p(xi)(t)$. Here e_j is the jth event, x_i is the ith news report. The maximization step is derived from this equation and this also produces the probability value of nth entity.

$$p(w_n|e_j)(t+1) = \left(\sum_{(i=1 \text{ to } M)} p(e_j|X_i)(t+1) * tf(i,n))/(N + \sum_{(i=1 \text{ to } M)} p(e_j|X_i)(t+1) + \sum_{(s=1 \text{ to } N)} tf(i,s) \right) \quad (2)$$

This probability of field token value in presence of all values can also be calculated as a posterior probability value based on prior, likelihood of the value given the news reports about the event and the prior probability of occurrence of event in the reports as described in the context of combination and calibration of methods for the purpose of forecasting of events appearing in [133].

$$f(p_t|n_t) = f(p_t) * f(n_t|p_t)/f(n_t) \quad (3)$$

where $f(n_t|p_t) = (n_t C_m) * p_t^{nt} * (1 - p_t)^{(m-nt)}$

This probability value can also be calculated by a beta binomial method. This can also be calculated as an optimal score approach where the conditional probability of the field token value is calculated based on defined number of past observations, and observations from news reports with application of appropriate weight-age to each that minimize a posterior log likelihood measure as described in [133].

This is described as follows,

$$E(p_t|n_t) = (T_p + w_{nt})/(T + w_m) = w_1 * (n_t/m) + w_2 * p \quad (4)$$

where $w_1 = w_m/(T + w_m)$, and $w_2 = T/(T + w_m)$, where p_t is the probability associated with entity of interest (posterior), nt is the occurrence count of the entity, m is the total occurrence count of all entities, T is the sample size and p(prior) is the sample mean.

The weight w is determined with the purpose of maximizing the log-likelihood value,

$$L(w) = \Pi_{t=1}^{T} E(p_t^*|\, n_t).$$

This w value determines the posterior probability of entity of interest.

Thus if w_i is the weight associated with ith message record, and T_i is the global trust value associated with source of record i, the composite trust in the event is calculated as

$$COMPT = \sum_{i} (w_i * T_i) \tag{5}$$

The fusion value of data is calculated using data fusion rule specified in [134]. This value is determined as

$$\mathrm{DV} = \sum\nolimits_{\mathrm{i}} (\mathrm{w_i} * \mathrm{T_i}) \Big/ \sum\nolimits_{\mathrm{i}} \mathrm{T_i} \tag{6}$$

where $\Sigma_i (w_i * T_i)$ is the composite trust in the message produced from this node and T_i is the Trust in the Trusted Partners associated with ith message record input to this node

$$\text{Composite Weight of Source Record} = (\mathrm{w_A} * \mathrm{u_A}) \tag{7}$$

where w_A is weight associated with source record in "A" and U_A is the utility associated with the source record from source at state "A" for the chosen forwarding action to B based on current network congestion situation. The utility measure u_A is defined as,

$$\mathrm{u_A} = \frac{(\text{Frequency count of successful transfer of packets from "A"})}{(\text{Frequency count of sum total of all packets transferred from "A"})} \tag{8}$$

The reward received at node B is $\mathrm{COMPT_{AB}}$ where the composite trust in A after interaction with "B" [40] which has been derived using Eq. 10 from a choice condition determined from application of Eq. 9.

The unit step reward for corresponding $p_{AB}(n)$ is valid for $n = 1$.

However multi step reward calculated for such node pair $(A,B) = p_{AB}(n)$ for $n \geq 1$, which is symbolized as p_{ABt} and is derived using application of Eqs. 13, 14, 15 and 16 appearing later in this document.

A similarity measure Sim(i, j) between two records i and j is defined where each of these records are described by k attributes and attr_in_merge(j) = 1, indicates if the jth attribute values of the records have similarity value > 0, else attr_in_merge(j) = 0.

The net similarity measure is defined as

$$Sim(i, j) = \sum_{\alpha(a=1 \text{ to } k)} (attr\_in\_merge(a))(wa * Sima(i, j)) \tag{9}$$

where Sima (i,j) is the similarity of the ath attributes of records i and j and w_a is the normalized probabilistic weight associated with the ath attribute of the merged record which is based on application of Eqs. 2, 3 and 4.

The procedure of cumulating occurrence counts of token or evidence for merging information from multiple sources has also been discussed in [135]. The choice of routing or forwarding of information based on reward calculated on aggregation based on data correlation and gain in reward or diminishment of distance measure to the goal node has been described in [136].

This net similarity measure calculated using Eq. 9, is used for determining the appropriate probability disjunction strategy. Here the probabilistic trust values pA and pB associated with individual source records participating in the merge are calculated from applying Eqs. 1 and 7.

If p_B is the probabilistic importance or trust value associated with record at node "B" and pA is the probability of source record at "A" forwarded to node "B" situated in the provenance graph is determined from application of Eqs. 1 and 7. Here the weight of the relevant record is multiplied with the probabilistic trust value measuring the trustworthiness of the agent at the node to produce the probabilistic value presented to the probabilistic disjunctive strategy described for temporal probabilistic databases [12, 137].

$$\text{The Net\_trust} = \text{measure defined on net similarity} \tag{10}$$

Where if the *net similarity measure* > *high threshold,* probabilistic disjunctive strategy [137] is adoped for the positively correlated case and the Net_trust is *min(1, pA* + *pB).* else if the *net similarity measure* = *0.5,* the probabilistic disjunctive strategy [12] is adopted for the independent case and the Net_trust is $pA + pB - pA * pB$. else if the *net similarity measure* < *low threshold and* > *minimum threshold*, probabilistic disjunctive strategy is adopted for the negatively correlated case [12] and the Net_trust is *min(1, pA* + *pB).* else If *net similarity measure* < *minimum threshold,* probabilistic disjunctive strategy is adopted for the ignorance case and the Net trust is *max(pA, pB).*

This net trust value $COMPT_{AB}$ or the reward received at node A is the composite or net trust in the event after merging the message records from pair of sources using method as described in Eq. 10 of this section. This measure considers the importance of similar sources in a cluster alongside sources with the news report "contribution" [138] or measure of distinctive relevance of sources in the reward value computation.

Rationale for application of Q Learning algorithm in this trust based representation

The agent selects an action probabilistically based on Boltzmann distribution defined using Q value probabilities associating states with actions [139]. The Q values are associated with policy of actions and thus justifying the use of provenance path probability values in determining action probabilities.

The transition probability p_{ij} between states i and j = exp ((f(j) − f(i))/t) has been described in [55]. The cost functions associated with states i and j are represented as f(i) and f(j) representing their energy states. The value of a policy is the sum of rewards obtained from execution of actions described in the policy. The Q Learning algorithm selects a policy with a higher reward. The difference (f(j) − f(i)) has been calculated from the optimum value of the policy connecting these states i and j. As a policy with higher reward (dominant policy) has always been selected over less rewarding policy [121] the procedure described for calculating reward value (Eq. 16) from p_{AB} (d) is appropriate here. Thus this procedure favours moves between similar or relevant nodes or states.

Only sub-topic sub-goal states of each cluster have been identified using application of method described in [24] and are retained for the re-computation of probabilistic weight connecting pair of nodes in the reduced clustered state space. The detected node pairs defining the provenance graph in the reduced cluster space are utilized to learn the policy path with application of Q Learning method. The Q Learning approach adopted is unique to our work in the news reports modelling application as the approach provides more weight-age to rewards from moves taken when the search space is less localized and distant from the goal while calculating a discounted sum of rewards present in the trust path to goal. The rewarding moves between nodes or states in a reduced

clustered space using a Q Learning approach provides more weight-age to early moves based on relevance or contribution of information associating or linking these nodes as this information is of more value as compared to that in proximity with the goal where the learning or the search has already been localized. Also final steps or edges in the path are associated with excessive granular representation of events which are discounted or attached a lesser weight in computing the weighted average value of reward from start node to goal node.

Application of Q Learning algorithm for learning provenance path

The Q Learning procedure is described as follows. Initialize V(s) for all states using initial network congestion conditions and using procedures described in Eqs. 1, 7 and 10. The network nodes are considered to be connected by links based on probabilities determined from current congestion situation as described in Eq. 8. Here s is the set of nodes, that are base database [140] relation sites, where record merging operation is applied on records arriving from two or more network sites.

Loop until chosen policy is good
{
For all s in S {
For all a in A
{ Step 1 :

$$Q(s, a) = R(s,a) + \nu^* \sum_{s' \text{ in } S} T(s,a,s') * V(s')$$

Step 2 : $V(s) = \max_a (Q(s,a))$

(11)
} /* end for all a in A */
} /* for all s in S)
Revise the V(s) values according to the updated Policy

}

Here the transition probability T(s,a,s') is determined from application of Eq. 16, and R(s,a) is the trust in the component of information contributed by state s in performing the transformation step at a state s' reached from forwarding action of data or information packet and this is determined from application of Eq. 16 after reaching state s' as described earlier.

The policy learned is considered to be "good" if the magnitude of the updated Value attributes are less than an acceptable threshold different from the magnitude of these attribute value prior to update. The greedy policy finds the optimal policy in finite number of steps sometimes earlier than convergence of iterated value function. Here on arriving at a better optimal policy, all the previously learned policy paths including the latest that are covered are ignored and a new policy which is sub-optimal but is presently optimal is learned. This procedure continues until all policy paths satisfying the goal are learned.

Thus a message is transferred through a selected policy path of nodes in a provenance network for maximizing the iterated value described for a node in the graph.

A composite trust value of a node "A" is calculated from its content importance, which includes the trust of the subject associated with the node and importance of the node. The change of importance measure between a node "A" of interest and active node "B" in the provenance path after traversing n steps in the provenance path is calculated where node B is active after executing n provenance steps.

A provenance trust score is calculated for every data "d" passing through this node "A". This is calculated as minimum or average of the trust values associated with nodes in the provenance path. The provenance trust scores of all data passing through this node "A" are averaged to produce a measure of trust associated with node "A".

The subject node "B" is included in the same cluster as the cluster identified for subject node "A" if the following conditions are satisfied.

The weight of source record at "A" after satisfying constraints associated with merging with record at "B" is w_{AB} and is calculated using Eq. 7. The weight measure is further normalized by maximum w_{AB} calculated for all links describing the provenance graph.

Also prior weight associated with record at "A" is $w_{ABt} = w_{AB}/\max(w_{AB})$

$$p_{AB}(1) = \max(w_{ABt}, 0.5) \tag{12}$$

and for "B" reachable from "A" in 1 step, where p_{AB} (1) is the Transition probabilistic trust value between nodes or the strength of trust connecting these nodes and also where in the absence of prior information this probability is initialized to 0.5.

If A is connected to B with a provenance path through one intermediate node C for a run of the workflow:

$$p_{AB}(2) = p_{AC}(1) * p_{CB}(1) \tag{13}$$

Alternatively $p_{AB}(2)$ can also be calculated as the average or the weighted average which is of more relevance in the adopted of Q Learning procedure of the trust transitions $p_{AC}(1)$ and $p_{CB}(1)$ if we adopt a procedure described in [44].

The transition probability p_{ABt} which is the probabilistic link value connecting any connected pair of nodes in the provenance graph has been defined in [27] and is calculated using a procedure Cluster_Node_1 described as follows.

Algorithm Cluster_Node_1

{

Nodes "A" and "B" are elements of the same cluster if

$$P_{ABt} = \text{avg}\left(p_{AB}(\text{"d1"}), p_{AB}(\text{"d2"}), \ldots, p_{AB}(\text{"ds"})\right) \tag{14}$$

where d_i s are distinct data input constituting data stream s and

$$p_{AB}(d) = \max\left(p_{AB(Path1)}(d), p_{AB(path2)}(d), \ldots, p_{AB(pathk)}(d)\right) \text{ and } p_{AB}(d) \tag{15}$$

}

If P_{ABt} > threshold value these nodes are positioned in the same cluster P_{ABt} which is the derived trust value of "A" to "B" obtained using distinct provenance paths ($path_i$) for a data input "d" to the stream.

Alternatively this trust value can be calculated as the weighted average of the trust values computed along the alternative paths where the weights are determined from the path length. Here this may be noteworthy that the links nearer to the start node contribute heavily to the net path trust value connecting any pair of start and goal nodes in the provenance path.

This is also described as to the procedure of combination of path trust values in [44].

The normalized derived probability of trust link A to B for action taken at state A

$$p_{ABc} = P_{ABt} \Big/ \sum\nolimits_{x} (P_{AXt}). \tag{16}$$

After every updated policy identification round, the revised trust values of agents/actors associated with nodes in the provenance network are propagated and updated according to the trust propagation rules described for social networks in [141].

Our solution to clustering nodes in provenance graph

The approach adopted in this work for Clustering of nodes in a provenance graph includes the following steps.

1. Identify new node of high centrality importance based on its connectedness or between-ness. These nodes may also be marked as split or merge nodes in the provenance graph describing application of news reports topic modelling. This is sub-goal node for a cluster.
2. Grow the cluster starting from a node of high importance which may be nodes of high measure value of degree or between-ness.

    2.1. The cluster node expansion and cluster edge expansion measures are considered for growing a cluster using a derived cluster node expansion capability measure. This method considers the nodes in the cluster and the complement of the nodes in the cluster remaining in the graph for defining the expansion measure. Also here if the link trust probabilities are fluctuating, a measure based on relative dependence is used for identifying the stability of clustering. The distance measure between any two nodes in the cluster is defined using a Joint Entropy distance value as specified later in Eq. 17. Pair of nodes are only considered to be members of the same cluster if the activity or task steps or units corresponding to these node are within a predefined threshold distance measure defined using Joint Entropy distance measure between node pair. Also the similarity or distance measure is refined using similarities of roles of Agents associated with these nodes. This distance measure is defined based on length of category or role classification tree path separating the nearest ancestor nodes of agent node pair situated in this topic-subtopic role hierarchy tree. The agent similarity or distance measure is provided higher weight-age when compared to trust link measure in calculating the similarity or distance measure between event node pair. Here an agent role may be represented as an attribute in the event record associated with a network node. This leverages the nodes associated with the same agent role for membership in the same cluster. The entity words and the

topic words are important in calculating similarity or distance measure for documents or news reports for these nodes to be positioned in the same topic or sub topic cluster. This similarity measure is represented using the probabilistic trust weight link connecting the nodes positioned in the provenance network. The distance measure calculation from this weight measure has been described later in this section.

2.2. An average distance measure is calculated for every pair of nodes in the identified cluster. Only those nodes are retained in the same cluster whose pair-wise distance is less than a defined threshold measure different from this average distance.

3. Continue Steps 1 to 3 until a cluster is identified for all nodes in the graph.
4. Identify all the cluster components in the graph.

4.1. Alternatively, node pairs with very high trust link weights may be removed from the provenance graph, and thus the clusters of connected graphs may be detected. The high trust link weighted edges may then be introduced, to establish links to sub-goal node(s) of a cluster.

The adjacent clusters/modules are merged based on the following procedure. A module score or a cluster score is defined for each module group in a workflow [45] based on distances of module element pairs contained in the module group. The workflow score is defined as the sum of the module group scores. A greedy strategy is adopted which merges two adjacent module groups with best workflow score. This process continues until only a predefined number of module groups remain. The adjacent cluster components are merged if these identified sub-goal nodes corresponding to the clusters are the same. A concept of role of macro action is applicable where this macro action satisfies a sub-goal and where a reward function may be used which is particular to the sub-topic sub-goal. The roles corresponding to the macro actions satisfies the same sub-topic or sub-goal role and have correlated topic models describing the connected event nodes present in the sub-goal or sub-topic cluster. This role can be interpreted as the semantic role or users of the same role having the same or similar "structural signatures" or users having a trusted group [41] role corresponding to the sub-topic cluster associated with tracking a particular topic or sub-topic or sub-goal of the story. All agents associated with a macro action may be interpreted as playing an informing role to the evaluating agent associated with the sub-topic sub-goal agent. As a consequence of merging of clusters, a hierarchical clustering of sub-topic sub-goal nodes is achieved where sub-goal nodes inferred from recursively defined merged clusters may indicate split or merge points in the goal topic or whole story description. Here a merging of clusters may aim at forming a merged cluster which have nodes in the merged cluster of similar Q values or Reward values as calculated from application of Eq. 16. The actions connecting the sub-topic sub-goal nodes of finally produced clusters are the macro actions on which the Q Learning procedure is re-applied. The distance measure defined on agent roles and trust links in combination with the reward based Q value approach produces a clustering where every cluster has nodes with same or similar agent roles [64].

This condition for merging of clusters is applied with the module score approach to obtain the final clustering. The probabilistic trust weight connecting the sub-goal nodes has been described as a macro action for identifying clusters which are updated using principles described earlier. A role is associated with interpreting a story theme or topic or subtopic or aspect of a story which may be derived from input vector of aspects.

An alternative clustering or classification approach has been described as follows. The examples associated with a state can be classified into one or more identified categories. The full set of states is subdivided into smaller set of states to reduce the entropy measure defined on categories. An action is described on a state with an attribute which is used to further classify the subset of examples in the state. The Information Gain functions that are used are Entropy based, Gini Index based and Discriminant power function based [expected count/(total count)]. The leaf nodes of this decision tree must satisfy the constraint that all activity or tasks nodes which are members of any leaf cluster node are executed by agents with the same role.

Here for measuring the applicability of our approach, a module group score can be defined for a group of module element nodes based on Joint Entropy distance [139] separating every pair of nodes (A, B) defined on this derived probabilistic weight P_{ABc} as discussed in Eq. 16.

This Joint entropy value based distance measure

$$D_{ij} = p_{ij} * \log(p_{ij}) + (1-p_{ij}) * \log(1-p_{ij}) \tag{17}$$

This distance measure is further normalized by the maximum distance value max (D_{ij}) for the provenance graph model. Thus $D_{ij} = D_{ij}/\max(D_{ij})$.

Thus quality of identified grouping or clustering of nodes in a module group can be represented using this distance measure and the difference from the weighted average distance calculated between every pair of nodes in the identified cluster. If this difference is lesser than a threshold value, the nodes are considered as elements of the same cluster. The adjacent groups of modules can be merged that satisfies a workflow score constraint as discussed earlier for a pre-specified limit on the number of clusters/groups in workflow.

Rationale for provenance graph structure and classifier learning

Only those events have been considered that have one or more of the necessary seed terms that have been used to describe a topic of interest to the user. Here the topic specific words that remove other words commonly occurring across topics have been identified using procedure described in [83]. The identification of story clusters for snapshot intervals of time and procedure for linking these clusters have been described in [106]. The models thus identified may be merged based on their importance in description of the story. Here interpreting clustering of documents for representing model cluster and Bayesian probabilistic approach [142] of assignment of document to a model cluster as has described in [143] are relevant. The application driven text data source can be static or dynamic where a static source implies that the document collection as not having frequent updates, while other text streams can be characterized as having many updates [144]. The topic modelling techniques have also been adapted with respect to the temporal scale for narrowing down events to fine granularity [144]. Hierarchical topics provide

an overview of topics from one corpora [34]. The method suggested in [34] provides a full picture of topics from multiple corpora which can represent time updated versions of earlier corpora, where the hierarchical topic models have been merged based on graph matching methods [34], such as graph edit distance and other such methods [34]. A phrase reinforcement learning has been proposed in [145] where a starting phrase represents the topic for which generating a summary of tweets has been proposed, and this best partial summary represents the selection of path with the maximum sum of weights along the path [145]. The above description summarizes related work in the merging of topic models and representing summary policy path.

The classifier policy path has been detected from this structure by application of Q Learning algorithm for learning provenance path as has been described in an earlier section. Q Learning procedure has earlier been described to produce a plan [10] of event or topic executions [74] from the start topic or event [107] to the goal topic or event [107].

Provenance graph model(s) have been constructed for time window (s) which have been determined from window length [127] and refresh rate [127]. The time windows can also be defined based on a concept of sliding window [127] with overlaps between the models defined for these time windows.

The integer linear programming, constraint satisfaction, and other emerging algorithm based set covering problem solutions have been described as relevant for generating a cover set of classifiers for the models. A classifier in a model is compared with all classifiers in every other model learned and the best match is considered for calculating significance of a classifier [146]. Each one of these significant classifiers present in the minimum cover set and which are associated with one or more model (s) are compared with all such significant classifiers which are associated with other such model (s) and a similarity measure between these models has been computed from the classifiers present in these models using a correlation ratio measure as described in [147]. Alternatively, this similarity measure can be calculated as Pearson Product moment correlation measure. A principle of covariance measure defined between a single value and a vector of values is the sum of the covariance measures calculated between the single value and each element of the vector. Alternatively a distance or similarity measure can be described between a pair of these models using a graph edit distance measure as has been described in [125]. The models are clustered using pair-wise similarity or distance measures between a model pair described by their representative classifiers as described earlier. A hierarchical agglomerative Approach [148] can be adopted for this with merging of most similar model pairs at every hierarchy where a merged model can been derived using procedures as has been described later in Steps 9.2, 9.3 and 11. The above description summarizes the relevance of covariance or correlation based clustering methods to our work.

A partitioning approach can be adopted for this model merging using the K-medoids approach where a model has been selected as representative for the cluster of models and this clustering method is realized with a Partitioning around “medoids” (PAM) or Clustering Large Applications (CLARA). The algorithm described as PAM starts with randomly collected seed models and improves the clustering with a greedy strategy by randomly selecting a model as a “medoid” which reduces the measure associated with the absolute error criterion [148] representing the sum of dissimilarities. Here all models

present in a cluster are merged. Alternatively, models are merged from using their property of markov equivalence producing an essential model. These have similarities with ideas appearing in [35, 36]. The algorithm described in Step 11 has been utilized to complete the definition of merged model. PAM (partitioning around "medoids"), a medoid based clustering algorithm which has been cited in our work is less influenced by outliers. Our representation of record instances for a time interval with a provenance graph model reduces the cluster space and makes it feasible for application of PAM. For large data sets, a sampling based method called CLARA can be used where after sampling the clustering methodology PAM has been applied to detect the best "medoid". A "medoid" based approach to clustering of points described in PAM, CLARA or CLARANS where two clusters has been be merged based on the farthest distance between two points in the cluster pair and this tested merged cluster satisfies less than a certain threshold value for the diameter measure. The revised centroid point has been identified from merging of clusters. This centroid point has a maximum distance measure value to a point in the merged cluster and satisfies a magnitude less than a threshold value in radius measure. The above description summarizes the database clustering methods such as PAM, CLARA, CLARANS which are relevant to our work.

Also for the cause of merging models, the "medoid" model in a cluster has been selected as the initial model. The other present models in the cluster has been merged iteratively such that the intermediate model minimally increases the error defined based on a maximum likelihood measure as has been described in [149]. Alternatively a Bayesian Model Merging principle has been adopted such that product of model prior and likelihood measure associated with the models has a maximum value [149]. Also the models can be ordered based on their relevance to a subject or topic using principles of Bayesian Factor or Bayesian sampling approaches as has been described in [150] and the models can be merged in an appropriate order within a cluster. Here a unified provenance graph model is thus constructed where model unification procedure has been applied for merging element models from incrementally acquired information obtained at more recent time intervals as has been described earlier. A merging of time series data using principles of Dynamic Time warping (DTW) has been described in [151] where the time series pair participating in the merge have been optimally aligned using principles of Dynamic Programming if the lengths of the time series pair have not been observed as same. A cluster has been represented by time series data which have not been considered as similar in nature and a representative time series has been derived from the time series data present in the cluster that considers the DTW distance to identify the closest time series [151]. The above description summarizes related work in merging models within a cluster using Bayesian Factor, or using methods for Time Series Merging based on Dynamic Time warping.

A hierarchical agglomerative clustering approach appearing in [11] has been adopted with merging of most similar model pairs at every hierarchy where a merged model has been derived using a procedure described later in Steps 9.2, 9.3 and 11. BOAT uses attribute selection method like Gini index which has been used for constructing for Regression Trees.

Boosting is a method of combining ensemble classifiers created from a weighted version of learning sample, where weights have been adjusted at each step to provide

increased weight to cases misclassified earlier. Adapting re-sampling has been identified as the key to success with misclassified cases receiving larger weights in the next step. Bagging has been applied to larger trees in contrast to boosting that works well with stumps or slightly larger trees [152]. The idea of combining ensemble of classifiers using a weighted version of each as has been described in boosting and this has interpretation of relevance to our work where earlier steps of learning have been provided more weight-age than those derived later and the significant classifiers thus derived have more weight-age value. Also the start nodes describing a topic in the Phrase Reinforcement Learning in learning of topic description as has been described in [145]. The above description summarizes the relevance of Boosting over Bagging for combining ensemble classifiers.

A reinforcement learning approach has been described as providing a balance between pruning for generalization and growing deeper trees for accuracy [153]. A continuous U tree algorithm transfers traditional U tree algorithm to reinforcement learning and this U tree algorithm can be viewed as a Regression Tree algorithm for storing state values [154]. The regression clustering (CART) with splits satisfying a maximum gain measure as modelled by ginni coefficient describes a probabilistic measure modelling the fraction of points of the predecessor nodes that are present in the one or the other successor node. The above description summarizes the relevance of U tree algorithm and CART algorithm to our work.

Structural Regression Trees integrates the regression method of learning into inductive logic programming [155]. This method however produces a solution to the Relational Regression Problems which have been difficult to understand, and assumes that all features are equally relevant to all parts of instance space, and also does not have easy utilization of domain space [155]. The SRT method has a simple method of tree selection based on Minimum Description Length (MDL) [113] principle. The MDL algorithm measures the simplicity and accuracy of the theory and data. The theory description length has been derived from encoding of literals and encoding of the predicted values in leaves [155]. The data length has been derived from encoding of errors [155]. A model has been selected with minimum message length associated with the sum of theory message length and data message length of the model [155]. The balance provided by Regression Tree approach [153] with the methodology of error complexity pruning and growing deeper tree for accuracy has similarity with the MDL based approach to learning as has been described in [155]. The interaction network summarization has been described with independent topical events that are temporally and topically coherent and this interaction network has been summarized by large events [156]. A collection of k-events has been selected that maximizes the node coverage and this task maps to the finding the maximum set cover solution [156]. The above description summarizes related work on the MDL principle and maximum cover set solutions for representing models.

Our approach to provenance graph structure and classifier learning

Step 1: An event or news report may be associated with more than single story. The time order of occurrence of events and their similarities is derived from a joint entropy distance measure linking the events. This measure has been used to cal-

culate the dependency between these events. The provenance graph of events can thus be described. Only those events are considered that have one or more of the necessary seed terms used to describe a topic of interest to the user. Here the topic specific words that remove other words commonly occurring across topics are identified. The topic words and the entity words together qualify in determining topic or sub-topic association of documents or news reports. The path to the final goal content or event or topic node represents the learned path to the recognized goal topic node. Thus from the description in related work in topic or story or event and from the brief description that appears in this work we have provided a rationale for constructing a provenance graph from events from establishing links between these events or stories related to these events.

Step 2: This information is also utilized to define the probabilistic weight of link connecting any two states or nodes and a distance measure separating these nodes present in the graph from applications of Eqs. 12–17.

Step 3: The probabilistic weight connecting two nodes in a graph is utilized to cluster the nodes in the graph using application of Steps 1 to 4.

Step 4: Only sub-goal states of each cluster are identified using application of method described in an earlier section on our approach to clustering of nodes of this work.

Step 5: Bayesian methods utilize these probabilities defined for this revised graph obtained at step 4 to build the graphical structure satisfying a constraint such as mdl, bd, bic for all nodes present in a provenance graph from the dependencies describing the discovered event or topic or story model graph. Here more than one provenance graph models may be produced from the application of this procedure to a time window. Here interpreting clustering of documents to represent model cluster and bayesian probabilistic approach of assignment of document to a model cluster as described in [143] are relevant

Step 6: The classifier policy path is detected from this structure by application of Q Learning algorithm for learning provenance path as described in an earlier section. Q Learning procedure has earlier been described to produce a plan of event or topic executions from the start topic or event to the goal topic or event. Many such alternative plans may be generated from this procedure which may be optimal or suboptimal.

Step 7: Provenance Graph model(s) are constructed for a time window determined from window length and refresh rate. The time windows may also be defined based on a concept of sliding window with overlaps between the models defined for these time windows. The time stamp of stories describing an event may cause a more detailed definition of an event or a topic at a later time window sometimes with some overlap between these definitions of an event or topic causing an overlap between the provenance graph models described for these time windows.

Step 7.1: A collection of classifiers has been identified in every model with this training information. Calculate a distance measure from these policy classifiers in a model from all policy classifiers in every other model using the SEC measure (distance as reciprocal of similarity value) concept for paths. Calculate a weight measure associated with the

identified classifiers in every model based on these similarity measures, where classifiers more similar to others have stronger weights. This weight can also be determined from application of Eqs. 15 and 16. These identified classifiers which provide cover for all the classifiers [33] are considered for model merging.

Step 8: This classifier weight adjustment for classifying "difficult" data is also included in the procedure for reward calculation in this work. This procedure also can provide insights into split or split-merge in topic/story definition. The reward calculation procedure as described earlier may also be sufficient in determining "important" or "difficult" with identifying edges or links of distinctive relevance data for classifying purposes.

Step 9: The rationale for merging or linking provenance graph models has been provided in earlier steps.

Step 9.1: Use this weight measure to compare the significance of the collection of classifiers detected in each model. Only those policy path classifiers which are more or most significant or those providing a minimum cover set for all classifiers are retained. A classifier in a model is compared with all classifiers in every other model learned and the best match is considered for calculating significance of a classifier. The procedure is applied to generate a reduced collection of classifiers that provides a cover set for all classifiers. Each one of these significant classifiers present in the minimum cover set and which are associated with one or more model(s) are compared with all such significant classifiers which are associated with other such model(s) and a similarity measure is computed for these models from their classifiers using a correlation ratio. Alternatively, this similarity measure can be calculated as Pearson Product moment correlation measure between a feature vector of edges describing a classifier associated with a model with those of another model. This is derived as the ratio of the covariance measure calculated between these vectors of values and the product of variance measures calculated for the individual vectors of values. A covariance measure defined between a single value and a vector of values is the sum of the covariance measures calculated between the single value and each element of the vector. Also the covariance measure is defined to have a commutative property that is used for this calculation. Only those model pairs are candidate for merging where this similarity measure exceeds a certain threshold value. Alternatively a distance or similarity measure may be described between a pair of these models using a graph edit distance measure. The models are clustered using pair-wise similarity or distance measures between a model pair as described earlier. A hierarchical agglomerative approach can be adopted for this with merging of most similar model pairs at every hierarchy where a merged model is derived using procedure a described later in Steps 9.2, 9.3 and 11.

Step 9.2: A partitioning may be adopted for this model merging using the K-medoids approach where a model is selected as a representative for the cluster of models and this clustering method is realized with a partitioning around medoids (PAM) or CLARA or CURE. The algorithm which starts with randomly collected seed models and improves the clustering with a greedy strategy by randomly selecting a model as a medoid which reduces the measure associated with the absolute error criterion representing the sum of dissimilarities. Here all models present in a cluster are merged using procedure described in Step 11. The ideas described in this paper where the graph models are first partitioned using a partitioning approach like CLARA or PAM or CURE and thus identified representative models in each cluster are merged using a Bayesian scoring approach. Alternatively, models are merged from using their property of markov equivalence producing an essential model. The algorithm described in Step 11 is required to complete the definition of merged model.

Step 9.3: Please refer steps 9.3.1, 9.3.2 and 9.3.3 below.

Step 9.3.1: A system of similar models detected in Step 9.1 forming a cluster can be merged using application of procedure defined as follows on a Bayesian scoring approaches. Here tasks unit steps or activities are linked or connected that leads to maximum improvement in the selected Bayesian score metric value and the path describing the changes to the graphical structure is determined from application of hill climbing, or simulated annealing or TABU search as described in [102] with random restarts for avoiding the solution from getting trapped in local minima.

Step 9.3.2: Alternatively the identified system of significant models represented with provenance graphs can be merged using techniques briefly described here. Here for graphs that are Markov equivalent and hence similar, a composite graph is constructed as the essential graph. Here the essential graph thus constructed has trust weight link connecting common nodes in the models which are candidates for merging and this link weight is defined using a procedure described in Step 11.

Step 9.3.3: The best classifier with maximum probability of selection for the clustered model is identified from application of Eqs. 13—16. Here the probabilistic trust weight link connecting any two nodes in the merged provenance graph, is derived as the weighted average of probabilistic trust values associated with common links or edges connecting the same pair of nodes corresponding to similar classifiers where weights are determined from Steps 2 and 3 of this algorithm. This approach is also applicable for computing weighted average of expected Q values of nodes with weights determined from the sample or model from where this information is retrieved.

The candidate models participating in the procedure for this merging have policy path classifiers which are either only member of any set of similar classifiers or are members in the minimum cover set of such similar classifiers that are present in the cluster of models. Also for the cause of merging models, the medoid model in a cluster may be selected as the initial model. The other present models in the cluster may be merged iteratively such that the intermediate model minimally increases the error defined based on a maximum likelihood measure [113]. Alternatively a Bayesian Model Merging principle may be adopted such that product of model prior and likelihood measure associated with the model is maximized. Also the models may be ordered based on their relevance to a subject or topic using principles of Bayesian Factor or Bayesian sampling approaches. The models may be merged in this order within a cluster. Here a unified provenance graph model is thus constructed where model unification procedure may be applied for merging element models from incrementally acquired information obtained at more recent time intervals as has been described earlier.

Step 10: The final classifier accuracy obtained from the similar models participating in merging is the vector of weighted average of the selected classifier feature edge weights that are common to the classifiers that are getting merged. This measure indicates the probability of selecting a candidate classifier (accuracy) derived from the system of significant classifiers.

Step 11: Alternatively, previously calculated weights associated with features describing the nearest subtopic or sub-story or goal or topic node obtained using method described in this work can indicate a change of policy where previously suboptimal action can become optimal and vice versa. A Q Learning procedure is applied to the revised provenance graph model derived after merging of the significant models. If a newly calculated policy path is not destined to the same goal topic or event node then it is removed from farther consideration. Alternatively, the expected gain from executing action at a state is the difference between the expected Q value as reward value calculated from Eq. 16 from executing the changed action and earlier Q value associated with taking optimal action at the state. The Value of Perfect Information (VPI) associated with taking action at a state is the weighted sum of expected gain measure calculated for all discrete probabilities associating the state, action pair which separates the best classifier policy value and the considered classifier policy value. Here a strategy is selected that maximizes the sum of expected Q value for a state action pair and Value of Perfect Information associated with state action pair The alternatives to the best classifiers which when considered are re-ranked from this expected gain measure. The candidate set of classifiers which form the minimum cover set of classifiers or is the only member of a set of classifiers for a unified provenance graph

model are considered for specifying recognition paths for relevant goal topics where these paths may be both general or discerning.

Conclusions

This paper described a reinforcement learning based message transfer model for transferring news report messages through a selected path in a trusted provenance network with the objective of maximizing the reward values based on trust or importance based and network congestion or utility based cost measures. The reward values have been calculated along a dynamically defined policy path connecting start topic or event node to a goal topic or event or issue nodes for incrementally defined time windows for a given network congestion situation. A hierarchy of agents of trusted roles has been used to accomplish the sub-goals associated with sub-story or subtopic in the provenance structure where an agent role has assumed the semantic role of the associated sub-topic. The twitted news story thread or plan of events has been defined in this work from the starting topic or event node to the goal topic or event node for incrementally defined intervals of time. The graphs have been clustered into subtopic and these sub-goals or sub topic nodes of a topic node at every level of granularity are associated with cluster of news reports which describe activities associated with sub-goal or sub-topic events. The policy path in a topic or story graph model has been defined by applying reinforcement learning principles [26] on dynamically defined event models associated with evolution of topic definition observed from incrementally acquired samples of input training data spanning multiple time windows. We have provided a methodology for unifying similar provenance graph models for adapting and averaging the policy path classifiers associated with individual models to produce a reduced set of unified models derived during training. A minimum set cover of classifiers has been identified for the models and a clustering procedure of the models has been suggested based on these classifiers. The methodology described in this work has been detailed for news reports modelling application. We aim at developing this methodology for application to econometrics in our future work.

Earlier works had only considered the provenance, computational trust, topic or subtopic user roles and data importance issues in isolation while this work integrates the considerations into a unified approach. The approach presented in our paper is novel to the news reports description and modelling problem for the above consideration. The Q Learning approach adopted is unique to our work in the news reports modelling application. Our Record Merging approach and calculation of net trust value at a database relation site hosted at a node on records arriving from other network node sites together with the records already present at the site or on the new event influencing the record at the site is unique to the news reports description and modelling application. The derivation of macro action(s) and the computation of reward(s) associated with these actions is unique to this work in producing a reduced clustered space in the news reports modelling application. The approach described in merging of event or topic models for identifying similar or discerning threads of news reports or messages over incremental time periods is unique to our work.

Authors' contributions
The work has been carried out by SKM under the supervision of SB. Both authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no any competing interests.

Availability of data and materials
The manuscript proposes new research ideas with empirical proof and hence the availability of data and material is not applicable.

Consent for publication
Both the authors provide their consent for the publication.

Ethics approval and consent to participate
Sanjoy Kumar Mukherjee does not receive any fellowship from any source. The work also does not involve any specific data or materials. Hence the Ethics approval and consent to participate is not applicable.

Funding
Not applicable.

Publisher's Note

Received: 30 September 2016 Accepted: 11 October 2017
Published online: 26 October 2017

References

1. Smart. Knowledge-intensive fusion for situational awareness: requirement for knowledge filtered awareness. Technical Report, University of Southampton, Web and Internet Science ePrintId 261067; 2005.
2. Malik, Gehani, Tarique, et al. Sketching distributed data provenance. Book Chapter, Data Provenance and Data Management in e-science. Berlin: Springer; 2013.
3. Deutsch, Milo, Roy, Tannen. Circuits for DataLog Provenance. In: ICDT; 2014. p. 201–12.
4. Bates, Mood, Valafar, et al. Towards secure provenance based access control in cloud environments. In: Proceedings of CODASPY; 2013. p. 277–84.
5. Green, Karvounarakis, Ives, Tannen. Update exchange with mappings and provenance. In: Proceedings of VLDB; 2007. p. 675–86.
6. Nies, Taxidou, et al. Towards multi-level provenance reconstruction of information diffusion in social media. CIKM; 2015. p. 1823–6.
7. Magliacane, Groth. Repurposing Benchmark Corpora for reconstructing provenance. In: CEUR Workshop Proceedings, SePublica, vol. 994; 2013. p. 39–50.
8. Groth, Gill. Requirements for provenance on the web. IJDC, vol. 7(1); 2012. ISSN 1746–8256.
9. Steyvers, Smythe, et al. Combining background knowledge and learned topics. Topics in cognitive science 3.1; 2011. p. 18–47.
10. Boutilier, Dearden. Using abstractions for decision theoretic planning with time constraints. In: Proceedings of the 12th national conference on AI, vol. 2; 2014. p. 1016–22.
11. Zin. Knowledge based social network applications to disaster event analysis. In: International MultiConference of engineers and computer scientists, vol. 1; 2013.
12. Mukherjee SK, Neogy S. Storage & retrieval of trusted information: a temporal probabilistic database approach. In: IEEE, 2015 third international conference on computer, communication, control and information technology (C3IT); 2015. p. 1–15.
13. German Research Center for AI. Glocal: event based retrieval of network media. D 3.3, Global Dynamics; 2010.
14. Gao, Li, et al. Joint topic Modeling for event summarization across news and social media streams. CIKM; 2012. p. 1173–82.
15. Hu Y, John A, et al. ET-LDA: joint topic modeling for aligning events and their twitter feedback. In: AAAI; 2012. p. 59–65.
16. Brigadir, Green, Cunnigham. Adaptive representation for tracking breaking news on twitter. Cornell University Library, arXiv:1403.2923v3 [cs.IR]; 2014.
17. Atefeh F, Khreich W. A survey of techniques for event detection in twitter. Comput Intell. 2013;31(1):132–64.
18. Mukherjee SK, Bandyopadhyay S. Clustering to determine predictive model for news reports analysis and econometric modeling. In: 2015 IEEE 2nd International Conference on Recent trends in information systems (ReTIS); 2015. p. 302–9.
19. Lee, Mikolov. Distributed representations of sentences and documents. In: ICML; 2014. p. 1188–96.
20. Schalkoff C. Intelligent systems: principles, paradigms and pragmatics. Burlington: Jones and Bartlett; 2011.
21. Nallapati R, Feng A, et al. Event threading within news topics. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management 2004. p. 446–53.

22. Wang, Zang, Ru. Automatic online news issue construction in Web Environment. In: Proceedings of WWW; 2008. p. 457–66.
23. Park, Lee, Song. Aspect-level News browsing: understanding news events from multiple view points. In: Proceedings of IUI; 2010. p. 41–50.
24. Chiu, Soo. Sub goal identifications in reinforcement learning: a survey. In: Advances in Reinforcement Learning, INTECH; 2011.
25. Humphrys. Action selection methods using reinforcement learning, from animals to animats 4; 1996. p. 135–44.
26. Bagherjeriyan, Eick, Vilalta. Adaptive clustering: better representative with reinforcement learning. Houston: Department of Computer Science, University of Houston; 2005.
27. Lieprens G, Palmer M. Alternatives for classifier system credit assignment. Proc IJCAI. 1989;1:756–61.
28. Hsu KW. On adjustment functions for weight-adjusted voting-based ensembles of classifiers. JCP. 2014;9(7):1547–52.
29. Tsoumaakas G, Katakis I, et al. Effective voting of heterogeneous classifiers. ECML. 2004;2004:465–76.
30. Valdovinos, Sanchez. Combining multiple classifiers with dynamic weighted voting. HAIS; 2009. p. 510–6.
31. Jing Jiang. Trust mining and analysis in complex systems. Doctoral Thesis. Auckland University of Technology; 2014.
32. Kumar S, Toshniwal D, et al. A comparative analysis of heterogeneity in road accident data using data mining techniques. Evolving Systems. Berlin: Springer; 2016.
33. Balachandar SR, Kannan K. A meta-heuristic algorithm for set covering problem based on gravity. Int J Comput Math Sci. 2010;4(5):223–8.
34. Liu, Chen. TopicPanorama: a full picture of relevant topics. In: VAST; 2014. p. 183–92.
35. Kamvar, Klein, Manning. Interpreting and extending classical agglomerative clustering algorithm using a model based approach. Stanford; 2002.
36. Guha S, Rastoggi R. Shim: "CURE: an efficient clustering algorithm for large databases". ACM Sigmod Record. 1998;27(2):73–84.
37. Ceolin, Groth. Trust evaluation through reputation and provenance analysis. In: Proceedings of 8th international workshop on uncertainty reasoning for the semantic web, vol. 900; 2012. p. 15–26.
38. Huang, Huang. Optimized event storyline generation based on mixture-event-aspect model. EMNLP; 2013. p. 726–35.
39. Acar, Ahmed, et al. A core calculus for provenance. POST 7215; 2012. p. 410–29.
40. Sutcliffe A, Wang D. Computational modeling of trust and social relationship. J Artif Soc Soc Simul. 2012;15(1):3.
41. DuBois, Golbeck. Improving recommendation accuracy by clustering social networks with trust. Recomm Syst Soc Web. 2009;532:1–8.
42. Barry, Keilbling, Perez. Hierarchical solution of large markov decision processes. In: Association for Advancement of Artificial Intelligence; 2010.
43. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge: MIT Press; 1998.
44. Tavakolifard. Similarity based techniques for trust management. In: Web Intelligence and Intelligent Agents, InTech Published; 2010.
45. Kim, Kim. A peer-to-peer workflow model for distributing large scale workflow data onto Grid/P2P. J Dig Inform Manag. 2005;3:2.
46. Fu, Liu. Research on qos routing algorithm in adhoc networks based on Reinforcement Learning. Electronics and Electrical Engineering, ISSN 1392-1215, vol. 19(2); 2013.
47. Chang, Ho, Keilbling. Mobilized Ad Hoc networks: a Reinforcement Learning approach. In: Proceedings of the international conference on autonomic computing; 2004. p. 240–7.
48. Forouzan. Data communications and networking. 4th ed. New York: McGraw Hill; 2000.
49. Attelveldt, Welbers. LDA models topics… But what are 'topics'? Glasgow Big Data; 2014.
50. Thuc, Mejova, et al. A relevance based topic model for news event tracking. In: SIGIR; 2009. p. 764–5.
51. Zhou, Mao, Loo, Abadi. Unified declarative platform for secure networked information system. In: Proceedings of ICDE; 2009. p. 150–61.
52. Ibbotson. Provenance: enabling and supporting provenance in grids for complex problems. In: Normal Research Project. England: ECS, University of Southampton; 2006.
53. Lim, Moon, Bertino. Provenance based trustworthiness assessment in sensor networks. In: Proceeding of DMSN; 2010. p. 2–7.
54. Urbano. A situation aware and social computational trust model. PhD thesis. Informatics Engineering, University of Porto; 2013.
55. Gao, Wu, Shang. Clustering with XCS on complex structure dataset. AI Adv Artif Intell. 2008;2008:489–99.
56. Mahadevan S. Average reward reinforcement learning: foundations, algorithms, and empirical results. Mach Learning. 1996;22(1):159–95.
57. Singh SP. Transfer of learning by composing solutions of elemental sequential tasks. J Mach Learn. 1992;8(3–4):323–39.
58. Kwok C, Fox D. Reinforcement learning for sensing strategies. Intell Robots Syst. 2004;4:3158–63.
59. Marsh. Formalising trust as a computational concept. Ph.D. Thesis, Computing Science & Mathematics e-Theses. Stirling: University of Stirling; 1994.
60. Yan ZH, Li F. Thread labeling for news events. J SJ Univ (Sci). 2013;18(4):418–24.
61. Yan ZH, Li F. Thread labeling for news event. J Shanghai Jiaotong Univ. 2013;18(4):418–24.
62. Li, Wang, McCallum. A Continuous-time model of Topic Co-occurrence Trends. In: Defense Technical Information Centre; 2006.
63. Welser HT, Gleave E, et al. Visualizing the signatures of social roles in online discussion groups. J Soc Struct. 2007;8(2):1–32.
64. Taylor K, Murthy J. Implementing Role based access control for federated information systems on the web. Proc ACSW Front. 2003;21:87–95.

65. Gureth, Groth, Stadler, et al. Assessing link data mappings using network measures. In: The semantic web: research and applications; 2012. p. 87–102.
66. Li, Wang, et al. Exploring temporal relations for event tracking using hierarchical dirichlet trees. Wuhan University, arXiv: 1312.2244v3 [cs.CL]; 2017.
67. Kaur, Gupta. A survey of topic tracking system. IJARCSSE. 2012;2(5):384–93.
68. Wang W, Jiang M, Yuan D. A fast hierarchical topic detection method. J Converg Inform Technol (JCIT). 2012;7(22):517.
69. Sauper, Barzilay. Automatic aggregation by joint modeling of aspects and values. JAIR; 2013.
70. McGuiness, Zeng, et al. Investigations into trust for collaborative information repositories: a Wikipedia Case Study. MTW, vol. 190; 2006.
71. Ahmed A, Ho Q, et al. Unified analysis of streaming news. In: Proceedings of the 20th international conference on World wide web; 2011. p. 267–76.
72. De Smet W, Moens MF. An aspect based document representation for event clustering: sa-ot accounts for pronoun resolution in child language. LOT Occas Series. 2009;14:55–68.
73. Liu W, Wang D, et al. A sub-topic partition method based on event network. In: The seventh international conference on internet and web applications and services. 2012. p. 194–9.
74. Cardona-Rivera, Cassell, et al. Indexter: a computational model of the event-indexing situation model for characterizing narratives. In: Workshop on computational models and narratives; 2012. p. 34–43.
75. Bögel T, Gertz M. Time will tell: temporal linking of news stories. In: Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries; 2015. p. 195–204.
76. Lakshmi, Mukherjee. Using cohesion model for story link detection. IJCSNS. 2007;7(3):59–66.
77. Hu P, Huang ML, Zhu XY. Exploring the interactions of storylines from informative news events. J Comput Sci Technol. 2014;29(3):502–18.
78. Caswell D, Russell F, et al. Editorial aspects of reporting into structured narratives. In: Proceedings of the 2015 Computation + Journalism Symposium; 2015.
79. Wang, Li. Story link detection based on event words. In: LNCS 6609; 2011.
80. Zheng W, Zhang Y, et al. Topic tracking based on keyword dependency profile. Information Retrieval Technology; 2008. p. 129–40.
81. Kumaran, Allan. Using names and topics for new event detection. In: HLT; 2005. p. 121–8.
82. Mei, Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: KDD; 2005. p. 198–207.
83. Yang, Zhang, Carbonell, et al. Topic conditioned novelty detection. In: ACM SIGKDD; 2002. p. 688–93.
84. Wang C, Thiesson B, Meek C, Blei D. Markov topic models. In: Artificial intelligence and statistics; 2009. p. 583–90.
85. Rajani, McArdle, et al. Extracting topics based on authors, recipients and content in microblogs. ACM SIGIR; 2014. p. 1171–4.
86. Shah C, Eguchi K. Improving document representation for story link detection by modeling term topicality. IPSJ Online Trans. 2009;2:27–35.
87. Tang, Wu, et al. Sketch the storyline with charcoal: a non parametric approach. IJCAI; 2015. p. 3848.
88. Kaur K, Gupta V. A survey of topic tracking techniques. Int J. 2012;5.
89. Liu, Wang, et al. A sub-topic partition method based on event network. ICIW; 2012. p. 194–9.
90. He X, Rekatsinas T, et al. Hawkestopic: a joint model for network inference and topic modeling from text-based cascades. In: International conference on machine learning; 2015. p. 871–80.
91. McGovern, Sutton, Fagg. Roles of macro action in accelerating reinforcement learning. In: Proceedings of the 1997 Grace Hopper Celebration of Women in Computing, vol. 1317; 1997.
92. CrossNo, Wilson, et al. TopicView: understanding document relationship using latent dirichlet association methods. United States, National Nuclear Security Administration; 2011.
93. Dalamagas. NHS: a tool for the automatic construction of News Hyper Text. BCS-IRSG; 1998.
94. Kowsika, Maheswari, et al. Context specific event models for news articles. Cornell University Library, arXiv: 1308.0897v1 [cs.CL]; 2013.
95. Menache I, Mannor S, Shimkin N. Q-cut-dynamic discovery of sub-goals in reinforcement learning. ECML. 2002;14:295–306.
96. Simmons MP, Adamic LA, et al. Memes online: extracted, subtracted, injected, and recollected. ICWSM. 2011;11:17–21.
97. Goel S, Huber M. Subgoal discovery for hierarchical reinforcement learning using learned policies. In: FLAIRS conference 2003. p. 346–50.
98. Rusch T, Hofmarcher P, et al. Model trees with topic model preprocessing: an approach for data journalism illustrated with the wikileaks afghanistan war logs. Ann Appl Stat. 2013;7(2):613–39.
99. Pal, Gilliam. Set based similarity measurement and ranking model to identify case of journalistic reuse. CL!NSS task at FIRE; 2013.
100. Chakraborty. Big data analytics for development: event, knowledge graphs and predictive models. PhD Thesis. Computer Science, New York University, Pro Quest Dissertation Publishing; 2015. p. 3740799.
101. Zhu Z, Wang P, et al. Network topic detection model based on text reconstruction. Ljublijana: Informatica; 2013. p. 367–72.
102. Gendreau, Potvin. Tabu Search. Hand book of MetaHeuristics, vol. 57. Springer Science and Business Media; 2006.
103. Himelboin, Smith, et al. Classifying twitter topic networks using social network analysis. In: Social Media + Society, 3.1; 2017. p. 2056305117691545.
104. Ahmed, Xing. Timeline: a dynamical hierarchical Dirichlet process model for recovering birth/death and evolution of topic in topic stream. UAI; 2010. p. 20–9.
105. Wang, Chen, et al. Targeted topic modeling for focused analysis. In: KDD; 2016. p. 1235–44.
106. Obispo. Tspoons: tracking salience profiles of online news stories. Master's Thesis. Computer Science, California Polytechnic State University; 2014.

107. Sato, Akaishi, Hori. Topic bridging by identifying the dynamics of spreading topic model. In: Intelligent interactive multimedia: systems and services; 2012. p. 619–27.
108. Rossi. Data mining: methodology and algorithms. In: Data mining concepts, models, methods and algorithms, ISBN 978-0-470-89045-5. New York: Wiley; 2011.
109. Najafabadi MM, Villanustre F, et al. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1.
110. Narayanan. Knowledge based action representation for metaphor and aspect. PhD Thesis, Engineering: computer science, University of California Berkley; 1997.
111. Knights, Mozer, et al. Detecting topic drift with compound topic model. ICWSM; 2009. p. 242–5.
112. Kanoun, Schaar. Big-data streaming applications scheduling with online learning and concept drift detection. In: Proceedings of DATE; 2015. p. 1547–50.
113. Buntine. A guide to the literature on learning probabilistic networks from data. IEEE Trans Knowl Data Eng. 1996;8(2):195–210.
114. Buntine. Operations for learning with graphical models. J Artif Intell Res. 1994;2(1):159–225.
115. Heckerman, Geiger, Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. J Mach Learning. 1995;20(3):197–243.
116. Chickering. Learning equivalence classes for Bayesian network structures. J Mach Learn Res. 2002;2:445–98.
117. Chickering, Heckerman, Meek. A Bayesian approach to learning Bayesian network with local structure. UAI; 1997. p. 80–9.
118. Vincente Il. An emergent architecture for scaling decentralized communication system. PhD Thesis. Electrical Engineering, Columbia University; 2011.
119. Atiya, Parlos, Kingber. A reinforcement learning method based on adaptive simulated annealing. In: IEEE 46th midwest symposium on Circuits and systems, 2003; 2013.
120. Gaama J, Zliobaite I, Bifet A, et al. A survey on concept drift adaptation. ACM Comput Surv (CSUR). 2014;46(4):44.
121. Guo, Liu, Malec. A new Q Learning algorithm based on the metropolis criterion. SMCB Part B. 2004;34(5):2140–3.
122. Partalas I, Tsoumakas G, et al. Pruning an ensemble of classifiers via reinforcement learning. Neurocomputing. 2009;72(7):1900–9.
123. Chickering DM. Personalizing influence diagrams: applying online learning strategies to dialogue management. J User Model User Adapted Interaction. 2007;17(1):71–91.
124. Buezas. Constraint based modeling of minimum set covering: application to species differentiation. Mater Thesis. Lisboa: European Master of Computational Logics; 2010.
125. Justice D, Hero A. A linear formulation of the graph edit distance for graph recognition. Ann Arbor. 2005;1001:48109.
126. Caswell, Russell, et al. Editorial aspects of reporting intro structured narratives. In: Computational + Journalism Symposium; 2015.
127. Brigadir, Greene, et al. Adaptive representations for tracking breaking news on twitter. Cornell University, arXiv:1403.2923v3 [cs.IR]; 2014.
128. Ahmed, Ho, et al. Online inference for the infinite topic-cluster model: storylines from streaming text. In: PMLR, 15; 2011. p. 101–9.
129. Keane, Yee, Zhou. Using topic modeling and similarity thresholds to detect events. In: Workshop on EVENTS at the NAACL-HLT; 2015. p. 34–42.
130. Schanda, Sanderson, et al. Examining new event detection. Australasian Document Computing ACM; 2014. p. 26.
131. Zhao, Mao, et al. Unified declarative platform for secure networked information systems. ICDE; 2009. p. 150–61.
132. Li Z, Wang B, Li M, Ma W-Y. A probabilistic model for retrospective news event detection. Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. SIGIR '05, Salvador, Brazil. New York: ACM; 2005. p. 106–13.
133. Primo C, Ferro CA, Jolliffe IT, Stephenson DB. Combination and calibration methods for probabilistic forecasts of binary events. Month Weather Rev. 2009;137:1142–9.
134. Connely FM, Candinin DJ. Stories of experience and narrative enquiry. SAGE J. 1990;19(5):2–14.
135. Hang CW, Wang Y, Singh MP. Operators for propagating trust and their evaluation in social networks. AAMAS. 2009;2:1025–32.
136. Hao, Wang. Sensor networks routing via Bayesian exploration. In: LCN; 2006. p. 954–5.
137. Dekhiyar A, Ross R, Subramaniam. Probabilistic temporal databases 1. ACM Trans Database Syst. 2001;26(1):41–95.
138. Chou TC, Chen MC. Using incremental PLSI for threshold resilient online event analysis. IEEE Trans Knowl Data Eng. 2008;20(3):289–99.
139. Goshtasby A. Similarity and dissimilarity measures Image registration. London: Springer; 2012. p. 7–66.
140. Anand, Bowers, Ludascher. Database support for exploring scientific workflow provenance graph. In: SSDBM; 2012. p. 343–60.
141. Ziegler CN, Lausen G. Propagation models for trust and distrust in social networks. Inform Syst Front. 2005;7(4–5):337–58.
142. Zhou, Zu, He. An unsupervised bayesian modeling approach to storyline detection from news articles. In: EMNLP; 2015. p. 1943–8.
143. Zhang, Gharamani, et al. A probabilistic model for online document clustering with applications to novelty detection. In: Advances in neural information processing systems; 2005. p. 1617–24.
144. Wanner, Stoffel, et al. State of the art report of visual analysis for event detection in text data streams. In: Computer graphics forum, vol. 33. 2014.
145. Shaikh GR, Padulkar DM. A survey of template based abstract summarization of twittter topic using ensemble SVM with Speech act. IJERT. 2013;2(11):37–47.
146. Netralova. Modelling and simulation of trust evolution. Technical Report No. DCSE/TR-2006-02, Computer Science & Engineering, University of West Bohemia, Czech Republic May; 2006.

147. Anand, Bowers, Ludascher. Database support for exploring scientific workflow provenance graph. SSDBM; 2012. p. 343–60.
148. Han J, Pei J, Kamber M. Data mining: concepts and techniques. London: Elsevier Inc; 2012.
149. Stolcke A, Omohundro SM. Best-first model merging for hidden Markov model induction. arXiv preprint cmp-lg/9405017; 1994.
150. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009;46(4):1004–17.
151. Kumar S, Toshniwal D. A novel framework to analyze road accident time series data. J Big Data. 2016;3(1):8.
152. Sutton. Classification and regression trees, bagging and boosting. In: Handbook of Statistics Volume 24; 2005. p. 303–29.
153. Garlapati, Ragunathan, et al. A reinforcement learning approach to online decision tree learning. Cornell University Library, arXiv: 1507.06923v1 [cs..LG]; 2015.
154. Uther, Veloso. Tree based discretization for continuous state space reinforcement learning. Aaai/iaai; 1998. p. 769–74.
155. Kramer. Structural regression tree. AAAI/iaai, vol. 1; 1996. p. 812–9.
156. Xia, Rozenntstein, et al. Discovering topically and temporally-coherent events in temporal network. In: ECML PKDD; 2016. p. 690–05.