IJWET
| Volume XX | Issue XX | Month XXXX (2019) ISSN : XXXX-XXXX

Semantic Lifting and Reasoning on the Personalised Activity Big Data Repository for Healthcare Research

Hong Qing Yu and Feng Dong
School of Computer Science and Technology
University of Bedfordshire
United Kingdom

Abstract

The fast growing markets of smart health monitoring devices and mobile applications provide opportunities for common citizens to have capability for understanding and managing their own health situations. However, there are many challenges for data engineering and knowledge discovery research to enable efficient extraction of knowledge from data that is collected from heterogonous devices and applications with big volumes and velocity. This paper presents research that initially started with the EC MyHealthAvatar project and is under continual improvement following the project's completion. The major contributions of the work is a comprehensive big data and semantic knowledge discovery framework which integrates data from varied data resources. The framework applies hybrid database architecture of NoSQL and RDF repositories with introductions for semantic oriented data mining and knowledge lifting algorithms. The activity stream data is collected through Kafka's big data processing component. The motivation of the research is to enhance the knowledge management, discovery capabilities and efficiency to support further accurate health risk analysis and lifestyle summarization.

Keywords: Big Data, Knowledge Discovery, Semantic Web, Ontology, Data Engineering, Data processing, Healthcare.

I. INTRODUCTION

THIS paper illustrates an innovative framework that manages and integrates multiple health-related data resources from wearable sensors, mobile and web applications. The aim of the research is to form an efficient backend platform and technology packages for mining personalised health knowledge which will exert influence on the future direction of people's self-care empowerment, disease prevention and importantly promote better lifestyles.

Currently the smart devices and mobile applications can collect enough data allowing big advantages in shifting medical care from institutions to the home environment, and to transform healthcare from a system that is largely reactive – responding mainly when a person is unwell – to one that is much more proactive in supporting patients in self-management [1]. Recent research evidence shows that patients with chronic conditions who are more actively involved in their own healthcare receive better health outcomes [2]. In addition, self-management skills can be developed and strengthened, even among those who are initially less confident, less motivated or have low levels of health literacy [3]. Therefore, ene of the key factors of success in self-healthcare empowerment is to allow the patient to gain valuable and understandable knowledge from their own data, bringing them tangible benefits.

There has indeed been growing interest in the 'initiative of self-monitoring', evidenced by the sharp market expansion in life-logging devices and apps. The sensors used are capable of constantly monitoring personal health behaviours and activities (e.g. walking, calories, heart rate, and diet), leading to unprecedented opportunities in self-care. Correspondingly, significant research effort has started to harvest and integrate the sensors for long-term health data collections – examples include MyHealthAvatar [4] and MyLifeHub [5]. Such long-term data collection is extremely valuable to individualised disease prediction and prevention, and to promoting healthy lifestyles. Although a wide range of data can be collected from different devices or applications, it is currently still quite difficult for people to discover correlations about themselves. Even for advanced tools like the Withings scale and Fitbit pedometer used to track daily weight and step count it is still not possible to get a clear indication of trends between the two. Nor is it possible to see how they interact within specific timeframes, i.e. days of the week, weekends versus weekdays, month to month, etc, without transposing the data into understandable knowledge [6, 7]. To achieve knowledge understanding and management, the highly heterogeneous and dynamic nature of the data brings new challenges [8, 9].

The major contributions of the paper are:

1) A hybrid framework for combining advanced big data technologies such as Kafka, NoSQL database and MapReduce with Semantic Web technologies such as Ontology and semantic rule language. The hybrid framework supports an efficient backend solution for integrating, processing and mining healthcare related data.
2) A domain specific Ontology is composed with an aim of mapping the raw data to semantics.
3) MapReduce-based event data mining algorithms are introduced for lifting data into higher level semantic repository.
4) Two types of knowledge reasoning rules are also illustrated working on top of this framework to do health condition checking and lifestyle summarisation.

The evaluation processes contain four steps of (1) organising the user focus group, (2) collecting the data from the users, (3) evaluating performance of the system with different settings, (4) analysis the evaluation results (will be discussed in Section V).

1) User focus group: the users are organised by the project partners from different EU countries where the MyHealthAvatar research partners are located. The partners includes healthcare research institutes and hospitals who selected the user group members with considerations of data privacy and security issues. Note that this paper will not touch the privacy or security research topics. However, every user who joined the user focus group has signed the consent form to allow us to processing their data for research purpose.
2) Collecting the data: the final evaluation selected 100 users within the user focus group to provide devices and applications that collected data for the system over 36 weeks.
3) Evaluation settings: For the semantic lifting scalability evaluation, we tested the performance time on the semantic mining/lifting algorithms on two scenarios with different settings of the Hadoop MapReduce environment: (1) increasing the data size (number of weeks) form 0 to 36 weeks but just one user's record and (2) increasing the number of users, but fixed 12-weeks records for all of them.
4) The evaluation results will be discussed in Section V.

The rest of the paper is organized as follows. Section II discusses our research background, research motivations and the literature review on related work. Section III explains the overall architecture of the proposed framework. Section IV introduces the ontology and mining algorithms defined for the semantic data retrieval and reasoning. Section V finally presented the platform interfaces, evaluations and outlines potential future work.

II. BACKGROUND, MOTIVATION AND RELATED WORK

A. Project background

Health and lifestyle-related data can be collected through sensor or mobile application data APIs or directly embedded in the healthcare system. Especially, in the last decade, Body Sensor Networks (BSNs) [10] have been developed to remotely collect data and upload vital statistics to servers over the internet mainly because of the high demands for efficient health monitoring have forced the health and wellness industry to embrace modern technological advances [11]. BSNs can efficiently provide monitored and recorded data, when communicated to suitable systems. The integration of the heterogeneous data is the foundation to support more advanced healthcare-related knowledge discovery and reasoning. The MyHealthAvatar (MHA) framework is a proof–of-concept EU-funded 3 million euros project for providing a digital representation of patient health status. It is designed as a lifetime companion for individual citizens that facilitates the collection of, and access to, long-term health-status information which includes social and sensor data, together with major data resources from traditional healthcare organisations. One of the most important tasks in the project is to efficiently manage the multi-data resources combined into a large dataset and enable discovery of hidden knowledge of the individual user's health condition and lifestyle.

B. Healthcare-related data integration and mining framework

Traditionally, the data integration task is mainly identified as a data warehousing process problem [13]. There are two typical approaches of wrappers and mediators. The goal of a wrapper is to access a source, extract the relevant data, and present such data in a specified format [14]. The role of a mediator is to collect, clean, and combine data produced by different wrappers (or mediators), so as to meet a specific information need of the integration system [15]. However, these techniques are struggling to deal with integration requirements on flexible data structures and are less feasible for datasets that are frequently updated, which is mostly the case since Web Services/Web APIs are mostly applied as data providers nowadays.

The paper [12] presents the Mobile Health Mashups system, a mobile service that collects data from a variety of health and well-being sensors and presents significant correlations across sensors in both a mobile widget as well as a mobile web application. The work focuses on analysing and detailing the technical solution, such as: integration of sensors, how to create correlations between two data sets, and the presentation of the statistical data as feeds and graphs. However, the system only mashes up two data resources – from Fitbit and Withings. The mashup simply federates the two data sets together without any advanced filtering and semantic mapping.

In recent years, the work in data integration research concerns the semantic mapping problem. For example, in the healthcare research domain, [16] proposes to build an interoperable regional healthcare system among several levels of medical treatment organizations, which includes the ontology based approach as the methodology and technological solution for information integration. This research work supports the interoperable regional healthcare system with functions of modularization and expansibility, thus the stability of the system is enhanced by the hierarchy structure. However, the complexity and the size of the data have not been significant reduced in the semantic layer to utilize further knowledge discovery and reasoning because of the graphical data repository features. It suggests that the semantic data integration should have pre-processing steps and only significant data should be lifted into the RDF repositories instead of transforming all the original raw data into semantic layer due to complexity nature of graph data query.

In conclusion, current data integration and mining research work has following drawbacks:

1. Difficult to handle data flexibilities and unstructured data.
2. Difficult to cope with big volume of data by only adding semantic layer or just dynamically mashup the data.
3. Difficult to deal with stream data and mapping them to semantic ontology layer.

C. *Healthcare-related ontology*

There are many medical ontologies that aim to support different medical research tasks of clinical research, trail investigation and biomedical investigation. Clinical Trial Ontology (CTO) and Ontology of Clinical Research (OCRe) [17] are developed to describe methods for binding to external information standards (e.g. BRIDG) and clinical terminologies (e.g. SNOMED CT [19]). These standards allow the indexing of research studies across multiple clinical trials and observational studies, interventions/exposures, outcomes, and health conditions. With such indexing, investigators interested in the evidence pertaining to a particular question (e.g. what is the effect of A on B in people with C) will be able to locate relevant research studies more easily across disparate data sources.

The Ontology for Biomedical Investigations (OBI)[1] is an open access, integrated ontology for the description of biological and clinical investigations. OBI provides a model for the design of an investigation, the protocols and instrumentation used, the materials used, the data generated and the type of analysis performed on it. In OBI the common formal language used is the Web Ontology Language (OWL).

PRotein Ontology (PRO)[2] has been designed to describe the relationships of proteins and protein evolutionary classes (ontology for ProEvo), to delineate the multiple protein forms of a gene locus (ontology for protein forms), and to interconnect existing ontologies. PRO provides an ontological representation of protein-related entities by explicitly defining them and showing the relationships between them. Each PRO term represents a distinct class of entities (including specific modified forms, orthologous isoforms, and protein complexes) ranging from the taxon-neutral to the taxon-specific.

Disease-Treatment Ontology (DTO) [18] is developed to model and represent treatment information found in the abstracts of medical articles. The aim of the DTO is to develop an automatic extraction system to extract treatment information from medical abstracts retrieved from the Medline database, to support information retrieval, question answering, summarization, and knowledge discovery. The purpose of the ontology is to serve as a knowledge base to store the extracted information and support these functions.

Translational Medicine Ontology (TMO) [19] is developed as a unifying ontology to bridging the gap between different terminologies of chemical, genomic and proteomic data with disease, treatment, and electronic health records. TMO demonstrates the usages of Semantic Web technologies for integrating patient and biomedical data, and reveal how such a knowledge base can be applied to support physicians in providing personalized care and recruiting of patients into active clinical trials. Thus, patients, physicians and researchers may explore the knowledge base to better understand therapeutic options, efficacy, and mechanisms of action. The major contribution of TMO is to build semantic links between traditional patient health record (PHR) ontologies and the semantic knowledge based on linked data cloud such as DO [21] and SNOWED CT. However, the TMO still focuses on managing the knowledge of patients' formal health record data without considering the user's daily activity data, which is also important for discovery healthcare knowledge of the individual patient, or healthy user.

III. OVERALL ARCHITECTURE

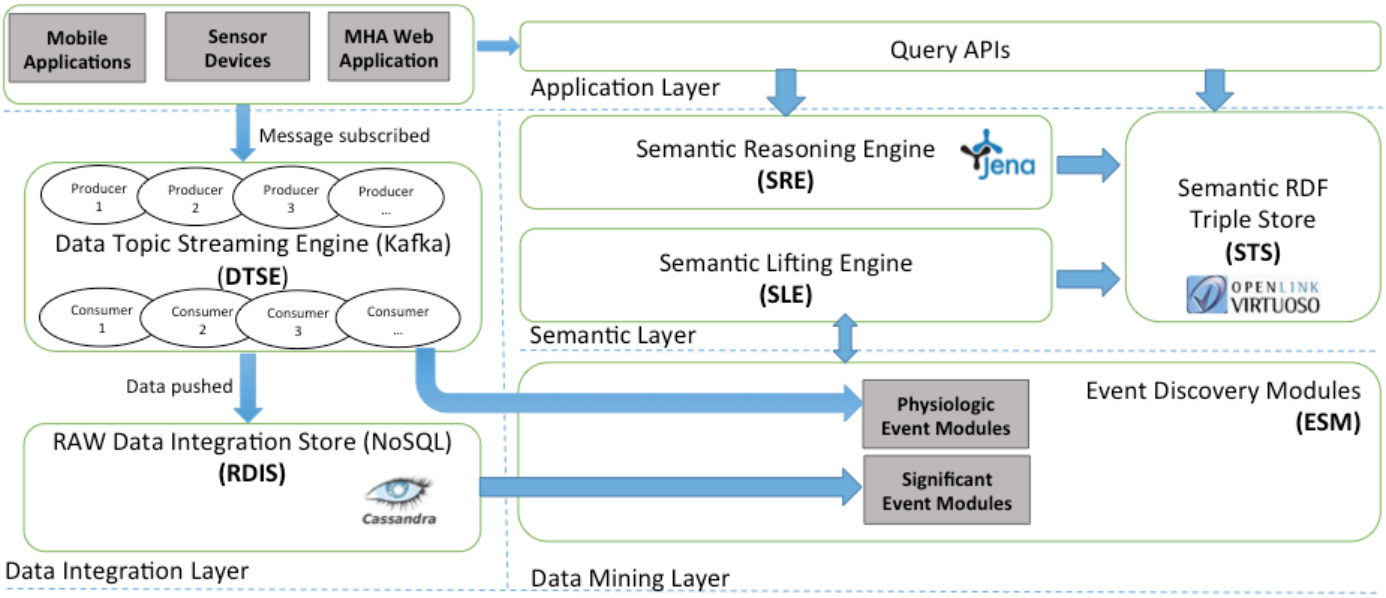


Fig. 1. Overall Framework Architecture

The overall architecture of the MHA semantic lifting and reasoning process can be divided into four layers: Data Integration Layer, Data Mining layer, Semantic layer and Application layer. All layers are introduced in this section, however, since the

[1] http://obi-ontology.org/page/Main_Page
[2] http://pir.georgetown.edu/pro/pro.shtml

components in other layers are standard open source frameworks, i.e. Apache Kafka[3] streaming framework and Cassandra NoSQL repository [23], the major focus of the paper is on the data mining and semantic layers that are discussed in more detail in sections IV, V and VI.

Table.1. Data integration sources and type

Data Collection	Sources	Data type
Steps	Fitbit/Moves/ MHA app	Count
Travelling & activity type	Fitbit/Moves/ MHA app	Minutes & transport type
Location	Moves/ MHA app	Coordination
Diet	MHA app	Calories & food category
Health Profile	MHA app – profile input	PHR like records
Weight	Withings scale	Grams
Body fat	Withings scale	Grams
Blood pressure	Withings BPM	mmHg
Slept hours	Fitbit/Withings	minutes
Awoken times	Fitbit/Withings	Count
Social activity	MHA app – calendar	Description

A. Data Integration

Collecting and meaningfully integrating heterogeneous data resources is a longstanding problem in data management and engineering research areas. In our research, we collect the desired data from multiple data resources including mobile applications (Moves), wearable sensors or digital measuring devices (Fitbit [22, 23] and Withings[4]) and the MHA platform. Each different data resource provides different and useful data information, as Table 1 shows. The data collection process applies Web API technologies following the OAuth security protocol[5]. The core of the data integration process includes two major components: (1) a data topic stream engine that is implemented by Apache Kafka streaming framework to efficiently deal with real-time big data extraction and transform; (2) a column-based NoSQL database (Cassandra) [24] has been developed to mash up the heterogeneous data as whole. Each column family stores a group of rows that contains a set of individual columns in a specific data structuring requirement. For example, one row in the 'activities column' groups all the data columns that store activity type, step counting and duration data elements. The other row in the family can store the location information that the user has travelled to or plans to visit. The 'profile' column family completely focuses on managing basic user profile information such as name and contact.

[3] https://kafka.apache.org/
[4] http://www.withings.com/
[5] http://oauth.net/2/

B. Data mining process

The integrated data contains much noisy data in the data streams, especially the data collected from sensor devices. For example, the positon data is updated frequently according to just a couple of meters difference in the user's movement (even if the user is actually still inside the 'same' place). Therefore, the major purposes of a data mining process are (1) to provide a much smaller amount of data that is more significant and meaningful to understand the user's conditions, and (2) to allow efficient storage of data in the semantic layer for supporting further advanced knowledge discovery and reasoning. We define the 'event' concept to refer to a data group that describes a fact derived from the integrated data repository. The events require discoverability based on the available data resources, which covers two aspects:

1) Significant activity events (SAE) include travelling to unusual or healthcare places, sport exercises, high calorie consumption activity and social activity. Normally, these kind of events requires historical data summarization such as month and year's data. Therefore, SAE detection is suitable to apply batch data processing mechanism which implemented by Cassandra MapReduce data processing algorithm.
2) Physiological (symptom) events (PSE) mainly refer to real-time reacting and well defined symptoms such as low/high blood pressure, unusual heart rate, poor sleeping and significant weight/fat changes. Because the real-time detection requirement, the PSE detection majorly directly analyse the data stream derived from Kafka partitioned topic streams.

The algorithms for mining the events are detailed in Section V.

C. Semantic Layer

The semantic layer comprises three components: Semantic repository, Semantic lifting engine and Semantic reasoning engine.

Virtuoso RDF repository [25] has been deployed on our private cloud server as centralised point to store the semantic triples and OWL ontology schema. Virtuoso automatically provides SPARQL endpoint and JDBC update connections to client applications and server-side developers. We use RSA encryption protocol to protect the data from the system and make sure only the user can see his or her own data with a private key.

We have developed our own semantic lifting engine that contains two major semantic mapping and RDF triple generation algorithms: significant event mapping and symptom event mapping. Section V illustrates both algorithms in depth.

The semantic reasoning engine aims to further mine the data based on the advantages of using data semantics for:

1) Detecting the user's lifestyle pattern
2) Detecting the semantic connections between life patterns or activities and health conditions/symptoms
3) Summarizing long-term user healthcare stories.

Apache Jena RDF semantic reasoning framework is applied to develop our semantic reasoning engine. Three sets of SWRL [25] based rules are defined in order to achieve our reasoning aims, which are described in Section VI.

Since the core of data mining and semantic reasoning processes relies on the semantic ontology design, the MHA H-Event Ontology is explained in the next section.

IV. SEMANTIC MODELLING AND LIFTING

A. Ontology design

The core concepts of the MHA H-event ontology includes 10 major terms, as Figure 2 presents. The ontology extends TMO terminologies with some existing semantic concepts from well-known domain ontologies and our defined personal activity together with treatment terms. The important concepts are listed here:

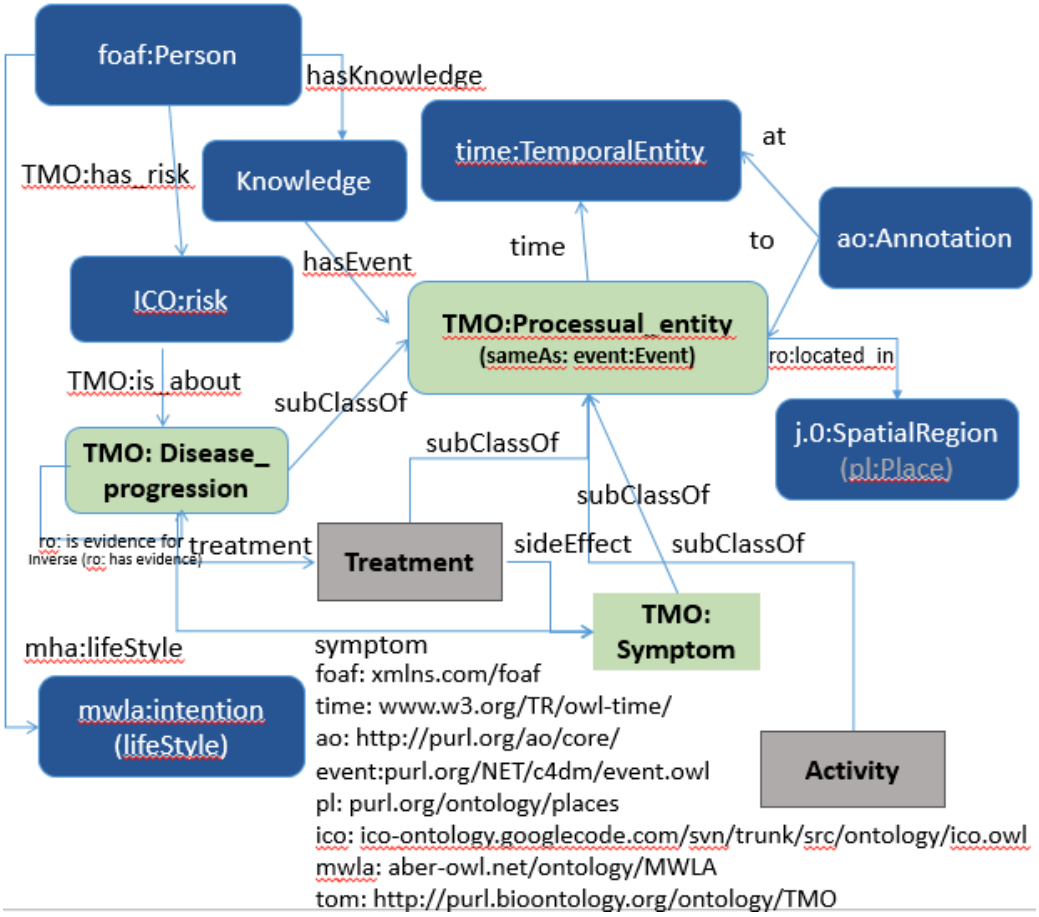
- **Event** is defined as the same as **TMO.Processual_entity** is a super concept to classify an interesting event that is related to the health of an individual user. The event is the super class of (discover a) **Symptom**, (taking a) **Treatment**, (diagnosed a) **TMO.Disease_progression** and (having a) **significant activity**. Each event associates with a particular time point on the user's time.TemporalEntity. In addition, Event is the central point of the whole ontology, which can be detected from the data mining layer.

- **Significant activity** is a subclass of the **Event** concept to identify the activity that is more significant to the user, rather than including all daily activities. In general, all the significant events should be related to understand the user's health situation or lifestyle. The activity type can be grouped by exercise type such as 'Running', 'Driving' or 'Shopping', but also can be categorized by the places and social activity type. Each significant activity should also record the duration, place, and possibly with distance, calorie consumption and steps.

- **Symptom** imported from TMO is a subclass of the Event concept to present unusual health-related conditions that are detected and concluded from the user's data. The same as all other events, the symptoms have a time stamp and place. Currently, the subclasses of Symptom include low/high blood pressure, unusual heart rate, poor sleeping and significant weight/fat changes. Other non-sensor symptoms can also be added but have to rely on the user's manually inputs.

- **Risk defined by ICO** is used as a concept to evaluate the possibility or progress levels to a particular health condition.

- **Lifestyle** is imported using the Intention concept in the MWLA ontology[6] that defines 25 lifestyle instances but can be enriched in the future.



- Fig. 2. Top layer MHA H-Event Ontology

In this paper, we focus on the knowledge discovery on the sides of activity and symptoms that can be dynamically mined from the monitoring data. However, the designed ontology is more general and comprehensive and can be used for integrating real clinical data such as PHR.

Based on the MHA H-Event Ontology [26], the semantic layer needs to lift two types of event: significant activity and physiological symptom. Currently the ontology is fixed terminology which has not considered the dynamic update issue which can adapt new suitable ontologies to the infrastructure.

B. Semantic significant activity event mining

SLAi is defined as the significant level of the *i*th activity detected from the lower-level data stream and Equation 1 is the calculation function:

$$SLAi = \prod Wi_j = Wi_{at} \cdot Wi_{loc} \cdot Wi_{dur} \qquad (1)$$

Where Wi_{at} presents the activity type weight of the *i*th activity, Wi_{loc} presents the activity location weight of the *i*th activity and Wi_{dur} presents the activity duration weight of the *i*th activity. Therefore, the final level score is the $\prod$ of the three weight values. The range of each weight is [0.1, 1], therefore $0 < SLAi \leq 1$. Table 2 shows the weight value distributions of activity type and location type:

Table. 2. Weight values from 0 to 1 for different activities and locations

Activity type	Weight	Location type	Weight
home/work	0.1	home/ work place	0.2
Walk	0.3	shop/ restaurant	0.4

[6] https://bioportal.bioontology.org/ontologies/MWLA

Transport	0.6	entertaining/ sports/ social places	0.6
Social	0.8	transport	0.8
exercise/ healthcare	1	other place	1/times been the place in the month +1

The activity duration weight is defined as Equation 2:

$$Wi_{dur} = TD \cdot (1/36000) \quad (2)$$

Where TD is the duration (seconds) of the activity and 36,000 is the number of seconds in 10 hours.
Finally, the significant threshold for lifting the activity as semantic knowledge to the semantic repository is set as:

$$SLAi \geq 0.02$$

C. Semantic physiological (symptom) event mining

The physiological event mining methods are defined based on medical measurement guidelines. At the moment, we concentrate on detecting four symptoms high/low blood pressure, unusual heart rates. All these three symptoms are well-defined in medical guidelines.

For example, the blood pressures have systolic blood pressure which measures how hard the heart's left ventricle contracts to circulate blood through the body. Diastolic blood pressure measures the pressure in the blood vessels when the heart's chambers are relaxed and filling with blood. UK National Health Service (NHS) guidelines[7] indicates that normal adult blood pressure should be between 90/60 and 140/90, where the top (first) number is the systolic pressure and the diastolic is the bottom (second) number. In addition, readings higher than 140/90 can be defined as high blood pressure and lower than 90/60 as low blood pressure.

The other mining methods are defined here based on similar UK NHS guidelines. Heart rate range should generally be in [60, 100], otherwise it is too slow or too fast. The sleep hours should generally be between six and nine hours.

D. Overall semantic mining algorithms

Based on previous discussed event mining methods, the big data oriented MapReduce algorithms to efficiently distribute the mining computations on the cloud nodes are explained here:

Input: txt = A Json output from Cassandra query or directly from Kafka topic, id= UserId // Normally a Kafka topic is formatted as Json document

Algorithm 1 *Significant activity event mining*

Output:
Septet <A, V1, V2, …V6> [] SA where A = activity type, V1 to V6 = the actual values describe the activity based on the ontology terms (activity_group, duration, place/location, distance, steps and destination_group)

```
//parsing the objects from input
Array Object [] r = JsonParser(txt);
//Mapper distributed procedure
MAPPER(_key, CalculateSignficant(r[‘id’])>0.02):
_key = r['id'].get(“activity”);
//value_list is a two-dimensional array storing the matched object row id to the key where significant value larger than 0.02 and each has a value 1
emit (_key, value_list)
//Shuffle procedure to combining different _key values to different reducers
COMBINER(_key, value_list):
emit(_key, value_out);
//Reduce procedure
REDUCER(_key, value_out):
   record_num = 0;
   value_sum = 0;
   for (i=0; i<value_out[length]; i++) {
     value : value_out[i]
     record_num += value[0];
```

Algorithm 2 *Physiological event mining*

Output:
Quartet <S, V1, V2, V3> [] QS where S = symptom name, V1 to V3 = the actual values describe the symptom (value, time and place/location)

```
//parsing the objects from input
Array Object [] r = JsonParser(txt);
//Mapper process
MAPPER(_key, CalculateSymptom (r[‘id’])==1):
_key = r['id'].get(“physiology”);
emit (_key, value_list)
//Shuffle process
COMBINER(_key, value_list):
emit(_key, value_out);
//Reduce process
REDUCER(_key, value_out):
   record_num = 0;
   value_sum = 0;
   for (i=0; i<value_out[length]; i++) {
     value : value_out[i]
     record_num += value[0];
     value_sum += value[1];
     QS[i].set(“S”)= _key;
     QS[i].set(“V1”) = r[record_number].get
((“value1”)+”,”+(“value2”));
```

[7] http://www.nhs.uk/NHSEngland/thenhs/about/Pages/overview.aspx

```
    value_sum += value[1];
    //set the semantic septet for lifting preparation
    SA[i].set("A")=_key;
    SA[i].set("V1") = r[record_number].get ("duration");
    SA[i].set("V2") = r[record_number].get("location");
    ...
  }
  emit(_key, value_sum SA) to semantic repository;
// Lifting to semantic layer is explained in Section VI
Semantic_Lift (_key, SA);
End
```

```
    QS[i].set("V2") = r[record_number].get("time");
    QS[i].set("V3") = r[record_number].get("location");
  }
  emit(_key, value_sum, QS) to semantic repository;
// Lifting to semantic layer is explained in Section VI
Semantic_Lift (_key, QS);
End
```

E. Semantic lifting

The semantic lifting process is a generalization of RDF triples based on the proposed ontology, which includes two steps of semantic mapping: domain mapping and property mapping with range assignment.

Step 1: Domain mapping

In the first step the domain matching algorithm is applied to identify the domain element; this is the simplest algorithm in these three steps. According to our JSON structure composing the summary data analysis, only activity type or symptom name elements are suitable candidates for the domain that can be lifted as subject elements of the instance RDF triples. If the element is under the activity's JSON structure, then a URI will be generated and specified as an Activity class defined in the OWL ontology. A similar process is generated for the symptom event.

Step 2: Property and range mapping

According to the JSON input structure, the property and range mapping are performed together based on the pre-defined mappings in Tables 3 and 4.

Table. 3. Semantic Mapping for Significant Activities

JSON structure syntax	Mapped property defined in the ontology	Range value
ranking value	mha: rank	(0,1]
duration,	mha: time	Seconds
destination	mha: located in	Annotation text or unknown
Distance	mha: distance	Metres
Step	mha: step	Count
activity_group	mha: hasEvent	Activity type

Table. 4. Semantic Mapping for Significant Physiological Symptom

JSON structure syntax	Mapped property defined in the ontology	Range value
Value	mha: hasValue	Text value with unit
time	mha: time	Time spot/date information
location	mha: located in	Annotation text or unknown

F. Semantic reasoning

The final goal to have the data lifted into the semantic repository is to enable mining the data further to discover hidden knowledge about the user, getting the benefit of the smaller but more machine understandable data representations – RDF triples based on well-defined semantic ontology. The specific objectives of the semantic reasoning process that are:

1) Understanding the user's lifestyles. It should be possible to reason some interesting lifestyle activities by our reasoning engine according to Medical Web Lifestyle Aggregator (MWLA)8 ontology developed by another EU-funded research project – CARRE.
2) The semantic relations between the user's activities and symptoms as well as the links between the lifestyle and the health conditions. In the long-term, the user's life-long health situation can be interpreted to help disease prediction and prevention.

In order to achieve the reasoning objectives, we have developed a semantic reasoning engine using the Jena semantic framework which supports SPARQL and SWRL (Semantic Web Rule Language) [27] and can be integrated into a Virtuoso RDF repository. There are two types of reasoning rules. The first one is based on SPARQL queries that can define the reasoning formulas at the ontology level (T-box). The second one is based on SWRL rules that cannot be specified at the ontology level, rather at the instances level (A-box). We represent two reasoning scenarios here to illustrate how these two different reasoning methods can be applied for inference of lifestyle pattern and linked to certain health conditions or symptoms.

[8] http://aber-owl.net/ontology/MWLA

Example 1: Travel/long commute lifestyle for the past month.

Definition: Long transport (more than 2 hours) activity events have been lifted in to the RDF at least four times in the last month. The reasoning process will be:

Construct (SPARQL query) the last month activity event RDF memory-model based on the ontology retrieved from the triple storage. Specify SPARQL query first. If the query returns a value, then it means the user satisfies the defined reasoning query. Then we can construct the semantic links between the **Person** to the **Travel** term defined in the MWLA as new knowledge to the semantic repository (Code 1).

```
PREFIX mha: <http://myhealthavatar.org/ontology/>
mwla: < http://purl.bioontology.org/ontology/MWLA>
   CONSTRUCT (?p mha:lifestyle mwla:Travel)
   SELECT (COUNT(?numberOflongTravelling) AS ?howMany ?p)
   WHERE { ?p mha:hasEvent ?e . ?e  rdf:type mha:transport . ?e  mha:time ?d . FILTER (?d >= 7200)
   }HAVING ( ?howMany > 4 )
```

CODE 1

Example 2: SWRL-based health-condition risk alarm reasoning.

Definition: If a person lacks activity and has age > 60, then there is a risk of high blood pressure.

The rule can be defined as Code 2 in the Jena rule engine.

```
   [rule: (?p mha:lifestyle mwla:noActivity), (?p foaf:age ?i), greaterThan(?i, 60) -> (?p
tom:has_risk ?x), (?x tom:is_about ?d), (?d rdf:type tom:High_Blood_Pressure)]
```

CODE 2

V. Evaluation and Future Work

Figure 3 presents the system interface of integrating data resources from third party applications through authenticated OATH protocol with user agreement. The data integration include 4 parts of Fitbit data, Moves data, Withings data and the data from our project applications. The raw data from these data resources are synchronised into our NoSQL database whenever the user logs into the system and clicks the synchronisation button (see step 3 in Figure 3). Finally, users can share some information from our application to their Twitter accounts. Figure 4 is the MyHealthAvatar mobile app that can collect user activity data, profile data and event planning data. The app can also set goals and virtualise the analysis result from our semantic mining process.

For our experimental research, three applications of Fitbit, Moves and Withings are integrated with the MyHealthAvator application (details of data usages are presented in Table 1). The test dataset are 100 mockup users and their 36 weeks period activity records.

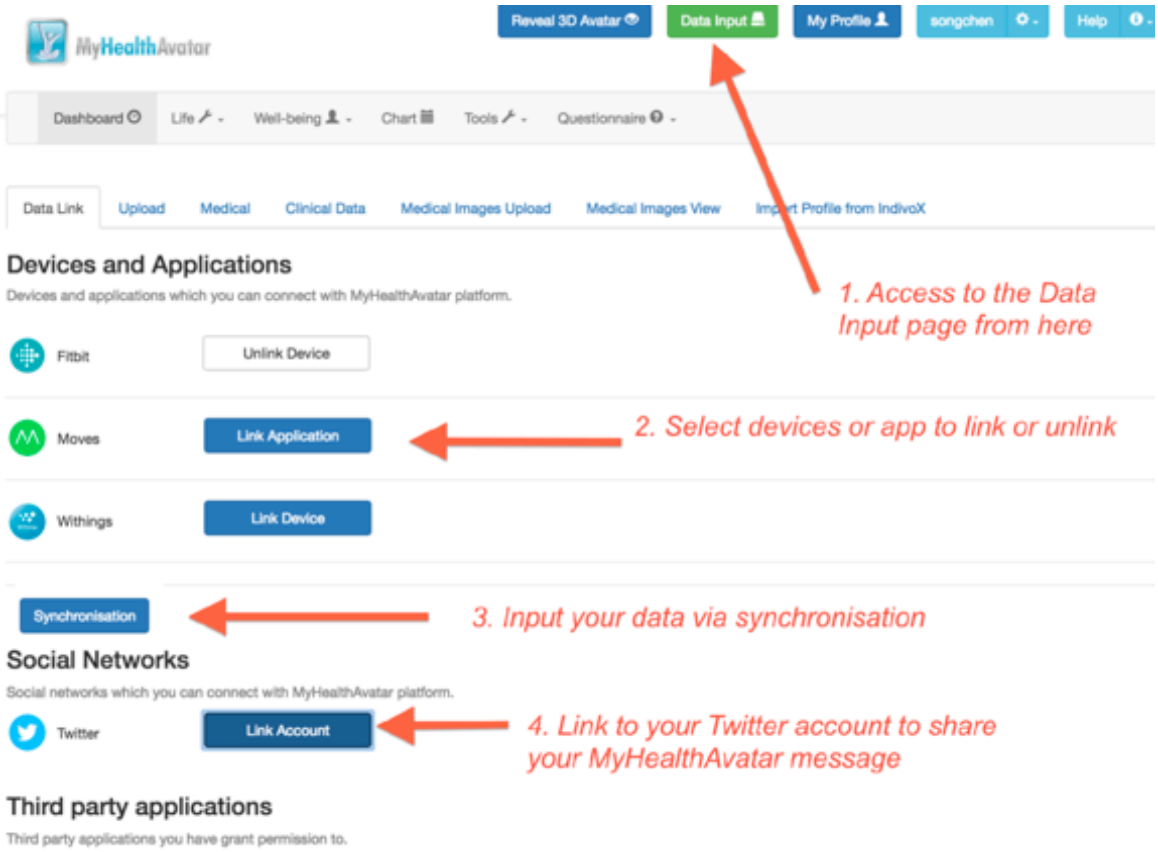


Fig. 3. Data integration interface

Fig. 4. MyhealthAvatar Mobile App

For the semantic lifting scalability evaluation we tested the performance time on the semantic mining/lifting algorithms on two scenarios with different settings of the Hadoop MapReduce environment: (1) increasing the data size (number of weeks) form 0 to 36 weeks but for just one user's record, and (2), increasing the number of users, but with a fixed 12-weeks of records for all of them. The evaluations are made in the environment of virtual machines, mainly in a Linode cloud-cluster environment with an Apache Hadoop configuration. The configuration settings are 1 node, 2 nodes, 3 nodes and a maximum of 4 nodes. The hardware

is an Ubuntu Linux Server 14.04, LTS 64-bit operating system for the repositories. Tomcat 7/8 over Java 8 64-bit provides the runtime environment for the repository interface. The operating system has all the latest security patches applied. Data repositories are Cassandra version 2.1.5 with CQL spec 3.2.0 and native protocol v3, and Virtuoso Open Source Edition v7.10.3207.

Figure 5 shows the first evaluation case and clearly demonstrates that increasing parallel level on MapReduce algorithm (adding more notes) can fast improve the performance on semantic mining/lifting tasks. The improvement rate is relatively close to 1× (number of nodes) which matches our expectation. The notable phenomenon is that larger data size can be processed in a reasonable time by adding more nodes.

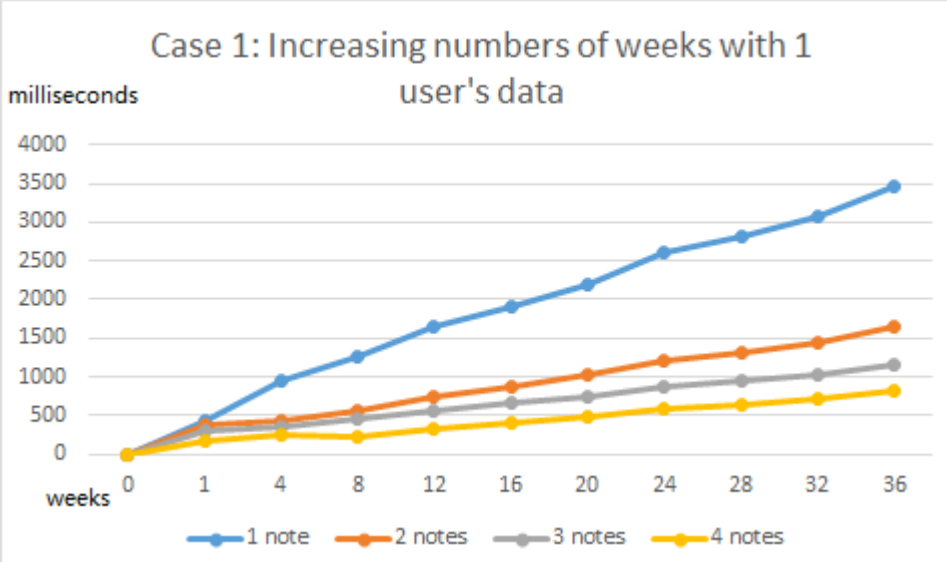


Fig. 5. Performance evaluation case 1

Figure 6 shows the second evaluation case. The performance dramatically decreases comparing to one user's 12 weeks data by increasing numbers of the users for all different environment settings. The major reason is due to the data size increasing. The other important reason is that data I/O communication delays the time by switching from different users, which isn't the case for the first test. The second evaluation result also illustrates increasing number of notes can improve the performance dramatically.

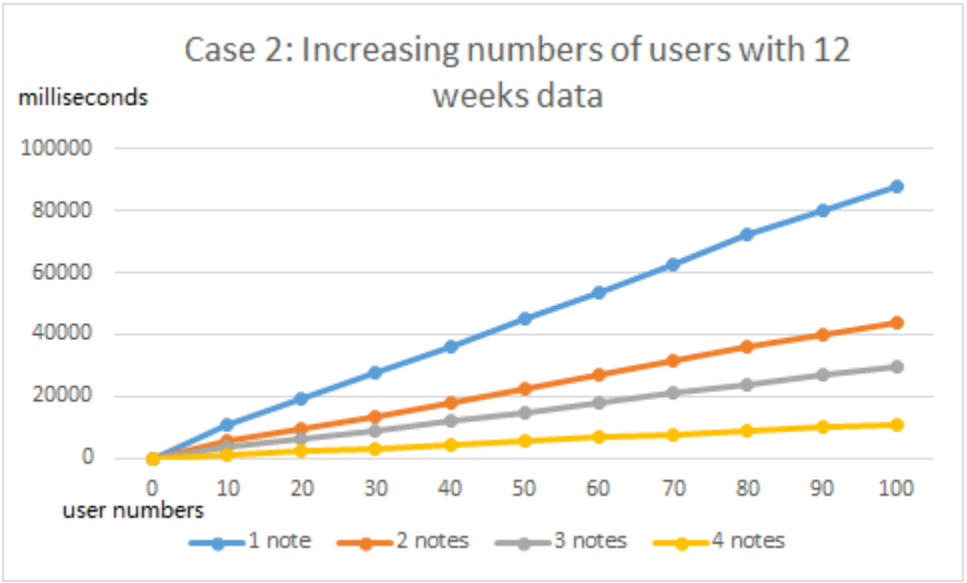


Fig. 6. Performance evaluation case 2

At the moment, the data used for this project are collected exemplar data rather than real data from an individual patient due to security and privacy issues, this will be a major research focus in the future and very important to address in the research, and finally apply commercially. However, this research's direction requires not only to involve engineering and scientific work but also, importantly, to develop policy-level agreements and standards.

The other important issue we have not fully explored is the data aggregation in the integration process. Aggregation research problems include automatically filling in missing data, correctly refining data and handling uncertainty in data that is gathered from different resources but for the same semantic terms, e.g. step counting value can be collected from multiple devices and mobile applications. A set of methods can now be investigated, including Prediction Mean Matching Imputation, KNN and Regression methods [28], Attribute Selection, Smart Tokens and Probabilistic Noisy Identification methods for removal of noisy data [29, 30]. Also, there exists a range of methods for representation and manipulation of uncertainty, such as probability density function, fuzzy sets, belief function, or interval sets, uncertainty propagation and sensitivity analysis, etc. [31, 32].

REFERENCES

[1] Wagner, E.H., Chronic disease management: what will it take to improve care for chronic illness?, Effective Clinical Practice, vol 1, no 1, pp 2–4, 1998.

[2] Greene, J. and Hibbard, J.H., Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. Journal of General Internal Medicine, vol 27, no 5, pp 520, 2012.

[3] Hibbard, J.H. and Greene, J., What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs, Health Affairs (Millwood), vol 32, no 2, pp 207–14, 2013.

[4] H. Kondylakis, M. Spanakis, S. Sfakianakis, V. Sakkalis, M. Tsiknakis, K. Marias, Z. Xia, H. Q. Yu, F. Dong, Digital Patient: Personalized and Translational Data Management through the MyHealthAvatar EU Project, International Conference of the IEEE Engineering in Medicine and Biology Society of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, Milan, Italy.

[5] P. Yang, M. Hanneghan, J. Qi, Z. Deng, F. Dong and D. Fan, Improving the Validity of Lifelogging Physical Activity Measures in an Internet of Things Environment, *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM),* Liverpool, 2015, pp. 2309-2314.

[6] Li, I., Dey,A. and Forlizzi, J. (2011) Understanding My Data Myself: Supporting Self-Reflection with Ubicomp Technologies. In Proc UbiComp 2011.

[7] Li, I. Personal Informatics and Context: Using Context to Reveal Factors that Affect Behavior. PhD Thesis. 2011.

[8] The Top Challenges in Big Data and Analytics, Lavastorm analytics, 2013, http://docplayer.net/1377052-The-top-challenges-in-big-data-and-analytics.html

[9] Hu, B., Carvalho, N., Laera, L., and Matsutsuka. T., Towards big linked data: a large-scale, distributed semantic data storage. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS '12)

[10] Garg, M. K., Kim, D., Turaga, D. S. and Prabhakaran, B., Multimodal analysis of body sensor network data streams for real-time healthcare. In Proceedings of the international conference on Multimedia information retrieval (MIR '10). ACM, New York, NY, USA, 469-478.

[11] Bauschlicher, D., Bauschlicher, S., and ElAarag. H., Framework for the integration of body sensor networks and social networks to improve health awareness. In Proceedings of the 14th Communications and Networking Symposium (CNS '11). Society for Computer Simulation International, San Diego, CA, USA, 19-26.

[12] K. Tollmar, F. Bentley and C. Viedma, "Mobile Health Mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device", *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, San Diego, CA, 2012, pp. 65-72.

[13] Diego Calvanese and De Giacomo, Giuseppe and Maurizio Lenzerini and Daniele Nardi and Riccardo Rosati, Data Integration in Data Warehousing, Int. J. of Cooperative Information Systems, vol. 10, no. 3, pp. 236-271, 2001.

[14] G. Wiederhold, Mediators in the architecture of future information systems, IEEE Comput. 25, 3 (1992) pp. 38-49.

[15] J. D. Ullman, Information integration using logical views, Proc. ICDT'97, LNCS (Springer-Verlag, 1997) pp. 19-40.

[16] Yang, H. and Li, W., An ontology-based approach for data integration in regionally interoperable healthcare systems. In: 11th International Conference on Informatics and Semiotics in Organisations (ICISO 2009), 11-12 Apr 2009, Beijing, China, pp. 93-96.

[17] Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, Wittkowski KM. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. J Biomed Inform. 2014 Dec;52:78–91.

[18] Scheuermann, R. H., Ceusters, W., & Smith, B. (2009). Toward an Ontological Treatment of Disease and Diagnosis. Summit on Translational Bioinformatics,2009, 116–120.

[19] Luciano JS, Andersson B, Batchelor C, et al. The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. Journal of Biomedical Semantics. 2011;2(Suppl 2):S1. doi:10.1186/2041-1480-2-S2-S1.

[20] Lee, Dennis et al. "Literature Review of SNOMED CT Use." Journal of the American Medical Informatics Association : JAMIA 21.e1 (2014): e11–e19. PMC. Web. 11 Apr. 2016.

[21] Schriml, Lynn Marie et al. "Disease Ontology: A Backbone for Disease Semantic Integration." Nucleic Acids Research 40.Database issue (2012): D940–D946. PMC. Web. 11 Apr. 2016.

[22] Fitbit Ltd, Fitbit Healthy Futures Report, 2013, http://www.trajectorypartnership.com/wp-content/uploads/2014/02/Fitbit-Healthy-Futures-Report-September-2013.pdf

[23] MACKINLAY, Molly Zellweger. Phases of Accuracy Diagnosis: (In) visibility of System Status in the Fitbit. Intersect: The Stanford Journal of Science, Technology and Society, [S.l.], v. 6, n. 2, jun. 2013. Date accessed: 27 Apr. 2016.

[24] DATASTAX Corporation, Introduction to Apache Cassandra – White Paper, July 2013. http://www.datastax.com/wp-content/uploads/2012/08/WP-IntrotoCassandra.pdf

[25] Erling, O. and Mikhailov, I., Chapter RDF Support in the Virtuoso DBMS in book of Networked Knowledge - Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems, Springer Berlin Heidelberg 2009.

[26] H. Q. Yu, X. Zhao, Z. Deng and F. Dong, "Semantic Lifting and Reasoning on the Personalised Activity Big Data Repository for Healthcare Research," 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and

Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, 2017, pp. 818-823.
[27] I. Horrocks, P. F. Patel-Schneider, S. Tabet, B. Grosof and M. Dean, SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission, 2014, http://www.w3.org/Submission/SWRL/
[28] Suthar, B., Patel, H., Goswami, A., A Survey: Classification of Imputation Methods in Data Mining, International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 2, Issue 1, January 2012.
[29] Tamilselvi, J.J. and Saravanan, V., "Handling Noisy Data using Attribute Selection and Smart Tokens," International Conference on Computer Science and Information Technology, pp.770-774, Aug. 29 2008-Sept. 2 2008.
[30] Kubica, J. and Moore, A, "Probabilistic noise identification and data cleaning". Third IEEE International Conference on Data Mining, vol., no., pp.131-138, 19-22 Nov. 2003.
[31] Halpern, J.Y., Reasoning about uncertainty. The MIT Press, October, 2003.
[32] Frey, H. and Patil., S., Identification and review of sensitivity analysis methods, Risk Analysis, 22(3):553-578, 2002.