IJCI
Menoufia University
Faculty of Computers and Information

International Journal of Computers and Information

Volume 5 No. 1

July 2016

Faculty of Computers

IJCI

and Information

INTERNATIONAL JOURNAL OF COMPUTERS AND INFORMATION

Faculty of Computers and Information, Menoufia University

Journal Secretary

Miss. Fainan Nagy El Sisi ijci@ci.menofia.edu.eg

Instructions Format

Paper Title* **(**use style: ***paper title***)

Subtitle as needed ***(paper subtitle)***

Authors Name/s per 1st Affiliation (*Author*)
line 1 (of *Affiliation*): dept. name of organization
line 2-name of organization, acronyms acceptable
line 3-City, Country
line 4-e-mail address if desired

***Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, or Math in Paper Title or Abstract.* (*Abstract*)**

Keywords—component; formatting; style; styling; insert (key words)

I. Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. Ease of Use

A. Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file "MSW_USltr_format".

B. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive."
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter," not "webers/m2." Spell units when they appear in text: "...a few henries," not "...a few H."
- Use a zero before decimal points: "0.25," not ".25." Use "cm3," not "cc." (*bullet list*)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$a + b = \gamma \qquad (1)$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ..."

D. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o."
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset," not an "insert." The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively."
- In your paper title, if the words "that uses" can accurately replace the word using, capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect," "complement" and "compliment," "discreet" and "discrete," "principal" and "principle."
- Do not confuse "imply" and "infer."
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al."
- The abbreviation "i.e." means "that is," and the abbreviation "e.g." means "for example."

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

An excellent style manual for science writers is [7].

IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

1) For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

a) Selection (Heading 4): Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select "1 Column" from the selection palette.

c) Deletion: Delete the author and affiliation lines for the second affiliation.

2) For author/s of more than two affiliations: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the "Columns" icon from the MS Word Standard toolbar and then select "1 Column" from the selection palette.

c) Highlight author and affiliation lines of affiliation 1 and copy this selection.

d) Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

e) Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select "1 Column". If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in one columns.

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is "Heading 5." Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract," will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1," "Heading 2," "Heading 3," and "Heading 4" are prescribed.

C. Figures and Tables

1) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1," even at the beginning of a sentence.

TABLE I. TABLE STYLES

Table Head	Table Column Head		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

Fig. 1. Example of a figure caption. *(figure caption)*

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization," or "Magnetization, M," not just "M." If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization (A (m(1)," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

ACKNOWLEDGMENT (HEADING 5)

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g." Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Table of Contents

Vol. 5 – No. 1

Specification -based Test Cases Generation for Multi-Level Service Composition

Shymaa Sobhy, Mahmoud Hussein and Ashraf B. El sisi
Faculty of Computers and Information,
Menofia University, Egypt ,
shaymaa.abdelaal@ci.menofia.edu.eg , mahmoud.hussein@ci.menofia.edu.eg and
ashraf.elsisi@ci.menofia.edu.eg

***Abstract*- Testing is the traditional validation method in the software industry. To ensure the delivery of high quality and robust service-oriented applications, testing of web services composition has received much attention. These services have become more and more complex, where they have to cope with strict requirements of business processes and their dynamic evolution, and interactions among different companies. In this context, the analysis and testing of such services demand a large amount of effort. To reduce the effort required for web-services testing, in this paper, we propose a specification-based approach to automatically generate test cases for web services composition that is modeled at different levels of abstraction. This approach specifies a service structure as multi-level models. To generate the test cases, it checks if the first level of the model has a parallel execution or a decision table to be solved by an algorithm that solves Chinese postman problem. Then, it identifies paths for last level of the model and relates the results of all levels with each other. To evaluate our approach, we applied it to four cases study using our developed tool. Compared to exiting approaches, our approach reduces testing cost and execution time, and increases testing reliability.**

Keywords--Service-oriented Applications, Web Services Composition, Model-based Approach, and Event-driven Model.

I. INTRODUCTION

Service-oriented architectures (SOAs) and web services have been used to enable loosely-coupled, distributed applications by using independent and self-contained services [1]. These services can be combined in a workflow that characterizes a new composite service. The resulting composite service is also called a web service composition (WSC [2, 4]. Such services have complex communications where the service behavior depends not only on the composition but also on the integrated services. Therefore, the testing process becomes complicated.

A common problem in testing any kind of application is to automatically generate meaningful test cases [3, 5]. The strategy of using models for test case generation is known as model based testing [6, 7 and 8]. Web applications are evolving rapidly, as many new technologies, languages, and programming models are used to increase the interactivity and the usability of web applications [8, 14]. This inherent complexity brings challenges to modeling, analysis, testing, and verification of this kind of applications.

To reduce the complexity of designing large composite services, the designers apply service decomposition where services are model at different levels of abstraction [10]. Many techniques have been introduced to support automatic testing of composite services. For example, a multi-observer architecture is proposed to detect and locate faults in composite web services [6], and a new model to describe a service choreography that manipulates data flow by means of XPath queries is introduced [18]. In addition, a model-based integration testing for service choreography using a proprietary model, called message choreography model (MCM) is proposed [24]. But, these techniques do not support testing service compositions that are modeled at different levels of abstraction.

In this paper, we propose an approach for generating test cases for composite web services that are modeled at different levels of abstractions. We use a model based technique called "Event Sequence Graph for Web Services Composition" to generate cost-effective test cases for service compositions that modeled at one level [9]. We improve that technique to generate test cases for web services modeled at a multiple level of abstraction. Our approach check if the first level has a parallel execution or decision tables to be solved using an algorithm that solves Chinese postman problem for a directed graph to identify paths by generating a Euler network. Then it identifies paths for last level of the service model by same algorithm. Finally, the approach relates results of all levels with each other. We evaluate our approach by applying it to four case studies that have different complexity. We also developed a tool to generate test cases for the case studies.

The remainder of this paper is organized as follows. Related work is analyzed in Section 2. Proposed approach in Section 3. Section 4 experimental results. Section 5 presents conclusion.

II. Related Work

This section introduces work that is related to our approach. We also explain in detail an existing technique that we use as a base for our approach. Web service testing has been studied intensively in the last years [1, 3, and 4] with a particular effort on formal testing (for a systematic review of the literature, see [17]). The service testing provides the reliability analysis and formal reviews to the service. In the following, we describe some of the existing approaches.

A model that describes a service choreography that manipulates data flow by XPath queries is introduced [18]. In the choreography, XPath queries can handle different XML schema files. This work is focused on test case generation for web service composition but modeled at one level only. Benares [6] proposes a multi-observer architecture to detect and locate faults in composite web services. The proposed architecture is composed of a global observer and local observers that cooperate to collect and manage faults found in the composite service. Their approach aims at testing the service composition that modeled as one level. Wieczorek [24] proposes a model-based integration testing for service choreography using a proprietary model, called message choreography model (MCM). The work is close to the proposed approach, with difference that, our approach uses a more abstracted model compared to MCMs, which form a domain-specific language created to design service choreography. Fevzi Belli and Christof Budnik proposed an approach for generation and selection of test cases based on statecharts [22]. This approach with scalable way uses regular expressions and regular expressions are used in the test process. But here we introduce event sequences graph that generate tests for multi-level graphs. JanTretmans proposed an overview of formal, model-based testing in general and of model-based testing for labeled transition system models in particular [23]. Also, it introduces the same concept of model-based testing but this still not deal with multilevel graphs. Belli and Endo [9] have proposed an event-based model, named (ESG4WSC) Event Sequence Graph for Web Services Composition that represents the request and response messages exchanged between services involved in a WSC. This approach is for generate tests for graph that modeled at one level only. We will improve it to deal with multilevel graphs. Table 1 shows algorithm for deriving CESs from an ESG4WSC.

III. The Proposed Approach

This section introduces firstly background secondly present a running example thirdly the proposed approach.

A. Background

This section introduces formal notions and algorithms that are relevant to the proposed approach. It also presents the underlying fault model, test case generation and minimizing test set [9, 15 and 21]. Event Sequence Graph for Web Services Composition: Event Sequence Graph (ESG) for Web Services Composition represents the request and response messages exchanged between services involved in a service composition. When a given event is refined by input parameters that determine the next events, decision tables (DTs) are associated to augment this representation. Decision tables are widely employed in information processing and are also traditionally used for testing. Table 2 shows decision table for xloan case study. ***Definition 1:*** A (simple/binary) decision table DT ={C, E, R} represents events that depend on certain constraints, where: C is nonempty finite set of constraints (conditions), which can be evaluated as either true or false, E is the nonempty finite set of events, and R is the nonempty finite set of rules each of which forms a Boolean expression connecting true/false configurations of constraints and determines executable or waited events. ***Definition 2:*** An ESG for web service ESG4WSC = {V, E, M, R, DT, f, Ξ, Γ} is a directed graph, where:

- V is a nonempty finite set of vertices representing events;
- E⊆V x V is a finite set of arcs (edges);
- M is a finite set of refining Event Sequence Graph for Web Services Composition models;
- R⊆V x M is a relation that specifies which Event Sequence Graph for Web Services Compositions are connected to a refined vertex;
- DT is a set of DTs that refine events according to function f;
- F: V→DT {ε} is a function that maps a decision table dt ∈DT to a vertex v ∈ V. If v ∈ V is not associated with a DT, then f(v)=ε;
- Ξ , Γ ⊆V are finite sets of distinguished vertices with ξ ∈ Ξ and γ ∈ Γ called entry nodes and exit nodes, respectively, wherein for each v ∈ V there exists at least one sequence of vertices (ξ ,v0, . . .,vk) from ξ ∈ Ξ to vk=v and one sequence of vertices (v0,…,vk, γ) from v0=v to γ∈ Γ with (vi,vi+1) ∈E for i=0,….,k-1and v≠ ξ, γ. ***Definition 3:*** Let V be as in Definition 2. Then, the set of vertices V is partitioned into V_e, $V_{refined}$,

V_{req}, and V_{resp} that is, $V = V_e \cup V_{refined} \cup V_{req} \cup V_{resp}$ and V_e, $V_{refined}$, V_{req}, and V_{resp} are pairwise disjointing, where:

- Ve is a set of generic events,
- Vrefined = $\{v \in V \setminus \exists m \in \bigwedge (v, m) \in R\}$ is a set of vertices refined by one or more Event Sequence Graph for Web Services Compositions. A refinement with more than one Event Sequence Graph for Web Services Compositions represents behavior running in parallel,
- Vreq is a set of vertices modeling a request to its own interface/operations (public) or an invoked service (private), and Vresp is a set of responses to a public or private request. Therefore, it is also remarked as public or private. ***Definition 4:*** Let DT be defined as in Definition 2. Then, the set of decision tables is partitioned into DT_{seq} and DT_{input}, where: DTseq is the set of DTs that model the execution restrictions for following events and DTinput is the set of DTs that model constraints for input parameter of invoked operations.

B. Fault Model: An ES (see Definition 5) describes a specific execution of a WSC that has to be enforced during testing. Thus, it is expected that exactly those events in the specified order are executed [9].

C. Test Case Generation: To cause and control a specific CES of the WSC, it is often inevitable to take control of partner services because they communicate with the system under test (SUT). And the flow of the WSC might depend on a returned response. The modeled constraints of DTs enable to validate the data passed to the service operations [13, 15 and 19].

D. *Minimizing the Test Sets*: The total number of CESs with minimal total length that cover the ESs of a required length is called Minimal Spanning Set of Complete Event Sequences (MSCES). Eentire walk occurs when the CES contains all EPs at least once [28]. The Chinese Postman Problem is expected to have a higher degree of complexity than MSCES problem introduced here as the edges of the ESG are not weighted [11, 12, 23 and 20].

E. Running example: The example involves three services: LoanService (LS), BankService (BS), and BlackListInformationService (BLIS). LoanService represents the business process xLoan. It has three operations: request, cancel, and select. BankService represents the financial agency that approves (or not) loans, and provides loan offers to its clients. The operations used in the example are approved, offer, confirm, and cancel. The BlackListInformationService provides an operation checkBL to check if a client has debits with other financial organization. The example is extended to add parallel flow in the process by including CommercialAssociationService (CAS). Similar to BlackListInformationService, CAS provides operations to check whether a client has debits with other commercial organization. In the extension, both services are supposed to be called in parallel. If the client has debit according to one of them, the client needs the bank approval [9]. The multi-level Event Sequence Graph for the running example (xloan) is in Figure 1.

F. *The proposed approach*: When the input is multi-level graphs, we apply our proposed approach. For increase efficiency, we apply the CPP algorithm only to two levels the first level and the last level. And then relate the results to each other. First, we apply (CPP) to first level. Second, we check if last level is a parallel execution. If true, we find the CESs for it. Then, we identify the valid successor for each CES with respect to the DT. Thirdly, we find the CESs of the inner sublevels. Then, we make replacing operation. Table 3 shows the algorithm for the proposed approach.

Table 1. An algorithm for deriving CESs from an Event Sequence Graph

Input: An Event Sequence Graph for Web Services Composition (one level).
Output: CESs.
1. **Foreach** (vertex = refined vertices) **do**
  Go to step 3 to Generate CESs for the refined vertices first.
2. **Add** multiple edges (representing EPs) to Event Sequence Graph for Web Services Compositions:
  If (refined vertex has a DT restricting the ongoing execution)
    Identify the valid successor for each CES with respect to the DT.
    Add an edge from the refined vertex to the allowed successor.
  Else
    Add an edge from the refined vertex to the successor (there should be only one) for each CES.
3. **Generate** CESs according to the CPP algorithm (i.e., cover all EPs by CESs of minimal total length).
4. **Replace** refined vertices in the resulting CES set of Step 3 with the CESs derived in Step 1 with respect to their allowed successors.
Return CESs

Table 2. The decision tabel for check Refined

Dtcheck	***R1***	***R2***	***R3***	***R4***
Event :BLIST: inBlist happen	T	T	F	F
Event :BLIST: NotinBlist happen	F	F	T	T
Event :CAS:DebetorTrue happen	T	F	T	F
Event :CAS:DebetorFalse happen	F	T	F	T
BSoffer				√
BSapproveBank	√	√	√	

Table 3 . The proposed approach

Input: Event Sequence Graph for Web Services Compositions that decomposed into different levels of abstraction (multiple levels).
Output: CESs.
Step 1: **Apply** (CPP algorithm) to (Frist level) to generate CESs for it and save it to an array data structure.
If (last level = parallel execution) **Go to** step 2
If (refined vertex has a DT restricting the ongoing execution)
Identify the valid successor for each CES with respect to the DT.
Else
Put the sequence from the refined vertex to the successor (there should be only one) for each CES.
Step 2: **Apply** (CPP algorithm) to (last level) to generate CESs for it and do and operation on results.
Step 3: **Get** the CESs of the inner sublevels manually.
Step 4: **Replace** the abstracted frist level node in CESs from Step 1 with other sublevels nodes result from step 3.
Step 5: **Replace** CESs from step 4 with CESs result from Step 2.
Return CESs

I.

IV. EXPERIMENTAL RESULTS

In this section, we show applying the proposed approach to the xloan example then show the evaluation by applying it to other three cases study

A. Applying proposed approach to the xloan example:

We will now apply the proposed approach for the xloan case study. Table 4 apply CPP algorithm to the first level that list two only sequences [16]. Table 5 shows the application of CPP algorithm to the last level to get the CESs and shows four sequences. Table 6 shows applying the AND operation to them. Finally, we go to replacing operation to relate the results. Table 7 shows the CESs of level 1 after replacing by level 2. Table 8 shows replacing results with results from table 6 related to DT constrains. The new propose approach is applied only at two levels rather than calling it three time as done at the old approach and its result is related to each other. This increase efficacy by reduce execution time and number of iteration of CPP algorithm.

B. Evaluation

We evaluate our approach by applying it to three other cases studies. The first one called Travel Agent Service. It provides a set of facilities to query and book a trip. It interacts with two services: ISELTA-hotel and Airlines services. It combines these two services to provide operations for searching and booking a travel involving flight and hotel reservations.

As the flight ticket and hotel reservation are essential in any travel, a successful booking using this service guarantees hotel and flight reservations [9]. The second called ABC services case study [25]. This service interacts with three other services: PartnerService01 (PS01), PartnerService02 (PS02), and PartnerService03 (PS03). It focuses on the flow of messages triggered by the operations of ABCService. The third called BCS-05 service which is a version from Business Connectivity Services. We have applied the old and proposed approach to these three cases studies (for the details, see [26]). In the following, we give the information about the case studies including test model information for higher length, and the execution time and number of iteration for the old algorithm and the proposed approach.

Table 4. CES1,2 after apply CPP for level 1of multi level graph

CES1	Start LSrequestLoan BSapproveBank BSapproved BSoffer BSOffers LSreplyOffers LSSelectOffers LSWrongOffer LSSelectOffers BSconfirmBank LSreplySelect End
CES2	Start LSrequestLoan **check** BSapproveBank BSapproved BSoffer BSOffers LSreplyOffers LSSelectOffers LSWrongOffer Timeout >2h BScancelBank End

Table 5. CESG after apply cpp for last level

CES1	[BLIS:checkBL , BLIS:inBList]

Tabel 6. CESG after AND operation for last level

CES1	[BLIS:checkBL BLIS:inBList] || [CAS:inDebtorsList CAS:debtorsTrue]

Table 7. CES1,2 after replacing level 2 for the proposed approach

CES1	No replace
CES2	Start LSrequestLoan **checkBLIS** BSapproveBank BSapproved BSoffer BSOffers LSreplyOffers LSSelectOffers LSWrongOffer Timeout>2h BScancelBank End
	Start LSrequestLoan **checkCAS** BSapproveBank BSapproved BSoffer BSOffers LSreplyOffers LSSelectOffers LSWrongOffer Timeout>2h BScancelBank End

Table 8. Replace with the original node from step one by DT Constrains

CES1	No Replace
CES2	Start LSrequestLoan **BLIS:checkBL BLIS:inBList CAS:inDebtorsList CAS:debtorsTrue** BSapproveBank BSapproved BSoffer BSOffers LSreplyOffers LSSelectOffers LSWrongOffer Timeout>2h BScancelBank End

C. Generating Event Sequences

A phenomenon in testing interactive systems that most testers seem to be familiar with is that faults can be frequently detected and reproduced only in some context. Further, the coverage criteria can be made more powerful by increasing the value of length coverage to be obtained thereby further reducing any negative effect of reducing the length of tests on the fault detection effectiveness. This makes a test sequence of a length greater than 2 is necessary since repetitive occurrences of some subsequences are needed to a failure to occur. Summary of the results are in Table 9. The Xloan service has execution time 280 ms for length 3 and 276 ms for length 4 while the number of test cases for length 3 is 8 and for length 4 are 11. The Travel agent service has execution time 276 ms for length 3 and 350 ms for length 4 while the number of test cases for length 3 is 15 and for length 4 are 17. The ABC services has execution time 359 ms for length 3 and 400 ms for length 4 while the number of test cases for length 3 is 9, and for length 4 are 10. The BCS-05 Service has execution time 372 ms for length 3 and 500 ms for length 4 while the number of test cases for length 3 is 15 and for length 4 are 18.

D. The Execution Time and Number of Iteration

The execution time and the number of iterations of our approach and the old algorithms are summarized in Table 10. First, the Xloan service has execution time of 414 ms for old algorithm and 260 ms for proposed approach while the number of iterations for old algorithm is 16 and for proposed approach are 10. Second, the travel agent service has execution time 860 ms for old algorithm and 607 ms for proposed approach and number of iterations for old algorithm is 34, and for proposed approach are 18. Third, the ABC service has execution time of 285 ms for old algorithm and 209 ms for proposed approach while the number of iterations for old algorithm is 10 and for the proposed approach are 8. Third, the BCS-05 service has execution time 332 ms for old algorithm and 274 ms for proposed approach, number of iteration for old algorithm is 17, and for proposed approach are 15. We found that when applying the old algorithm to the decomposed graphs this will require repeated algorithm N time equals to N levels of graphs which increases the execution time and the number of iterations of algorithm. Figure 2 shows execution time in millisecond and Figure 3 shows number of iterations for the old and the proposed approaches.

Table 9. Test model information for higher length

Length greater than 2	Execution time		Test cases	
	Length (K=3)	*Length (K=4)*	*Length (K=3)*	*Length (K=4)*
Xloan service	280ms	377ms	8	11
Travel agent	276ms	350ms	15	17
ABC services	359ms	400ms	9	10
BCS-05 Service	372ms	500ms	15	18

Table 10. The execution time and number of iteration for the old algorithm and our proposed approach

Cases	Execution time		Iterations	
	Old	*proposed*	*Old*	*proposed*
Xloan service	414	260	16	10
Travel agent service	860	607	34	18
ABC services	285	209	10	8
BCS-05 Service	332	274	17	15

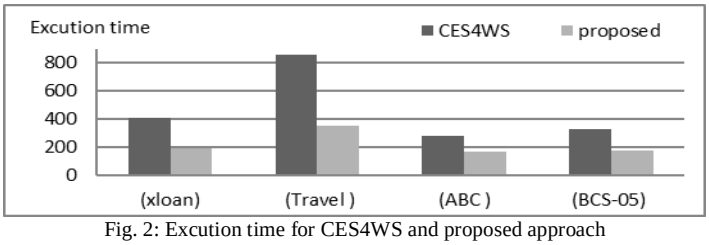


Fig. 2: Excution time for CES4WS and proposed approach

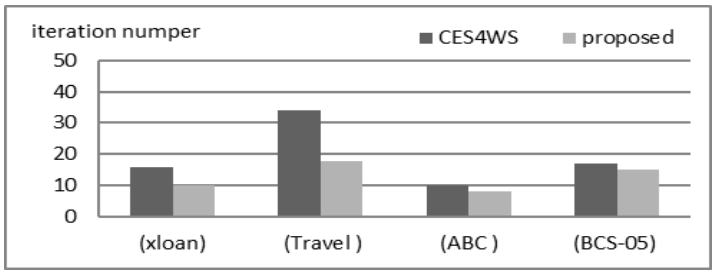


Fig. 3. Iteration time for CES4WS and proposed approach

V. CONCLUSION

Testing is the most critical and expensive phase of the software development life cycle. In this paper, we have improved a technique called Event Sequence Graph for Web services compositions to generate cost-effective test cases for WSCs that decomposed at different levels of abstraction. We found that when the input is multiple graphs, the old approach generates unrelated test cases and takes more iteration and execution time. When our approach is applied, it works properly with less iteration and less execution time and gives related test cases. We introduce algorithms to generate test cases from multi-level graphs. We also generate test cases for length greater than 2 that is necessary since repetitive occurrences of some subsequences are needed to a failure to occur/reoccur. We have evaluated our approach by four case studies with different complexity, parallel execution and decision tables. In the future, we will make our approach holistic to perform positive testing as we do here and also to test undesirable situations (i.e. negative testing) based on the service model. We will also perform other testing stages such as test cases execution. Finally, it is essential to conduct experimental comparisons with other approaches, such as structural testing for web service compositions.

REFERENCES

[1] G. Canfora, and M. Penta, "Service-oriented architectures testing: a survey", In Software Engineering: International Summer Schools (ISSSE), Springer, 2009.

[2] M.Papazoglou,W. Heuvel," Service oriented architectures: approaches, technologies and research issues", The International Journal on Very Large Databases The VLDB journal,vol. 16(3),pp. 389-415,2007.

[3] D. Kung, C.Liu, and P. Hsia, "A model-based approach for testing Web applications", In: Proc. of Twelfth International Conference on Software Engineering and Knowledge Engineering, Chicago, July, 2000.

[4] M.Schmidt, B.Hutchison,and P.Lambros, "The enterprise service bus: making service-oriented architecture real", IBM Systems Journal, vol. 44(4), pp.781-797, 2005.

[5] F. Lars, T.Jan, and R. de Vrie., "Towards model –based testing of web services", International Workshop on Web Services Modeling and Testing, 2006.

[6] A. Benharref, R.Dssouli, R. Glitho,and M.Serhani,"Towards the testing of composed Web services in 3rd generation networks", In IFIP International Conference on Testing of Communicating Systems (TESTCOM), Vol. 3964, pp. 118-133,2006.

[7] F.Belli,and B.Christof, and W.Lee,"Event-based modeling, analysis and testing of user interactions: approach and casestudy", Software Testing, Verification and Reliability, vol.16(1), pp.3-32 ,2006. Van der Aalst WMP. Formalization and verification of event-driven process chains. Information and Software technology,1999;

[8] F.Robert, and B.Rumpe,"Model-driven Development of Complex Software", A research roadmap, Future of Software Engineering,IEEE Computer Society,pp. 37-54,2007.

[9] F.Belli, A.Endo, M. Linschulte, and A.Simao, "A holistic approach to model-based testing of Web service compositions", Software – Practice and Experience, vol. 44(2), pp. 201-234 ,2014.

[10] F.Belli,and B.Christof, and W.Lee,"Event-based modeling, analysis and testing of user interactions: approach and casestudy", Software Testing, Verification and Reliability, vol.16(1), pp.3-32 ,2006.

[11] F.Belli,and M. Linschulte,"Event-driven modeling and testing of real-time Web services", Service Oriented Computing and Applications, 4(1), pp.3-15, 2010.

[12] F.Belli, A.Endo, M.Linschulte,and A.Simao,"Model-based testing of Web service compositions", Service Oriented System Engineering (SOSE), pp. 181-192, IEEE, 2011.

[13] Z.Hong, P.Hall, and J.May,"Software unit test coverage and adequacy", ACM Computing Surveys (CSUR),vol. 29(4), pp.366-427 1997.

[14] A.Paul, and J.Offutt, "Introduction to software testing", Cambridge University Press, 2008.

[15] F.Belli, and C.Budnik,"Minimal spanning set for coverage testing of interactive systems",In First International Colloquium on Theoretical Aspects and Computing (ICTAC),Springer, pp. 220-234, 2004.

[16] Y. Lin, and Z.Yongchang,"A new algorithm for the directed Chinese postman problem" ,Computers & operations research ,vol.15(6), pp. 577-584,1988.

[17] A.Endo, A.Takeshi, and A.Simao,"A systematic review on formal testing approaches for Web services", Brazilian Workshop on Systematic and Automated Software Testing, International Conference on Testing Software and Systems, pp.89, 2010.

[18] L.Mei, C.W, and T.Tse, "Data flow testing of service choreography", Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering, pp. 151-160, 2009.

[19] V.Stoyanova, P.Dessislava, and I.Sylvia, "Automation of test case generation and execution for testing web service orchestrations",Service Oriented System Engineering (SOSE), IEEE, pp. 274-279, 2013.

[20] F.Belli, and C.Budnik,"Towards optimization of the coverage testing of interactive systems", Computer Software and Applications Conference, Vol. 2, pp. 18-19, IEEE, 2004.

[21] W.Douglas " Introduction to graph theory. ",Vol. 2,Upper Saddle River: Prentice hall, 2001.

[22] F.Belli and A.Hollmann, "A holistic approach to testing of interactive systems using statecharts". InProceedings of 2nd South-East European Workshop on Formal Methods (SEEFM 05), South-Eastern European Research Center SEERC 2005 (pp. 1-15).

[23] J.Tretmans ,"Model-based testing and some steps towards test-based modelling". InFormal Methods for Eternal Networked Software Systems 2011 (pp. 297-326). Springer Berlin Heidelberg.

[24] A. Cavalli, TD.Cao, W.Mallouli,"Webmov: A dedicated framework for the modelling and testing of web services composition",InWeb Services (ICWS), 2010 IEEE International Conference,2010 Jul 5 (pp. 377-384),IEEE.

[25] A.Endo,"Using models to test web service-oriented applications.", an experience report, 2012.

[26] https://www.docdroid.net/cbR9qpv/version7detalid.pdf.html, Last accessed: July 2016.

DTSRS: A Dynamic Trusted Set based Reputation System for Mobile Participatory Sensing Applications

Hayam Mousa [a, b], Sonia Ben Mokhtar[b], Lionel Brunie[b], Osama Younes [a], Mohiy Hadhoud [a]
[a] Faculty of Computers and Information, Menoufia University, Egypt
[b] LIRIS, INSA de Lyon, France

***Abstract*—Participatory sensing is an emerging paradigm in which citizens voluntarily use their mobile phones to capture and share sensed data from their surrounding environment in order to monitor and analyze some phenomena (e.g., weather, road traffic, pollution, etc.). Participating users can disrupt the system by contributing corrupted, fabricated, or erroneous data. Different reputation systems have been proposed to monitor participants' behavior and to estimate their honesty. There are some attacks that were not considered by the existing reputation systems in the context of participatory sensing applications including corruption, collusion, and on-off attack. In this paper, we propose a more robust and efficient reputation system designed for these applications. Our reputation system incorporates a mechanism to defend against those attacks. Experimental results indicate that our system can accurately estimate the quality of contributions even if collusion is committed. It can tolerate up to 60% of colluding adversaries involved in the sensing campaign. This enables our system to aggregate the data more accurately compared with the state-of-the art. Moreover, the system can detect adversaries even if they launch on-off attack and strategically contribute some good data with high probability (e.g. 0.8).**

Keywords—Participatory sensing; malicious; collusion attack; On-Off attack; reputation; trust

I. INTRODUCTION

Everyday, millions of people move around carrying a variety of handheld devices equipped with sensing, computing, and networking capabilities (e.g., smartphones, tablets, music players, GPS watches, in-vehicle sensors, etc.). The advancement and widespread use of such devices have contributed toward the emergence of a new kind of application called *participatory sensing* [1]. These applications exploit both the mobility of the participants and the sensing capabilities of their devices to construct opportunistic mobile sensor networks [2].

In participatory sensing, participants capture sensed data from their surrounding environment using a variety of sensors (e.g., GPS, camera, microphone, accelerometer, gyroscope, digital compass, etc.) embedded in their devices. Then, they share their collected observations with a backend server, which processes the received data to monitor, map, or analyze some incidents or phenomena of common interest.

Participatory sensing systems can be applied to serve many of our daily life needs, including monitoring health [4], traffic [5], noise [3], weather, commerce, as well as many other applications [6].

In these applications, no restrictions are usually imposed about the participants' experience, concern, trustworthiness, and interest. In addition, they are not usually paid for their participation in the sensing campaign. Thus, they usually do not have strong motivations to comply with the tasks' requirements. That is, they are not concerned about some parameters which may improve the quality of their contributions (e.g. time, location and/or the position of the device during the sensing process). As a consequence, participatory sensing applications are vulnerable to *erroneous* and *malicious* participants. We define erroneous and malicious participants as those who mislead and disrupt the system measurements by reporting false, corrupted or fabricated contributions either intentionally or non-intentionally. Non-intentional (i.e. erroneous) corruption may originate from a malfunctioning sensor while intended (i.e. malicious) corruption is deliberately committed to alter the system measurements in a specific location. For instance, an adversary can put his device in a non-appropriate position. Alternatively, a participant can modify a contribution before sharing it. Malicious participants may further launch various types of attacks such as Sybil, collusion, on-off attack, etc. Some of these attacks are discussed in Section III. Consequently, the need arises for approaches that try to detect erroneous participants and deter or mitigate malicious ones in order to evaluate the veracity and accuracy of participants' contributions and therefore to build robust and reliable application systems [7].

Among the classical solutions to deal with erroneous and malicious users is the notion of *trust* and *reputation* systems [8]. Some of these systems depend on the *reputation* of entities for assessing the trust of their behaviors.

Reputation is defined as the probability that the past contributions of the participant were correct. Thus, assessing the trust and reputation of participants permits the system to evaluate their expected behavior for their future interactions.

Different reputation systems have been proposed for participatory sensing applications. We have studied, classified, and compared those systems in [9]. It is evident that, those systems are in their infancy, and have several limitations. One of these limitations is the estimation of the quality of contributions in the existence of collusion attack. A few researchers have addressed the estimation of the quality of contributions for example, Huang et al. [10], Wang et al. [11], and Manzoor et.al. [12]. However, such systems are not resistant against malicious colluding adversaries. These systems exploit some consensus and outlier detection algorithms (e.g. [13], [14]) to evaluate the consistency of each contribution. Subsequently, such systems are biased if a good participant is surrounded by a number of colluding adversaries. That is, it ends up getting a good participant defined as malicious and vice versa.

In this paper, we propose a novel and efficient reputation system to estimate the trustworthiness of participants' contributions. The system also adopts a methodology to detect adversaries even if there is a large number of colluding adversaries. It also incorporates other novel parameters, including a proximity factor and users' feedback, to assign a trust score to each contribution. These trust scores give the system the ability to aggregate more accurate data which may reflect the ground truth more precisely compared with the state-of-the-art. The parameters exploited by the system are collected by most existing participatory sensing applications (e.g. data, location, etc.). Thus, our system is applicable to most of typical participatory sensing applications (e.g. noise, pollution, weather, traffic, etc.).

The rest of this paper is organized as follows: Section II states the previous work and its limitations. We then give an overview about the participatory sensing and its threat model in Section III. We describe and discuss in details our proposal in Section IV. The experimental results of our reputation system are discussed in Section V. Finally, we conclude this paper in Section VI.

II. Related WoRk

Different reputation systems have been proposed in literature for different participatory sensing applications. A reputation system for noise monitoring application system is presented by Huang et al. in [10]. This system adopts a robust average algorithm through a watchdog module to measure the quality of the recorded noise samples provided by each participant. In [12], Manzoor et al. measure the quality of participant contribution through a Gaussian membership function. Wang et al., in [I1], use a similarity factor to measure the consistency of each contribution compared with the others. All these systems adopt some outlier detection or consensus algorithms to measure the deviation of each contribution from the common consensus (e.g. [13], [14]). Thus, the results of these systems disrupt if a large number of malicious or colluding adversaries is involved in the sensing campaign.

Other reputation systems have been proposed earlier for social participatory sensing applications in [16], [17], and [18]. These systems mainly depend on some social parameters for estimating the trustworthiness of participants. These parameters include friendship duration, interaction time gap, familiarity, etc. However, these parameters are not usually available in all participatory sensing applications. Thus, these systems are not applicable with the wide range of participatory sensing application.

Complementary to reputation based trust systems, researchers suggest to equip smartphones' sensors with an embedded Trusted Platform Module *TPM* [19], [20] [21], and [22]. Such a module ensures the authenticity of participants' contributions. Furthermore, some TPM based systems can protect data from unauthorized access through applying some authentication and hardware cryptography mechanisms. Although, TPM solutions have some merits, they also suffer from a number of limitations. A major limitation of TPM-based solutions is that they only consider data authenticity regardless of the participant's sincerity and honesty. TPM cannot detect contributions from malicious participants who deliberately initiate sensing actions that cause distortion of their contributions (e.g. putting the device in non-appropriate position). In addition to erroneous contributions that originate from a malfunctioning sensor.

For more details about reputation systems in participatory sensing, its classification, their merits and limitations, and different research directions in this domain, please refer to our survey presented in [9].

III. System Model

In this section, we establish a framework that allows us to analyze the reputation system presented in Section IV.

A. Definitions

Trust and reputation have been defined earlier in the context of participatory sensing by Wang et al. in [11] as follows:

Definition 1: **Trust of a contribution**: The trust of a contribution C, denoted as $Trust(C)$, is the probability of C being correct, as perceived by the server.

Definition 2: **Reputation of a Participant**: The reputation of a participant p_i, denoted as $\hat{R}_{p_i}$, is the synthesized probability that the past contributions sent by p_i are correct, as perceived by the server.

Definition 3: **Participant's Behavior**: The participant p_i is identified as a good participant if he is assigned a reputation score $\hat{R}_{p_i}$ that exceeds a predefined minimum threshold τ. Otherwise, he is identified as a malicious or adversary. The following equation describes this concept.

$$behavior = \begin{cases} Good & if\ \hat{R}_{p_i} \geq \tau \\ Malicious & if\ \hat{R}_{p_i} < \tau \end{cases} \quad (1)$$

B. Threat Model

Participatory sensing applications are vulnerable to a number of attacks. We have defined these attacks in [9]. Below, we define the attacks that are mainly considered along this work (e.g. Corruption, collusion, on-off attacks). We treat them here in the context of participatory sensing applications for the first time.

- **Corruption attack** leads to an erroneous contribution. It may arise as a result of a malfunctioning sensor of a participant's device. In addition, the adversary can deliberately contribute corrupted or forged data. A local processing module can also be used by an adversary for modifying the sensed data before sharing it. An adversary can also initiate sensing actions which may corrupt the sensed data by putting his device in non-appropriate positions. In air quality mapping system, the adversary may put his device beside a cigarette flame. The system should have strong capabilities to identify correct contributions in order to identify and exclude corrupted ones.
- In **Collusion attack**, malicious colluding participants coordinate their behavior in order to provide unified false, corrupted contributions, and/or false feedback. Multiple malicious participants acting together can cause more damage than each one acting independently. If the majority of participants collude they can mislead the system measurements and decisions. In order to attain robustness against such attack, systems should not rely on consensus algorithms to define good and bad contributions. Otherwise, the system measurements and decisions are biased.
- In **On-Off attack**, an adversary alternates between normal and abnormal behaviors. Specifically, the adversary provides false data randomly and irregularly with a probability. The adversary can keep his trust above the required threshold by alternating his behavior as required. This makes it difficult to be detected. To defend against this attack, the system should keep the history of participants and should have a good capability to define their instantaneous trust. The behavior of an on-off attacker is usually unstable along time.

IV. DTSRS: A DYNAMIC TRUSTED SET BASED REPUTATION SYSTEM

A. Overview

In the context of participatory sensing, evaluating the quality of contributions provided by participants is a crucial task. By the term *quality*, we mean how much a contribution is close to the ground truth in the sensing area. In the state of the art, authors measure the consistency of a contribution with the other contributions provided by other participants. However, this measure is usually disrupted especially when there are a large number of adversaries involved in the sensing campaign. In different contexts, the systems rely on a trusted third party that can provide her with the ground truth. However, in the context of participatory sensing, this third party is not available. Thus, we try to propose a more efficient and robust mechanism for evaluating contribution quality depending on a *Trusted Participant* set (TP). We define this set such that *it involves the participants with the higher reputation scores* $\hat{R}_{p_i}$. Those participants are usually more trusted and have higher probabilities to submit good data. Relying on this set of reliable contributions to evaluate the rest of contributions gives our system better idea about which contribution is correct and which is false. Therefore, our system has better capabilities to detect adversaries who contribute bad data. The trusted set is *dynamic* such that it is updated after each campaign in order to base on the most recent reputation information. Thus, we refer to the proposed system as a **Dynamic Trusted Set based Reputation System** (***DTSRS***). In the following, we describe how this trusted set is constructed and updated. Then, we illustrate how this set is exploited to assess trust through the sensing campaign.

B. Trusted Set Management

The trusted set management process is depicted in Fig. 1. This process is composed mainly of two phases, *Initialization phase* **I** and *Main phase* **M**. Sub-figures a and b depict both these phases. First, the *initialization phase* which is carried out for n times when the application is started and there is no available reputation data concerning each participant (step 1I). Therefore, we adopt a methodology to use the most consistent contributions in order to evaluate the others (step 2I). Through this phase, the involved participants are assigned reputation scores that identify them either as good or malicious participants (step 3I, 4I). The details of this phase are illustrated in the following subsection. After a number of iterations (e.g. n), the system can move to the *main phase* (step 1M) and update the trusted set according to the reputation scores of the participants (step 2M). Then, this set is used to evaluate the current contributions and subsequently calculate new trust and reputation scores (step 3M, 4M). Below, we describe in more detail both the initialization and update of the trusted set.

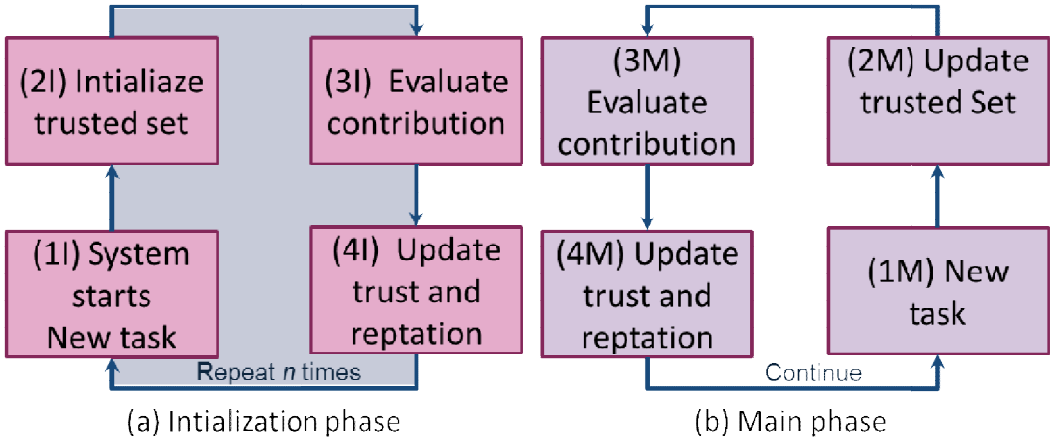


Fig. 1. Trusted Set Management Process

1. The Initialization Phase

The Initialization Phase is illustrated in the algorithm depicted in Fig. 2. This phase starts by determining the size of the trusted participants TP as an input. First, we measure the similarity between each two different contributions (C_{p_i}, C_{p_j}) using some similarity measures as the ones introduced in the field of data mining in [15]. The output of this measure ranges from -1 for completely conflicting contribution to +1 for contributions which are exactly the same. We then calculate the average consistency of each contribution (line 8). Hereafter, the available contributions are arranged according to this average (line 11). Finally, the first contributions that have higher average of consistency are selected and considered as the initial trusted contributions, (lines 12-16).

2. Update the Trusted Set

In this phase, the trusted set is updated according to the current reputation score of the participants. This process is illustrated in the algorithm shown in Fig. 3. The current reputation scores of participants are reported in some way depending on the system methodology (lines 1-10). If the sensing campaign is non-anonymous (e.g. the identity of participants is known), the reputation scores of those participants are retrieved either from a common database or reputation queries are sent to a reputation server. Otherwise, reputation scores are demonstrated using some anonymous demonstration (e.g. anonymous reputation certificate), if the sensing campaign is anonymous. That is reputation scores become accessible at the application server by some way. The contributions of the current task are then arranged according to the reputation scores of their providers (line 11). Finally, the first TP contributions are selected as the trusted set (lines 12-16).

Algorithm 1 Intialiaze the trusted set

Input:

np ▷ %: The number of participants%
TP ▷ %: The size of the trusted set %
C_{p_i} ▷ %: The set of contributions for the task t %

while (C_{p_i} is not the end of the contributions) **do**
  $ASF(C_{p_i}) \leftarrow 0$ ▷ %: Initialize the similarity to 0%
  while (C_{p_j} is not the end of the contributions) **do**
    Measure $S(C_{p_i}, C_{p_j})$ ▷ %: [15] %
    Accumulate $ASF(C_{p_i}) + \leftarrow S(C_{p_i}, C_{p_j})$
    $j \leftarrow j + 1$
  end while
  $SF(C_{p_i}) \leftarrow ASF(C_{p_i})/np$ ▷ %: Calc. average%
  $i \leftarrow i + 1$
end while
Sort C_{p_i} in descending order according to SF
$k \leftarrow 0$
while $k < TP$ **do**
  add C_{p_k} to the trusted set TP
  $k \leftarrow k + 1$
end while

Fig. 2. Initialize Trusted Set Algorithm

Algorithm 2 Update the trusted set

Input:

np ▷ %: The number of participants%
TP ▷ %: The size of the trusted set %
C_{p_i} ▷ %: The set of contributions%

while (C_{p_i} is not the end of the contributions) **do**
  if (p_i is a new participant) **then**
    Set $\hat{R}_{p_i} \leftarrow 0$ ▷ %: Initialize $\hat{R}$ of a new p%
  else if non-anonymous campaign **then**
    Retrieve $\hat{R}_{p_i}$ from reputation database
  else if anonymous campaign **then**
    Get $\hat{R}_{p_i}$ using anonymous demonstration
  end if
  $i \leftarrow i + 1$
end while
Sort C_{p_i} in descending order according to $\hat{R}$
$k \leftarrow 0$
while $k < TP$ **do**
  add C_{p_k} to the trusted set TP
  $k \leftarrow k + 1$
end while

Fig. 3. Update Trusted Set Algorithm

C. Trust Assessment and Reputation Update

In this subsection, we provide an overview about the methodology exploited for trust assessment and reputation in our DTSRS system. Fig. 4 depicts the main trust parameters exploited in our system and identified as blue boxes in the figure. While white boxes are the information sources and the brown ones refer to the modules where these parameters are aggregated for assessing the trust and reputation of each contribution. We target to assess the trust of contributions in a way that the assigned trust scores reflect the consistency of the contributions with ground truth rather than the consistency of them with each other as considered in the state-of-the art. This provides our system with much more resistance against collusion. A brief definition of the function of each module is described as follows:

First, a *contribution evaluation module* evaluates the quality of a participant's current contribution. It measures the deviation of each contribution from the mean of the set of contributions which are provided by the most trusted participants. First, contributions that belong to the same task $Task_t$ are grouped together (step 1 and 2). The trusted set is defined according to the methodology described in the previous subsections. Then, the deviation of each contribution from the mean of those trusted contributions is calculated and assigned a score θ_{p_i} (Step 3).

Second, sensed data are published through a public server. Thus, end users query these data. Those users are themselves a subset of the participants who are involved in the sensing area. They are sometimes permitted to provide a feedback for the received contributions. An accurate feedback usually reflects how much the rated contribution agrees with ground truth as perceived by the user. User q report a feedback about the contribution of a participant p_i noted as $F_q(p_i)$ (step 4). A user may report a feedback that does not reflect his genuine opinion about the target contribution. This is considered as unfair rating attack defined by Jqsang in [25]. Thus, the feedback is evaluated to mitigate the effect of such attack and aggregated to assign a feedback score α_{p_i} to the target participant (step 5).

Third, participatory sensing is usually interested in a specific sensing area. Additionally, sensed data are affected by the distance from the sensing area. For instance, a noise sample recorded by a participant is significantly affected by a nearby sound source such as train station, crowd, etc. This noise is considered to attenuate by going away from its' source [26]. Subsequently, the closer a participant to the sensing area, the more accurate his contribution is considered. Here, we propose to define a proximity factor δ_{p_i} that measures the vicinity of a participant to the center of the sensing area. The contribution is subsequently assigned a score which reflects its possible decay according to the nature of the application (step 6).

Finally, the reputation score $\hat{R}_{p_i}$, which is previously assigned to the participant according to his previous contributions, is also considered (step 7). This score describes the historical behavior of the participant. Thus, it gives an indication of the participant expected behavior during the subsequent tasks. Incorporating the historical reputation score of the participant enables to trace the behavior of the participant and help to detect the ones who launch on-off attack.

In the trust mapping module, the collected measures including θ_{p_i}, α_{p_i}, δ_{p_i}, and $\hat{R}_{p_i}$, concerning the contribution of the target participant p_i are integrated to assign a $Trust$ score to his current contribution (step 8). The reputation score $\hat{R}_{p_i}$ of the participant p_i is then updated to R_{p_i} (step 9). In the following subsection, we discuss the details of these modules.

1) Contribution Evaluation

Consider np is the number of participants who joined the sensing campaign and p_i is one of them who submits a contribution C_{p_i} such that ($i \in \{1,2,3,\dots,np\}$). First, the trusted set TP is defined according to the methodology illustrated above. The mean of those trusted contributions ($C_j \in \{C_1, C_2, C_3, \dots, C_{TP}\}$) is calculated and is noted as $\mu(C_{TP})$ (Equation 2). The higher the similarity of a contribution $C_{p_i}$with this mean, the more reliable it is considered. Thus, the deviation of each contribution, from this mean, is calculated and is noted as *Contribution Deviation* d_{p_i} as depicted in Equation 3.

$$\mu(C_{TP}) = \frac{\sum_{j=1}^{j=TP} C_{p_i}}{TP} \quad (2)$$

$$d_{p_i} = abs\left(C_{p_i} - \mu(C_{TP})\right), i \in \{1, 2, 3, \dots, np\} \quad (3)$$

Where $abs(.)$ is the absolute function, np is the total number of contributions provided by np participants for the considered task.

d_{p_i} ($\forall\, i \in \{1,2,3,\ldots,np\}$) is then normalized to the range [0,1]. The normalized deviation is noted as $d_{p_i}^{no}$ as shown in Equation 4. $d_{p_i}^{no}$ of 0 means that the contribution of p_i is exactly the same as the mean of the trusted contributions. Whereas, 1 means it completely contradicts with this mean. The participant's contribution C_{p_i} is assigned a score θ_{p_i} which reflects its quality by feeding the normalized deviation as an input to an exponential distribution. The output of this distribution is depicted in Fig. 3. Using this distribution, participants are assigned scores which commensurate with their normalized deviation. For example, a contribution with a normalized deviation of 0 is assigned the maximum score 1. The output of the distribution is defined according to Equation 5 and depicted in Fig. 3.

$$d_{p_i}^{no} = \frac{d_{p_i} - \min\{d_{p_i}\}_{i=1}^{np}}{\max\{d_{p_i}\}_{i=1}^{np} - \min\{d_{p_i}\}_{i=1}^{np}} \tag{4}$$

$$\theta_{p_i} = e^{-d_{p_i}^{no}} \tag{5}$$

We also normalized the input (i.e. the deviation) to the range [0,1]. Thus, we have the quality score θ_{p_i} output in the range$[e^0, e^{-1}] \sim \rightarrow [1,0.37]$. We run the experiments different times to determine the most suitable range for normalization. Using the exponential distribution, we found that the normalization of the input to the range [0, 1] and getting an output in the range [1, 0.37] allows for more accurate data aggregation. That is calculating the weighted sum of the available contributions according to these scores is more close to the ground truth.

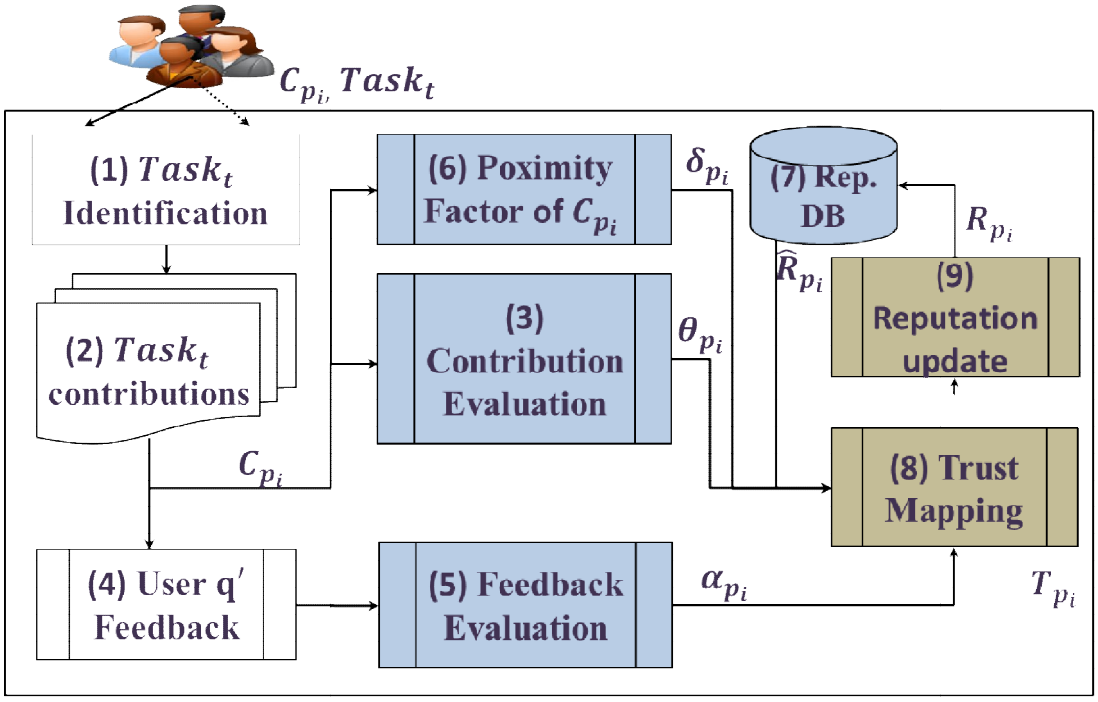


Fig. 4. The Framework of our DTSRS system

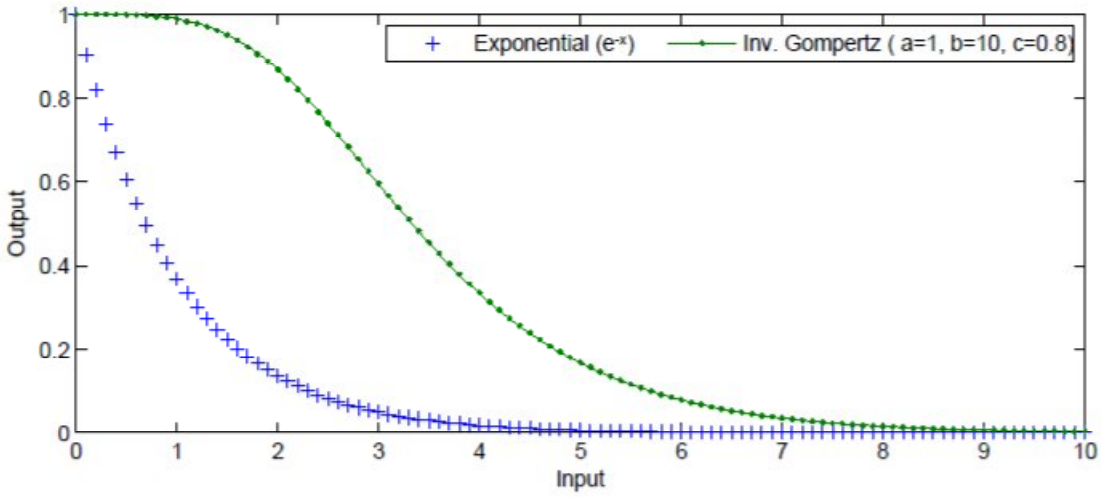


Fig. 5. The output of both exponential and Inv. Gompertz

2) Feedback Processing

This module targets to evaluate the user's feedback $F_q(p_i)$ which lies in the range [0, 1]. For this, if the reputation score of the rater q exceeds the reputation score of the target participant p_i (i.e. $\hat{R}_q > \hat{R}_{p_i}$) the reputation $\hat{R}_q$ of the rater q is used as a weight for his provided rating, as depicted in Equation 6. Otherwise, the rater's feedback is excluded. Consequently, the rate provided by a poor reputation user is less considered, and vice versa. Different feedback scores which assigned for the same contribution are aggregated. An average feedback score α_{p_i} is then calculated according to Equation 7, where FP is total number of feedback providers. The aggregated feedback score lies also in the range [0, 1].

$$F_{q_{Evl}}(p_i) = F_q(p_i) \times R_q \quad if \quad \hat{R}_q > \hat{R}_{p_i} \ \forall\, q \in \{1,2,3,\dots,F\} \tag{6}$$

$$\alpha_{p_i} = \frac{\sum_{q=1}^{q=FP} F_{q_{Evl}}(p_i)}{FP} \tag{7}$$

3) A Proximity Factor

The proximity factor, as we mentioned earlier, measures the vicinity of a participant to the sensing area. As a first step towards the calculation of this measure, the distance between the center of the *Target Sensing Area* **TSA** and the *Sensing Location* (SL_{p_i}) where the contribution is captured by the participant p_i is calculated. Here, we adopt the Euclidean distance as a simple and common distance measure (Equation 8).

$$\beta_{p_i} = \sqrt{(TSA_x - SL_x(p_i))^2 + \left(TSA_y - SL_y(p_i)\right)^2} \tag{8}$$

Where TSA_x and TSA_y and $(SL_x(p_i), SL_y(p_i))$ are the coordinates of the TSA and $SL(p_i)$ respectively

The proximity score depends on the considered phenomenon and its dispersion rate. Some phenomena are location *sensitive* such as noise, pollution, traffic, etc. Other phenomena are more *stable* in the sensing area such as temperature and precipitation. Thus, for this measure, the administrator of the application server has to classify the application according to its sensitivity to the sensing location (i.e. sensitive or stable). The class of the considered phenomena defines the way in which the proximity factor is calculated.

The calculated distance is used to assign a proximity score according to the class of the application. Firstly, we consider a stable phenomenon which is steady in different locations in the sensing area. The same weight is assigned to all contributions which are captured inside the sensing area. Subsequently, a participant is assigned a proximity score δ_{p_i} which is either 1 or 0 to indicate his existence either inside or outside the sensing area respectively, as depicted in Equation 9 where r is the radius length of the sensing area.

$$\delta_{p_i} = \begin{cases} 0 & where \ \beta_{p_i} \geq r \\ 1 & where \ \beta_{p_i} < r \end{cases} \tag{9}$$

Alternatively, for location sensitive applications, we use the calculated distance β_{p_i} as an input to the inverse Gompertz function to calculate the proximity factor δ_{p_i}, as depicted in Equation 10. The output of this function is depicted in Fig. 3. Through this function, participants are assigned maximum proximity scores when they are at the center of the sensing area. These scores decrease gradually as they go away from the center. For instance, a participant of distance zero ($\beta_{p_i} = 0$) is exactly in the center of the sensing area. Such participant is assigned the upper proximity score which is 1. Thus, this score allows the application server to trust more the contributions which originate nearby the center of the sensing area.

$$\delta_{p_i} = 1 - a \times e^{-be^{-c\beta}} \tag{10}$$

Where, a is the upper asymptote. b controls the displacement of the output along the x axis and c adjusts the growth rate of the function. a, b, and c is selected such that the function output matches the radius of the sensing area.

4) Trust Mapping

The calculated parameters are aggregated to calculate $Trust$ of the considered contribution according to the definition of trust presented earlier. These parameters include the current contribution evaluation θ_{p_i} evaluated by the contribution evaluation module, the aggregated feedback α_{p_i}, the proximity factor δ_{p_i}, and the reputation score $\hat{R}_{p_i}$ assigned to the participant through the previous campaign, see Equation 11.

$$Trust(C_{p_i}) = W_1 \times \alpha_{p_i} + W_2 \times \alpha_{p_i} + W_3 \times \delta_{p_i} + W_4 \times \hat{R}_{p_i} \tag{11}$$

Where $\sum_i^4 W_i = 1$, $\hat{R}_{p_i}$ of a new participant is set to 0 in order not to give a new participant the ability to inject bad data to the system unless he behaves correctly for a period of time.

5) Reputation

The value of the reputation score $\hat{R}_{p_i}$ of the participant p_i is update to a new value R_{p_i}. The reputation update process depends on the quality score assigned to the participant contribution θ_{p_i}. If this score is greater than a predefined threshold τ, the participant is rewarded by increasing his reputation score with ε_r such that the output reputation does not exceed 1. Oppositely, if the contribution quality score is below this threshold, the participant is penalized by decreasing his reputation score with ε_p such that the reputation score is not less than 0. We set $\varepsilon_r, \varepsilon_p$, this makes adversaries aggressively penalized while reputation is built gradually. The reputation update process is formulated in the following equation.

$$R_{p_i} = \begin{cases} min\{\hat{R}_{p_i} + \varepsilon_r, 1\} & if\ \ \theta_{p_i} \geq \tau \\ max\{\hat{R}_{p_i} - \varepsilon_p, 0\} & if\ \ \theta_{p_i} \geq \tau \end{cases} \tag{12}$$

The reason behind the exploitation of the contribution score θ_{p_i} to calculate the reputation of participants and not the trust score of his contribution is apparent for different reasons. Firstly, the trust score incorporates the proximity factor. Whereas, the calculation of this factor depends on the location of the participant which is constrained by the participant's habits and his daily activities. Thus, this factor only affects the reliability of the contribution but not the honesty of the participant. Thus, it should not affect the participant's reputation score. Secondly, using the score of the current contribution allows us to update the reputation score of the participant such that it reflects the most recent behavior of the participant.

V. Experimental Evaluation

We implemented our scheme with a MATLAB simulation to measure the accuracy of our reputation and trust assessment method. Since the communication is not our concern, we implemented both the server and participants on the same machine.

A. The application

In this simulation, we consider a noise monitoring application. Thus, we generated the data in accordance with a real noise levels described in [27]. We consider a sensing area where the mean μ of the noise data at the center of this area is 60 db. The noise waves are considered to attenuate due to scattering and absorption. The amplitude of the attenuated wave is calculated according to Equation 13.

$$A = A_0 . e^{-\sigma Z} \tag{13}$$

Where A_0 is the unattenuated noise wave at the center of the sensing area, A is the reduced amplitude after the wave has traveled a distance Z, while σ is the attenuation coefficient of the signal traveling in the Z direction. We

consider $\sigma = 0.0023$. The term e is the exponential (or Napier's constant). The units of the attenuation value in Napier per meter can be converted to decibel/meter by dividing by 0.1151. We also consider a sensing area of radius 300m.

Good participant always send correct sensing data which commensurate with their location. However, we assume that, adversaries launch on-off attack. They send correct data to gain a high reputation scores, then they randomly send false sensing data. The probability by which an adversary sends correct data is referred to as its *nature*. We set the mean of false data to deviate from the correct data such that this mean corresponds to a different level of noise $\mu + \mu/3$ (i.e. 80 db). This means that an adversary contributes data which correspond to a completely different level of noise. Thus, only one false report has an impact on the measurements. Furthermore, we assume that, all false reports support each other. Hence, we consider the worst case when all adversaries collude to cause the biggest possible disturbance to the system. However, this case can be hardly met in realistic systems, but it enables us to evaluate our system under the most difficult circumstances. We generated a random sensing location for each participant such that they are uniformly distributed along the sensing area. Table 1 lists our default parameter settings.

B. The System Parameters

In this test, we measure the effect of using different values of the trust parameters' weights exploited in Equation 11, to see how they affect the calculated trust of contributions. In this test, we set the weights of both the feedback and the proximity factor to 0. We test different values of W_1 versus W_4 which correspond to the weight of the contribution quality and the weight of the reputation of the contribution provider respectively.

TABLE I. TABLE I: DEFAULT PARAMETER SETTINGS

Parameter	Value
Number of participant for each task NP	100
The correct noise amplitude at the center	60db
The value of adversary noise amplitude $\mu + \mu/3$	80db
W_1	0.4
W_2	0.0
W_3	0.2
W_4	0.4
r	300m
a; b; c	1, 10, 0.3
ϵ_r	0.02
ϵ_p	0.5

We run the experiment and measure the effect of the weight parameter on the trust score assigned to good participants' contributions as depicted in Fig. 6. It is observed that good contributions are not trusted for a number of tasks while the weight of the reputation score has much more strength compared with the weight of the contribution quality ($W_4 > W_1$). This is because the trust score is affected by the historical reputation which is just initialized to 0. This effect is released when W_1 increases and W_4 decreased. Thus, good contributions are trusted from the first task even if the reputation score of their provider is not so high. That is W_1 increases the effect of the instantaneous behavior over the historical behavior. Whereas, the increase of W_4 supports the effect of the participants' historical behavior on the calculated trust. To this end, we set the weights of both reputation and the quality of contribution (W_1 and W_4) to equal values in the other experiments in order for the trust to reflect the behavior of the participant through previous campaigns and to indicate also the quality of his current contribution as well.

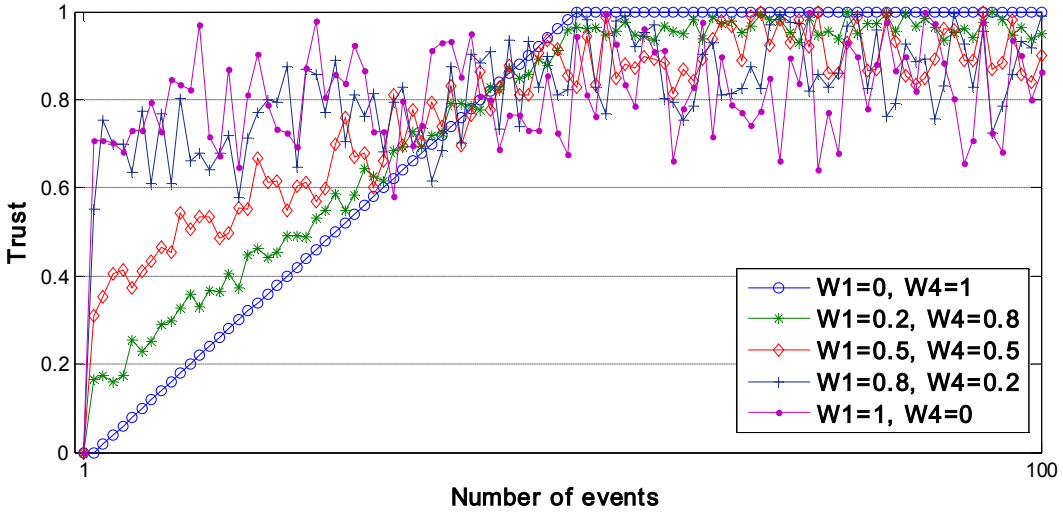


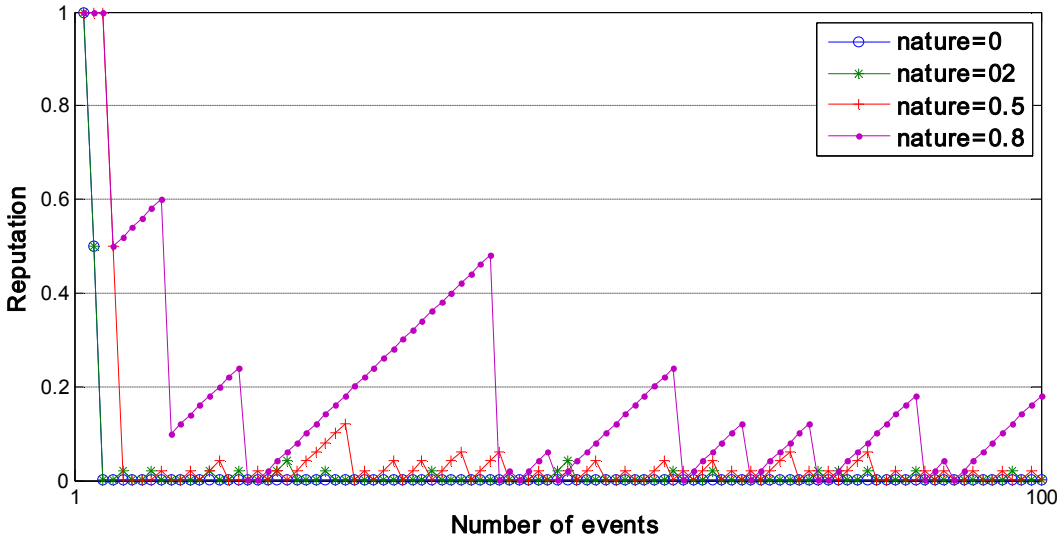
Fig. 6. Impact of using trust weights on trust of good contributions

C. On-off Attack

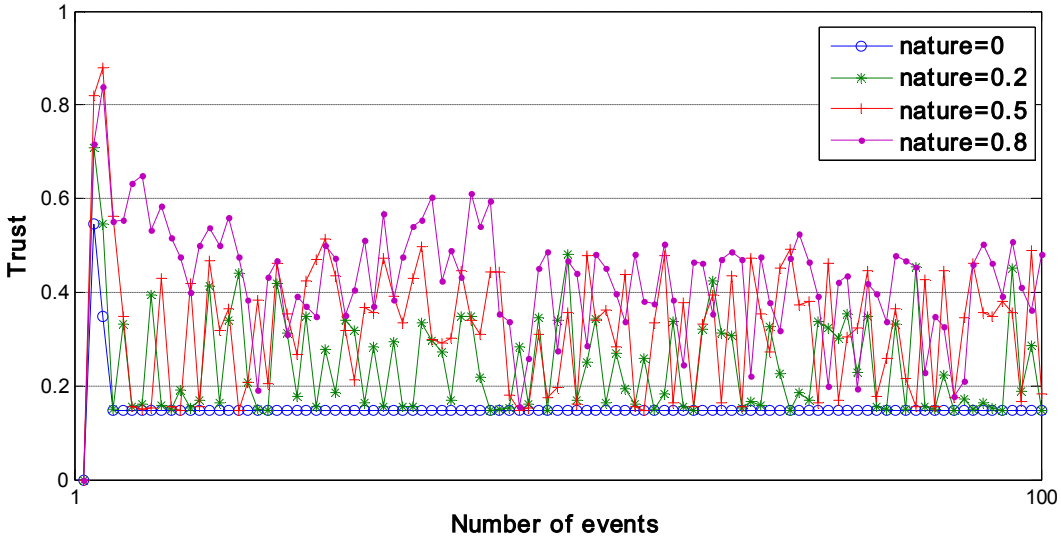
In this experiment, we test the system robustness against on-off attack. We measure how the reputation and trust scores of an on-off attacker are affected by his nature. We study the behavior of five adversaries with different values of nature 0, 0.2, 0.5, 0.8. The nature represents the probability by which an on-off attacker sends correct data. To test the worst case, we assume that the five adversaries behaved in good manner until their reputation scores have reached 1 before the test. We then run this experiment for 100 tasks.

In Fig. 7 (a), we can see that the reputation score of an adversary degrades. The reputation scores of adversaries with nature 0, 0.2, and 0.5, drop down very quickly until it reaches 0. While the reputation scores of adversaries with higher nature (e.g. 0.8) still drop down more slowly. These scores drop to a very low level even if the adversary sends correct data with a very high probability (i.e. 0.8 in this case). An adversary is severely punished for each bad transaction but rewarded gradually for good ones. Thus, bad transactions have larger influence on the reputation score.

We examine the computed trust scores assigned to reports sent by those adversaries. Fig. 7 (b) depicts these results. It is obvious that, the trust score of reports received from adversary with nature 0 are usually assigned a score around 0.2 the minimum possible value of trust. This value result when the reputation score is 0 and contribution quality is very bad $d_{p_i} = -1 \rightarrow$ (i.e. θ_{p_i}~0.37). Thus, $Trust \rightarrow (0.4 \times 0.37 + 0.4 \times 0 + 0.2 \times \delta_{p_i})$, where $\delta_{p_i} \in [0,1]$. $Trust \rightarrow [0.15,0.25]$. While the trust of reports from adversaries with nature 0.2, 0.5, and 0.8 fluctuates much more since they sometimes send correct data. However, their trust scores do not usually exceed 0.5. That is the system does not trust an adversary even if he sends correct reports with high probability (e.g. 0.5 and 0.8).



(a) Reputation



(b) Trust

Fig. 7. Impact of the adversaries' nature on the reputation and trust in the proposed DTSRS system

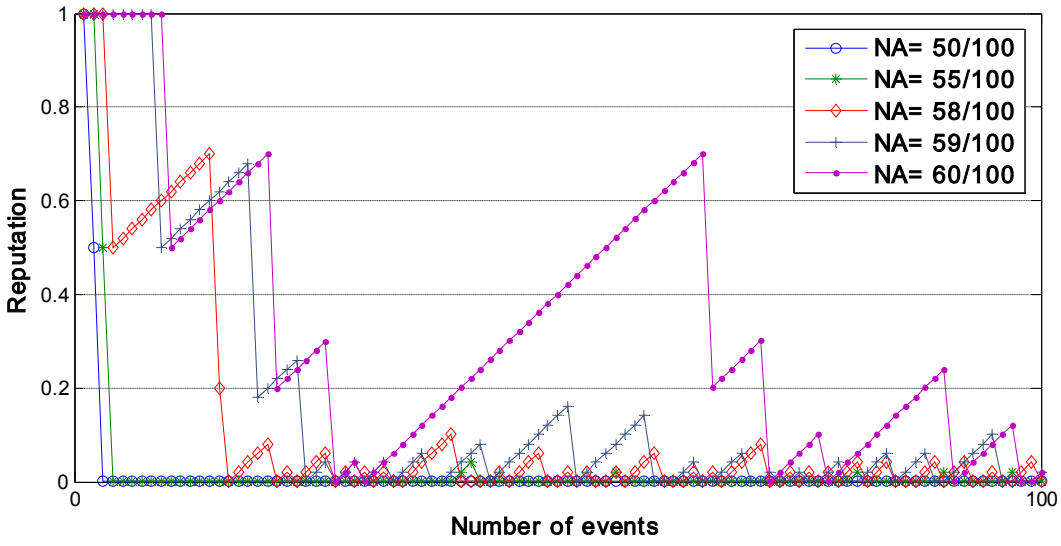
D. Collusion attack

In the following experiment, we measure the resistance of our system under collusion attack. So, we need to define the number of adversaries NA with which the system fails to detect the adversaries under the predefined test setup. In this test, we set the nature of all adversaries to be 0.

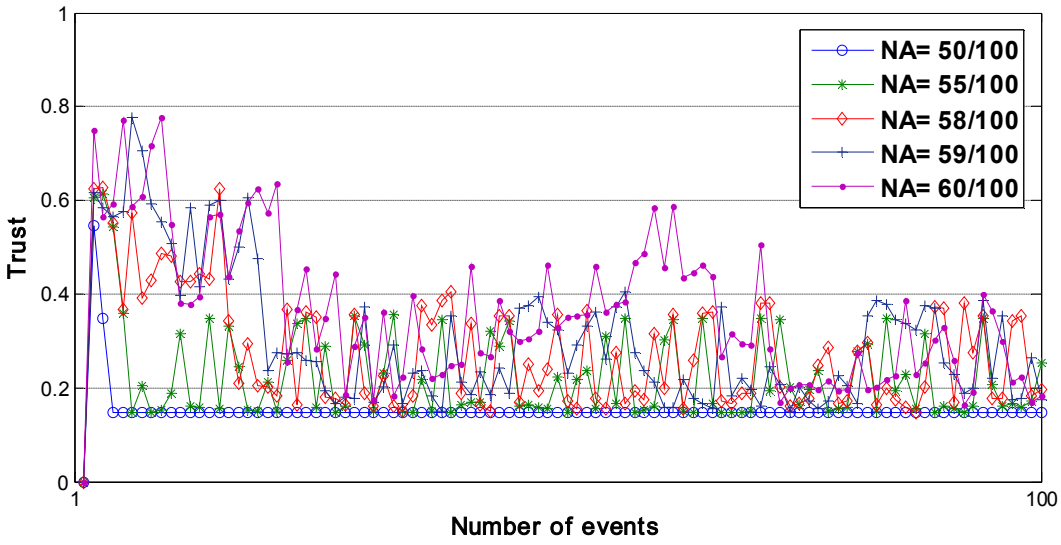
Intuitively, the system perfectly detects adversaries as long as the total number of adversaries is less than the size of the trusted set. In this case, there is no intersection between the trusted set and adversaries. We need to test to which extent the trusted set can involve some adversaries and still properly detect adversaries. Thus, we vary the number of adversaries NA in the campaign as 50, 55, 58, 59 and 60 where the trusted set size is 60. That is the trusted set involves 10, 15, 18, 19, and 20 adversaries respectively.

Fig. 8 (a) depicts the results of this test. It is evident that, the reputation scores of adversaries rapidly drop down to reach zero. However, this drop becomes slow with the increase in the number of adversaries from 58, 59, and 60. However, the system fails to identify adversaries while the number of adversaries NA surpasses 60 adversaries.

Fig. 8 (b) shows the trust of the contributions of those participants. It is clear that, adversaries' contributions are usually assigned low trust scores (i.e. [0.2, 0.4]) even if the number of adversaries reaches 60% of the total number of participants. However, when the number of adversaries reaches 60% of the total number of participants, the trust of adversary's contribution fluctuates. This means that the DTSRS proposed system does not trust adversaries' contributions (i.e. $Trust < 0.5$), where the number of adversaries reaches 60% of the total number of the participants under the current test setup. This reflects that our system is resistant under collusion attack.



(a) Reputation



(b) **Trust**

Fig. 8. Impact of the number of adversaries on reputation and trust of an adversary in the proposed DTSRS system

E. Comparison

We measure the accuracy of the aggregated data and how much it agrees with the ground truth. In this experiment, we evaluate the usage of the exponential distribution as a mapping function compared with the functions used in the state-of-the-art such as Gompertz in [10], and Gaussian function in [12]. We compute the scores $Trust$ assigned to each contribution C_{p_i} according to each function f. These scores are then used to calculate the average of the collected contribution of each task as shown in Equation 14. We run the experiment for 100 tasks.

$$v_{f,t} = \frac{\sum_{i=1}^{NC}(Trust_f(C_{p_i}) \times C_{p_i})}{NC} \quad (14)$$

Where f is the reputation function (e.g. f can be exponential, Gompertz, or Gaussian), i is the contribution number and NC is the total number of contributions available for the task t.

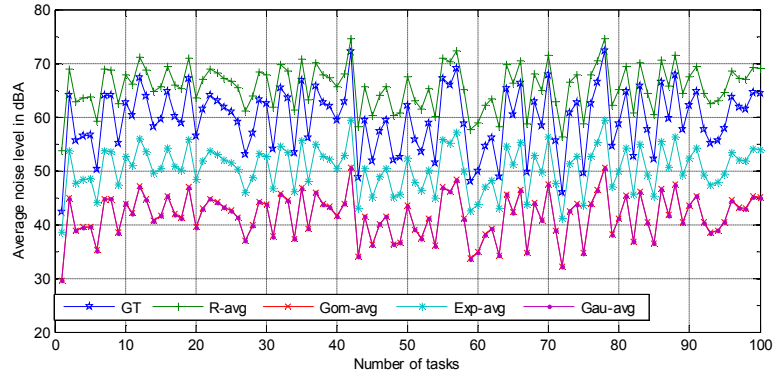
We also include the raw average; it is calculated by averaging the collected contributions for the same task without incorporating any additional scores. We run the experiment while the number of adversaries is 30/100 with nature 0. We consider false data with four different mean values 80, 100, 120, and 150 dBA, each one is considered in a separate run. The large values of false data cause more disruption for the aggregated data. We run the experiment four times while each run contains 100 tasks and there are 100 contributions for each task. We measure the deviation, of the average calculated according to Equation 14 in each task, from the correct ground truth which has a mean of 60 dBA.

The results of this experiment are shown in Fig. 9 subfigues(a, b, c, and d). The closeness of the calculated average to the ground truth' plot indicates the accuracy of the mapping function for assigning appropriate trust score to each contribution. As it can be observed, the raw average is significantly different from the ground truth since all contributions are equally considered. By looking at the raw average data in the different sub-figures, it becomes worse with the increase of the mean value of false data in sub-figures a, b, c, and d. Both Gompertz and Gaussian averages also deviate significantly from the ground truth. By looking at the sub-figures a, b, c, and d, the Gompertz and Gaussian based averages nearly achieve the same deviation from the ground truth whatever the mean of false data. On the other hand, the average calculated based on exponential distribution not only approximates the ground truth more closely in all sub-figures a, b, c, and d. Additionally, the performance of exponential distribution based average enhances from a to b and c, and it has the best performance in d. That is the exponential based average system can perform better when the mean value of false data significantly deviate from the correct ones.

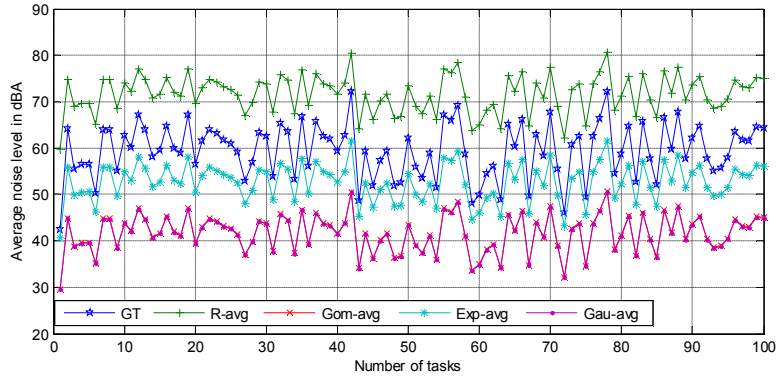
This means that our system has better capabilities to reflect the nature of the ground truth data even if the system faces a massive disruption. This is because our proposed exponential distribution has much sharper degradation which allows it to highly consider contributions which have a slight deviation from the correct ones (i.e. the ones that have a deviation more close to 0). In addition, it allows to aggressively assigning bad scores to the ones which have much more deviation. Thus, they are less considered. Therefore, the calculated data average has better capabilities to reflect the ground truth data than the Gompertz and Gaussian.

VI. CONCLUSION

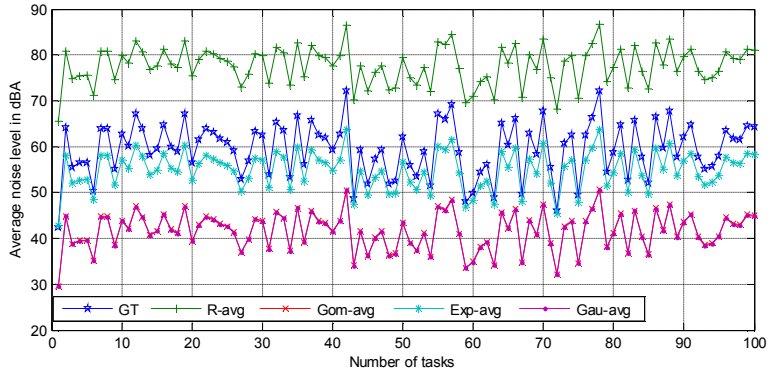
In this paper, we propose the DTSRS reputation system for participatory sensing applications. The system depends on a dynamic trusted set of participants to identify the good data in each campaign. DTSRS system also incorporates other parameters such as the vicinity to the sensing area and the users' feedback to calculate a trust and reputation score for each participant. We experimentally evaluated the system by incorporating it within a simulated system for noise monitoring participatory sensing application. The results indicate that DTSRS system accurately assesses the quality of participants' contributions. It exposes the average of the aggregated data to the minimum possible noise. In addition, the system clearly identifies adversaries even if the number of colluding adversaries reaches 60% of the total number of participants in the campaign. Furthermore, adversaries who launch on-off attack are clearly identified even if they contribute good data with high probability (e.g. 0.8). Therefore, the proposed DTSRS reputation system can defend against corruption, On-Off, and collusion attacks which are not considered in literature. In a future work, we target to manage both the conflicting objectives of trust assessment and privacy preservation of participants in participatory sensing environment. Thus, we are going to incorporate the DTSRS reputation system within a privacy preserving framework.



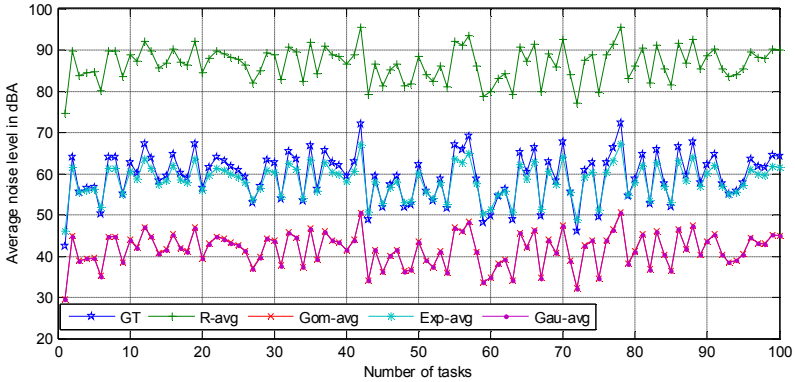
(a)false data 80 dBA



(b)false data 100 dBA



(c)false data 120 dBA



(d)false data 150 dBA

Fig. 9. Average aggregated noise level where the mean of a correct data is 60 dBA and the mean of a false data is 80, 100, 120, 150 depicted in sub-figures a, b,c, and d respectively

REFERENCES

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in Workshop on World-Sensor-Web (WSW 06): Mobile Device Centric Sensor Networks and Applications, 2006, pp. 117–134.

[2] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," IEEE Communications Magazine, vol. 48, no. 9, pp. 140–150, Sep. 2010.

[3] E. Kanjo, J. Bacon, D. Roberts, and P. Landshoff, "Mobsens: Making smart phones smarter," IEEE Pervasive Computing, vol. 8, no. 4, pp. 50–57, Oct 2009.

[4] L. Nachman, A. Baxi, S. Bhattacharya, V. Darera, N. Kodalapura, V. Mageshkumar, S. Rath, and R. Acharya, "Jog falls: A pervasive healthcare platform for diabetes management," in Pervasive Computing, vol. 6030, May 2010, pp. 94–111.

[5] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "Greengps: A participatory sensing fuel-efficient maps application," MobiSys, 2010, pp. 151–164.

[6] W. Khan, Y. Xiang, M. Aalsalem, and Q. Arshad, "Mobile phone sensing systems: A survey," IEEE Communications Surveys Tutorials, vol. 15, no. 1, pp. 402–427, First 2013.

[7] A. Kapadia, D. Kotz, and N. Triandopoulos, "Opportunistic sensing: Security challenges for the new paradigm," COMSNETS'09, 2009, pp. 127–136.

[8] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," Communications of the ACM, vol. 43, no. 12, pp. 45–48, 2000.

[9] H. Mousa, S. B. Mokhtar, O. Hasan, O. Younes, M. Hadhoud, and L. Brunie, "Trust management and reputation systems in mobile participatory sensing applications," Computer Networks, vol. 90, no. C, pp. 49–73, Oct. 2015.

[10] K. L. Huang, S. S. Kanhere, and W. Hu, "On the need for a reputation system in mobile phone based sensing," Ad Hoc Networks, vol. 12, pp. 130–149, 2014.

[11] X. O. Wang, W. Cheng, P. Mohapatra, and T. Abdelzaher, "Enabling reputation and trust in privacy-preserving mobile sensing," IEEE Transactions on Mobile Computing, vol. 99, p. 1, 2014.

[12] A. Manzoor, M. Asplund, M. Bouroche, S. Clarke, and V. Cahill, "Trust evaluation for participatory sensing," in MobiQuitous, 2012, pp. 176–187.

[13] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "Loci: fast outlier detection using the local correlation integral," in International Conference on Data Engineering, 2003, 2003, pp. 315–326.

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," SIGMOD Rec., vol. 29, no. 2, pp. 93–104, May 2000.

[15] C. Beecks, M. S. Uysal, and T. Seidl, "A comparative study of similarity measures for content-based multimedia retrieval," in ICME, 2010, pp. 1552–1557.

[16] H. Amintoosi and S. S. Kanhere, "A reputation framework for social participatory sensing systems," MONET, vol. 19, no. 1, pp. 88–100, 2014.

[17] ——, "A trust framework for social participatory sensing systems," in MobiQuitous, 2012, pp. 237–249.

[18] R. R. Kalidindi, K. Raju, V. V. Kumari, and C. S. Reddy, "Trust based participant driven privacy control in participatory sensing," CoRR, vol. abs/1103.4727, 2011.

[19] A. Dua, N. Bulusu, W.-C. Feng, and W. Hu, "Towards trustworthy participatory sensing," in Proceedings of the 4th USENIX Conference on Hot Topics in Security, 2009.

[20] A. Dua, W. Hu, and N. Bulusu, "Demo abstract: A trusted platform based framework for participatory sensing," in IPSN, 2009 [21] P. Gilbert, L. P. Cox, J. Jung, and D. Wetherall, "Toward trustworthy mobile sensing," in HotMobile '10, 2010, pp. 31–36.

[21] P. Gilbert, J. Jung, K. Lee, H. Qin, D. Sharkey, A. Sheth, and L. P. Cox, "Youprove: authenticity and fidelity in mobile sensing," in SenSys, 2011, pp. 176–189.

[22] C. Marforio, A. Francillon, S. Capkun, S. Capkun, and S. Capkun, " Application collusion attack on the permission-based security model and its implications for modern smartphone systems," Department of Computer Science, 2011.

[23] H. Alzaid, E. Foo, J. G. Nieto, and E. Ahmed, "Mitigating on-off attacks in reputation-based secure data aggregation for wireless sensor networks," Security and Communication Networks, vol. 5, no. 2, pp. 125–144, 2012.

[24] A. Jøsang and J. Golbeck, "Challenges for robust of trust and reputation systems," in (STM), 2009.

[25] L. L. Beranek and I. L. Ver, "Noise and vibration control engineering-principles and applications," John Wiley & Sons, vol. 1, 1992.

[26] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, "Fundamentals of acoustics," 4th Edition, by Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, James V. Sanders, pp. 560. ISBN 0-471-84789-5. Wiley-VCH, December 1999., vol. 1, 1999.

A Comparative Study for Arabic Text Classification Based on BOW and Mixed Words Representations

Rouhia M.Sallam
Faculty of Applied Sciences, Taiz University, Yemen
Email: rohiya79@yahoo.com
Hamdy M. Mousa and Mahmoud Hussein
Faculty of Computers and Information, Menoufia University, Egypt
Email: {hamdimmm@hotmail.com and mahmoud.hussein@ci.menofia.edu.eg}

***Abstract*- This paper compares two methods for features representation in Arabic text classification. These methods are bag of words (BOW) that mean the word-level unigram and mixed words representations. The mixed words use a mixture of a bag of words and two adjacent words with different proportions. The main objective of this paper is to measure the accuracy of each method and to determine which method is more accurate for Arabic text classification based on the representation modes. Each method uses normalization and stemming. The results show that the use of mixed words in features representation achieves the highest accuracy by 98.61% when normalization is used.**

Keywords—Arabic Text Categorization, Frequency Ratio Accumulation Method, Term and Document Frequency, Features Selection, bag of words and Mixed Words.

.

I.INTRODUCTION

Text Categorization (TC) is an automatic process for grouping documents based their contents into pre-defined categories that are known in advance [1]. There are a tremendous number of Arabic text documents that are available online which are growing every day. As a consequence, text categorization becomes very important and a fast growing research field. The developments of such text classification systems for Arabic documents are a challenging task because of the complexity of the Arabic language. The language has a very complex morphology and high inflection. It consists of 28 letters and is written from right to left. In addition, most of the Arabic words have a tri-letter root [2]. However, there is still a limited research for the Arabic text categorization due to the complex and rich nature of the Arabic language compared to other languages [3, 4].

There are several different techniques for automatic text classification including Support Vector Machines (SVM), K- Nearest Neighbor (KNN), Neural Networks (NN), Decision Trees (DT), Maximum Entropy (ME), Naïve Bayes (NB), and Association Rules [5- 8]. Most of these techniques have complex mathematical and statistical models and power consuming and do not usually lead to accurate results for the categorization [9].

In this paper, we compare two methods that use for represented the features in Arabic text classification. These methods are bag of words (BOW) and the mixed words which is a mixture of a bag of words and two adjacent words with different proportions. Also used Term Frequency (TF) technique in features selection. In addition, a simple mathematical model is used which called Frequency Ratio Accumulation Method (FRAM). Normalization and stemming approaches are also used.

This paper is organized as follows. In Section 2, an overview of the related work is presented. Section 3 introduces the proposed comparative process for the two representations. Section 4 presents the experimental results. Finally, conclusions and future work are put forward in Section 5.

II. RELATED WORK

Arabic text documents available online are growing every day. Arabic language has complex internal word structures and the complicated construction of Arabic words from their roots.

Al-Shargabi [10] has compared three techniques for Arabic text classification based on stop words elimination. These techniques are: Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Naïve

Bayesian (NB), and J48 [6]. He have used vector space techniques in features representation, and Arabic data set contains 2363 documents divided into six categories: Sport, Economic, Medicine, Politic, Religion, and Science, where 60% of them is used for training and the rest "40%" are used as testing. The results of accuracy using these techniques achieved 94.8%, 89.42% and 85.07% respectively.

Wahbeh A. et al. [11] introduced an approach for automatic Arabic text classification with and without the stemming application. Using Support vector machine (SVM), Decision Trees (C4.5), Naive Bayes (NB) Classifiers, unique words were extracted in representing the features. They collected a corpus from different trusted websites. The corpus consists of 1100 documents that classified into nine categories: agriculture, art, economics, health and medicine, law, politics, religion, science, and sports. They used accuracy, recall, precision and F-measure to evaluate experiments. The results achieved are 87.79%, 88.54% for the accuracy with SVM and Naïve Bayes respectively when stemming is used. On the other hand, the results achieved when stemming is not applied have lower accuracy (i.e. 84.49% and 86.35%).

Nezreg et al. [12] introduced multiple methodologies for the categorization of English text automatically, these methodologies combine Bag-of-Words, Bag-of-Concepts, and Bag-of-Words with Bag-of-Concepts in representation patterns. Three classifiers were used: SVM, decision trees and KNN and two corpus were used that have 11 categories of reuters-21578 articles and 7 categories of 20 newsgroup. They used precision for evaluating the classification. The results with decision trees give better results than SVM and KNN for their semantic aspect in the classification. The improvements that happened is 22.85% for 20 newsgroups corpus and 4.65% for the Reuters-21578 corpus.

Mesleh et al. [13] proposed an approach for Arabic text classification using SVMs compared to other classification methods (SVMs, Naïve Bayes and k-NN). They collected a corpus from Al-Jazeera, AlNahar, Al-hayat, Al-Ahram, and Al-Dostor. The corpus consists of 1445 documents that classified into nine categories (Computer, Economic, Education, Engineering, Law, Medicine, Politics, Religion and Sport). They used the vector space representation to represent the Arabic documents. CHI square method was used as a feature selection. They used normalization but not stemming because it is not always beneficial for text categorization since many terms may be conflated to the same root [14]. The experimental results show that classification effectiveness is 88.11% using SVMs.

Hadni et al. [15] introduced a new method for Arabic multi word terms (AMWTs) extraction based on a hybrid approach. They used linguistic AMWTs approach to extract the candidate MWTs based on Part Of Speech (POS). A statistical approach is also used to incorporate the contextual information by using a proposed association measure based on Term-hood and Unit-hood for AMWTs extraction. They used three statistical measures: C-Value, NC-Value and NTC-Value [16] for evaluation by two steps (i.e. reference list and validation).

Diab [17] has used multi-word features in Arabic document classification and two similarity functions: the cosine and the dice similarity functions. He also applied inverse document frequency (IDF) to prevent frequent terms from dominating the value of the function, and used different light stemmers on multi-word features. The dataset was collected from well-known Arabic websites that contain 300 documents. The results show that unordered pairs produce 2% improvement compared to ordered pairs while ordered triples produce bad results.

Zhang et al. [18] used multi-word features representation with support vector machine as a classifier to improve document classification. They proposed a method based on the adaptation of mutual information (MI) and context dependency for compound words extraction from very large Chinese Corpus, and they report that their method is efficient and robust for Chinese compounds extraction. Two strategies were developed based on the different semantic level of the multi-words. The first is the decomposition strategy using general concepts for representation and the second is combination strategy using subtopics of the general concepts.

Suzuki and Hirasawa [19] proposed a new classification technique called the Frequency Ratio Accumulation Method (FRAM). N-gram character and the word N-gram are used as feature terms. The performance for FRAM outperforms the Naive Bayes method (baseline method).The technique is evaluated through a number of experiments using newspaper articles from Japanese CD-Mainichi 2002, and English Reuters-21578. The classification accuracy is the highest when word N-grams is used as feature terms, the results of accuracy are 87.3% for Japanese CD-Mainichi 2002 and 86.1% for English Reuters-21578.

A lot of the approaches in text classification treat documents as a bag-of-words with the text represented as a vector of a weighted frequency for each of the distinct words or tokens. This simplified representation of text has been shown to be quite effective for a number of applications [6].

III. EXISTING APPROACH

Arabic language has vowel diacritics that are written above or under letter that give the desired sound and meaning of word. Due to the increase of availability of digital Arabic documents and the important need of automated text categorization, many approaches are proposed. But, they did not achieve researchers' satisfaction and have high computation cost. The main steps/stages of our approach for each method are BOW and mixed words for features representation in Arabic text classification, Fig.1 shows the stages that include: Arabic text pre-processing, normalization, stemming, and feature representation and selection. These stages are used in both: training and testing phases. In the following, we describe these stages in detail.

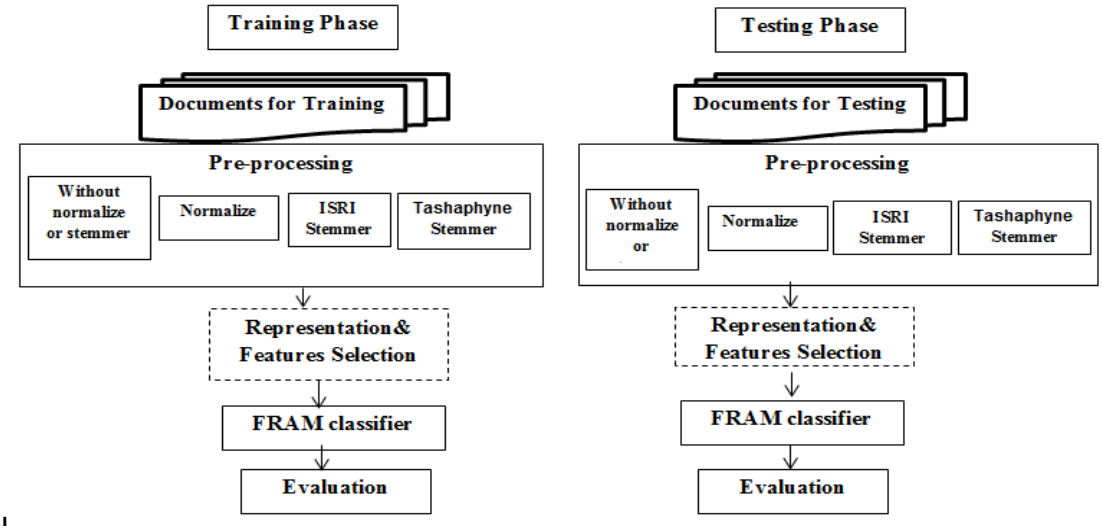


Fig. 1: Arabic Text Categorization

A.Text Preprocessing

This stage necessary due to the variations in the way that text can be represented in Arabic. First, the text documents are converted to UTF-8 encoding. Then, The Arabic stop words are removed. Some Arabic documents may contain foreign words, special characters, numbers [20, 21]. Finally, words with length less than three letters are eliminated; often these words are not important and are not useful in TC. The preprocessing stage includes also normalization and stemming.

A. Normalization

For normalization, a very efficient normalization technique is applied (i.e. Tashaphyne normalization) [22]. Normalization of some Arabic letters such as "ة" to "ه" and "ي" to "ى" and "آ ,أ ,إ " to "ا". In addition, diacritics such as "تَشْكِيلُ" to "تشكيل" and elongation "حاســـــــب" to "حاسب" are performed.

B. Stemming

We have applied two efficient stemming algorithms: Information Science Research Institute's (ISRI) stemmer and Tashaphyne stemmer. Because they have better performance in comparison with other stemmers [23, 24].

a) ISRI Stemmer

The Information Science Research Institute's (ISRI) Arabic stemmer shares many features with the Khoja stemmer [25]. However, it does not employ a root dictionary for lookup. In addition, if a word cannot be rooted, it is normalized by the ISRI stemmer (e.g. removing certain determinants and end patterns) instead of leaving the word unchanged. Furthermore, it defines sets of diacritical marks and affix classes. The ISRI stemmer has been shown to give good improvements to language tasks such as document clustering [26].

b) Tashaphyne Light Arabic Stemmer

The Tashaphyne stemmer normalizes words in preparation for the "search and index" tasks required by the stemming algorithm. It removes diacritics and elongation from input words [27]. Then, segmentation and stemming of the input is performed using a default Arabic affix lookup list for various levels of stemming and rooting [27].

Tashaphyne Light Arabic Stemmer provides a configurable stemmer and segmented for Arabic text.

B. Representation and Features Selection

The representation "Bag-Of-Word" BOW is the most popular document representation scheme in text categorization. In this model, a document is represented as a bag of the terms occurring in it and different terms are assumed to be independent of each other. BOW model is simple and efficient [28]. In addition, a lot of work has been done to extract MWT in many languages. Many of researchers use AMWTs features to improve Arabic document classification [15, 16, 17, 18].

We are dealing with a huge feature spaces. Therefore, a feature selection mechanism is needed. The most popular feature selection method is term frequency [29].

a) BOW Representation

In the first method, BOW is used in the features representation step. First, the frequencies for every term in all categories are calculated and sorted according to the largest frequency. Second, we take the top 25% of the features when normalization and stemming are not used and take 50% of the features otherwise. These two percentages have been defined experimentally.

b) Mixed words Representation

In the second method, first, the frequencies for every term in all categories are calculated and sorted according to the largest frequency. Second, when BOW is applied, we take the top 25% of the features when normalization and stemming are not used and take 50% of the features otherwise. Third, when mixed words are applied, we take the top 50% of the features from BOW, and take the top 3% from two adjacent words in all experiments. These percentages have been defined experimentally. Finally, for both the two methods, the frequency ratio (FR) is calculated by the FRAM classifier in each category as follows [9]:

$$FR(t_n, c_k) = \frac{R(t_n, c_k)}{\sum c_k \in C\, R(t_n, t_k)} \qquad (1)$$

Where, the ratio (R) of each feature term for each category is calculated by:

$$R(t_n, c_k) = \frac{f_{ck}(t_n)}{\sum t_n \in T\, fc_k(t_n)} \qquad (2)$$

Here, $f_{ck}(t_n)$ refers to the total frequency of the feature term t_n in a category ck. Thus, in the training phase, the FR of all feature terms are calculated and supported in each category. Then, the category evaluation values or category scores are calculated which indicates the possibility that the candidate document in the testing phase belongs to the category as follows:

$$E_{di}(C_l) = \sum_{tn \in di} FR(t_n, t_k) \qquad (3)$$

Finally, the candidate document di is classified into the category $C_{\wedge k}$ for which the category score is the maximum, as follows:

$$C_{\wedge k} = \operatorname{argmax} c_{k \in c} E_{di}(c_k) \qquad (4)$$

IV. Experimental Results

The proposed methodology is implemented using Python 3.4.2 [30, 31]. In addition, our experiments are conducted on a laptop with the following specifications: 2.5 GHz Intel core i5 processor with 4 GB of RAM, and windows 8 enterprise.

A. Evaluation Metrics

Four standard evaluations are used: accuracy, recall, precision, and F-measure. The categorization accuracy of the approaches is computed by the equation [32]:

$$\text{Accuracy} = \frac{\text{Number of correctly identified documents}}{\text{Total number of documents}} \quad (5)$$

Precision, Recall and F-measure are defined as follows [33]:

$$\text{Recall(R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{Precision(P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{F} - \text{measure(F1)} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \quad (8)$$

Where:

- TP: number of documents which are correctly assigned to the category.
- FN: number of documents which are not falsely assigned to the category.
- FP: number of documents which are falsely assigned to the category.
- TN: number of documents which are not correctly assigned to the category.

Three different data sets (i.e. Dataset 1, Dataset 2, and Dataset 3) are collected from the website: www.aljazeera.net [34, 35]. They are used to evaluate the efficiency of our proposed approach.

Dataset1 consists of 1800 documents that are separated into six categories: art, health, religion, law, sport, and technology.

Dataset2 consists of 1500 documents separated into five categories: arts, economic, politics, science and sport.

Dataset3 has 1200 documents which are separated into four categories: international, literature, science and sport.

The datasets are divided into 70% of the documents are used for training while 30% of the documents are used for testing. These percentages are defined experimentally.

B. The Results of Experiments using BOW

In Table 1, the results show that the highest precision achieved for Dataset1 is 100% in sport category when Tashaphyne stemmer is used. Also the results shows that the highest recall, precision and F-measure achieved when normalization and stemming are not used is 98.9% with sport category, and it is the same percentage when normalization is used with art and sport categories. In case of the use of stemmers, the highest recall is 98.9% in sport category when ISRI and Tashaphyne stemmers are used [36].

Table 1: Results of Recall, Precision and F1 for Dataset1

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.978	0.979	0.978	0.989	0.978	0.983	0.900	0.988	0.942	0.978	0.989	0.983
Health	0.978	0.946	0.962	0.967	0.956	0.961	0.888	0.954	0.919	0.922	0.933	0.927
Law	0.933	0.966	0.949	0.967	0.936	0.951	0.944	0.833	0.885	0.989	0.839	0.908
Religion	0.956	0.935	0.945	0.956	0.977	0.966	0.922	0.902	0.912	0.922	0.988	0.954
Sport	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.967	0.978	0.989	1.000	0.994
Technology	0.944	0.966	0.955	0.933	0.965	0.949	0.933	0.956	0.944	0.900	0.976	0.936
Average	96.30	96.32	96.29	96.67	96.69	96.66	92.96	93.29	93.01	95.00	95.42	95.06

The results in Table 2 (for Dataset2) gives that the highest recall and precision achieved when normalization and stemming are not used is 100 % with economic, science , sport and politics categories, and it is the same percentage when normalization is used with economic, science and sport categories. When stemmers are used, the highest recall and precision is 100% in economic, politics and Science categories when ISRI stemmer is used. Also when Tashaphyne stemmer is used achieved highest recall is100% in sport category.

For Dataset3, in Table 3, the results shows that the highest recall achieved when normalization and stemming are not used is 98.9% with sport category, and it is the same percentage the highest precision when normalization is used with science and sport categories. When stemmers are used, the highest precision is 98.9% in science category when ISRI stemmer and the highest recall is 100% in sport category when Tashaphyne stemmer is used.

Table 2: Results of Recall, Precision and F1 for Dataset2

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.989	0.979	0.984	0.979	0.979	0.979	0.867	0.934	0.902	0.978	0.967	0.972
Economic	1.000	0.918	0.956	1.000	0.938	0.968	1.000	0.677	0.807	0.989	0.881	0.932
Politics	0.856	1.000	0.922	0.889	0.988	0.936	0.678	1.000	0.808	0.822	0.961	0.887
Science	1.000	0.978	0.989	1.000	0.989	0.994	0.922	1.000	0.960	0.978	0.989	0.983
Sport	1.000	0.978	0.989	1.000	0.978	0.989	0.978	0.978	0.978	1.000	0.978	0.989
Average	96.89	97.06	96.82	97.33	97.40	97.29	88.89	91.87	89.09	95.33	95.53	95.26

Table 3: Results of Recall, Precision and F1 for Dataset3

	Without normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
International	0.900	0.976	0.937	0.956	0.978	0.967	0.933	0.955	0.945	0.956	0.935	0.945
Literature	0.956	0.945	0.950	0.967	0.957	0.961	0.911	0.891	0.901	0.933	0.966	0.949
Science	0.978	0.934	0.957	0.989	0.989	0.989	0.966	0.989	0.976	0.944	0.966	0.955
Sport	0.989	0.978	0.983	0.989	0.978	0.983	0.911	0.901	0.906	1.000	0.979	0.989
Average	95.56	95.88	95.67	97.50	97.51	97.49	93.05	93.39	93.21	95.83	96.11	95.95

From overall experiments for Arabic text classification by using BOW, the results investigate that normalization can enhance categorization process of documents and gives better evaluation than without normalization and stemming.

C. The Results of Experiments using Mixed Words

In Table 4, for Dataset1, the results show that the highest recall and precision achieved is 100% in art and sport categories when normalization and stemming is used. Also, the results show that the highest recall, precision and F-measure achieved when normalization used is 100% with sport category. The highest recall is 100% in sport category when ISRI stemmer is used. Recall, precision and F-measure achieved is 98.8% in Art and sport when Tashaphyne stemmer is used.

Table 4: Results of Recall, Precision and F1 for Dataset1

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	1.000	0.947	0.973	1.000	0.957	0.978	0.878	0.975	0.924	0.878	0.988	0.929
Health	0.989	0.947	0.967	0.989	0.947	0.967	0.833	0.974	0.898	0.922	0.943	0.933
Law	0.933	0.988	0.960	0.967	0.978	0.972	0.967	0.853	0.906	0.978	0.779	0.867
Religion	0.967	0.978	0.972	0.967	1.000	0.983	0.933	0.875	0.903	0.889	0.976	0.930
Sport	1.000	0.989	0.995	1.000	1.000	1.000	1.000	0.947	0.973	0.989	0.978	0.983
Technology	0.944	0.988	0.966	0.956	1.000	0.977	0.933	0.944	0.939	0.922	0.965	0.943
Average	97.22	97.29	97.21	97.97	98.03	97.97	92.40	92.81	92.38	92.96	93.80	93.10

The results in Table 5 indicate that the highest recall, precision and F-measure for Dataset2 achieved when normalization and stemming are not used is 100% in sport category and it is the same percentage when normalization is used with sport category. When stemmers are used, the highest recall is 100% in economic and sport categories when ISRI stemmer is used. Also, when Tashaphyne stemmer is used, the highest recall and precision of 100% in art, economic, politics and sport categories is achieved.

For Dataset3, Table 6 shows that the highest recall and precision achieved when normalization and stemming are not used is 100% with literature and sport categories, and it is the same percentage when normalization is used with literature, science and sport categories. When stemmers are used, the highest precision is 100% in international and science categories, and the highest recall is 100% in sport category when Tashaphyne stemmer is used.

Table 5: Results of Recall, Precision and F1 for Dataset2

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.978	0.978	0.978	0.989	0.978	0.983	0.733	0.970	0.835	0.744	1.000	0.854
Economic	0.967	0.978	0.972	0.967	0.967	0.967	1.000	0.643	0.783	1.000	0.709	0.830
Politics	0.944	0.955	0.950	0.933	0.965	0.949	0.633	0.983	0.770	0.711	1.000	0.831
Science	1.000	0.978	0.989	1.000	0.978	0.989	0.878	0.988	0.929	0.933	0.988	0.960
Sport	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.867	0.92	1.000	0.841	0.914
Average	97.78	97.77	97.77	97.78	97.77	97.76	84.89	88.98	84.91	87.78	90.76	87.76

Table 6: Results of Recall, Precision and F1 for Dataset3

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
International	0.922	0.988	0.954	0.978	0.988	0.983	0.833	1.000	0.910	0.956	0.925	0.940
Literature	1.000	0.968	0.984	1.000	0.957	0.978	0.756	0.944	0.840	0.911	0.965	0.937
Science	0.978	0.946	0.961	0.978	1.000	0.989	0.889	1.000	0.941	0.956	0.956	0.956
Sport	0.989	1.000	0.994	0.989	1.000	0.994	0.989	0.669	0.798	1.000	0.978	0.989
Average	97.22	97.55	97.34	98.61	98.66	98.62	86.67	90.34	87.20	95.56	95.58	95.54

From the previous results for Arabic text classification by using mixed words, best results have been achieved using normalization. Then, in the second place is the results achieved when both normalization and stemming are not used. Finally, the third place is for the results with stemming applied, where Tashaphyne stemmer achieved better results than ISRI stemmer.

D. Discussion

Table7and Fig. 2 show a comparison between the two representations: BOW and mixed words. The results show that the highest accuracy achieved by used mixed words when normalization is 98.61% with Dataset2, while in BOW method by used normalization achieved 97.50% in the same dataset. The mixed words methods showed the highest accuracy in all datasets and all experiments, except for the use of stemming the results are a significant decrease but in some categories have achieved 100% accuracy as shown in Tables 4, 5 and 6.

.

Table 7: Comparison between the two representations BOW and mixed words

	Without normalization or stemming		Normalization		ISRI stemmer		Tashaphyne stemmer	
	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words
Dataset1	96.30%	97.22%	96.67%	97.96%	92.96%	92.41%	95.0%	92.96%
Dataset2	96.89%	97.78%	97.33%	97.78%	88.89%	84.89%	95.33%	87.78%
Dataset3	95.56%	97.22%	97.50%	98.61%	93.06%	86.67%	95.83%	95.56%

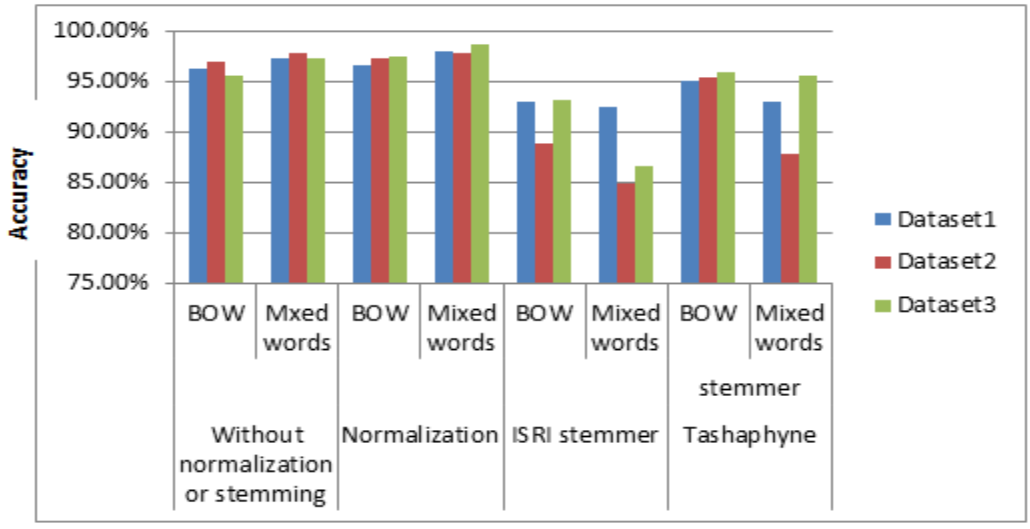


Fig. 2: Comparison between the two representations BOW and mixed words

Table 8 shows the results for execution time for all stages of classification with different datasets and the four experiments for two methods. The first method by using BOW took less execution time in all experiments and all datasets. It was less time for execution with the ISRI stemmer where it took is 46 seconds in Dataset3. While the second method by using mixed words took 78 seconds with Dataset3 (See Fig.3 that shows the comparison between BOW and mixed words in the execution time).

Table 8: Results the execution time of for the two methods

	Without normalization or stemming		Normalization		ISRI stemmer		Tashaphyne stemmer	
	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words
Dataset1	133s	500s	125s	488s	71s	147s	80s	158s
Dataset2	88s	341s	85s	266s	56s	120s	64s	127s
Dataset3	77s	205s	74s	198s	46s	78s	51s	86s

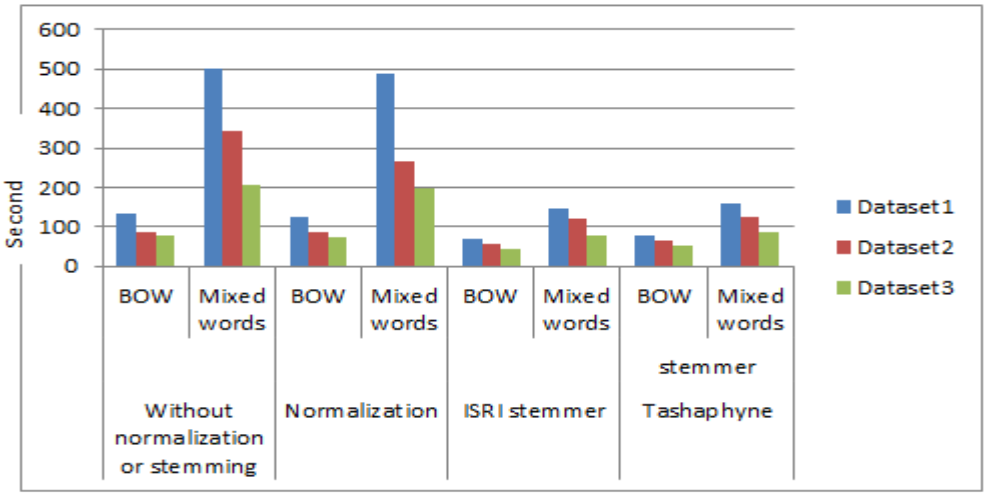


Fig. 3 Comparison between the two methods in the execution time

By analyzing the previous results for the two methods for features representation in Arabic text classification (i.e. BOW and mixed words), we observed the following. On the one hand, the results in mixed words method showed the highest accuracy in all datasets and all experiments. But, it takes more execution time. On the other hand, the use of BOW achieves less accuracy and takes less execution time in all experiments.

V. CONCIUSION

In this paper, we have compared two methods in features representation for categorizing Arabic text. The first method applies BOW while the second method applies technique in the features representation (i.e. mixed words). Each method uses a simple efficient technique for features selection. Also, we have applied Frequency Ratio Accumulation Method classifier with normalization and two stemming mechanisms: ISRI and Tashaphyne stemmers are used.

The results show that the use of mixed words achieves the highest classification accuracy of 98.61% with normalization, while the use of BOW achieves 97.22% with normalization. In addition, the use of BOW method has less execution time in all experiments.

In the future work, several approaches that have been applied to English and other languages will be used for improving Arabic text categorization. In addition, new techniques for features representation and selection will be introduced.

REFERENCES

[1] N.Tripathi, "Level Text Classification Using Hybrid Machine Learning Techniques" PhD thesis, University of Sunderland, 2012.

[2] Laila, K.," Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study" Conference on Data Mining | DMIN'06 |, ,pp.78-82, 2006.

[3] R. Al-Shalabi, G. Kanaan, and M. Gharaibeh "Arabic text categorization using kNN algorithm",2006, pp.1-9.

[4] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity "Automatic Arabic text classification",Journee's internationals d'Analyse statistique des Données Textuelles, pp.77-83,2008.

[5] F.Harrag, E.ElQawasmeh "Neural Network for Arabic text classification", Applications of Digital Information and Web Technologies, 2009. ICADIWT '09. Second International Conference, pp. 778 – 783, 2009.

[6] F.Sebastiani, " Machine learning in automated text categorization"ACM Computing Surveys,Vol. 34 number 1, ,pp.1-47, 2002.

[7] H.Sawaf, J.Zaplo, and H.Ney"Statistical Classification Methods for Arabic News Articles" Workshop on Arabic Natural Language Processing, ACL'01, Toulouse, France, July 2001.

[8] Y.Yang and X. Liu" Re-examination of Text Categorization Methods"Proceedings of 22nd ACM International Conference on Research and Development in Information Retrieval,SIGIR'99, ACM Press, New York, USA, 1999,pp. 42-49.

[9] B.Sharef, N.Omar, and Z.Sharef "An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation" The International Arab Journal of Information Technology, Vol. 11, No. 2, March 2014, pp.213-221.

[10] B. Al-Shargabi, W. AL-Romimah and F. Olayah " A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination" ,ACM, Amman, Jordan 978-1-4503-0474-0/04,2011.

[11] A. Wahbeh, M. Al-Kabi, Q. Al-Radaidah, E. AlShawakfa and I. Alsamdi, "The Effect of Stemming on Arabic Text Classification: An Empirical Study", In International Journal of Information Retrieval Research (IJIRR, pp.54-70, 2011.

[12] H. Nezreg, H. Lehbab and H. Belbachir," Conceptual Representation Using WordNet for Text Categorization", International Journal of Computer and Communication Engineering, Vol. 3, No. 1, January 2014.

[13] A. Mesleh. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System", Journal of Computer Science, pp. 430-435. 2007.

[14] T. Hofmann, "Introduction to Machine Learning", Draft Version 1.1.5, November10, 2003.

[15] H. Meryem, S. Ouatik A. , Lachkar, " A Novel Method for Arabic Multi-Word Term Extraction", International Journal of Database Management Systems (IJDMS) Vol.6, No.3, pp.53-67. , June 2014.

[16] H. Meryem, A. Lachkar, S. Ouatik, " Multi-Word Term Extraction Based On New Hybrid Approach for Arabic Language" ,Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014 ,pp. 109-120,2014.

[17] D. Abuaiadah, "Arabic Document Classification Using Multiword Features",International Journal of Computer and Communication Engineering, Vol. 2, No. 6, pp.659-664. November 2013.

[18] W.Zhang, T. Yoshida and X. Tang, "Text classification based on multi-word with support vector machine" Elsevier, pp.879-886. 2008.

[19] M.Suzuki, S.Hirasawa," Text Categorization Based on the Ratio of Word Frequency in Each Categories", In Proceedings of IEEE International Conference on Systems Man and Cybernetics, Montreal, Canada, 2007, pp. 3535-3540.

[20] R.Al-Shalabi,G.Kanaan, J.Jaam, A.HasnahandE.Hilat "Stop-word Removal Algorithm for Arabic Language"Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications,IEEE-France,2004,pp.545-550,CTTA'04 .

[21] M. El-Kourdi, A. Bensaid and T. Rachidi "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm"20th International Conference on Computational Linguistics. August, Geneva, 2004.

[22] https://pythonhosted.org/Tashaphyne/Tashaphyne.normalize-module.html.

[23] Sh. Oraby, Y. El-Sonbaty and M. El-Nasr "Exploring the Effects of Word Roots for Arabic Sentiment Analysis" International Joint Conference on Natural Language Processing, 471–479, Nagoya, Japan, 14-18 October 2013.

[24] A.Ezzeldin, Y.El-Sonbaty and M.Kholief"Exploring the Effects of Root Expansion "College of Computing and Information Technology, AASTMT Alexandria, Egypt, 2013.

[25] T. Kazem, E. Rania, and C. Je.rey"Arabic Stemming Without A Root Dictionary" Information Science Research Institute, USA, 2005.

[26] A. Kreaa, A. Ahmad and K. Kabalan "Arabic Words Stemming Approach Using Arabic WordNet" International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.4, No.6, November 2014.

[27] https://pypi.python.org/pypi/Tashaphyne/Vol.4, No.6, November 2014.

[28] W.Pu, N.Liu "Local Word Bag Model for Text Categorization" Seventh IEEE International Conference on Data Mining, 2007, pp.625-630.

[29] O.Garnes, "Feature Selection for Text Categorization" Master thesis, Norwegian University of Science and Technology, June 2009.

[30] https://www.python.org/downloads/.

[31] http://www.nltk.org/_modules/nltk/stem/isri.html

[32] M. Turk, and A. Pentland."Eigenfaces for recognition. Journal of Cognitive Neuroscience" vol. 3, no. 1,1991, pp. 71 -86.

[33] R.Elhassan, M.Ahmed “Arabic Text Classification on Full Word” International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 5, May 201 5, pp.114-120.
[34] http://diab.edublogs.org/dataset-for-arabic-documentclassification/
[35] https://sites.google.com/site/mouradabbas9/corpora
[36] R. Sallam, H. Mousa, M. Hussein "Improving Arabic Text Categorization using Normalization and Stemming Techniques,” International Journal of Computer Applications (0975 – 8887) Volume 135 – No.2, February 2016.

Semantic-based Approach for Solving the Heterogeneity of Clinical Data

Basma Elsharkawy, Rashed Salem, and Hatem Abdel Kader
Information Systems Department
Faculty of Computers and Information,
Menoufia University, Shebin Elkom, Egypt.

Abstract: **Clinical records contain massive heterogeneity number of data types, generally written in free-note without a linguistic standard. Other forms of medical data include medical images with/without metadata (*e.g.*, CT, MRI, radiology, etc.), audios (*e.g.*, transcriptions, ultrasound), videos (*e.g.*, surgery recording), and structured data (*e.g.*, laboratory test results, age, year, weight, billing, etc.). Consequently, to retrieve the knowledge from these data is not trivial task. Handling the heterogeneity besides largeness and complexity of these data is a challenge. The main purpose of this paper is proposing a framework with two-fold. Firstly, it achieves a semantic-based integration approach, which resolves the heterogeneity issue during the integration process of healthcare data from various data sources. Secondly, it achieves a semantic-based medical retrieval approach with enhanced precision. Our experimental study on medical datasets demonstrates the significant accuracy and speedup of the proposed framework over existing approaches.**

Keywords-Schema data integration, Heterogeneity, Image retrieval, Semantic ontology, OWL, RDF, XML.

I. INTRODUCTION

Big data is a collection of datasets in a large variety of domains [15], healthcare is one of such domains. There are different types of data including structured, semi-structured, and unstructured. Statistically, 80% of medical data are unstructured, which further complicates the management of these data [11, 21]. The major source for healthcare applications is patient records. Data integration is the task of combining different data sources, and providing a unified view of the data. Such integrated data are needed to be standardized and kept in a repository, *i.e.*, data warehouses, for ease of retrieval and analytics later [16].

However, integrating data from a variety of sources is not a trivial task, due to the large volumes of heterogeneous data during mapping, ranking, and key matching [9]. Moreover, structural and semantic heterogeneity is another problem that faces data integration [10, 13]. In this paper, we address the problem of resolving structural and semantic heterogeneity for healthcare applications. While structural heterogeneity addresses schema conflicts, semantic heterogeneity addresses meaning conflicts, *e.g.*, synonyms and homonyms conflicts. Fortunately, semantic Web can be exploited to resolve semantic heterogeneity issue. Using semantic Web, the same concepts which given by several words, *i.e.*, synonyms, as well as the different concepts given by the same word, *i.e.*, homonyms, can be defined. Semantic technologies (*e.g.*, Ontology) provide major solutions to semantic interoperability in healthcare systems. Moreover, ontologies can deliver solutions for image retrieval. They seek to map the low-level image features with high level ontology concepts.

Compared with content-based and keyword-based image retrieval, ontology-based retrieval concentrates on capturing semantic content. Furthermore, ontology plays an important role to represent the knowledge as a set of concepts within a domain and the relationships between pairs of concepts. Ontologies can be used to support a variety of tasks in different domains including knowledge representation, natural language processing, information retrieval, database integration, digital libraries, *etc*.

This paper proposes a framework in which different medical data types are merged into a unified format. The proposed framework tackles heterogeneity issue during data integration process. By the proposed framework physicians can build a patient's history record, and thus helps physicians in decision making. The proposed framework keeps all patient history without losing any data. Furthermore, this paper proposes a semantic-based framework for medical data retrieval.

This paper is organized as follows; section 2 presents background of semantic schema mapping and integration approaches. Section 3 introduces a literature review of clinical data management. The proposed semantic-based

framework for resolving heterogeneity and retrieval of healthcare data is discussed in section 4. Section 5 presents how to handle several cases of medical data. Implementation and discussion of the proposed framework are provided in section 6. Section 7 concludes the work and highlights the future work.

II. Semantic Schema Mapping Approaches

Schema mappings are expressions that specify how an instance of a source database can be transformed into an instance of a target database. In recent years, they have received an increasing attention both from the research community and the commercial tools market.

A schema-mapping system is used to support the process of generating and executing mappings in practical scenarios. It typically allows users to provide an abstract specification of the mapping as a set of correspondences among schema elements, specified through a friendly user-interface. Based on such specification, the mapping system will first generate a number of mappings – usually under the form of tuple generating dependencies (tgds) that correlate source tables with target tables; then, based on these mappings, an executable transformation, *i.e.*, a runtime script in SQL or XQuery, can be practically used to run the mappings and generate solutions.

Christian Bizeret al. [29] introduced the mappings process on the Web and a composition method for chaining partial mappings from different sources based on a mapping quality assessment heuristically. By introducing R2R mapping language that designed to fulfill each of the vocabulary cherry picking and interlinking and discovery, whereas every term must be identified with its own dereferenceable URI in order to enable mappings to be interlinked with RDFS or OWL vocabulary term definitions and mappings by RDF links.

A. Semantic Schema Matching Approaches

The semantics of schema concepts acts a critical role in the determining mappings/ matching process between different data sources. Identifying both the implicit and explicit meaning of schema label, the semantic correspondences among the elements of different schemas have been defined. This identification requires the development of a method for lexical annotation, *i.e.*, finding the meanings of a schema label in a reference lexical database. Several methods are connected with this problem by using lexical knowledge in different ways.

In healthcare environments, Lee, C. Y., et al. [28] proposed an attribute matching algorithm to resolve semantic conflicts and interoperability problems, which does the semantic matching over two steps; first step is checking the attribute similarity with domain knowledge. The second step is checking word relatedness through overlapped phrases, hyponyms and hyponyms.

Partyka, J., et al. [24, 25] addressed semantic heterogeneity challenge between different data sources. One of traditional methods is N-gram method that often fails. Fundamentally, it depends on discovering the similarity among shared instances, that results in an overestimation of semantic matching between independent attributes. They proposed an approach initially depends on choosing similarity among value attributes, then examining the instances between them which is known as an entropy-based distribution (EBD). Then, they compared the N-gram method and the new T-Sim method for calculating EBD.

Chena, N., et al. [26] mentioned that the syntactic schema matching method cannot identify possible semantic mapping relationships; for example, in healthcare domain, element ‘diagnosis’ and element ‘prescription’ have identical semantics, until this time they cannot be identified by the syntactic method. They proposed the Node Semantic Similarity (NSS) method based on conjunctive normal forms and a vector space model. They designed a hybrid algorithm based on label meanings and annotations for computing the relationship between concepts of label. Then, the semantic relationship is translated between nodes into a propositional formula which confirms the validity of this formula to confirm the semantic relationships. Firstly, the algorithm calculates the label and node concepts, secondly it computes the conceptual relationship. The Zhao, C. [30] has proposed a multilayer schema matching approach with many layers. The first layer connected with semantic similarity. The second layer verifies the functional dependency to formulize structural information of schemas. A third layer proposes a probabilistic factor. The last layer confirms the mapping element pairs process with reasonable depending on each layer’s results. In general, the semantic similarity measure works on data preprocessing, then it does the lexicographic similarity and generates the filtered matching sets.

Islam, A. and Inkpen, D. [27] addressed the text similarity challenge to solve semantic heterogeneity as a critical challenge in any data sharing integration system, a distributed database system, a web service, or a one-to-one data management system. The Semantic Text Similarity (STS) method has been recommended, that discovers the similarity of two texts in terms of semantic and syntactic information (by common-word order method). They use three similarity functions in order to extract more general text similarity approach. At the beginning, string similarity and semantic word similarity are considered. Then, they introduce common-word order similarity function to

combine syntactic information. Finally, the text similarity is derived by merging string similarity, semantic similarity and common-word order similarity with normalization.

In all mentioned data integration research, the semantics of the transformations are strongly linked to the implementation method. The intention is that the integrated database be implemented as a view of the component databases, and that queries against the integrated database be executed by translating them into queries against the component databases and then combining the results. The semantics of the individual transformations are given by their effects on queries. However, the lack of any independent characterization of their semantics makes it difficult to prove properties of the transformations, or to use any alternative implementation of the methodology.

B. Schema Integration Techniques

The schema integration derives from two tasks: database integration, and integration of user views, which occurs during the design phase of a database when constructing a schema that satisfies the individual needs of each of a set of user groups. However, they fail to note that these two kinds of schema integration are fundamentally different. For database integration, instances of each of the source databases are transformed into instances of the merged schema. Moreover, when integrating multiple user views, instances of the merged schema must be transformed back into instances of the user views. A good schema-integration method should therefore take account of its intended purpose and include semantics for the underlying transformations of instances [24].

III. RELATED WORK

There are many approaches proposed in the literature for managing clinical records including, chunking, data-driven, free-text assignment codes, content-based image retrieval, and semantic-based image retrieval. The chunking approach is proposed to identify non-recursive words and base noun phrases in the text, *i.e.*, a key issue in symptom and disease identification as a term. Thus, it extracts structured data from clinical records easily. Chunking handles data annotated by medical domain using a chunk annotation scheme with extra credibility, which involves symbols as noun phrases (NPs), main verb (MVs), and a common annotation for adjectival and adverbial phrases (APs) [19]. However, there is a very limited amount of annotated text of this kind available for health-care systems [22].

In data-driven approaches, a driving data element that is an independent variable should be selected. The independent variable is used to determine the other linked patient information. Syndrome/sub-syndrome classification and 3-digit ICD-9 final diagnosis code are used to determine the driving data element. The data element that is used to realize the patient's record would have a clear and easily recognized relationship. The mapping from data element is well defined if the patient has been grouped by a single value of this data element. The challenge is not only in data storage and access, but also in scalability of healthcare sources [4].

Medical image retrieval can help physicians in finding information that assist them in decision making. Medical image retrieval systems extract features as color, texture, shape and spatial relationships. Image features are extracted from the full image and then are indexed. The variety of medical image types makes the process of retrieval is a non-trivial task. For instance, radiology images face many difficulties [5]. Particularly, such radiology images contain rich information and specific features that need to be carefully recognized for medical image analysis.

Image retrieval systems are generally classified into two major approaches. The first approach searches local or global image features such as color or texture. The other approaches add key words to images as an annotation. Content-based Image Retrieval (CBIR) approach is considered a rapidly advancing research area. It depends on searching similarity of image features from a database based on the color, shape and texture [5]. Images are presented as a query against image database. The similarity between image features in the database is retrieved with the help of indexing images [7, 20]. The indexing of images provides a rapid path for searching image databases [3]. However, there is still a "Semantic gap" between what users need and what CBIR systems can achieve. In particular, there are no sensible means by which queries can be presented to CBIR systems [14, 18].

The Semantic-based Image Retrieval (SBIR) systems include several components of information extraction such as a textual description and visual feature, and semantic image retrieval. The extraction process of SBIR is based on low level features of images to identify objects. Open issues are the nature of digital images, as well as descriptions of images, *i.e.*, high-level concepts such as rat and dogs. However, the main problem is the semantic gap discrepancy between low-level features and high level concepts [12]. Moreover, different users at a different time may give different interpretations for the same image [1, 17]. TABLE I provides a summarized comparison among medical retrieval approaches and our proposed framework.

TABLE I MEDICAL DATA TYPE RETRIEVAL APPROACHES

	CBIR	SBIR	Chunking	Data-driven	Proposed Framework
Data type	Image	Image	Free text	Free text	Image, Free text, audio, and video
Data missing	Lossy	Lossy	Lossy	Lossy	Loss-less
Challenges	1)Semantic Understanding of media is visual 2)Integrating, Searching, Selecting	1)The different Forms of images 2)Lack of relation between objects and the meaning	1)Identification Of medical concepts 2)Clarification of medical concepts relations	1)Data scalability 2)Data access	1)Grew up of Global ontology
Precision	In accurate with medical images	In accurate with medical images	In accurate in medical concepts	Inaccurate due to data missing	Accurate
Performance	Degrading with large database	Degrading with general concepts	Degrading with medical concepts	Degrading with large database	High Performance
Scalability	Low scale	Low scale	Low scale	Low scale	High scale

IV. PROPOSED SEMANTIC-BASED FRAMEWORK FOR INTEGRATING MEDICAL DATA

Clinical data are represented in structured and unstructured form. Surgical producers, treatment and drugs data are examples of structured data. Structured data can be computerized and allow performing analysis of data, queries and aggregation for patient records. They are organized in a mightily mechanized and manageable structure. Structured data are prepared for seamless integration into a database or well-structured file format. Structured data needs to stay comparatively simplistic and uncomplicated [11]. Moreover, structured data depend on a data model. The data model specifies how data will be generated, stored, processed and accessed [6]. Structured data are generated through constrained choices in the form of data entry, which overall drop-down menus, check boxes, and pre-filled templates. This type of data is easily searchable and aggregated, can be analyzed and reported, and is linked to other information resources. The high cost and performance limitations of storage, memory and processing allows relational databases and spreadsheets using structured data are the only way to effectively manage data [6].

Unlike, unstructured clinical data may contain free clinical notes and multimedia contents such as medical images and voice. These data may have an interior structure, nevertheless they are still considered as an unstructured form because the data which they contain don't care appropriate sorted into a database. The concept of "big data" is widely associated with unstructured data. Big data denote to extremely large datasets that are difficult to analyze with traditional tools [2, 6].

There is a variety of challenges for handling clinical notes including ungrammatical, short phrases and abbreviations. The proposed framework helps in solving these challenges of clinical notes, and the heterogeneity of clinical record. In addition to structured medical data, the proposed framework merges medical images, clinical notes and audio data type into a unified framework. Then, physicians can perform a "DL language" or "SPARQL" query to access the different data types. The proposed medical image retrieval framework handles medical images as *url* and gives a label for each medical image. In the case of medical reports, which consist of both medical images and text, the process of text extraction is performed firstly before performing medical image process.

In this paper, the Semantic Web ontology are used to solve data integration challenges in healthcare system. Ontology makes the relationships between patient data clear, the data can be derived from various resources. For example, if two different patient infected with the same diagnosis, whereas they are represented by different names or different identifiers, ontologies are used to map these names to the same diagnosis descriptor. Due to heterogeneity and complexity of data integration in healthcare system, the hybrid ontology approaches will be used [13, 32]. The semantic ontology of each patient is described by local ontology. In order to build the global ontology, all local ontologies are merged to one with shared all vocabulary and physician's concepts.

A. Technical process of integrating medical record

The data type in patient record is not unified, it is consisting of various media types. Therefore, the mapping document can be generated using local ontologies that defines the semantics of the source data then merging them with the global one. The free text is transformed to XML. The structure of an XML node is basic building part of mapping ontology and the relationship between the elements. Each XML node has a specific element in the OWL ontology. For example, the following XML node maps the general part of the ontology.

<XMLNode ="globalOWL:patient1">
</XMLNode>

The OWL class element determines the class to be mapped into output RDF file. The OWL class mapping element may contain multiple OWL class elements that are arranged in the ontology. Thus, the OWL class element consists of defined element containing fixed data, which specifies the class names to be constructed in the RDF document. The OWL property element contains specific relations between classes.

The RDF document is created from the input XML document, OWL ontology and the medical record. The main class in the OWL ontology is owl:Thing and all classes are subclasses of it. The semantic relationship has been defined between patient's data in the output RDF file. The ontology also defines the meaning of each element. The RDF is based on the idea of identifying things using Web identifiers (URIs), and describes resources by giving them labels. The RDF statement defines patient's label and asserts that some relationships, indicated by the predicate, holds between other patient records. As an example of a resource on the Web, we can have the following statement. The web page whose URI is "http://www.patient1.org/boons.html"is referred by an URI as "http://wiki.hip.fi/xml/ontology/BonesDEPT.owl". Generally, an ontology is a description of concepts and relationships that can be found in medical records. In the context of this paper, the language in which these statements are written is Description Logics "DL" and queries are performed using SPARQL [22].

B. System Architecture

The architecture of the proposal approach according to previous methodology is shown in Fig. 1. It contains three layers: physical layer, semantic service layer, and application layer.

1. The physical layer consists of data sources in healthcare system.
2. The semantic layer involves the ontology base, semantic query and reason service.
3. The application layer is designed to access ontology from remote locations and at different platforms.

V. Multiform Medical Data Handling

The proposed framework focuses on integrating data from heterogeneous medical, data sources. Here in, we introduce how the proposed framework handle unstructured medical data such as clinical notes, medical images, physician's reports and audio data type. In Fig. 2, patient record has been handled either as clinical note, medical image, audio type or patient report. In the proposed framework a local ontology will be merged with the global one. The physicians or administrator allow doing any query on the global ontology.

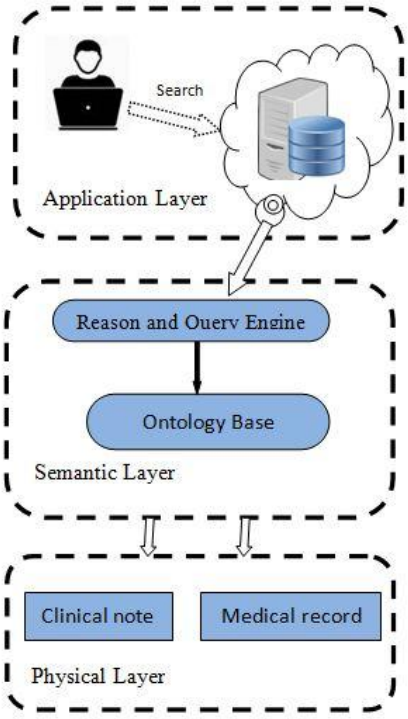


Fig. 1 The architecture of proposal approach

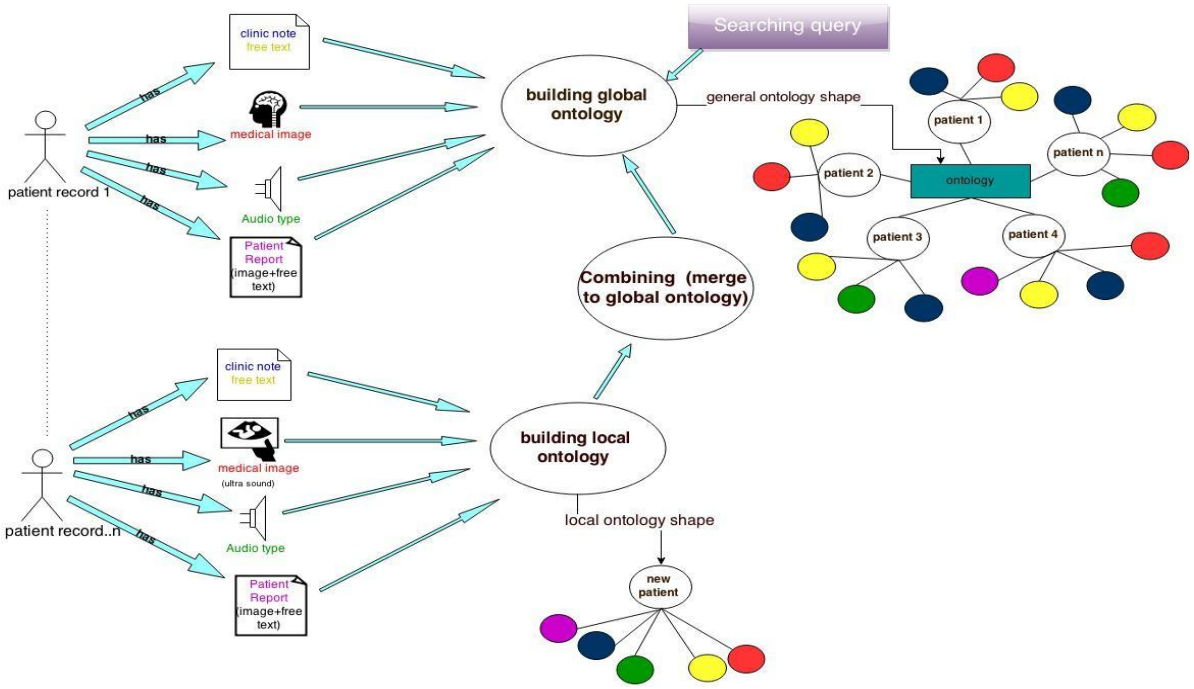


Fig. 2 The proposed framework

A. *Clinical text - free form*

Clinical text has wealthy detailed information of great possibility usage to scientists and health service researchers. Text schema describes the structure of a text file and how a text document is read or written in a raw format. The structure of text stream defines either fixed column widths or columns which are separated by delimiters. To convert a text schema to XML Schema, a specified separator, field separator and the field names of the text file should be determined. In other side, XML Schemas contain annotations for providing additional information, such as medical information. The conversion process between XML and CSV has been done automatically by detecting all repeated elements in XML that are used for splitting data to rows.

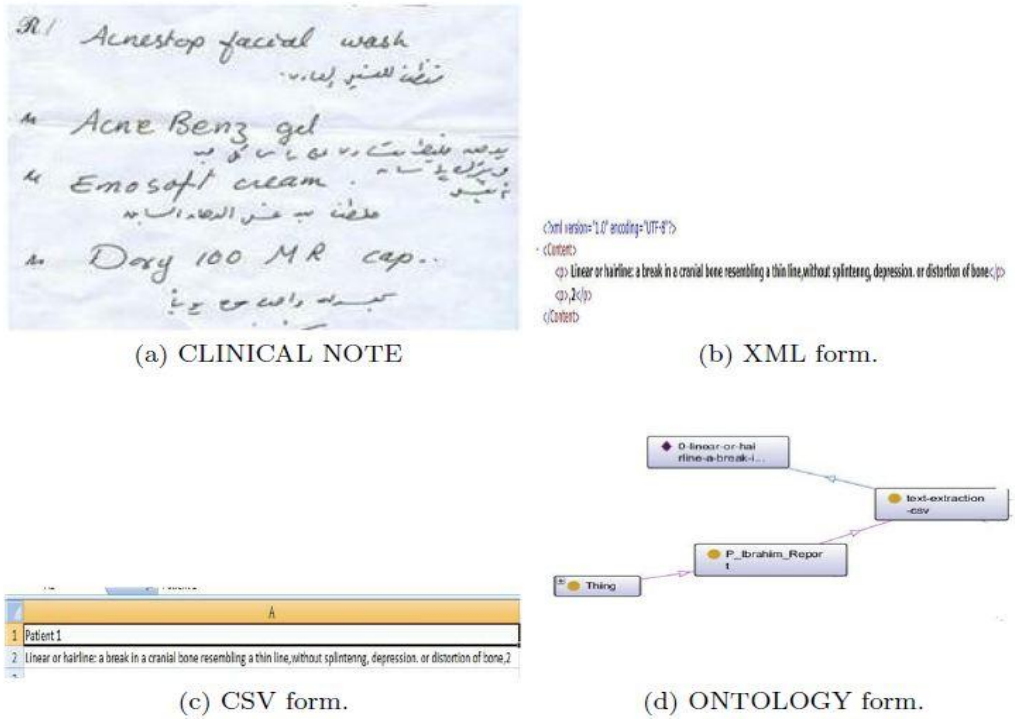
The proposed framework can handle free text as a block. Clinical note is extracted from clinical notes of physician's prescriptions. While the input is free text, the expected output merges all patient data and performing query to get all data merged for the patient. There are three processes in managing clinical notes. Firstly, extracting clinical notes process has done from physician's prescriptions and the output of this process is in unstructured form. Secondly, transforming process for unstructured form has done to get semi-structured form as to facility the dealing when building ontology. Finally, transforming XML form to get a structured form, as in "CSV file" which allow building RDF "ontology" file where queries can be carried out to access the medical data, see Fig. 3.

B. *Medical image data type*

The second form is medical images in healthcare application. The major challenge in this case is how to handle these large numbers of images with their various formats and then merging them with other medical data types. The proposed framework helps in solving this challenge by building ontology for these images and accesses any of images by its label through the global ontology. Physicians can get medical image in ontology by its label or URL, see Fig. 4(a). DICOM images as example consist of two parts, i.e., text header and binary image. Medical images metadata as well as field names or image's URL are indexed and transformed to XML elements.

C. *Audio media data type*

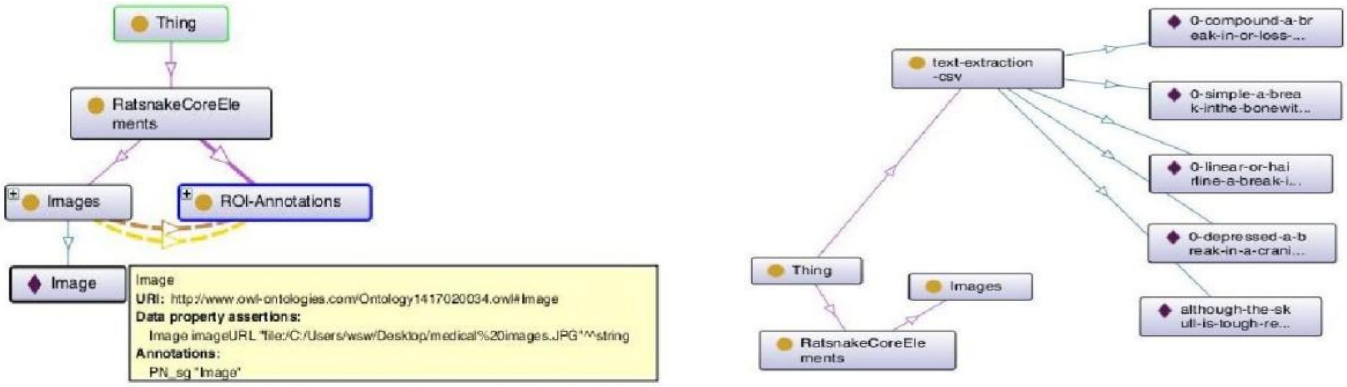
Audio media type is represented in health care as medical diagnosis from physicians a broad or encounters between physicians and patients. There are five processes to complete this stage. The first process is performing speech to text process. The second process is acting as the first process in free text data type. Output from the first process should be converted to semi-structured form as XML form. Then a structured form represented in CSV file has been created and global ontologies have been built, which we can perform query simply to access all media types.



(a) CLINICAL NOTE (b) XML form.

(c) CSV form. (d) ONTOLOGY form.

Fig. 3 Clinical note processing cycle



(a) Ontology of Medical image (b) Ontology of patient report

Fig. 4 Both ontology of medical images and patient report

D. The medical patient report

Output report may contain free text and medical images. Thus, images are extracted firstly from reports and then ontology is created for both image and extracted free text. This case involves handling both free text and medical images as discussed previously. The proposal framework deals with patient report depending on the extracted free text of the report and creates ontology for free text see Fig. 4(b). Medical images have been handled as a normal image, but also we can get pure image and build its ontology.

VI. Implementation and Discussion

The proposed framework is implemented with freely and open-source tools, *i.e.*, semantic Web technologies. The proposed ontology is configured as follows: number of classes, number of individuals and number of properties are 21,58 and 68, respectively. Moreover, maximum depth, maximum number of children and average number of

children are 5,75 and 4, respectively. Indeed, a local ontology for each media type of medical information is created, and then such local ontologies are merged into global one. The global ontology being built is efficiently scalable; new patient records can be added and merged easily. Thus, all patients' medical information is integrated into their history without loss of any data. The global ontology merges other healthcare domain ontology such as accounts, hospital budget, geographical places analogies, etc. By merging all these information, a background for patient history is complete, which can help physicians in decision making. Furthermore, the technician can apply queries against the global ontology using either URI, labels, DL language or SPARQLE. Searching inside the ontology is tested several times, the searching process is simple, and the results are returned quickly and accurate as shown in Fig. 5. The main challenge in retrieval system was the correlation among medical concepts. Fortunately, the proposal framework tackles this challenge by adding rules and relations among concepts. In Fig. 6, three patient have a problem in lung, but two of them have cancer in lung. Therefore, physicians get alarm that patient 3 may has cancer in lung. Accordingly, the proposal framework can help physicians in decision making.

To implement CBIR approach for comparison with the proposed framework, dataset of around 90 medical images is used. Three different categories are used including several datasets. The size of the first dataset is 4.6M including "20 image" and the size for each one is about 232 K. The size of the second dataset is 6.1M including "23 image" with size for each one is 266 K. The third dataset is 12.9M including "25 image" and the size for each one is 516 K. The size of the fourth dataset is 13.4M including "25 image" with size for each one is 548K. Note that, the main problem facing retrieval systems is semantic gap and relationships among medical concepts [8]. The proposed framework tackles this issue including semantic gap and related concepts. CBIR systems retrieve medical images according to the distance similarity between the query image and the dataset as shown in Fig. 5. By the experiments, CBIR query is an image showed in Fig. 5 (a). The results of CBIR system depend on the distance similarity between the query image and the dataset. Fig. 5(b) shows the result from CBIR. However, it can be noticed easily that query is a chest medical image has cancer and the results show different images including broken hand, thus it leads to low accuracy.

(a) Query of CBIR (b) Result of CBIR

Fig. 5: CBIR-based retrieval

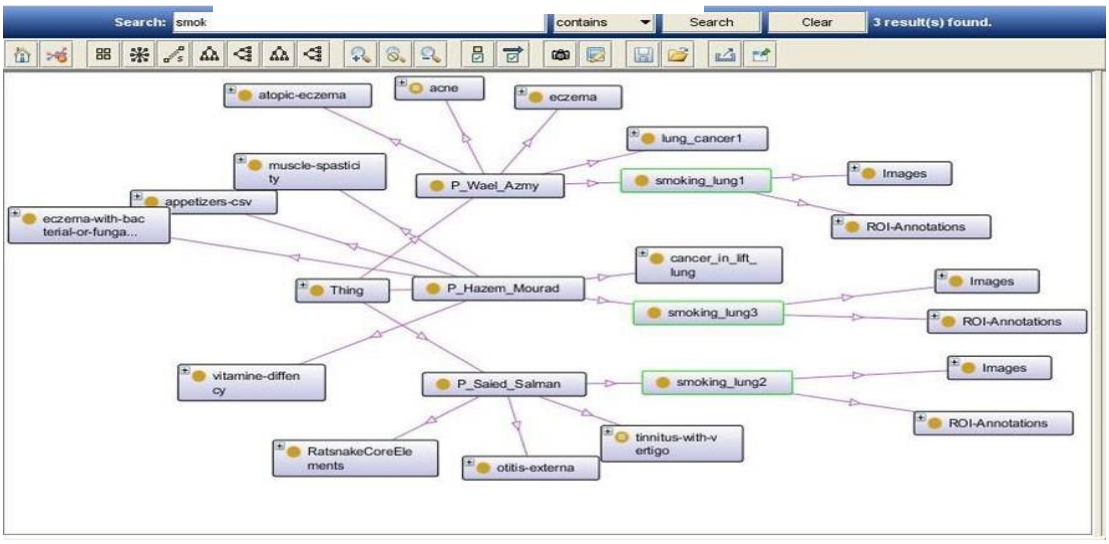


Fig. 6 Relationships of medical concepts

However, the proposed framework retrieves all medical image according to their labels. For example, if we apply searching against the ontology using label or annotator "smoking lung", physicians retrieve medical images of three patient have a smoking lung. Moreover, physicians can observe that patient 1and 2 have a ray cancer from the retrieved results. Thus, they conclude that patient 3 may be infected by cancer as shown in Fig. 3. Therefore, the proposed framework exploits related concepts and helps physicians in decision making. To calculate the accuracy of the proposed framework against CBIR, Fig. 7 shows the recall, precision and consumed time for transforming images into XML format. Generally, results indicate that the proposed framework outperforms CBIR. Precision is the fraction of retrieved images that are relevant to physician's information need. The precision of the proposed frame work is better than CBIR, although the proposal framework takes into account the human errors. Moreover, there call of the proposed framework outperforms the CBIR due to challenges of CBIR which tackled by the proposed framework. Generally, results in TABLE II indicate that the proposed framework outperforms CBIR. Precision is the fraction of retrieved images that are relevant to physician's information need.

Annotation approach implementation - Now, we highlight some implementation observations and differences between the proposed framework and other literature approaches. For instance, NLP approaches apply POS tagging to get annotation for all tokens such as NN, MV, etc. However, when applying POS tagging, massive data are lost particularly the medical concepts in addition to the lack of relationship information among such concepts. Compared with the proposed framework, there is no data loss. The annotation approach had to find syntactic structure as possible. The main units of annotation are adopting chunk. The Harvey corpus is used to deal with free text. The Harvey corpus is a chunk-annotated corpus. Chunking tends to solve this challenge using part of speech (POS) tagging. The main problems are caused by unknown tokens which caused obscurity due to neglected words or phrases. Generally, the main challenge in the annotation approach is to get the similarity between encoding enough information to improve the research, and realizing clarity, accuracy, and conciseness [18]. The two typical examples (without chunk annotation for clarity) and example (3) below illustrates the chunking annotation process.

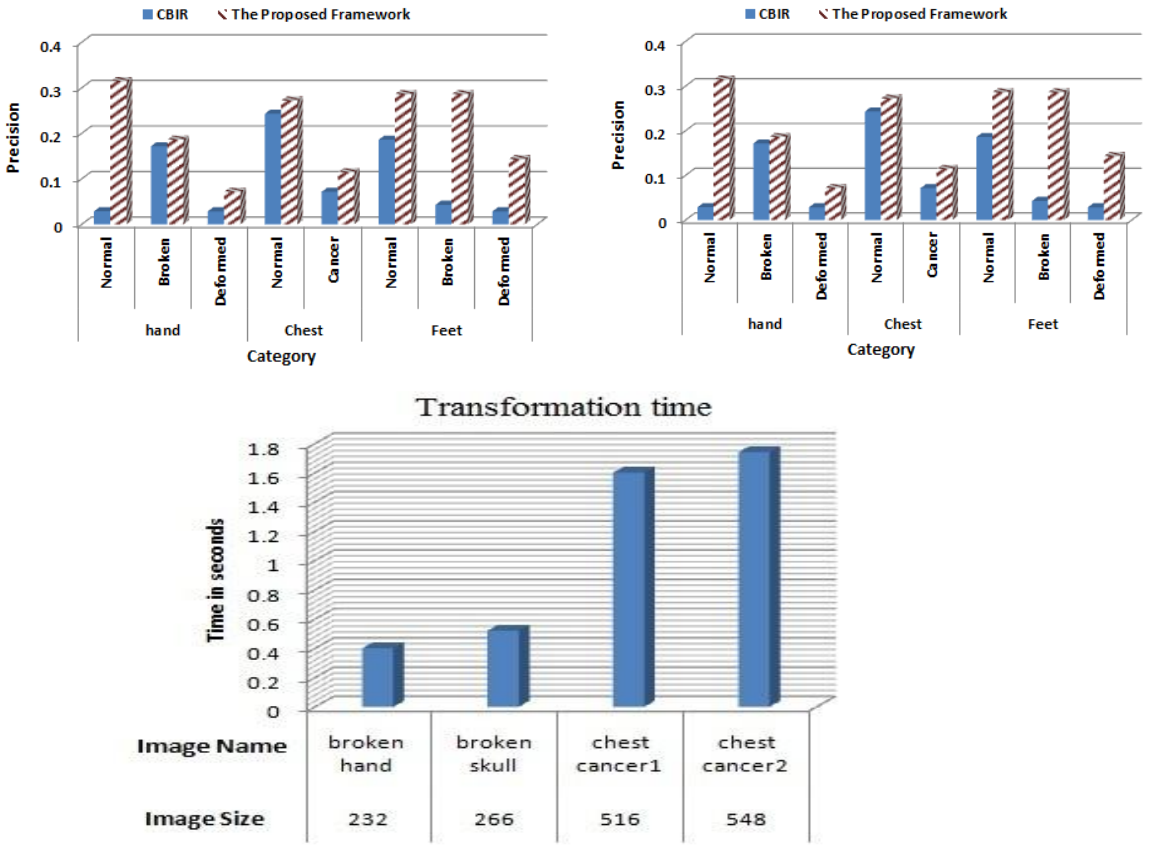


Fig. 7 Precision and Recall of the proposed framework vs. CBIR system

TABLE II RESULTS OF PRECISION AND RECALL FROM THE PROPOSAL APPROACH

Category		CBIR Color & Shape (Histogram)		The proposed approach	
		Precision	Recall	Precision	Recall
Hand	Normal	0.2857	0.6667	0.3142	0.7333
	Broken	0.1714	0.35	0.1857	0.65
	Deformed	0.0285	0.1	0.0714	0.5
Chest	Normal	0.2428	0.5	0.2714	0.6333
	Cancer	0.0714	0.3	0.1142	0.7
Feet	Normal	0.1857	0.65	0.2857	1
	Broken	0.0428	0.6	0.2857	1
	Deformed	0.0285	0.4	0.1428	1
Average		**0.1321**	**0.4458**	**0.2089**	**0.7770**

VII. CONCLUSION

Managing unstructured clinical data is one of the major problems in healthcare systems. The heterogeneity of clinical data is considered as a critical roadblock to achieving integration and interoperability between systems. In this paper, we proposed a semantic-based framework for managing heterogeneous medical data including free clinical notes, audio data, and medical images. It tackles the heterogeneity by unifying different medical data types into a unified form and building ontology. The ontology keeps all patient history and enables queries for patient and diseases history. The manipulation of the proposed framework is demonstrated by the different medical cases.

The future step is to handle medical videos, which can be transformed into medical images, and integrate them with different medical data types. Moreover, we plan to enrich the framework with spatial and temporal information of patients to discover new insights from analytics.

REFERENCES

[1] M.Alkhawlani, M.Elmogy, and H.El Bakry. Text-based, content-based, and semantic-based image retrievals: A survey. International Journal of Computer and Information Technology (ISSN: 2279,0764) Volume04?Issue 01, January 2015.

[2] Belle, R. Thiagarajan, S. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian. Big data analytics in healthcare. Biomed research international, 2015.

[3] D. P. Bhamare and S. A. Abhang. Content based image retrieval: A review. International Journal Of Computer Science And Applications, 8(2), 2015.

[4] A. L. Buczak, S. Babin, and L. Moniz. Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making, 10(1):59, 2010.

[5] R. Chaudhari and A. Patil. Content based image retrieval using color and shape features. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 1(5), 2012.

[6] N. Grover. 'big data'-architecture, issues, opportunities and challenges. IJCER,3(1):26–31, 2014.

[7] L. Haldurai and V. Vinodhini. A study on content based image retrieval systems. International Journal of Innovative Research in Computer and Communication Engineering , Vol. 3, Issue 3., March 2015.

[8] R. Jobay and A. Sleit. Quantum inspired shape representation for content based image retrieval. Journal of Signal and Information Processing, 5(02):54, 2014.

[9] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq. Challenges of data integration and interoperability in big data. In Big Data (Big Data), 2014 IEEE International Conference on, pages 38–40. IEEE, 2014.

[10] L. Kang, L. Yi, and L. Dong. Research on construction methods of big data semantic model. In Proceedings of the World Congress on Engineering, volume 1,2014.

[11] A. Katal, M. Wazid, and R. Goudar. Big data: Issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404–409. IEEE, 2013.

[12] H. Kaur and K. Jyoti. Survey of techniques of high level semantic based image retrieval. International Journal of Research in Computer and Communication technology, IJRCCT, ISSN 2278-5841, Vol 2,, 2(1):015–019, Issue 1, January 2013.

[13] R. Kienast and C. Baumgartner. Semantic data integration on biomedical data using semantic web technologies. INTECH Open Access Publisher, 2011.

[14] P. Kulkarni, S. Kulkarni, and A. Stranieri. A novel architecture and analysis of challenges for combining text and image for medical image retrieval. International Journal for Infonomics (IJI), 2014.

[15] S. J. Pooja, Reema Gupta. Big data: Advancement in data analytics. International Journal of Computer technology and applications, 2014.

[16] K. Priyanka and N. Kulennavar. A survey on big data analytics in health care. International Journal of Computer Science and Information Technologies 5(4):5685–5688, 2014.

[17] R. Rahimzadeh, A. Farzan, and Y. F. Fathabad. A survey on semantic content based image retrieval and CBIR systems. International Journal on "Technical and Physical Problems of Engineering" (IJTPE) Published by International Organization of IOTPE, March 2014.

[18] S. Sasikala and R. S. Gandhi. Efficient content based image retrieval system with metadata processing. International Journal for Innovative Research in Science and Technology, 1(10):72–77, 2015.

[19] A. Savkov,J. Carroll, and J. Cassell. Chunking clinical text containing non- canonical language. ACL 2014, page 77, 2014.

[20] D. S. Seema H. Jadhav1, Dr.Sunita Singh. Content based image retrieval system with semantic indexing and recently retrieved image library. International Journal of Advanced Computer Technology (IJACT), 2012.

[21] J. Sun and C. K. Reddy. Big data analytics for healthcare. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1525–1525. ACM, 2013.

[22] O. Uzuner, M. Yetisgen, and A. Stubbs. Biomedical/clinical NLP. COLING 2014, pages 1–2, 2014.

[23] C. Y. Lee, H. Ibrahim, M. Othman, and R. Yaakob. Reconciling semantic conflicts in electronic patient data exchange. In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, pages 390–394. ACM, 2009.

[24] J. Partyka, L. Khan, and B. Thuraisingham. Semantic schema matching without shared instances. In Semantic Computing, 2009. ICSC'09. IEEE International Conference on, pages 297–302. IEEE, 2009.

[25] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Qing Yu, D. Giordano, I. Marenzi, and B. Pereira Nunes. Interlinking educational resources and the web of data: A survey of challenges and approaches. Program, 47(1):60–91, 2013.

[26] N. Chen, J. He, C. Yang, and C. Wang. A node semantic similarity schema matching method for multi-version web coverage service retrieval. International Journal of Geographical Information Science, 26(6):1051–1072, 2012.

[27] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data(TKDD), 2(2):10, 2008.

[28] D. Ramesh and C. Kumar. Schema integration based merging and matching algorithm for agricultural HDDBs. Arabian Journal for Science and Engineering, 40(9):2555–2569, 2015.

[29] C. Bizer and A. Schultz. The R2R framework: Publishing and discovering mappings on the web. In Proceedings of the First International Conference on Consuming Linked Data-Volume 665, pages 97–108. CEUR-WS. org, 2010.

[30] L. C. Keung, S. Niukyun, J.-F. Ethier, L. Zhao, V. Curcin, and T. N. Arvanitis. The integration challenges in bridging patient care and clinical research in a learning healthcare system. 2014