IntechOpen

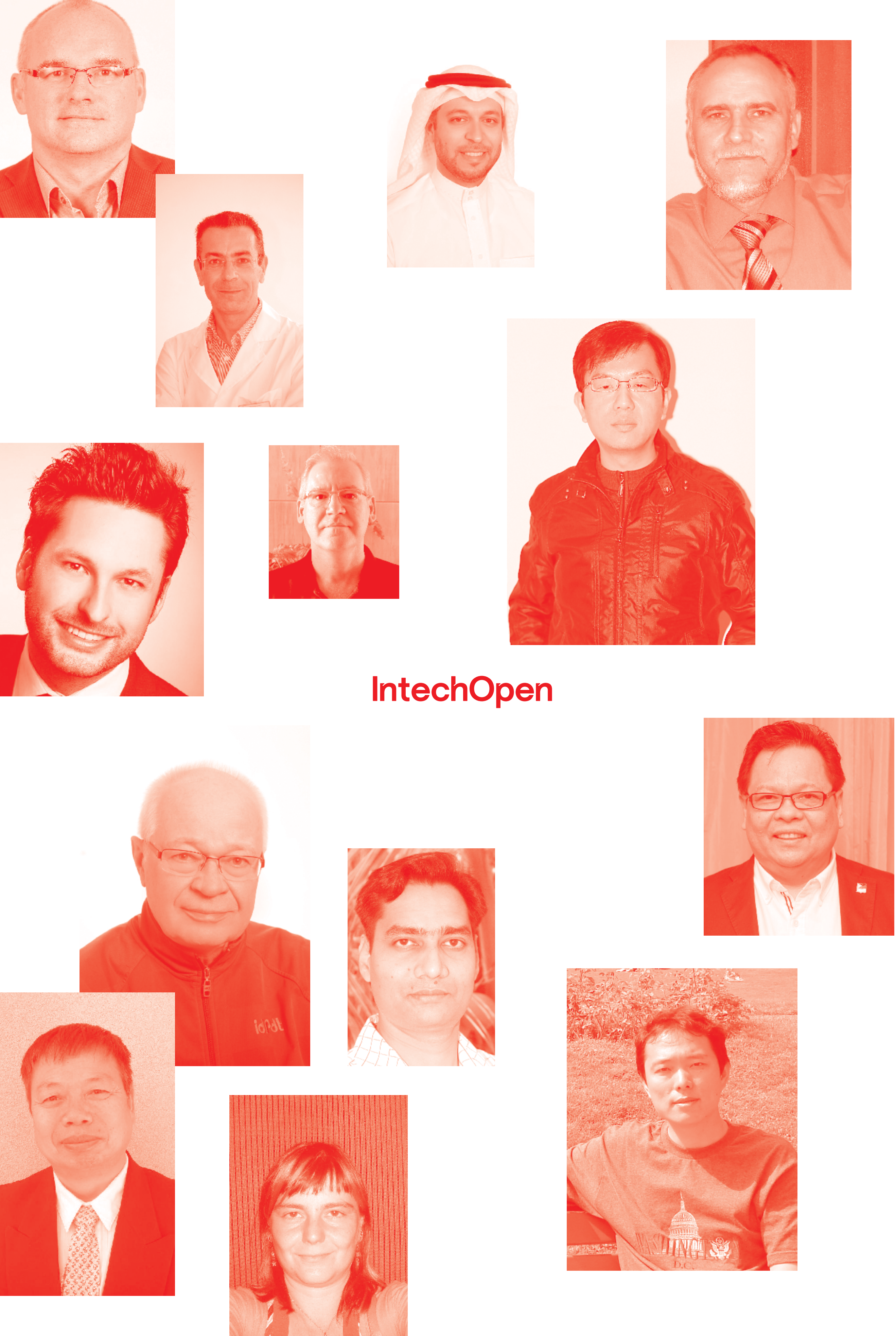
Open Data

Edited by Vijayalakshmi Kakulapati

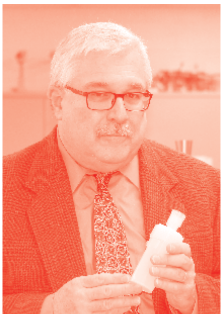
Open Data

Edited by Vijayalakshmi Kakulapati

Published in London, United Kingdom


IntechOpen

Supporting open minds since 2005



Open Data
http://dx.doi.org/10.5772/intechopen.94663
Edited by Vijayalakshmi Kakulapati

Contributors
Monika M. Wahi, Natasha Dukach, Omer Hassan Abdelrahman, Farah Jemili, Hajer Bouras, Vijayalakshmi Kakulapati, Kannadhasan Suriyan, Nagarajan Ramalingam

Notice
Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2022 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Open Data
Edited by Vijayalakshmi Kakulapati
p. cm.
Print ISBN 978-1-83968-315-2
Online ISBN 978-1-83968-316-9
eBook (PDF) ISBN 978-1-83968-317-6

Meet the editor

Prof. Vijayalakshmi Kakulapati received a Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad. She is currently a professor in the Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad. She has twenty-six years of industry and teaching experience and is a member of various professional bodies, including the Institute of Electrical and Electronics Engineer (IEEE), Association for Computing Machinery (ACM), Computer Science Teachers Association (CSTA), LMISTE, LMCSI, International Association of Computer Science and Information Technology (IACSIT), FIETE, and more. She has more than 110 publications in national and international journals and conferences, 30 book chapters, and 2 books to her credit. She has received numerous awards, including Excellence in Research, Best Reviewer, appreciation awards, and more. Her areas of research include theoretical and practical information retrieval problems as well as machine learning applied to large-scale textual applications. Her research has focused on retrieval models, query/document representations, term weighting, term proximity models, and learning to rank (machine-learned ranking functions). She is also passionate about seeing research problems applied to real-world problems, especially those dealing with large, complex data sets. Along these lines, she is working with evaluating and designing novel search algorithms for web search and summarization. Currently, Dr. Kakulapati is working with big data analytics, health informatics, the Internet of Things, deep learning, artificial intelligence, and data sciences.

Contents

Preface

Open data is freely usable, reusable, or redistributable by anybody, provided there are safeguards in place that protect the data's integrity and transparency. Analysis and processing of public open data (POD) repositories in order to obtain relevant information from query log data.

This book describes how retrieved data can improve different learning qualities of digital networking, particularly performance and reliability. The book also describes developing artificial intelligence (AI) and machine learning or related models, knowledge acquisition problems, and feature assessment by incorporating data sources (blogs, search query logs, document collection) as well as interactive data (images, videos, and their explanations, multi-channel handling data).

The search query log created by manual intervention with the POD repository is a good source of knowledge. The data in the search query log is generated from users who interact with online communities. However, there is an understanding of the concept with economic models in specific sectors, for example, the telecom sector, where prices are appropriately designed and implemented. There is a significant gap between recently evolved extraction methodologies of POD repositories and their applicability across numerous organizational processes.

This book is useful for undergraduates, scientists, and professionals working in open data. It includes five chapters.

Chapter 1: "Knowledge Extraction from Open Data Repository"

This chapter analyzes how researchers retrieve data from POD repositories. The increase in the number of affluent online platforms, social media, and collaboratively related web resources has amplified the evolution of socio-technical systems, resulting in domains that demonstrate both the conceptual model of the required system approaches and the collaborative form of their participants. The POD repositories at impressive levels and retrieve the information from query log data to investigate these factors' effects. This investigation aims to maximize the quality of a POD repository from a new perspective. First, we offer a unique query recommender system that can help consumers reduce the length of their querying operations. The goal is to discover methods that will allow users to engage with the open data repository quickly and with fewer requests.

Chapter 2: "Open Government Data: Development, Practice, and Challenges"

This chapter focuses on the principle of open data, emphasizing Open Government Data (OGD). It discusses the context and features of OGD and identifies dangers, barriers, and problems. It also examines the benefits of OGD, as well as perceived risks, obstacles, and challenges.

Chapter 3: "Framework to Evaluate Level of Good Faith in Implementations of Public Dashboards"

Public dashboards (PDs) must be measured by how often they satisfy customer demands. This chapter provides a methodology for assessing the amount of contracting parties' development of PD. It begins by looking at the problems governments face when sharing files with good conscience, even though OGD laws and regulations are being implemented worldwide. The chapter provides a use case in which scientists investigate a PD in their environment that looks to be adopting OGD but is not doing so in good conscience and designs an equivalent approach.

Chapter 4: "Intrusion Detection Based on Big Data Fuzzy Analytics"

Intrusion detection (ID) is an essential technology for improving network security by identifying assaults or anomalous accesses. Most conventional ID systems have several drawbacks, including high dropout frequencies and poor malware detection. Managing scattered and massive information presents a problem for ID systems. Working with inaccurate data is also problematic. This chapter uses a fuzzy c-means (FCM) algorithm to cluster and classify pre-processed learning data.

Chapter 5: "Artificial Intelligence and IoT: Past, Present and Future"

This chapter examines how artificial intelligence (AI) can assist healthcare practitioners in making optimal treatment decisions. The increased usage of patient records and the development of big data analysis techniques have resulted in reliable and efficient applications of AI in health services. As guided by proper diagnostic inquiries, advanced AI algorithms may find critical clinical data in massive data, assisting in treatment decisions. The chapter also discusses the Internet of Things (IoT), a system that connects physical devices to the Internet using near field communication (NFC) and wireless sensor networks (WSNs).

Vijayalakshmi Kakulapati
Sreenidhi Institute of Science and Technology,
Department of Information Technology,
Yamnampet, Ghatkesar, Hyderabad, Telangana, Inida

Section 1

Fundamental Aspects of Open Data

Chapter 1

Knowledge Extraction from Open Data Repository

Vijayalakshmi Kakulapati

Abstract

The explosion of affluent social networks, online communities, and jointly generated information resources has accelerated the convergence of technological and social networks producing environments that reveal both the framework of the underlying information arrangements and the collective formation of their members. In studying the consequences of these developments, we face the opportunity to analyze the POD repository at unprecedented scale levels and extract useful information from query log data. This chapter aim is to improve the performance of a POD repository from a different point of view. Firstly, we propose a novel query recommender system to help users shorten their query sessions. The idea is to find shortcuts to speed up the user interaction with the open data repository and decrease the number of queries submitted. The proposed model, based on pseudo-relevance feedback, formalizes exploiting the knowledge mined from query logs to help users rapidly satisfy their information need.

Keywords: Data Mining, Query, Public Open Data, Social Network, Knowledge Extraction

1. Introduction

SNS (Social networking services) is online services, platforms, or sites designed to support the development of Internet-based communities or community links between, for instance, individuals who often regularly interact with hobbies, experiences, and emotional interactions. SNS includes an account of every member and its community ties and a range of additional capabilities (typically a biography). A significant number of SNS are social media and allow consumers to communicate Online like e-mail and automatic messages. While SNS is often an individual-centered service in a broad context, social media facilities are a team. Social media platforms may be regarding constitute SNS. Online Communities enable individuals inside individual unique systems to exchange opinions, tasks, experiences, and goals.

Social networks communicate to interact in many innovative approaches, such as shows, hashtags, perform and engages electronically, revealing further cooperation and projected benefit that could scarcely imagine only a short time before. Online communities can play a significant role in the organizational processes as well as helping to develop company concepts and emotional responses and give up different prospects for the examination of social interactions and social behavior.

Presently, people rely on social media and its vast and diverse wealth and have progressively penetrated each human living area. Increasingly individuals prefer to engage valuable time on social media to develop a significant social entertaining community and again try to communicate with each other so often that the interaction around them is robust. POD repository analytics is perhaps a commonly used scientific and commercial approach for investigating the social media of interpersonal, organizational, and corporate links. The necessity for solid knowledge in DPO analysis has lately increased with ready availability to computational power and the rise of social popular social networking platforms such as LinkedIn, Twitter, Netlog, and more.

Twitter social network by study the contents of the tweets and the links between the tweets to extract knowledge from log data. By selecting buzzwords, began the 'Twitter review and then collecting all Twitter posts (Tweets) correlated to the keywords. It is a social-economical problem in India. Mining the query log based on social networks like Facebook, Twitter, etc. Study and address the discovery, access, and citation of POD repositories like Twitter data sets; and strengthening educational programs of academics of current and future generations specializing in such areas. This is an auspicious time for extracting useful information from social media query log data. Substantial efforts to decipher large amounts of data are steps towards complete search log records integrating POD repository analysis. From these data sets, we extract valuable knowledge.

The search log obtained by user actions with the Public Open Data (POD) database is an excellent data collection for improving its efficiency and the effectiveness of the online community. The data in the user input logs are gathering from individuals who communicate on online platforms. The search log assessment is complicating due to the variety of customers and diverse resources. As a result, numerous scientific articles written about query log analysis.

The word "data set" can also describe the data in a set of specifically relevant tables that correlate to a specific investigation or occurrence. Records generated by satellites testing hypotheses using devices aboard communications satellites are one instance of such a category. A data source is the standard measurement for data provided in a POD repository in the open data domain. Over a quarter a million datasets are gathering on the European Open Data platform. Alternative interpretations were presented in this area, although there is presently no accurate statement. Various difficulties (relevant data resources, non-relational datasets, etc.) make reaching a compromise more challenging. The utilization of query logs for knowledge discovery improves the speed of the POD repositories and improves the use of open data source capabilities.

POD repository analysis and mining for valuable extract knowledge from query log data. We perform on knowledge discovery, ML or similar approaches, challenges connected to pre-processing and model assessment, for data sets (web usage log files, query logs, collection of documents), and collaborative data (images, videos, and their explanations multi-channel handling data). We summarize the fundamental results concerning query logs: analyses, procedures used to retrieve knowledge, the outstanding results, most practical applications, and open issues and possibilities that remain to be studied. We discuss how the retrieved knowledge can be utilized to progress different social media class features, mainly its effectiveness and efficiency.

In addition, several concurrent inquiries of multiple distinct users are addressing by business social networks. The query stream has simultaneously been defining by a stop-time rate, making it impossible for the POD repository to generate massive query load times without over-sizing [1]. Web Search engines Query Logs Social

Network Analysis [2], Web search engine quality approaches developed for query logs, November 2013.

1.1 Motivation

To observe and comment on Twitter users, Twitter is used to share channels for personal information (intimate and confidential life). Since this proportion of Twitter users' Tweets increases significantly and the clickthrough rate, the number of RSs (recommender systems) changes its methodological approaches for the same query-based RSs system. Hence, user perception and feelings include effectively user tweets connected to user's decision making. It is hard to complete and comprehend the retrieval of the needed words from the customer. ML classification approaches from the content evaluation represent the finest and most helpful approach for trend analytics and sentiment modeling "learn the user's query patterns and generate the query functions with good predictions." ML-based RSs are used to classify user tweets and make suggestions. The feature selection and classification responsibilities in the ML method of customer tweets are significant in designing effective RS. But there is no assessment interpretation for ML-based RSs. The primary goal of the work for RS is to give awareness based on recommendations, choices, and choices by enhancing the effectiveness of their suggestions through a suitable system model.

A Recommender system for dynamic Tweets offers content based on user preferences and desires by evaluating the chronological, current, and user tweets posting pertinent data. Interactive tweets that are most suited to specific users and tweet logs data are proposing. The excess issues of information for data analysis and collection are reducing, and the search publishes user tweets in a dynamically and personalized fashion and makes precise and accurate demands.

DTSRS (Dynamic Tweets Status Recommender system) retrieves relevant data from public tweets that allow an adequate content-aware comprehension, accessible to most appropriate tweets for redundant user tweets. The advantage of the developed system DTSRS is that the user does not have to consume enough attention-seeking tweets, helping to make an effort too minimal and testing the tweets' dynamics to the user, reducing and increasing the user's happiness.

User-Query Centered Recommender System (UQCRS) is applied to exploit different measures to demonstrate the efficiency of recommendations delivered. The proposed algorithm exhibits an effective result to the search shortcuts issues.

1.2 Challenges in twitter content query-based recommendation

1. Retrieving the collection of Twitter tweets that match with one or more content keywords of user-query.

2. Ranking the query within the text.

3. Develop a method for user-query-centered knowledgebase integration.

4. Predict the outcome of the automatic query-analyzer of Twitter tweets concerning the recommendations

These challenges can be solved using query expansion and semantic models. In query expansion, the reformulation of the query is made based on the vocabulary mismatch among the query and content retrieved. Through semantic models, similar words of the user query are extracting.

1.3 Objectives

1. Scalability and Real-Time Performance Analysis.

2. Discovering the inherent variability in mining the POD repository using query log data.

3. Comprehensive quality measures analysis.

4. Algorithmic consistency across domains.

5. Valuable identification of information and domain services using POD repository data analysis.

CF (Collaborative Filtering) and CBR (Content-based recommender) systems are the predominant forms of recommendation. The content-based recommender system, which is our subject of interest? In the Twitter recommender system, Twitter tweets' essential nature is noisy and with less content for understanding because of the exact use of posting user [3]. Creating the strictly relevant content of the recommendation system with the account authority is considered in [4], with the learning to rank algorithm considered. The relevant content is a similar type of information retrieval [5] using the tweet contents posted by the user and user friends, which provides recommended set of tweets to the user. The Recommendation System [6, 7] is using to construct the query in the Twitter content query-base recommender system by identifying relevant Recommended tweets.

2. Literature survey

Social media connects next to each other individuals in various methods, such as web-based gaming, marking, earning, and socializing, showing effective technologies to collaborate and communicate that were unthinkable only recently. In addition, online communities contribute to the corporate strategy and assist in altering economic models, sentiments and introduce various opportunities for studying direct intervention and collaborative actions.

Several previous researchers suggested using the Internet search query records to derive linguistic relations between queries or terms. The idea that the web search query logs provide knowledge via clicks confirms the relation among searches and records selected by individuals. The writers relate questions and words in the information gathered based on these data. This technique has also been using to group requests from log files. The cross-reference text is linked to similarities depending on query information, proximity editing, and hierarchical resources to identify better clusters. Such clusters are utilizing to discover identical queries for querying systems.

Twitter is a massive amount of information social network, to perform an analysis on Twitter, a keyword-based search for possible and relevant posts [8], where such search keywords cover all the possible tweets of the user [9], which is a lengthy and time-consuming process. Typically, to reduce the complexity of searching the posts from a Twitter data source, a user search keyword identification is made [10] to reduce the manual effort. User search keywords extraction is developing on the target keywords instead of the general word phrase of the keyword selected. This keyword extraction process is iterative because of the user's regular interaction in the social network through a web search and advertising. So there is a need for a query recommendation and query expansion system to improve the keyword extraction process from Twitter tweets and provide recommendations for Twitter users. The keyword's frequency statistics and machine learning models recommend [11, 12] classifying the keywords and extracting them from the Twitter tweets. The search-based dataset is given in [13] to find the keyword topics and search the keywords in the dataset, but for the enormous tweets, this method results in relevant tweets or empty tweets because of a vast Twitter dataset. Therefore, keyword recommendation in search based through query suggestion is recommended [14, 15]. In keyword recommendation, the query system is designed based on the relevant keywords in the Twitter tweets through query log mining and search query suggestions [16]. The query expansion [17] from the original query is complete with expanding and improving query ranking for the searched tweets through query suggestions.

In recent years, POD repository analyses have received considerable interest mainly because of the increase in online microblogging and media disseminating websites and the generating substantial of an extensive POD repository. Furthermore, despite increasing attention, the significant financial uses of online communities in extraction are not very well understood. While there is a great deal of study on various issues and methodologies of POD repository extraction, there is a difference between approaches produced and used in practical situations by the researchers. Thus, such approaches are yet relatively unknown to their corporate development implications.

Though there is a significant difference between the recently established POD deposition extraction technologies and their application. Some sectors, such as the telecommunications business, whereas charges are appropriately structured to indicate the names and the persons as operators), POD results indicate POD results substantial market orientation synergy. Moreover, most POD repository analytics studies have focused on generic issues previously mentioned rather than considering particular commercial uses. As a result, the possible use of the POD repository evaluation and analysis in the industry is hardly known.

1. The future generation information processing must make it much easier to routinely analyze diverse data sets such as extensive reading, videos, and knowledge created by users like blogging.

2. Innovative approaches are essential for the data analysis of numerous contributing inputs, for example, contextual performance, clinical development research findings, previous warnings, incidence history.

3. Novel approaches require decentralized systems where the related items are searching, and the similarity's reliability needs to be measured.

4. In observational predictive analytics, topic specialists require sophisticated higher transmission. That applies to advanced methods such as cartoons and non-traditional ones such as poetic analytics.

5. A recurrent topic seems to be the necessity to integrate consumer demands into any novel computational technology, system, or technique as a type of data with the involvement of domain knowledge.

Mining methods workflow contains a group of mining data and models, with an utmost data operator work to set the parameters of the mining model used. In mining, the data is not expressing indirect form, but it is unseen in the model connectors. The user provides the indirect form of data and applies a model on the indirect form of data, generating the direct form of data. During this process, mining techniques should distinguish between components, are data model, operators, and parameters. Enable the user in designing such mining for the web. There is a need for the development of online data workflow through concepts and categories. Online data refers to a frequent visitor of a group or several web pages in social networks to cover and gather the complete user required information through locating the web page and fetching the desired user valid information.

Web pages are complete application-specific to fetch the user's desired target information through the user-defined keywords using a constrained specific web application to provide up-to-date information through the Online Social Networks (OSN). OSN is protecting billions of active and passive web user's knowledges. The rapid change in social networking sites has proven an exponential growth in user information and knowledge exchange rate. According to [18], two-thirds of the online users browse a social network or an eCommerce website, with an average of 10% of all internet utilization time. By covering such a large amount of helpful information exchange, OSNs through social media become an excellent platform for mining techniques and research in data analysis.

The method allows data on social media user tweets for goods and commodities. In [19], a RES approach is recommending to give a level of precision compared to the previous approaches used in the tweet assessment of the consumer. Fasahte et al. [20] has presented the method to anticipate tweets by utilizing the Online Reviews dataset sorting procedures. It investigates the search engine extraction and training algorithm to collect data from the unstructured text in the available online content. In addition to the keyword-based evaluation, the data model on the Internet is connecting with complex searches. They utilized to locate tweets on various tweets while maintaining the surfing data operational inside the account location. The data collection, processing of data, and data sampling are all three aspects of tweets availability. Verma et al. [21] developed a dynamic analysis classification technique by implementing ML and evaluated the different variables in these learning approaches. A public repository response assessment technique [22] discussed huge data volumes on Twitter to create the emotional state of every message. Rosenthal et al. [23] describes the user opinion mining system used to extract similar users' views from the person's view using a moderate data analysis method. Ibrahim et al. [24] established an online emotional assessment that supplied many functional tweets out of interest to identify comparable personal data. The decision relates to extracting features, extensive conversion, and different recognition using machine learning techniques in many tweet solutions for the clustering techniques, correlating the query response pattern, relationship regulations for Twitter tweet extraction, and visualization in the Tweet API application.

The phrase retrieval from several texts is provided by [25–27] since the words should be user-specific, and the searching procedure should be preserving. Due to the powerful conventional method to all these, it is possible to investigate a method that relies on the recommendation, utilizes iteratively in the search engine and advertising searching.

Optimization techniques provide AI (Artificial Intelligence) and NLP (Natural Learning Processing) capabilities in order to deliver necessary assessed user suggestions interpretations in different networks/services of social network applications [28, 29]. Interface design such as mobile web apps permits various movies, cuisine, literature, YouTube, healthcare and more information related material. Films, culture, and entertainment are communal societies. Depending on the user's awareness of the material, the recommender system [30] has problems with confidentiality and protection. Thus, classic recommenders are becoming inevitable for current user ratings and Twitter posts to evaluate user-generated content [31, 32].

A social media micro-blogging process recommendation enables 140-character consumers to utilize tweets and retweets, known as individual tweet statuses [33]. Such tweets are related unidirectionally, as it posts the tweet, and others who tweet reciprocate follow a tweet. The predefined list solely concerns the retweeting user. Tweets contain a specific user interest context and content, for example, movies or music [34]. Tweets are connecting to the abovementioned subjects. The person who posts is the origin of news, and the customer who retweets is the follower data source, try to evaluate tweet content. Microblogging is a form of communication via the consumer Twitter medium of complete information. Twitter members are classifying into three categories: the first relates to posts, where people organize into a significant number of supporters. The latter concerns the tweet requesters, where the users submit a rarely post to comprehend and observe tweet material continuously. The last concerns user twitch [35] relationships, including friends/relatives in which all the posts are personalized.

These classes classify according to the post and retweet consumers, followers, and comparable results [36]. The rating is for multiple users, supporters, and retweets with the suggested providers and searchers of knowledge. Depending on these ordering, users and followers get the reputation of relatively large text containing signs of impact in the production of tweets [37]. Posts and repost individuals are classified first, social media posts and tweets are the subsequent most important for tweet users, and the third reciprocate relation to the discussion forums of these ratings [38]. Subsequently, the relevance rating and connectivity are estimated to identify the Twitter post system by giving the prominent tweet the correct weight, which affects the linked user's position for various stuff exchanging relations [39, 40].

3. Recommender system

In this paper, UQCRS proposes a concept of the recommendation system at the user-query level that mainly aims to find the correct tweet's information extraction through content to user requirements.

3.1 Proposed recommender system

Unlike the previous works of the recommendation system, the proposed UQCRS is a system capable of performing Twitter mining from a vast Twitter database through query logs and user tweets for understanding the user interaction in the Twitter tweets. UQCRS provides the search-based tweets content recommendation through the found user-query centered content in the short and long tweets to depict the intention of the user tweets. In the proposed system, the workflow is: firstly, the Twitter background knowledge is extracting for user-query-centered knowledgebase integration. Secondly, implementing the strategy of UQCRS on the Twitter knowledge repositories. And finally, evaluation and illustration through the discussion of the proposed URCRS system are made.

3.1.1 Content-based twitter tweets detection

Twitter users communicate with the recommender systems through a user interface like a web portal or a mobile app. Depending on the user availability, the user interacts with the social network to extract the information in the tweets on user interest, which is predicting by the tweets ranking method to provide the list of proposed tweets based on the content keywords of the user query. The UQCRS data system depends on the database it stores and updates the tweets based on the content and ratings of tweets through the query search. UQC recommendation system tweets content architecture depends on the profile and database of user Twitter profile that store query information and update the entities continuously through Twitter user customer recommendation, as represented in **Figure 1**.

With the content feedback and query-centered analysis, recommender systems are implementing in **Figure 1**, used within the e-commerce websites, to guide the Twitter customers through retrieving log data and Twitter mining originate themselves.

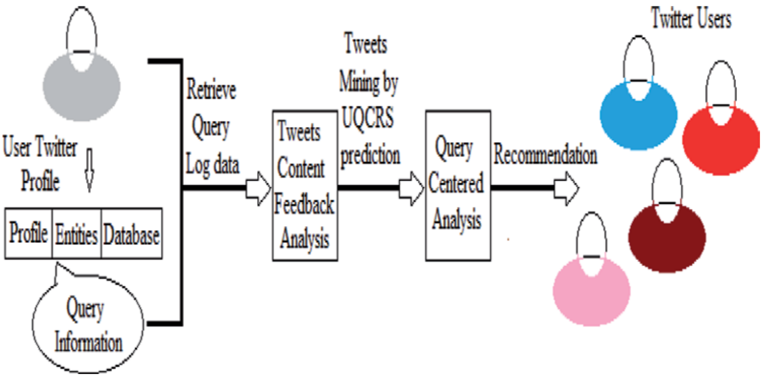


Figure 1.
Architecture of the UQCRS implementation.

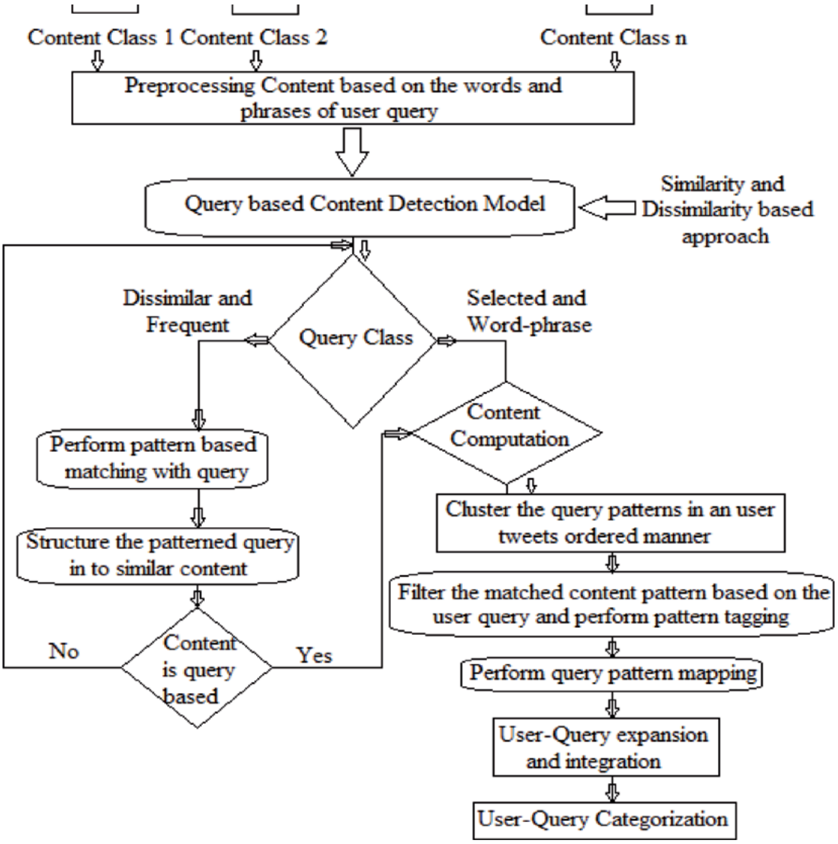
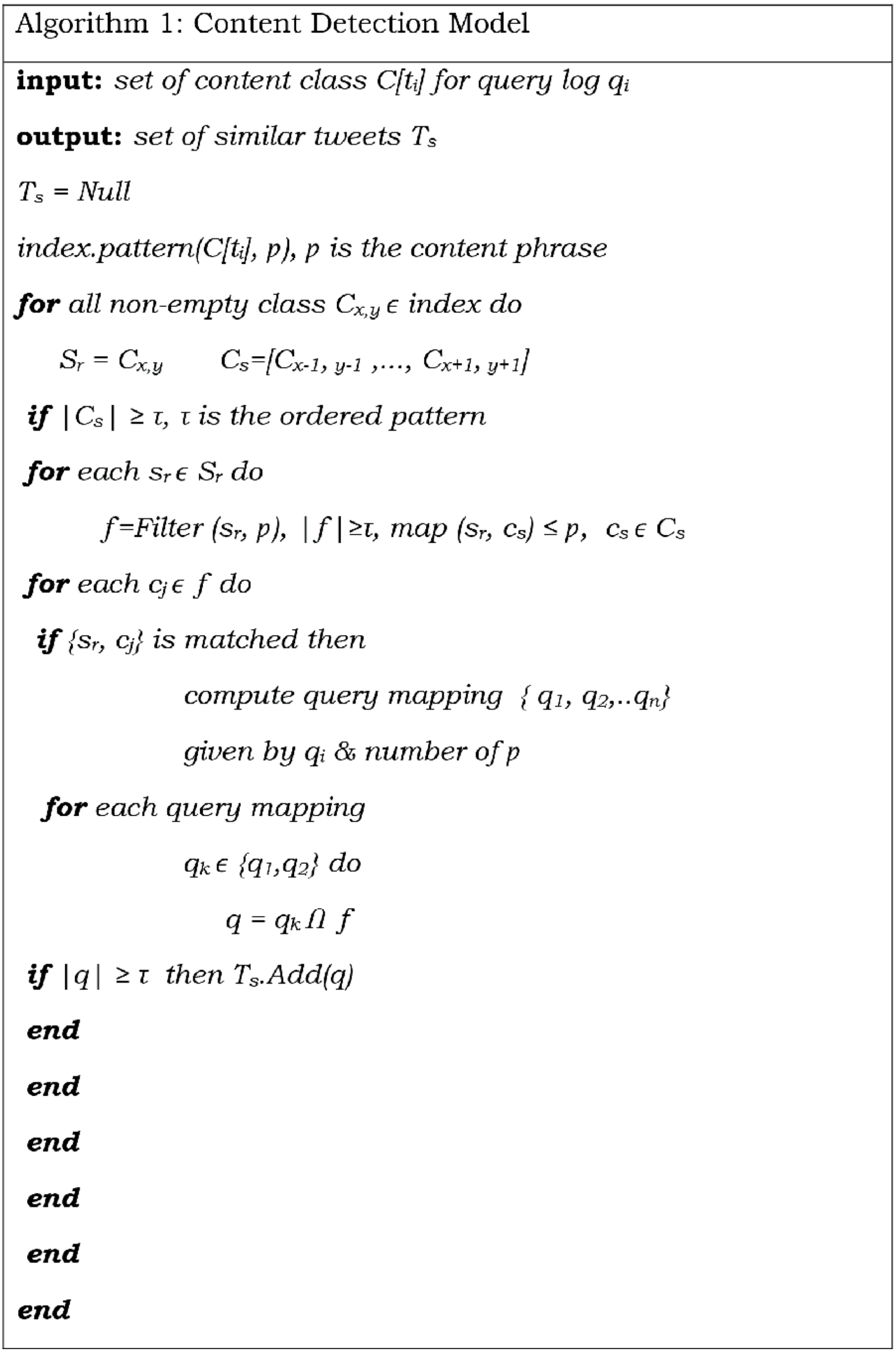

Content Class 1
Content Class 2
Content Class n
Preprocessing Content based on the words and phrases of user query
Query based Content Detection Model
Similarity and Dissimilarity based approach
Dissimilar and Frequent
Query Class
Selected and Word-phrase
Perform pattern based matching with query
Structure the patterned query in to similar content
Content is query based
No
Yes
Content Computation
Cluster the query patterns in an user tweets ordered manner
Filter the matched content pattern based on the user query and perform pattern tagging
Perform query pattern mapping
User-Query expansion and integration
User-Query Categorization

Figure 2.
User-query pattern categorization.

The tweets clustering process and the tweets' filtering are performing, shown in **Figure 2**, of the tweets as given by the user recommendation, which analyses at every instant of user interaction.

In the proposed system, each analysis is a service, and the Content Detection Model algorithm's operation is explaining in **Figure 3**, which is describing in two parts. The first content phrase built as a final set of query phrases per query log



```
Algorithm 1: Content Detection Model
input: set of content class $C[t_i]$ for query log $q_i$
output: set of similar tweets $T_s$
$T_s$ = Null
index.pattern($C[t_i]$, p), p is the content phrase
for all non-empty class $C_{x,y} \epsilon$ index do
    $S_r = C_{x,y}$    $C_s=[C_{x-1,\ y-1},\dots,\ C_{x+1,\ y+1}]$
 if $|C_s| \geq \tau$, $\tau$ is the ordered pattern
 for each $s_r \epsilon S_r$ do
        $f$=Filter ($s_r$, p), $|f| \geq \tau$, map ($s_r$, $c_s$) $\leq p$,  $c_s \epsilon C_s$
 for each $c_j \epsilon f$ do
  if {$s_r$, $c_j$} is matched then
              compute query mapping  { $q_1$, $q_2$,..$q_n$}
              given by $q_i$ & number of p
   for each query mapping
              $q_k \epsilon \{q_1,q_2\}$ do
                 $q = q_k \cap f$
 if $|q| \geq \tau$  then $T_s$.Add(q)
 end
 end
 end
 end
 end
end
```

Figure 3.
User twitter tweets detection.

brings together a maximum number of ordered patterns that make each filtered design generate enough matched tweets. In the second, query mapping is computing that shares similar tweets during consecutive query logs, exceeding the maximum ordered pattern.

3.1.2 User-query centered knowledgebase integration

User query analysis is complete by extracting knowledge from query log data, shown in **Figures 3** and **4**. Here, the scenario is that successive user tweets are removed based on the query log data, and the matched query is accessed

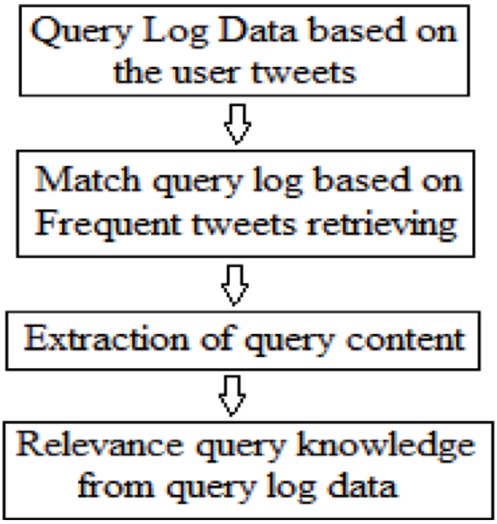

Query Log Data based on the user tweets
Match query log based on Frequent tweets retrieving
Extraction of query content
Relevance query knowledge from query log data

Figure 4.
Extracting the knowledge from query log data.

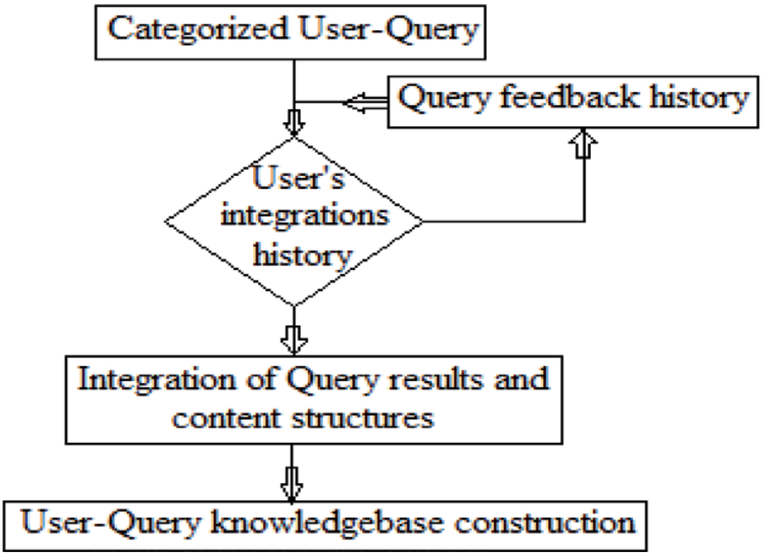

Categorized User-Query
Query feedback history
User's integrations history
Integration of Query results and content structures
User-Query knowledgebase construction

Figure 5.
User-query centered knowledgebase integration.

correspondingly through subsequent extraction. The accessed query content related to a Twitter account focuses on extracting the respective accounts retrieved from the tweets account. The latter tweet's accounts focus on the generic tweets with similar content, with identical query phrases. But during relevance query knowledge, similar user query search with less effective and provide the presence of user tweets with high similarity with different text types.

The scenario of user-query knowledgebase construction is showing in **Figure 5**. Here, from the previous methods, with the selected user-query procedure used as a tweet query category, user integration history is proposed, with knowledge extraction at every query feedback history loop-back. For query integration, the projected tweets from one day to many days are structured for continuous knowledge construction through the constant tweet's attribution and trace the future knowledge analysis, which evolves in the future based on the tweeted user query.

The model of the proposed recommendation system is shown in **Figure 6**, consists of:

a. a set T_U of N users, $T_U = \{T_U1, T_U2, T_U3,...., T_UN\}$

b. a set C of M items, $C = \{c_1, c_2, c_3,....., c_M\}$

c. a query cluster matrix Q_C, $Q_C = [qc_{mn}]$ where $m \epsilon T_U$ and $n \epsilon C$

d. a set f of N feature query sets, $f = [f_{mn}]$

e. a tweet knowledge weights $K_\omega = \{\omega_1, \omega_2, ... \omega_N\}$

User item set is associated with the number of feature vectors representing the tweet customers with different tweet phrases assigned to the user-query content model. In the recommender content model, the decision ranking prediction compares the users and Twitter queries in the categorized user-query item set and tweet's weights.

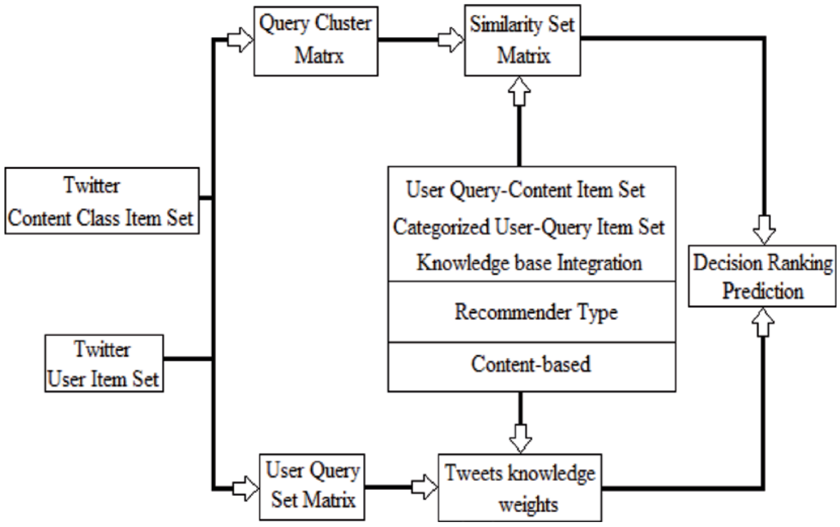

Query Cluster Matrx
Similarity Set Matrix
Twitter Content Class Item Set
User Query-Content Item Set
Categorized User-Query Item Set
Knowledge base Integration
Recommender Type
Content-based
Decision Ranking Prediction
Twitter User Item Set
User Query Set Matrix
Tweets knowledge weights

Figure 6.
The proposed user-query centered recommender system.

4. Results and discussions

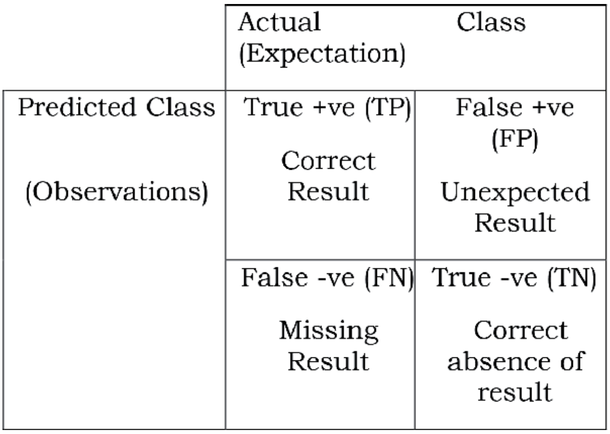
For experiments, a random public user Twitter dataset and real-time data using the API of Twitter is complete. The Twitter tweets containing the keywords "basket," "pencil," "work," "enter," and "formal" from the public domain are taking as the standard bag-of-words approach. Used this dataset for classification and collected 300 documents in each of the public domains.

For the classification of tweets, the true +ves, true -ves, false +ves, and false -ves constraints are utilized to equate the consequences of the classifier under the test with investigation techniques, which is illustrating in **Figure 7**.

The relations between TP, FP, FN, and TN are:

a. The relation $TP / [TP + FP]$ is describing as precision, which is the correctly classified metric.

b. The relation $TP / [TP + FN]$ is describing as Recall, which is the actual classified metric.



	Actual (Expectation) Class	
Predicted Class (Observations)	True +ve (TP) Correct Result	False +ve (FP) Unexpected Result
	False -ve (FN) Missing Result	True -ve (TN) Correct absence of result

Figure 7.
Classification matrix model for metric analysis.

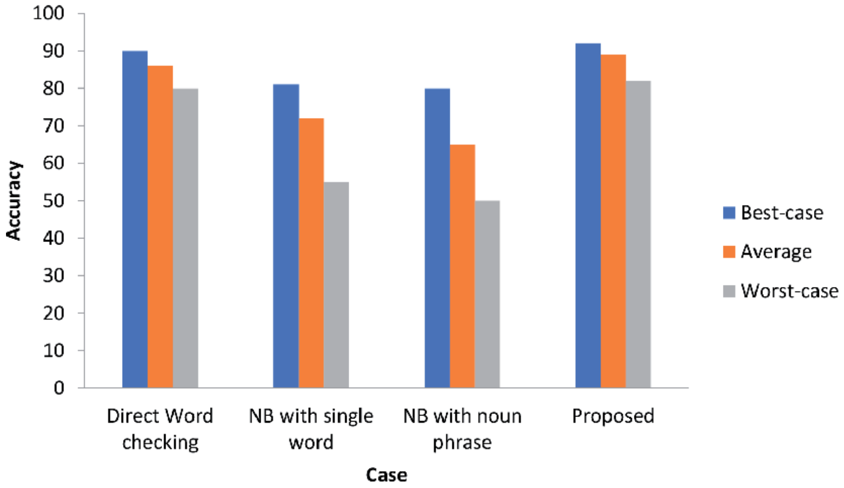


Figure 8.
Accuracy comparison of a classified tweet.

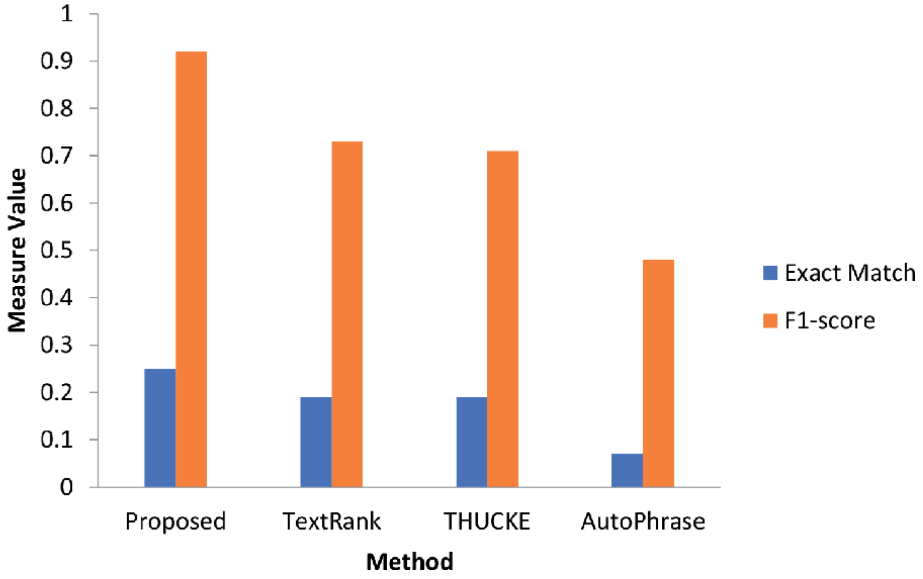


Figure 9.
Comparison of different algorithms for measured values.

c. Relation $2 * \left[{precision * recall}\big/{precision + recall} \right]$ is describing as the F-measure, which is a measure of precision and recall.

Figure 8 shows the accuracy of classifying a user query tweet in the users defined the recommended system. The highest accuracy is achieved through the proposed work, with an incredible number of word phrases, depending on the content of the user query and compared with Naive Bayes classifier (NB) [41].

Figure 9 compares the proposed system with different dataset approaches in terms of F1-score and the exact match. Because of the phrase-based content mining is made and tweets analysis is made accurately on two other datasets of different methods [42, 43] and proposed.

5. Conclusions

Social media connects connected individuals in numerous ways. For instance, online people can interact, communicate, collaborate, and socialize, showing new types of integration and interaction that were difficult to imagine only a short period before. Online communities also play a significant role in entrepreneurship, influence company feelings, and concepts and create many opportunities for unique investigation of person and team performance. In this chapter, a user query-centered recommendation system is deliberate to improve user-query analysis and tweets analysis for Twitter tweets. The proposed and implemented query categorization gives a satisfactory exact match performance in tweets categorizing. The content model is the most important model for the majority of tweets. Apart from finding the tweets, the phrases are identified and located in accuracy performance metric. As discussed, content integration is helpful for tweets match retrieval. For the given user tweet query, if the aim is to retrieve similar tweets from the public repository, retrieving the keywords is improved with the use of words and phrases.

Also, a novel algorithm based on content detection is using to extract the tweets using the bags-of-word method. Using the tweet's knowledge weights the proposed recommendation system avoids the dissimilar tweet's pattern identification problem. The above said three parameters are complete, which indicates that the proposed approach produces better accuracy results than the other methods.

6. Future scope

Future work focuses on The DTSR System may be extended along with additional user profiles such as film playlists, community groups, social media tweets, user emotion, user posts, and feature tweets to better the method recommended.

Author details

Vijayalakshmi Kakulapati
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India

*Address all correspondence to: vldms@yahoo.com

References

[1] References: Wayne Xin Zhao, et al. "Incorporating Social Role Theory into Topic Models for Social Media Content Analysis," IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 4, April 2015.

[2] Metin Turan, et al. "Automatize Document Topic and Subtopic Detection with Support of a Corpus," Social and Behavioral Sciences, Published by Elsevier, DOI: 10.1016/j. sbspro.2015.02.373, 2015.

[3] K. D. Rosa, et al. "Topical clustering of tweets," Proceedings of the ACM SIGIR: SWSM, 2011.

[4] Y. Duan, L. et al. "An empirical study on learning to the rank of tweets," in Proceedings of the 23rd COLING, 2010, pp. 295-303.

[5] M. Pennacchiotti, et al. "Making your interests follow you on Twitter," in Proceedings of the 21st CIKM, 2012, pp. 165-174.

[6] A. Pal et al., "Identifying topical authorities in microblogs," in Proceedings of the 4th ACM WSMINING. ACM, 2011, pp. 45-54.

[7] J. Weng, et al. "Twitter rank: finding topic sensitive influential Twitterers," in Proceedings of the 3rd ACM WSMINING, 2010, pp. 261-270.

[8] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. Information Retrieval 2(4):303-336.

[9] Zhao, W. X et al. 2011. Topical keyphrase extraction from Twitter. In ACL, 379-388.

[10] El-Kishky, A.; et al. 2014. Scalable topical phrase mining from text corpora. VLDB 8(3):305-316.

[11] Danilevsky, M.; et al. 2014. Automatic construction and ranking of topical keyphrases on collections of short documents. In Proceedings of MINING.

[12] Zhao, W. X.; et al. 2011. Topical keyphrase extraction from Twitter. In ACL, 379-388.

[13] King, G.; et al. 2014. Computer-assisted keyword and document set discovery from the unstructured text—copy at http://j.mp/1qdVqhx 456.

[14] Luke, T.; et al. 2013. A framework for specific term recommendation systems. In SIGIR, 1093– 1094.

[15] Bhatia, S; et al., 2011. Query suggestions in the absence of query logs. In SIGIR, 795-804.

[16] Zhang, Y.; et al. 2014. Bid keyword suggestion in sponsored search based on competitiveness and relevance. Information Processing & Management 50(4):508-523.

[17] Hahm, G. J.; et al. 2014. A personalized query expansion approach for engineering document retrieval. Advanced Engineering Informatics 28(4):344-359.

[18] Global Faces and Networked Places, A Neilsen report on Social Networking's New Global Footprint, March 2009. Neilsen company.

[19] Z. Tan et al. "An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle," IEEE Access, vol. 5, pp. 27211-27228, 2017.

[20] U. Fasahte, et al. "Hotel recommendation system," Imperial Journal of Interdisciplinary Research, vol. 3, no. 11, 2017.

[21] A. Verma et al. "A hybrid genre-based recommender system for movies

using genetic algorithm and kNN approach," International Journal of Innovations in Engineering and Technology, vol. 5, no. 4, pp. 48-55, 2015.

[22] H. Jazayeriy et al. "A fast recommender system for the cold user using categorized items," Mathematical and Computational Applications, vol. 23, no. 1, p. 1, 2018.

[23] Rosenthal, S., et al. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).

[24] M. Ibrahim et al. "Design and application of a multivariant expert system using Apache Hadoop framework," Sustainability, vol. 10, no. 11, p. 4280, 2018

[25] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. Information Retrieval 2(4):303-336.

[26] Zhao, W. X.; et al. 2011. Topical keyphrase extraction from Twitter. In ACL, 379-388.

[27] El-Kishky, A.; Song, Y.; Wang, C.; Voss, C. R.; and Han, J. 2014. Scalable topical phrase mining from text corpora. VLDB 8(3):305-316.

[28] H. Li, D. et al. "Disaster response aided by tweet classification with a domain adaptation approach," Journal of Contingencies and Crisis Management, vol. 26, no. 1, pp. 16-27, 2018.

[29] H.-T. Zheng, et al. "A deep temporal neural music recommendation model is utilizing music and user metadata," Applied Science, vol. 9, no. 4, p. 703, 2019.

[30] Java, A., et al.: Why we twitter: understanding microblogging usage and communities. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. pp. 56-65 2007.

[31] Joachims, T., et al.: Accurately interpreting clickthrough data as implicit feedback. In: Proc. of the 28th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'05). pp. 154-161 2005.

[32] Kwak, H., et al.: What is Twitter, a social network or a news media? In: Proc. of the 19th Int. Conf. on World Wide Web (WWW'10). pp. 591-600 2010.

[33] Liben-Nowell, D., et al.: The link prediction problem for social networks. In: Proc. of the 12th Int. Conf. on information and knowledge management. pp. 556-559. CIKM '03, ACM, New York, NY, USA 2003.

[34] M. Gao et al. A movie recommender system from tweets data, January 2019, http://cs229.stanford.edu/proj2015/299_report.pdf.

[35] M. Ibrahim et al. "Design and application of a multivariant expert system using Apache Hadoop framework," Sustainability, vol. 10, no. 11, p. 4280, 2018.

[36] M.-Y. Hsieh, et al. "Building a mobile movie recommendation service by user rating and APP usage with linked data on Hadoop," Multimedia Tools and Applications, vol. 76, no. 3, pp

[37] Muhammad Ibrahim, et al. " A Neural Network-Inspired Approach for Improved and True Movie Recommendations," Hindawi, Computational Intelligence and Neuroscience, Volume 2019, pp-1-19.

[38] O. Loyola-Gonzalez et al. "Fusing pattern discovery and visual analytics approach in tweet propagation,"

Information Fusion, vol. 46, pp. 91-101, 2018.

[39] Phelan, O. et al..: Using Twitter to recommend real-time topical news. In: Proc. of the 3rd ACM Conf. on Recommender Systems (RecSys'09). pp. 385-388 2009.

[40] S. Kumar, et al. "Movie recommendation system using sentiment analysis from microblogging data," 2018, https://arxiv.org/abs/1811.10804.

[41] Rada Mihalcea et al. 2004. Textrank: Bringing order into text. In EMNLP,

[42] Zhiyuan Liu, et al. 2011. Automatic keyphrase extraction by bridging the vocabulary gap. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. ACL, 135-144.

[43] Jingbo Shang, et al. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering 30, 10 (2018), 1825-1837.

Chapter 2

Open Government Data: Development, Practice, and Challenges

Omer Hassan Abdelrahman

Abstract

This chapter explores the concept of open data with a focus on Open Government Data (OGD). The chapter presents an overview of the development and practice of Open Government Data at the international level. It also discusses the advantages and benefits of Open Government Data. The scope and characteristics of OGD, in addition to the perceived risks, obstacles and challenges are also presented. The chapter closes with a look at the future of open data and open government data in particular. The author adopted literature review as a method and a tool of data collection for the purpose of writing this chapter.

Keywords: open data, open government data, OGD, OGD development, OGD principles, OGD practice, OGD barriers, open data portals

1. Introduction

The concept of Open Government Data (OGD) has been heavily debated during the last few years. It has drawn much interest and attention among researchers and government officials worldwide. Many of the developed and developing countries have launched open data initiatives with a view to harnessing the benefits and advantages of open government data. This chapter is dedicated to highlighting the various aspects of open data and open government data.

According to the Open Definition, "Open" in the context of data and content "means anyone can freely access, use, modify, and share for any purpose". There are many types of data that can be open and used or re-used by the public. These include data relating to culture, science and research, finance, statistics, weather, and environment [1, 2].

The Open Knowledge Foundation outlined key features of openness as the following:

- Availability and access: the data must be available as a whole, in a convenient and modifiable form and at a reasonable reproduction cost, preferably by downloading over the internet.

- Reuse and redistribution: the data must be provided under terms permitting reuse and redistribution, with the capability of mixing it with other datasets. This data must be machine-readable.

- Universal participation: the data should be available for everyone to use, reuse and redistribute without discrimination against fields of knowledge, or against persons or groups [2].

Features of open data also include the following aspects: Data should be primary and timely and accessed data must be available in non-proprietary formats and free to use with unrestricted license. Data should also be as accurate as possible. Although most of the data will not meet all of these criteria, data is only truly open if it meets most of them [3].

The earliest appearance for the term open data was in 1995. It was related to the disclosure of geographical and environmental data in a document written by an American agency. The scholarly community understood the benefits of open and shareable data long before the term open data was a technical object or political movement [4].

The Scholarly Publishing and Academic Resources Coalition (SPARC) defined open data from a research perspective as: "Open Data is research data that is freely available on the internet permitting any user to download, copy, analyze, re-process, pass to software or use for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" [5]. SPARC stressed the benefits of open data in that it accelerates the pace of discovery, grows the economy, helps ensure people do not miss breakthroughs, and improves the integrity of the scientific and scholarly record,

The current concept of open data and particularly open government data (OGD) started to become visible and popular in 2009 with a number of governments in the developed world who announced new initiatives to open up their public information records such as the USA, UK, and New Zealand. These initiatives were triggered by the mandate for transparency and open government from the then American President Barack Obama administration, thus kick starting the Open Government Data Movement [6, 7].

To legalize the use of the published public data, open data must be licensed. This license should permit people to freely use, transform, redistribute and republish the data even on a commercial basis. A number of standard licenses designed to provide consistent and broadly recognized terms of use are employed. These licenses include: Creative Commons (CC), Open Data Commons Open Database License (ODbL), and Open Data Commons Public Domain Dedication and License (PDDL). Some governmental organizations and international organizations have released their own tailored Open Data license such as The Worldbank Data License, French open Data License, and UK Gov. Data License. Standard licenses have many advantages over bespoke licenses, including greater recognition among users, increased interoperability, and greater ease of compliance [8, 9].

2. Development of the open government data concept

Open government data (OGD) is government-related data that is made open to the public. Government data usually contain various datasets, such as budget and finance, population, census, geographical data, parliament minutes, etc. It also includes data collected by public organizations or agencies such as data related to climate or pollution, public transportation, traffic, child care or education [10].

Open government data has been associated with Open Government which is perceived as a phenomenon encompassing a number of characteristics and dimensions such as information availability, transparency, participation, collaboration, and information technologies [11]. The concept of open government data can be traced back to the year 1966 when the USA federal government passed the Freedom of Information Act (FOIA). The coming of the internet and new information and telecommunications technologies contributed to the more recent interest and understanding of the value and benefits of government information for the sake of transparency, collaboration and innovation [12]. Two significant consequent developments contributed positively to the open government data; these are the launching of data.gov in the USA in May 2009 and the data.gov.uk in the United Kingdom (UK), in January 2010. It subsequently spread out to many other countries around the world, as well as to international organizations, including the World Bank and the Organization for Economic Co-operation and Development (OECD). Moreover, the concurrent advances in the information and telecommunications technologies also played a role in the development of open government data, coupled with the passing of open standard laws by many countries such as Canada, the USA, Germany and New Zealand, and the setting of policies on open data focusing on indexing government data holdings [13, 14].

In 2015 a number of governments, civil society members, and international experts convened with the purpose of representing an internationally-agreed set of norms for how to publish government and other public sector organizations data. They then formulated a set of principles called the Open Data Charter. They introduced these principles with the following statement:

"We, the adherents to the International Open Data Charter, recognize that governments and other public sector organizations hold vast amounts of data that may be of interest to citizens, and that this data is an underused resource. Opening up government data can encourage the building of more interconnected societies that better meet the needs of our citizens and allow innovation, justice, transparency, and prosperity to flourish, all while ensuring civic participation in public decisions and accountability for governments…" [15].

The conveners agreed to adhere to the following set of principles concerning access and release of government and public sector data. That data should be.

i. Open by Default;

ii. Timely and Comprehensive;

iii. Accessible and Usable;

iv. Comparable and Interoperable;

v. For Improved Governance and Citizen Engagement;

vi. For Inclusive Development and Innovation [15].

The scope of Open government data which is made available with no restrictions on its use, reuse, or distribution covers all data funded by public money excluding private, security sensitive, and confidential data.

3. Open government data practice

The Open Data Barometer, an international benchmark of how open data publishing is used by governments for accountability, innovation and social impact ranked, in 2018, 30 leading world countries, excluding the EU countries, according to their performance and commitment to the principles of open data movement. It measured the progress these 30 governments have made against three essential ingredients for good open data governance, defined as part of the Open Data Charter updates process, namely Open by Default, Data Infrastructure, and Publishing with Purpose. In other words, the Barometer ranked governments according to three criteria: readiness for open data initiatives, implementation of open data programs and impact that open data is having on business, politics, and civil society. The top ten ranking countries were Canada, UK, Australia, France, South Korea, Mexico, Japan, New Zealand, USA, and Germany, respectively [16]. On the other hand, the open data maturity assessment reported on data maturity in European countries for the year 2020. It provided insight into the developments in the open data field in European countries, including the 27 EU Member States, and the participating European Free Trade Association (EFTA) countries Liechtenstein, Norway, and Switzerland, including the Eastern Partnership countries Azerbaijan, Georgia, Moldova, and Ukraine, besides the United Kingdom. The assessment measured open data maturity with regard to four dimensions: policy, impact, portal, and quality. Maturity was scored against these dimensions, forming an overall score of open data maturity for each country. The countries were clustered into four groups, from the most mature to the least. Seven countries are labeled as trend setters according to their performance. They are Denmark, Spain, France, Ireland, Estonia, Poland, and Austria [17].

Public institutions are among the largest creators and collectors of data in many different fields or categories. These data categories include areas such as transportation, traffic, finance, environment, economy, government, weather, geographical information, tourist information, statistics, business, public sector budgeting, performance levels, and science and technology. Data about policies and inspection in fields such as education quality, safety, food… etc. is also included. In addition to this, international OGD sites have a specific characteristics and data patterns in terms of their OGD levels, data formats, and datasets. Top data formats used are CSV, PDF, RDF JSON, and XML. **Table 1** provides the definitions and examples of these file types. However, there are clear variations among world regions in terms of the number of data formats, datasets, and data categories [7, 23].

Boosting democratic control and political participation, fostering service and product innovation, and enhancing law enforcement are three primary motivations to publishing government data. In comparing the open data strategies of five countries, namely Australia, Denmark, Spain, the United Kingdom and the United States, It was found that the focus of the strategic plans differs. For example, the United States government focused on transparency for the purpose of increasing public engagement, Denmark emphasized the potentials that open data offers for the introduction of new products and services, whereas the United Kingdom focused on the use of open data for strengthening law enforcement [24].

Citizens use four types of OGD applications in order to engage with their governments for the purpose of open government. The first type of application focuses on using access to government information to weed out corruption in government. The second type of application focuses on the direct benefit to the public of access to legal materials, such as access to the law itself. The third application is related to informing policy decisions whereby information helps citizens to better understand their own communities. The fourth type of application is related to consumer products where applications are products that bring open government to a wide consumer audience [14]. Open Government Data can lead to a more effective and efficient government particularly regarding its relation with citizens. This can be achieved by increasing transparency and accountability, developing trust, credibility and reputation, promoting progress and innovation, encouraging public education and community engagement, and storing and preserving information

File type	Definition	Example
CSV	CSV stands for "comma-separated values". This type of file is a simple text file where information is separated by commas. These files are usually encountered in spreadsheets software and databases. They may also use other characters such as the semicolons to separate data. By using a CSV file format, complex data can be exported from one application to a CSV file, and then imported in that CSV file into another application.	Product, Size, Color, Price Shirt, Large, White,$15 Shirt, Small, Green $12 Trousers, Medium, Khaki, $35
JSON	JSON stands for "JavaScript Object Notation". This is a text file format for storing and transporting data. By using JSON, JavaScript objects can be stored as text. The string in the example defines an object with 3 properties: name, age, job and each property has a value.	'{"name": "Jack", "age":41, "job": accountant}'.
RDF	An RDF file is a document written in the Resource Description Framework (RDF) language. This language is used to represent information about resources on the web. It contains the website metadata. Metadata is structured information. RDF files may include a site map, an updates log, page descriptions, and keywords.	<?xml version = "1.0? > <rdf xmlns = "http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:s = "http://description.org/schema/"> <Description about = "https://www.xul.fr/Wells"> <s:author>The Invisible Man<s:author> </Description> </rdf>
XML	An XML is a file written in extensible markup language. It is used to structure data for storage and transport. In an XML file, there are tags and text. The tags provide the data structure, and the text in the file is surrounded by these tags, which adhere to specific syntax guidelines. The XML format is used for sharing structured information between programs, and between computers and people, both locally and across networks.	<part number = "1976" > <name> Windscreen Wiper</name> <description>The Windscreen wiper automatically removes rain from your windscreen, if it should happen to splash there. It has a rubber <ref. part = "1977" > blade</ref.> which can be ordered separately if you need to replace it. </description> </part>

Table 1.
File types definitions and examples [18–22].

over time [25]. Therefore, open data can lead to open government which is defined as: "..... a multilateral, political, and social process, which includes in particular transparent, collaborative, and participatory action by government and administration. To meet these conditions, citizens and social groups should be integrated into political processes with the support of modern information and communication technologies, which together should improve the effectiveness and efficiency of governmental and administrative action" [26].

3.1 Portals and the publication of OGD

According to the principles of OGD, data must be: complete, primary, timely, accessible, and machine-readable. It should also be non-discriminatory, non-proprietary and License-free. Furthermore, public institutions should publish all data they have if it would not violate security, privacy or other legitimate restrictions [27].

The World Wide Web Consortium (W3C) outlined three steps for publishing open data, which will help the public to easily find, use, cite and understand the data:

Step 1: Publishing the data in its raw form. The data should be well-structured to enable its use in an automated manner by the users of the data. Data may be in XML, RDF or CSV formats. Formats used should allow the data to be seen as well as extracted by the users.
Step2: Creating an online catalog of the raw data, complete with documentation, to enable users to discover published data.
Step 3: Making the data human readable as well as machine-readable [28].

Open data portals are a very important component of open data infrastructure. They connect data publishers with data users enabling the former to deliver open data and establish the necessary relationships for increasing transparency. Open data portals, which are essentially data management software, contain metadata about datasets so that these datasets could be accessed and utilized by the users. The open data portal includes the tools which help the users to find and harvest all relevant data from public sector databases. From the users' perspective, features of open data portals can be used to specify datasets they need and to request for datasets [29]. Thus, Open data portals play the role of interface between government data and citizens who use or reuse this data. Consequently, a portal should have user- friendly features such as a clean look with a search facility. The portal should also provide information about the responsible authority which hosts the portal written clearly and in a simple language. The portal's contents should be organized into categories and subcategories. It should also aim to engage citizens' ideas and feedback in addition to its basic function of making data available to stakeholders. Data quality and standards, and the language settings are very important elements in portals so that they can satisfy their users' needs [30].

The World Wide Web Consortium's (W3C) benchmark for publishing open government data and the World Bank's technical option guide outlined the necessary technical requirements for establishing efficient and modern OGD data centers. These requirements include, among other things, that:

i. Public datasets should be published in their raw state rather than in an analyzed form,

ii. Each dataset is accompanied by a well-documented metadata, and.

iii. Data is stored in multi formats – both human and machine- readable formats, such as CSV, XML, PDF, RDF JSON etc. to enable the users to easily access published data. It is expected that documental data are stored in either PDF, doc(x) or Excel, and geographical data are stored in Keyhole Markup Language (KML) or their equivalent alternatives [28, 31].

As for OGD portals' content and functionality requirements, these include the following:

- A number of Datasets.

- Links to External Websites.

- A number of data Categories such as data about education, weather, budget ...etc.

- Data Currency. Data should be current and up to date.

- Availability of metadata. Datasets should come with requisite information that adequately describes the data.

- Data Search. A 'search box' feature should be available to allow users to easily locate specific information by entering a search term.

- Availability of working social media plugins. This feature enables data users to share their experiences and suggest new datasets through comments in social media websites such as Facebook, tweeter…etc.

- Data Visualization Functionality. This allows users to visualize the data. Visualization features come in various forms such as tables, graphs, maps etc. [28, 31].

There are a number of additional requirements that contribute to making portals achieve sustainability, meet user needs and have an added value impact. These requirements are the following:

- Dataset should be organized for use and not only for the sake of publication.

- Should learn from the techniques used by recent commercial data market, share knowledge to promote data use, and adapt methods that are common in the open source software community;

- Invest in best practices related to discoverability.

- Enhance reuse by publishing high quality metadata;

- Ensure interoperability by adopting standards;

- Engage with more users and re-users by co-locating tools;

- Enhance value by linking datasets;

- Being accessible by offering both options for big data, such as Application Programme Interfaces (API), and options for more manual processing, such as comma separated value files, thus ensuring a wide range of user needs are met;

- Co-locating documentation to make it easy for non-expert users to understand the data;

- Assess how well they are meeting users' needs by being measurable [32].

A number of open source and commercial open data portals software exist. Some of the more widely used open source software are the following:

i. CKAN: This is an open source data portal designed to allow publishing, sharing and managing datasets; it has a number of functionalities to the managers and end-users such as full-text search, reporting tools, and multi-lingual support. It also provides an Application Programming Interface (API) to access the data.

ii. DKAN: compared to CKAN, this software has more data-oriented features including scrapping, data harvesting, visual data workflow, and advanced visualization. DKAN users are mainly government organizations and Non-Governmental Organizations (NGOs).

iii. Socrata: It has a number of powerful data management tools for database management, data manipulation, reporting, visualization with advanced options and customized financial analytics insights. Socrata has two licenses; an open source license for the community edition and commercial one for the enterprise edition.

iv. Dataverse: It is built to share and manage large data-sets. It helps its users to collect, organize, and publish their data-sets in a collaborative platform. Dataverse is employed around the world by Non-Governmental Organizations (NGOs), Government organizations and research centers [8].

3.2 Open data best practice

The European Data Portal published a report in the year 2020 highlighting the best open data practices implemented by the three top performing countries of the year 2019 assessment - Cyprus, France and Ireland. The reported practices may be applicable to other international contexts. The practices were categorized into four aspects relating to open data, namely, Open Data Policy, Open Data Portal, Open Data Impact, and Open Data Quality. **Table 2** shows the best practices associated with each one of these aspects.

Aspect	Best practices
Open Data Policy	• Setting up of open data policy and legislation and strategy.
	• Development of an implementation plan so as to have an actionable strategy and clear responsibilities.
	• Setting up of an open data liaison officers network and maintaining of close contact with them
Open Data Portal	• Inclusion of features that go further than enabling users to find datasets.
	• Focusing on interaction between data providers and data re-users through discussion forums, dataset-specific feedback and rating systems.
	• Provision of example cases of open data re-use.
Open Data Impact	• Monitoring and analysis of the re-use of open data.
	• Making sure of the availability of the right datasets at the right time,
	• Interaction with data re-users for the purpose of understanding their needs through open data events.
Open Data Quality	• Provision of manuals and technical guidelines for the purpose of responding to frequently asked questions.
	• Active training of data providres.
	• Integration of all measures, guides and training possibilities in one platform on the portal

Table 2.
Open data best practices [33].

4. Benefits of open government data

Open government data has a number of economic and political implications and benefits, particularly on the democratic aspect. They include better transparency, citizens' trust in the government and collaboration in governance, and economic development. These benefits of open government data utilization can be detailed in the following:

i. Political and social benefits. These include the following aspects: more transparency, democratic accountability; more participation and self-empowerment of citizens; creation of trust in government; public engagement; equal access to data; new governmental services for citizens; improvement of citizen services, citizens' satisfaction, and policy-making processes; allowing more visibility for the data provider; creation of new insights in the public sector; and introduction of new and innovative social services.

ii. Economic benefits. These include the following aspects: economic growth; stimulation of innovation; contribution towards the improvement of processes, products, and services; adding value to the economy by creating a new sector; and availability of information for investors and companies. When Open Data is used to produce new products or start new services, it can increase the demand for more data causing the release of more datasets and improvements in data quality [34].

iii. Operational and technical benefits. These include the following aspects: The ability of reusing data; optimization of administrative processes; improvement of public policies; accessing external problem-solving capacity; fair decision-making by enabling comparison; easier discovery of and access to data; creating new data based on combining and integrating existing data; validation of data by external quality checks; and avoidance of data loss [6, 23].

For OGD to be beneficial, it should meet the following conditions:

i. Quality of data. This includes four components; timeliness, availability of metadata, accuracy and usefulness.

ii. Legislation/policy. A clear legal framework should be put in place in order to organize the relationship and avoid uncertainties regarding copyright, privacy, personal data and data openness.

iii. Skills. Technical skills and knowledge about data on the part of users is essential in order for them to be able to use open government data, such as knowledge about statistics or programming.

iv. Infrastructure. Infrastructure is required for the purpose of facilitating the exchange of data between government institutions and users. Examples of such infrastructure are software for data analytics and discovery, and web-based platforms. The essential features of OGD infrastructure which have strong impact on its utilization are feedback mechanisms between supplier and users, and data processing capabilities,

v. Privacy. OGD policies should consider privacy issues by eliminating private-sensitive data and data related to national security issues. These policies should ensure compliance to confidentiality and privacy guidelines [35].

5. Open government data barriers, challenges, and risks

5.1 Barriers and challenges

Open government data faces a number of barriers and challenges that may impede its development and implementation. Some of these barriers are related to either the data providers or the data users, while other barriers can be attributed to both sides. Barriers that might face either side are outlined below:

- Prevalence of closed government culture and lack of open government data policy.
- Existence of privacy legislation that protects privacy violation leading to identification of persons, besides existence of conflicting laws about data access.
- Poor data quality. This includes lack of sufficient and accurate data and availability of obsolete and non-valid data.
- Difficulty in searching and browsing data due to lack of metadata or an index, complexity of available data formats and datasets, in addition to information overload and lack of open data user manuals.
- Lack of standardization of open data policy and lack of metadata standards and absence of standard software for processing open data [23, 24, 36].

There are some other barriers that are encountered by both open data publishers and users. These include the following:

- Lack of technical knowledge on metadata quality on the part of portal owners. This may lead to the publication of inappropriate metadata which in turn causes re-users to find it difficult to find the data they need.
- There are political, organizational, legal, technical and financial barriers. These instances can be improved by taking into consideration the specific needs of the users of open data.
- Geospatial data has its own specific barriers resulting from the use of different standards in relation to other types of open data. Dealing with this type of data requires specific technical knowledge and expertise [37, 38].

Challenges at the institutional level include the avoiding culture of governments whereby governments are reluctant to open public data, the time consuming procedure to access and reuse data, and the fact that governments do not take users' ideas into consideration in government administration. On the other hand, challenges at the users' level include lacking of advance search facilities, lacking of helpdesk facilities for the users, and lacking of expertise to analyze data on the part of the users [39].

There are some challenging factors which discourage institutions and governmental bodies to join an open government data initiative. These factors include, but are not limited to the following:

- Lack of awareness of open data.
- Lack of motivation and purpose to opening public data.
- Lack of open-mindedness about the application of open data and the focus on publishing of data regardless of its good quality or perceived value.
- Non allocation of budget for opening data because it is still a recent not fully understood concept.
- Absence of an institutional body that is dedicated solely to the task of opening data which results in lack of regular monitoring of the performance of the open data initiative [10].

5.2 Risks

Many risks confront and may consequently impede the proper implementation and utilization of open government data. They include the ones listed below:

- Difficulties in determining who owns the published data. This may be accompanied by unclear responsibility and accountability about publishing the data.
- Unintentional violation of privacy and violation of legislation may take place.
- Published data can be biased, misinterpreted or misused.
- Open data may have negative consequences for the government.
- Poor information quality may lead to wrong decision making.
- Embargo period may cause published data to be out of date.
- Others may profit from open data rather than the intended citizens.
- Data with little or no value may be published resulting in a waste of resources [40].

To avoid and mitigate OGD challenges and risks, a number of practical solutions can be designed to enhance the accessibility and reusability of open government data on the legal, institutional and technical levels. These solutions include:

- Creation of data portals and metadata;
- Simplification of licensing issues.
- Education of data users and data providers on what is technically and legally possible so they can develop their plans within these boundaries.
- linking of the discussion on technical and legal requirements so that the former may not end up being difficult to implement in national jurisdictions and the latter may not be unrealistic or unadjusted to technical developments and practice [41].

6. Future of OGD

Four key elements characterize future trends of open data, namely:

- Purposeful publication of the data, focusing on its impactful reuse.

- Strengthening of data collaboration and partnerships; expanding the circle of those involved in open data projects and enabling more direct collaboration between data holders and data users.

- Advancement of open data at the subnational level and emphasizing on building open data capacity and meeting open data demand at the subnational level, rather than only at the national level. This is achieved by publishing data held by the public sector and other institutions in cities, municipalities, states, and provinces.

- Prioritization of data responsibility and data rights; Potential bias in the analysis and use of certain open datasets or how open data initiatives might negatively impact the rights of citizens. Moreover, privacy issues should be taken into consideration by practitioners. These are key elements in any open data project [42].

A seminar held by Statisticians, civil society and private sector ahead of the 48th session of the UN Statistical Commission that took place in 2017 discussed new trends and emerging issues in open data in light of the 2030 Agenda for Sustainable Development. The outcome of this seminar included the following insights and recommendations for the purpose of making the world more open to open data:

- Providing free access and use of data by open data platforms for purposes of transparency, accountability and daily decision making;

- Ensuring that the principles of data rights and access are matched with strict ethical and security protocols;

- Facilitating and enabling the efforts to making data more open by advanced technologies and approaches to data architecture and management.

- Collaboration with civil society in issues related to open data, particularly in the areas of principles, readiness and evaluation of openness of data, and collaboration with academia and technology firms for building portable and interoperable common technology infrastructure as a public good [43].

7. Conclusion

This chapter explored the various aspects of open government data. The chapter opened by defining the concepts of openness, open data, and open government data (OGD). It then proceeded to explaining how OGD developed during the last few years and highlighted the most important cornerstones of this development. The chapter then explained the various requirements of OGD implementation and utilization. It highlighted the practice of OGD around the world. It also explained the role of portals in OGD implementation and utilization, outlining their various technical and functional requirements, besides introducing a number of open source portal software and applications. The chapter then elaborated on a number of benefits and advantages of open government data for the government and the citizens. Finally the chapter discussed the barriers, challenges and risks that are confronted by open government data initiatives. The chapter then closed by highlighting some perceived future trends of open government data.

Author details

Omer Hassan Abdelrahman
Department of Library and Information Science, University of Khartoum, Khartoum, Sudan

*Address all correspondence to: omhass@hotmail.com

References

[1] Open Definition: Defining Open in Open Data, Open Content and Open Knowledge [Internet]. 2021. Available From: https://opendefinition.org [Accessed: 2021- 07-14]

[2] Open Knowledge Foundation [Internet]. 2021. Available From: https://okfn.org/opendata/ [Accessed: 2021- 07-14]

[3] Yannis C, Anneke Z, Charalampos A, Marijn Thomas L, Enrico F. The World of Open Data: Concepts, Methods, Tools and Experiences. 1st ed. Springer International Publishing; 2018. 229 p. DOI: 10.1007/978-3-319-90850-2

[4] Simon Chignard. A brief history of Open Data [Internet]. 2013. Available From: http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data [Accessed: 2021- 07-15]

[5] SPARC. Open data [Internet]. 2021. Available From: https://sparcopen.org/open-data/ [Accessed: 2021- 07-14]

[6] Daniel D, Jonathan G, Tim M, Antti P, Rufus P, Julian T, Ton Z. Open Data Handbook [Internet]. 2012. Available From: https://opendatahandbook.org/guide/en/ [Accessed: 2021- 07-20]

[7] Rong Tang, Jie Jiang. Characteristics of Open Government Data (OGD) Around the World: A Country-based Comparative Meta-Analysis. Data and Information Management. 2021;5(1):11-26. DOI: https://doi.org/10.2478/dim-2020-0026

[8] Top 16 Open source Data Portal Solutions for Open Data Publishing [Internet]. 2019. Available From: https://medevel.com/15-data-portals-opensource/ [Accessed: 2021- 07-29]

[9] Open Data Essentials [Internet]. 2019. Available From: http://opendatatoolkit.worldbank.org/en/essentials.html [Accessed: 2021- 07-25]

[10] Judie A, Fabrizio O, Simon S, Sören A. A systematic review of open government data initiatives. Government Information Quarterly.2015; 32(4):399-418. DOI: https://doi.org/10.1016/j.giq.2015.07.006

[11] Ramon G, Mila G, Theresa A. Beyond Transparency, Participation, and Collaboration? A Reflection on the Dimensions of Open Government. Public Performance & Management Review. 2020; 43(3):483-502. DOI: 10.1080/15309576.2020.1734726

[12] Tasha A, David M, Mekala R, Matt R. Future of Open Data: Maximizing the Impact of the OPEN Government Data Act [Internet]. 2019. Available From: https://www.datafoundation.org/future-of-open-data-maximizing-the-impact-of-the-open-government-data-act [Accessed: 2021- 08-25]

[13] Kathleen J. Open Government Data and the Right to Information: Opportunities and Obstacles. The Journal of Community Informatics. 2012;8(2). DOI: https://doi.org/10.15353/joci.v8i2.3042

[14] Joshua T. Open Government Data: The Book. 2nd ed. 2014. Available From: https://opengovdata.io [Accessed: 2021- 08-25].

[15] International Open Data Charter [Internet]. 2016. Available From: https://opendatacharter.net/principles" [Accessed: 2021- 07-20]

[16] Open Data Barometer - Leaders Edition [Internet].2018. Available From: https://opendatabarometer.org/leadersedition/report/ [Accessed: 2021- 08-15]

[17] Open Data Maturity Report. Luxembourg: Publications Office of the European Union; 2020. DOI: 10.2830/619187

[18] Dave Johnson. What is a CSV file? How to open, use, and save the popular spreadsheet file in 3 different apps. [Internet]. 2021. Available From: https://africa.businessinsider.com/tech-insider/what-is-a-csv-file-how-to-open-use-and-save-the-popular-spreadsheet-file-in-3/4gbqn4b. [Accessed: 2021- 9 -12].

[19] JSON – Introduction. [Internet]. 2021. Available From: https://www.w3schools.com/js/js_json_intro.asp. [Accessed: 2021-9-12].

[20] Fileinfo.com.[Internet].2017. Available From: https://fileinfo.com/extension/rdf. [Accessed: 2021- 9 -12].

[21] RDF, Resource Description Framework, how to use.[Internet].2021. Available From: https://www.xul.fr/en-xml-rdf.php#complete-example. [Accessed: 2021- 9 -12].

[22] What is an XML file and How Do I Open One? [Internet]. 2021. Available From: https://www.indeed.com/career-advice/career-development/xml-file. [Accessed: 2021- 9 -12].

[23] Marijn J, Yannis C, Anneke Z. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management. 2012; 29(4): 258-268. DOI: 10.1080/10580530.2012.716740

[24] Noor H, T. Van den Broek. Open data: an international comparison of strategies. European Journal of ePractice. 2011; 12(1):1-13. http://unpan1.un.org/intradoc/groups/public/documents/UN-PADM/UNPAN046727.pdf

[25] Cara O. Strategic Planning: 5 Benefits of Open Government Data [Internet]. 2016. Available From: https://envisio.com/blog/5-benefits-of-open-government-data/ [Accessed: 2021- 08-25]

[26] Bernd W, Steven B. Open Government: Origin, Development, and Conceptual Perspectives, International Journal of Public Administration. 2015; 38(5):381-396. DOI: 10.1080/01900692.2014.942735

[27] Jędrzej W. Barriers to Using Open Government Data. In: Proceedings of the 3rd International Conference on E-Commerce, E-Business and E-Government (ICEEG 2019); June 2019; Lyon, France. 2019. p.15-20. DOI: 10.1145/3340017.3340022

[28] Publishing Open Government Data. W3C Working Draft 8 [Internet]. 2009. Available From: https://www.w3.org/TR/gov-data/ [Accessed: 2021- 08-30]

[29] Martin L, Anastasija N. Transparency-by-design: what is the role of open data portals? Telematics and Informatics. 2021; 61:101605. DOI: https://doi.org/10.1016/j.tele.2021.101605

[30] Renata M, Martin L. Evaluating the Quality of Open Data Portals on the National Level. Journal of Theoretical and Applied Electronic Commerce Research, Electronic Version. 2017; 12(1):21-41. DOI: 10.4067/S0718-18762017000100003

[31] Eric A, Anthony A. Open Government Data in Africa: A preference elicitation analysis of media practitioners. Government Information Quarterly. 2017; 4(2):244-255. DOI: https://doi.org/10.1016/j.giq.2017.02.005

[32] The future of open data portals [Internet].2020. Available From: https://op.europa.eu/en/publication-detail/-/publication/a1b8aa36-daae-11ea-adf7-01aa75ed71a1/language-en [Accessed: 2021- 08-15]

[33] Open Data Best Practices in Europe. [Internet]. 2020. Available From: https://data.europa.eu/en/highlights/open-data-best-practices-europe [Accessed: 2021- 9 -13].

[34] Open Data Essentials [Internet]. 2019. Available From: http://opendatatoolkit.worldbank.org/en/essentials.html [Accessed: 2021- 08 -15]

[35] Igbal S, Albert M, Stephan G. Utilization of open government data: A systematic literature review of types, conditions, effects and users. Information Policy. 2017; 22(1):1-24. DOI: 10.3233/ip-160012

[36] Deo S, Stuti S. Barriers to Open Government Data (OGD) initiative in Tanzania: Stakeholders' perspectives. Growth and Change. 2019; 50(1):470-485. DOI: 10.1111/grow.12282.

[37] Barriers in working with Open Data and Ways to cope with them [Internet]. 2017. Available From: http://aims.fao.org/news/barriers-working-open-data-and-ways-cope-them [Accessed: 2021- 08-25]

[38] Jorn B, Wendy C, Heleen V. Analytical Report 5: Barriers in working with Open Data. European Union; 2020. DOI: 10.2830/88151

[39] Open Data Challenges [Internet]. 2017. Available From: https://online-learning.tudelft.nl/articles/open-data-challenges/ [Accessed: 2021- 08-30]

[40] Anneke Z, Marin J. The negative effects of open government data - investigating the dark side of open data. In: Proceedings of the 15th Annual International Conference on Digital Government Research; June 2014; Aguascalientes, Mexico, New York: ACM; 2014. p.147-152. DOI: https://doi.org/10.1145/2612733.2612761

[41] Melanie D, Katleen J. Legal and Institutional Challenges for Opening Data across Public Sectors: Towards Common Policy Solutions. Journal of Theoretical and Applied Electronic Commerce Research. 2014. 9(3). DOI: http://dx.doi.org/10.4067/S0718-18762014000300002

[42] Stefaan V, Andrew Y, Andrew Z, Susan A, Ania C, Matt G. The Emergence of a Third Wave of Open Data. How to Accelerate the Re-Use of Data for Public Interest Purposes While Ensuring Data Rights and Community Flourishing. GovLab; 2020. https://opendatapolicylab.org/images/odpl/third-wave-of-opendata.pdf

[43] Understanding the future of open data [Internet]. 2017. Available From: https://www.un.org/pt/desa/understanding-future-open-data [Accessed: 2021- 08-30]

Chapter 3

Framework to Evaluate Level of Good Faith in Implementations of Public Dashboards

Monika M. Wahi and Natasha Dukach

Abstract

To hold governments accountable to open government data (GD) standards, public dashboards need to be evaluated in terms of how well they meet public needs. To assist with that effort, this chapter presents a framework and rubric by which public dashboards can be evaluated for their level of good faith implementation. It starts by reviewing challenges to governments sharing data in good faith despite increasing open government data (OGD) policies and laws being put in place globally. Next, it presents a use-case in which the authors explain how they examined a public dashboard in their local context that appeared to be following OGD, but not in good faith, and developed an alternative implementation that appeared to increase the level of good faith. The framework and rubric proposed were used to successfully compare and contrast the level of good faith of both implementations, as well as another public dashboard described in the scientific literature, and to generate recommendations to increase the level of good faith. In conclusion, the utility of this framework and rubric for evaluating and comparing good faith in public implementations of dashboards was demonstrated, and researchers are encouraged to build upon this research to quantify the level of good faith in public dashboards as a way of increasing oversight of OGD compliance.

Keywords: public reporting of healthcare data, quality of healthcare, cross infection, public health informatics, data visualization

1. Introduction

There has been a global trend for populations to increasingly hold governments accountable to open government data (OGD) standards [1]. Because of this, governments have undertaken open data projects, such as providing public access to government data through publicly-accessible dashboards [2, 3]. However, government actors also may have an incentive to hide or obscure data, so there are barriers to accessing data for public dashboards [1]. This chapter focuses on the specific problem where governments attempt to demonstrate compliance with OGD standards through the presentation of a public dashboard, while at the same time, appearing to hide or obscure the data it is supposed to represent through poor dashboard design.

Our motivation to tackle this topic comes from our own disappointing experience trying to use a public dashboard implemented as part of OGD standards established where we live, in Massachusetts in the United States (US). Currently, in general, no standard guidance or recommendations are in place as a process to follow for the development of OGD public dashboards, and no framework or rubric has been proposed to evaluate them. These challenges are barriers to assessing how well public dashboards meet public need, and holding governments accountable for this. The significance of our contribution is that we propose a framework and rubric on which to base the evaluation of how well these public dashboards meet public need. The implication is that the application of this framework and rubric can be further researched in terms of utility in evaluating public dashboards. From this starting point, globally, we can begin to develop scientific consensus on what attributes in evaluate to a good-faith public dashboard implementation, and what the public should rightfully expect from the implementation of an OGD public dashboard.

2. Guidance for the design of public dashboards

The COVID-19 pandemic brought to attention a longstanding need for well-designed dashboards in public health and medicine [4]. It also brought to light that there are no uniform guiding principles behind developing publicly-facing dashboards intended to serve public interests. As a prime example, a recent review of United States (US) government public dashboards for COVID-19 found that "states engaged in dashboard practices that generally aligned with many of the goals set forth by the Centers for Disease Control and Prevention, Essential Public Health Services" (from abstract) [4]. However, the results of this review do not address whether the public was adequately served by any of these dashboards that were funded with the public's money. Important questions not answered were: Did these dashboards meet the public's information needs? Did they meet the information needs of public health practitioners? Or more importantly – whose information needs were these dashboards supposed to meet, and what were these needs?

2.1 Philosophies behind public dashboard design

At present, there is no overarching philosophy behind public dashboard design, for public health or other topics [2]. Although individual projects will publish use-cases where they discuss their design philosophy [2, 5, 6], there has not been an overall effort by the professional informatics societies or other academic groups to assemble principles behind the design for dashboards intended to serve the public. This may be because such an effort would be daunting, and would require a relatively narrow scope. The scope should be aimed at addressing high-level requirements focused on ensuring that the public's needs are met by whatever dashboard solution is developed, regardless of the topic.

This chapter will attempt to summarize the literature into a framework that provides a general, generic rubric by which to evaluate how well a dashboard design for the public ensures that the public's needs are met through measuring their adherence to high-level requirements. The framework will also put forth a method by which to compare alternative dashboard solutions aimed at meeting similar public needs as to how consistent the solution is with the public's dashboard requirements.

The framework and rubric are intended to evaluate outcomes. Logically, a design process that adequately includes the public that the dashboard is intended to serve will inevitably produce a dashboard solution that meets these outcomes. Hence, there is no need to invest public funding in bloated efforts such as the Rapid Cycle Quality Improvement (RCQI) model, which is promoted by many health departments and organizations, and is extremely paperwork intensive [7, 8]. Part of what causes the RCQI model to be so effort-intensive is that it measures process outcomes. By contrast, the evaluation framework for public dashboards recommended in this chapter is streamlined, and focused on achieving a design solution, not a process solution.

Nevertheless, an optimal design solution will not be achieved without an adequate design process. Therefore, it is important to consider how the public should be involved in the process of designing public dashboards – especially those that are publicly-funded, and therefore have obligations to respond to the public's needs.

2.2 Dashboard design process

As stated previously, there is currently no agreed-upon best-practices design process for dashboards in general, and public dashboards specifically [2]. Each time a dashboard is developed, a different design process is used. But a generic, logical process can be summarized in **Figure 1**.

As shown in **Figure 1**, typically, before the dashboard is designed, some sort of design process is chosen, and this design process is followed to develop an "alpha prototype". The alpha prototype represents a working mock-up that exists for the purposes of getting feedback and working out an initial design. Next, the alpha prototype undergoes a testing process to inform developers as to modifications that are necessary before widespread testing is done. Once those modifications are made, a beta prototype exists, and can be launched for field testing.

As described in **Figure 1**, depending upon the project, there can be different components included in the design process for the alpha prototype. First, there will be iterative design processes as part of designing the alpha prototype, as well as the development of design documentation and the actual creation of the prototype. The details behind each of these components will vary by project. Once the alpha prototype exists, the process to convert it to the beta prototype involves some sort of user testing, and some sort of evaluation for adherence to standards. Granted, an alpha prototype may be released into the field without having undergone the beta prototype process, but that means it has not been user-tested or evaluated for adhering to standards.

This logical process can apply to any dashboard development effort. As one example, researchers aimed to design a dashboard for clinicians [9]. They wrote

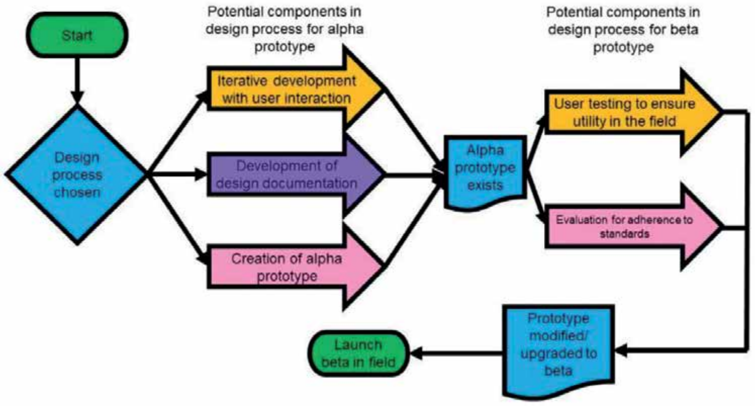

Start
Design process chosen
Potential components in design process for alpha prototype
Iterative development with user interaction
Development of design documentation
Creation of alpha prototype
Alpha prototype exists
Potential components in design process for beta prototype
User testing to ensure utility in the field
Evaluation for adherence to standards
Prototype modified/ upgraded to beta
Launch beta in field

Figure 1.
Generic logical dashboard design process. This design process produces an alpha prototype for initial testing, and a beta prototype for widespread field testing.

requirements and developed an alpha prototype, then worked with clinicians to gain feedback to guide the development of a beta prototype (which would presumably be developed in the future and field-tested) [9]. This article focused mainly on the feedback process to improving the alpha prototype, but the focus of articles can be on any part of the dashboard design process. Another article focused on the development of a beta prototype aimed at both the public and leaders for real-time decision-making related to traffic flow [2]. While the beta prototype was developed and appeared ready for testing, the article did not report any results, so the current final stage of this project was not evident in the article [2].

Although, this logical design process should theoretically involve the intended users of the dashboard, and prototypes should undergo iterative testing, this is not always the case with public dashboards. Because public dashboards often involve government agencies and leaders at some level, whether as data sources or as intended audiences, these forces can have unintended impacts on the dashboard design and quality.

2.3 Governmental data suppression and misrepresentation

As a general trend, consumers are demanding more data transparency, and calls are being made for governments to make data available for public oversight [1]. Likewise, there is an increasing trend toward using dashboards for empowering the public [2, 3]. Not only do dashboards of public data provide a mechanism for public oversight of leaders, but they also reduce *information asymmetry*, which refers to the circumstance in which one party (the government) has more information than another party (the public), thus disempowering them [2, 10, 11].

However, governments are not always keen to share the data for various reasons. It has been argued that government agencies will be more likely to comply with open government data (OGD) practices if they see it as an opportunity to showcase their agency's success [1]. However, if the agency believes the data will cast the agency in a negative light, the agency may be less likely to be inclined toward OGD practices. Ruijer and colleagues recommend that institutional incentives and pressure be created for OGD, because governments have a natural interest in suppressing data they think may be harmful to them in some way if analyzed [1].

However, data suppression is not the only method governments employ to prevent data use and interpretation. One limitation of legal requirements for OGD is that the agency may comply with the requirements in bad faith. During the COVID-19 outbreak in early 2020, a state epidemiologist in Florida said she was fired for refusing to manually falsify data behind a state dashboard [12]. Simply reviewing the limitations of big data can reveal ways to share big data in bad faith in a dashboard, such as visualizing too much data, visualizing incomprehensible or inappropriate data, and not visualizing needed data [13].

For this reason, in addition to holding governments to OGD standards, government efforts need to be evaluated as to whether or not they meet OGD standards *in good faith*. The framework presented here guides as to how to evaluate good vs. bad faith implementations of a public dashboard.

2.4 Dashboard requirements

The evaluation framework presented has six principles on which to judge the level of good or bad faith in a public dashboard: 1) ease of access to the underlying data, 2) the transparency of the underlying data, 3) approach to data classification, 4) utility of comparison functions, 5) utility of navigation functions, and 6) utility of metrics presented. These principles will be described below.

2.4.1 Access to underlying data

A dashboard is essentially a front-end, with data behind it being visualized [14]. Hence, once a dashboard is published, members of the public may want to access the underlying data for various reasons, including oversight of the dashboard. But governments resistant to data-sharing may use the dashboard in bad faith as a firewall between the public and the underlying data to prevent data access [1]. Hence, good faith OGD principles hold that public dashboards should not serve as barriers, but instead serve as facilitators to access the underlying data being visualized in the dashboard.

2.4.2 Transparency of underlying data

Although raw data are used for the dashboard, in the dashboarding process, they undergo many transformations to be properly visually displayed [9, 14]. The processing of the data can develop calculations that are then displayed in the dashboard. Therefore, to be transparent, the dashboard must not only facilitate access to the underlying raw data, but also to the transformations the data underwent in being displayed. A simple way to accomplish this kind of transparency is to use open source tools and publish the code, along with documentation [14]. This allows citizen data scientists an opportunity to review and evaluate the decisions made in the dashboard display.

2.4.3 Data classification

How data are classified in a dashboard can greatly impact the utility of the dashboard. As an example, developers of an emergency department (ED) dashboard that was in use for five years under beta testing found that after the ED experienced an outbreak of Middle East Respiratory Syndrome (MERS), major structural changes were needed to the dashboard [15]. Another paper about developing a visualization of patient histories for clinicians described in detail how each entity being displayed on the dashboard would be classified [16]. Hence, inappropriate classifications or ones deliberately made in bad faith can negatively impact data interpretation to the point that the dashboard could be incomprehensible to its users.

2.4.4 Comparison functions

Dashboards are used to inform decision-making, and therefore, being able to make needed comparisons is an important factor in a dashboard's usability [14, 17]. As an example, the public traffic dashboard described earlier presented visualizations of the ten most congested areas of the city, as well as textual feedback on the two most suitable routes between downtown and outlying areas, to provide optimal comparators to allow the public to make the most-informed route decision [2]. While ultimately, optimal design choices could be debated, it is easy to conceive of how agencies looking to maintain opacity could obscure data interpretation in a dashboard in bad faith by deliberately limiting the ability to make useful comparisons.

2.4.5 Navigation functions

Dashboards are typically at least somewhat interactive, providing the user the ability to navigate through the data display, which responds to actions by the user [14, 18, 19]. When operating in good faith, developers often conduct extensive usability testing to ensure that the dashboard is intuitive to use in terms of navigating through the data display, and that any interactivity is useful [15]. But when implemented in bad faith, a dashboard could be designed to deliberately confuse the user as to how to navigate and interpret the data in the dashboard.

2.4.6 Metrics presented

One of the main purposes of dashboards is to present metrics that represent statistics or visualizations meant to summarize a particular state of the data [14, 15]. For example, in the traffic dashboard, the metrics presented are intended to communicate traffic congestion to the users, while the metrics presented in the clinical dashboard are intended for healthcare workers to use in clinical decision-making [2, 16]. In a good-faith effort, developers may conduct extensive user-testing to ensure that the metrics presented are communicative to dashboard users, as is often done with dashboards developed to serve workers in healthcare [9, 20]. However, in a bad faith effort, the metrics presented could be deliberately confusing to the user, and serve merely to hide ugly truths in the underlying data.

3. Use-case: dashboard for hospital-acquired infection rates at Massachusetts hospitals

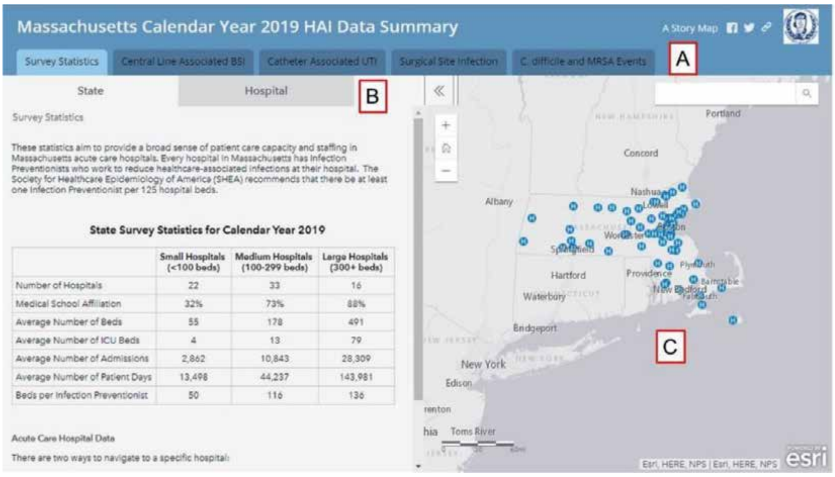
In the US, the Commonwealth of Massachusetts (MA) Department of Public Health (DPH) posts annual healthcare-associated infection (HAI) reports about MA hospitals on their web page [21]. The purpose of the reports is public data transparency by law [21]. Briefly, HAI, such as catheter-acquired urinary tract infection (CAUTI), central line-associated bloodstream infections (CLABSI), and surgical site infections (SSI), are serious issues because they are the fault of the hospital, and can lead to sepsis, which is a systemic infection that can end in death [22, 23]. Since catheterization happens in the intensive care unit (ICU) setting, typically hospitals track CAUTI and CLABSI as part of healthcare quality activities centered around ICUs [14]. By contrast, SSIs are tracked in association with specific operative procedures (e.g., colon surgery) [24].

In the US compared to other countries, rates of HAI are relatively high, likely because they are not required to be tracked at the federal level [14, 25]. Hospitals can opt into the federal voluntary tracking system called the National Health Safety Network (NHSN), but the NHSN does not have a publicly-facing dashboard, and the data are inaccurate, especially in undercounting severe events [14, 24, 26]. As HAI is a serious public health issue, there has been a call for greater data transparency, so the reports posted on the MA DPH web site represent MA's attempt to comply with state-level mandates for OGD.

Although summary reports are available for download on the MA DPH web site, it is not possible to access hospital-level reports directly from the web site. To download hospital-level reports, the user must access a dashboard presented on the web page in a link (**Figure 2**).

As per **Figure 2**, once inside the dashboard, individual PDF-style reports can be found through navigation to the hospital in question, and the reports appear to present formatted output from a database. One way to navigate to the hospital record is to locate it on the map ("C" in **Figure 2**) and click on its icon, causing the panel "B" to display hospital-level metrics and a link to the hospital's PDF report.

Each PDF report has a header displaying attribute data about the hospital (e.g., number of beds), followed by a series of ICU-, procedure-, and infection-level output. This mirrors the structure seen in the dashboard tabs and reports



Survey Statistics

These statistics aim to provide a broad sense of patient care capacity and staffing in Massachusetts acute care hospitals. Every hospital in Massachusetts has Infection Preventionists who work to reduce healthcare-associated infections at their hospital. The Society for Healthcare Epidemiology of America (SHEA) recommends that there be at least one Infection Preventionist per 125 hospital beds.

State Survey Statistics for Calendar Year 2019

	Small Hospitals (<100 beds)	Medium Hospitals (100-299 beds)	Large Hospitals (300+ beds)
Number of Hospitals	22	33	16
Medical School Affiliation	32%	73%	88%
Average Number of Beds	55	178	491
Average Number of ICU Beds	4	13	79
Average Number of Admissions	2,862	10,843	28,309
Average Number of Patient Days	13,498	44,237	143,981
Beds per Infection Preventionist	50	116	136

Acute Care Hospital Data

There are two ways to navigate to a specific hospital:

Figure 2.
MA HAI public dashboard landing page. Note: "A" labels a menu of tabs that can be used for navigation to view metrics on the various hospital-acquired infections (HAIs). In panel labeled "B", tabs can be used to toggle between viewing state-level metrics and hospital-level metrics. Hospitals can be selected for display using a map labeled "C".

(**Figure 2**, "A" and "B"). For the ICUs at each hospital, the report displays a set of tables summarizing CAUTI and CLABSI rates, followed by time-series graphs. For a set of high-risk surgical procedures, SSI rates and graphs for the hospital are displayed. Medication-resistant staphylococcus aureus (MRSA) and *C. Difficile* infections are serious HAIs that can be acquired in any part of the hospital and are diagnosed using laboratory tests [27]. Rates and graphs of MRSA and *C. Difficile* infections are also displayed on the report.

The underlying data come from the NHSN. This is not stated on the dashboard. Instead, there is a summary report and presentation posted alongside the dashboard on the web site, and the analyses in these files are based on NHSN data [21]. It seems that the DPH is using this NHSN data using as a back-end to the dashboard, and the dashboard is an attempt to comply with OGD laws.

Because the authors are aware of the high rates of HAI in the US, and because we both live in MA and we both are women who are cognizant that sexism in US healthcare adds additional layers of risk to women [28], we identified that we were in a state of information asymmetry. Specifically, we had the *information need* to compare MA hospitals to choose the least risky or "lethal" one for elective surgery or childbearing (planned procedures), but we felt this need was not met by this OGD implementation.

In this section, we start by evaluating the existing MA DPH HAI dashboard against our good vs. bad faith framework. Next, we propose an alternative dashboard solution that improves the good vs. bad faith features of the implementation.

3.1 Considering existing dashboard: design process and requirements

Figure 3 provides a logical entity-relationship diagram (ERD) for the data behind the dashboard.

As described earlier, the landing page (**Figure 2**) provides a map by which users can select a hospital, causing the metrics for the hospital to appear in a panel. The user chooses which measurement to view (e.g., CAUTI) through navigation using the tabs. This suggests the dashboard is aimed at individuals with a working knowledge of MA geography who are intending on comparing and selecting hospitals

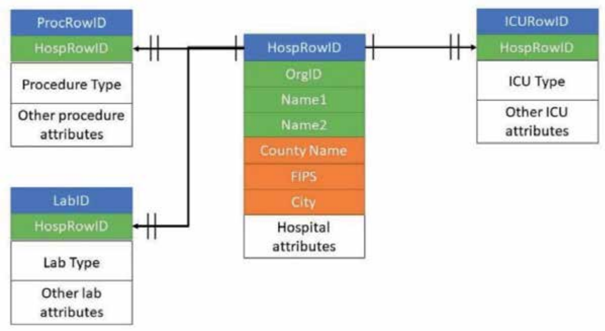
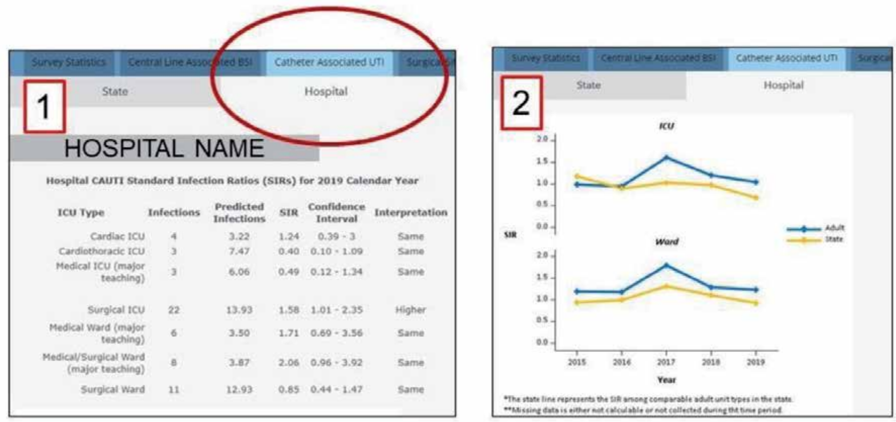


Figure 3.
Logical entity-relationship diagram for data behind dashboard. Note: The schema presented assumes four entities: The hospital entity (primary key [PK]: HospRowID), each intensive care unit (ICU) attached to a hospital which contains the frequency of infection and catheter days attributes to allow rate calculation (PK: ICURowID), each procedure type attached to a hospital (to support the analysis of surgical site infection [SSI], with PK: ProcRowID), and each other infection type at the hospital not tracked with ICUs (PK: LabID).

least likely to cause HAI for an elective procedure (e.g., childbearing), or to establish as their top choice of the local hospital should they ever need to be admitted. This interface makes it difficult to compare HAI at different hospitals, because metrics from more than one hospital cannot be viewed at the same time. Further, metrics about different HAIs at the same hospital are on different panels, so within-hospital comparisons cannot be facilitated. There appears to be no overall metric to use by which to compare hospitals in terms of their HAI rates.

Figure 4 shows an example of the metrics reported by each hospital on the dashboard reporting panel ("B" in **Figure 2**). The figure also shows one of the two tables and one of the two figures displayed on the CAUTI tab for the selected hospital. In all, two tables and two figures are displayed in portrait style in panel "B" (**Figure 2**), and **Figure 4** shows the top table and figure displayed. In the table displayed (labeled "1" in **Figure 4**), the metrics presented are the number of infections, predicted



Hospital CAUTI Standard Infection Ratios (SIRs) for 2019 Calendar Year

ICU Type	Infections	Predicted Infections	SIR	Confidence Interval	Interpretation
Cardiac ICU	4	3.22	1.24	0.39 - 3	Same
Cardiothoracic ICU	3	7.47	0.40	0.10 - 1.09	Same
Medical ICU (major teaching)	3	6.06	0.49	0.12 - 1.34	Same
Surgical ICU	22	13.93	1.58	1.01 - 2.35	Higher
Medical Ward (major teaching)	6	3.50	1.71	0.69 - 3.56	Same
Medical/Surgical Ward (major teaching)	8	3.87	2.06	0.96 - 3.92	Same
Surgical Ward	11	12.93	0.85	0.44 - 1.47	Same

*The state line represents the SIR among comparable adult unit types in the state.
**Missing data is either not calculable or not collected during the time period.

Figure 4.
*Dashboard metric display for each hospital. Note: To view hospital-acquired infection (HAI) rates at hospitals, a hospital is selected (**Figure** 2, panel "C"), then the user selects the tab for the HAI of interest. In **Figure 4**, a hospital has been identified, and a tab for catheter-associated urinary tract infection (CAUTI) has been selected (see circle). Two tables and two figures are presented in portrait format on the reporting panel for each hospital (**Figure** 2, panel "B"). **Figure 4** shows the first table and figure presented ("1" and "2"); the table reflects stratified metrics for CAUTI at each ICU at the hospital, and the graph reflects a time series of these metrics stratified by hospital vs. state levels, and intensive care unit (ICU) vs. ward ("ward" is not defined in the dashboard). The metrics provided in "1" are a number of infections, predicted infections, standard infection ratios (SIRs), a confidence interval for the SIR, and an interpretation of the level. In "B", the SIR is graphed.*

infections, standard infection ratios (SIRS), a confidence interval for the SIR, and an interpretation of the level. The figure (labeled "2" in **Figure 4**) displays a time-series graph of SIRs for the past five years. In the other table on the panel (not shown in **Figure 4**), ICU-level metrics are provided about catheter-days, predicted catheter-days, Standard Utilization Ratios (SURs) and their confidence interval, and an interpretation, and an analogous time-series graph of five years of SURs is presented (also not shown in **Figure 4**).

SIRs and SURs are not metrics used typically by the public to understand rates of HAI in hospitals. Risk communication about rates to the public is typically done in the format of *n per 10,000* or *100,000*, depending upon the magnitude of the rate [29]. Further, stratifying rates by ICU is confusing, as prospective patients may not know what ICU in which they will be placed. Because the hospital environment confers the strongest risk factors for HAI (e.g., worker burnout), HAI rates will be intra-correlated within each hospital [30]. Therefore, it is confusing to present all these rates and stratify them by ICU. **Figure 4** only displays 50% of the information available about CAUTI at one hospital. With each tab displaying similar metrics about SSI and other infections, the experimental unit being used is so small, it obfuscates any summary statistics or comparisons. Also, it is unclear how the "predicted" metrics presented were calculated.

Ultimately, the design process and requirements behind this dashboard are not known. There is no documentation as to how this dashboard was designed, and what it is supposed to do. It appears to be an alpha prototype that was launched without a stated *a priori* design process, and without any user testing or formal evaluation.

3.2 Alternative design: design process and requirements

We chose to redesign the dashboard into a new alpha prototype that met requirements that we, as members of the public, delineated. Consistent with the good faith principles proposed, our requirements included the following: 1) the dataset we use should be easily downloadable by anyone using the dashboard, 2) the documentation of how the dashboard was developed should be easy to access, 3) hospitals should present summary metrics rather than stratified ones, 4) different HAI metrics for the same hospital should be presented together, and 5) there needs to be a way to easily compare hospitals and choose the least risky hospital. To do this, we first obtained the data underlying the original dashboard. Next, we analyzed it to determine better metrics to present. We also selected open-source software to use to redeploy an alpha prototype of a new dashboard. Finally, we conducted informal user testing on this alpha prototype.

3.2.1 Obtaining the data from the original dashboard

Scraping was done in open-source *RStudio* and predominantly used packages *pdftools*, and *pdftables* [31, 32]. All the PDF reports from each hospital were downloaded and placed in one directory. As a first step, a loop was used with the *pdftools* package which crawled through each report extracting the data into memory. This was done in conjunction with the *pdftables* package, which is essentially an online application that applies structure to the unstructured tabular data placed in memory from *pdftools*. To use this online application, an application programming interface (API) key is issued from the *PDF Tables* web site, and is used in the *RStudio* programming to pass the data to the online application [33]. The code resulted in the data being processed into a series of unstructured *.xlsx files and downloaded locally. Then, in a final data cleaning step, data were transformed into the tables in the format specified in **Figure 3**.

3.2.2 Determining metrics to present

We chose to focus our inquiry on the data from the hospital and ICU tables, as CAUTI and CLABSI are by far the most prevalent and deadly HAIs [23]. Therefore, we scoped our alpha prototype to only display data from the ICU and the hospital tables (although we make all the data we scraped available in the downloadable dataset). This limited us to basing the dashboard on hospital- and ICU-level metrics only.

Next, we intended to present CAUTI and CLABSI frequencies and rates, whereby the numerator would be the number of infections, and the denominator would be the "number of patients catheterized". We felt that the dashboard's use of catheter-days as the rate denominator was confusing to the public, and appeared to attenuate the prevalence of patients having experienced a CAUTI or CLABSI. Although, "number of patients catheterized" was not available in the data, "annual admissions" was. Since the proportion of patients admitted annually who are catheterized probably does not vary much from hospital to hospital, we chose to use the number of admissions as the denominator and a proxy measurement.

Third, we needed to develop a way of sorting hospitals as to their likelihood of causing an HAI to allow easy comparisons by public users, so we decided to develop an equation to predict the likelihood of an HAI at the hospital. We did this by developing a linear regression model with hospital-level attributes as independent variables (IVs), and CAUTI rate in 2019 as the dependent variable (DV). We chose CAUTI over CLABSI after observing the two rates were highly correlated, and CAUTI was more prevalent.

Table 1 describes the candidate IVs for the linear regression model. The table also includes the source of external data that were added to the hospital data. We studied our IVs, and found serious collinearity among several variables, so we used principal component analysis (PCA) to help us make informed choices about parsimony [37]. The data predominantly loaded on three factors (not shown). The first factor included all the size and utilization variables for the hospital; these were summed into a Factor 1 score. The second-factor loadings included the proportion of those aged 65 and older and the non-urban flag (**Table 1**), so those were summed as Factor 2. Proportion non-White was strongly inversely correlated with Factor 2, so it was kept for the model, and county population did not load, so it was removed from the analysis. Factor 3 loadings included teaching status, for-profit status, and Medicare Performance Score (MPS). Rather than create a score, we simply chose to include the variable from Factor 3 that led to the best model fit to represent the factor, which was MPS. Then we finalized our linear regression model, and developed a predicted CAUTI rate ($\hat{y}$) using our model that included the following IVs: MPS, Factor 1 score, Factor 2 score, and proportion of non-White residents in hospital county.

Next, we used the regression equation to calculate $\hat{y}$ as a "lethality score" for each hospital. Of the 71 hospitals in the dataset, 21 were missing MPS and 8 were missing other data in the model. Therefore, only the 42 with complete data (IVs and DVs) were used to develop the regression model. As a result, the lethality score was nonsensical for some hospitals; where the residual was large, the lethality score was replaced with the 2019 CAUTI rate. If CAUTI data were missing, it was assumed that the hospital had no CAUTI cases, and therefore was scored as 0.

Once the lethality score was calculated, we chose to sort the hospitals by score, and divide them into four categories: least probable (color-coded green), somewhat probable (color-coded yellow), more probable (color-coded red), and most probable (color-coded dark gray). Due to missing CAUTI information and many hospitals having zero CAUTI cases, our data were severely skewed left, so making quartiles of the lethality score to divide the hospitals into four categories was not meaningful. To compensate, we sorted the data by lethality score and placed the

Variable	Source	Role	Source
Teaching hospital status	Original dashboard	Exposure	Scraped data from dashboard
Hospital profit status	Original dashboard	Confounder	Scraped data from dashboard
Measurements of hospital size (number of beds)	Original dashboard	Confounder	Scraped data from dashboard
Measurements of hospital utilization (number of admissions, number of patient days)	Original dashboard	Confounder	Scraped data from dashboard
Medicare Performance Score	Medicare performance score dataset [34]	Confounder	Medicare performance score dataset
Non-urban county	Rural health information hub [35], United States census [36]	Confounder	Scraped data from dashboard to determine county, then application of hospital rurality flag developed from public data based on county
Hospital county population size	United States census [36]	Confounder	Census measurements
Hospital county population proportion below the poverty level	United States Census [36]	Confounder	Census measurements
Hospital county population proportion below the poverty level	United States Census [36]	Confounder	Census measurements
Hospital county population proportion of non-White residents	United States Census [36]	Confounder	Census measurements
Numerators for rates – infection frequencies	Original dashboard	Create outcome	Scraped data from dashboard
Infection rates	Original dashboard	Outcome	Calculated as the numerator for rates divided by admissions

Table 1.
Conceptual model specification.

first 23 hospitals (32%), which included all the hospitals with zero cases, in the least probable category. We placed the next 16 (23%) in somewhat probable, the next 16 (23%) in more probable, and the final 16 (23%) are most probable. We chose to use this classification in data display on the dashboard to allow for easy comparison between hospitals of risk of a patient contracting HAI.

3.2.3 Choice of software and display

R is an open-source analytics software that allows for user-developed "packages" to be added *a la carte* to the main application [38]. RStudio was developed to be an integrated development environment (IDE) for R that allows for advanced visualization capabilities that support dashboard development [39]. In RStudio, web applications like dashboards can be placed on a host server and deployed on the internet such that as users interact with the web front-end, it can query and display data from the back-end hosted on the server.

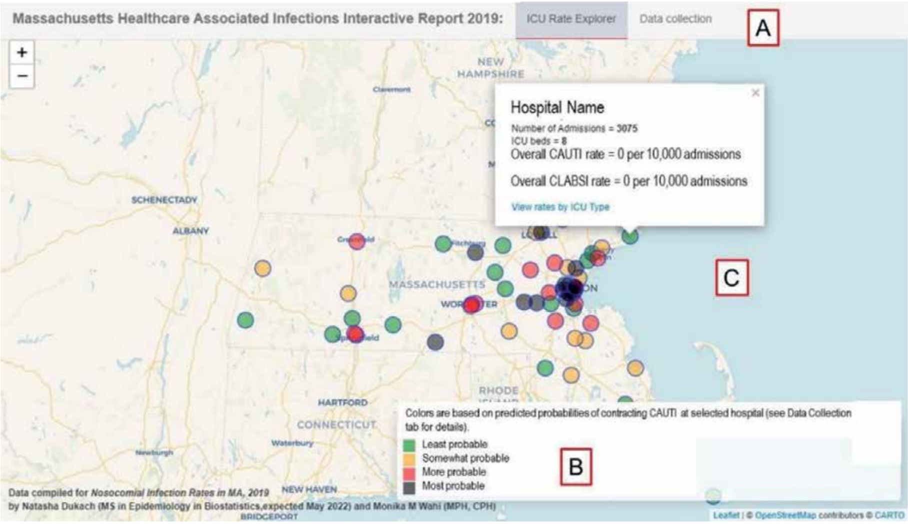
The package *Rshiny* [40] was developed to support dashboarding in RStudio, and can work with other visualization packages depending upon the design goals of the dashboard. In our newly designed dashboard, the package *leaflet* was used for a base map on which we placed the hospital icons (like the original dashboard), and add-ons were made to display other items. These add-ons were adapted from other published codes [41]. JavaScript with wrapper DT (data table) was used to display stratified ICU rates, and CSS was used for formatting.

The dashboard we developed was deployed on a server (https://natasha-dukach.shinyapps.io/healthcare-associated-infections-reports/) and code for the dashboard was published (https://github.com/NatashaDukach/HAI-MA-2019). When accessing the link to the dashboard, the user initially sees a map with icons (in the form of dots) on it indicating hospitals. The icons are color-coded according to the lethality score described previously. Clicking on an icon will expand a bubble reporting information about the hospital (**Figure 5**).

As shown in **Figure 5**, like the original dashboard, this one has a map for navigation. Unlike the original, it only has two tabs: "ICU Rate Explorer" (the one shown in **Figure 5**), and "Data Collection", which provides documentation and links to original data and code (see "A" in **Figure 5**). The hospital icons are placed on the map and coded according to our color scheme (see legend in **Figure 5** by "B"). This allows for easy comparison between hospitals. When clicking on an icon for a hospital, a bubble appears that contains the following hospital metrics: Number of admissions, number of ICU beds, overall CAUTI rate, and overall CLABSI rate. There is also a link on the bubble where the user can click to open a new box that provides CAUTI and CLABSI rates stratified by ICU. Future development plans include adding other overall rates (e.g., for SSI), and adding in data from previous years to allow for the evaluation of trends.

3.2.4 User testing

Informally, members of two potential user bases were queried as to their reactions to the differences between the two dashboards: members of the academic public health space, and members of the MA public. When the dashboard redesign



Hospital Name
Number of Admissions = 3075
ICU beds = 8
Overall CAUTI rate = 0 per 10,000 admissions
Overall CLABSI rate = 0 per 10,000 admissions

Colors are based on predicted probabilities of contracting CAUTI at selected hospital (see Data Collection tab for details).

🟩 Least probable
🟨 Somewhat probable
🟥 More probable
⬛ Most probable

Data compiled for *Nosocomial Infection Rates in MA, 2019* by Natasha Dukach (MS in Epidemiology in Biostatistics,expected May 2022) and Monika M Wahi (MPH, CPH)

Figure 5.
Alternative dashboard solution. Note: In our new version, two tabs are created (see "A"). The figure shows the first tab titled "ICU Rate Explorer". The second tab, titled "data collection", has information about the design of the dashboard and links to the original code. Each of the hospitals is indicated on the map by a color-coded icon that can be clicked on to display a bubble. The legend by "B" displays our color-coding scheme. When clicking on a hospital icon, hospital-level metrics are shown in a bubble, and there is a link that leads to the display of intensive care unit (ICU)-level metrics (see "C").

was pitched as a project to public health academics, it was dismissed as an unimportant escapade for various reasons. Some reasons cited were lack of agreement on terminology and patient safety priorities, the challenges with undercount of HAIs in NHSN data, and differential reporting accuracy in teaching vs. non-teaching hospitals. Academics also acknowledged that the system for tracking, addressing, and preventing HAIs is hopelessly broken in the US, and therefore it seems a waste of time to prop up such a system when it produces inaccurate data.

A few members of the MA public who are familiar with technology also provided informal feedback about the utility of the dashboard from a patient standpoint. They reported that the alternative solution was more intuitive than the original, and did a better job of representing the highly limited data from the NHSN.

These differences in reactions underscore the challenge of OGD and ensuring that public dashboards are developed and deployed in good faith. Those from the public health field expressed that since the system is broken and the data are inaccurate, they should be dismissed, while those in the public felt that since the data existed, it should be accessible, even if it was not completely accurate. It not only highlights the differing perspectives of those on either side of information asymmetry, it glaringly illustrates how those who are being held accountable by the usage of the data see dashboarding differently than those who are using the data to do oversight and accountability.

4. Application

We wanted to compare the original HAI dashboard with the one we developed based on the good faith principles described earlier. We started by creating the framework presented in **Table 2**, which guides as to the good faith and bad faith characteristics of public dashboards.

Using this framework, we applied a rating system. We chose zero to represent "neither bad faith nor good faith", −5 to represent "mostly bad faith" and + 5 to represent "mostly good faith". Then, based on our experience and available information, we rated the original MA HAI dashboard and our alternative dashboard solution to compare the ratings. To experiment with applying our framework to another public dashboard, we used the information published in the article described earlier to rate the traffic dashboard which was in Rio de Janeiro [2]. Our ratings appear in **Table 3**.

As shown in **Table 3**, using **Table 2** as a rubric and our rating scale, we were able to rate each dashboard and assign a score. We were also able to define in the comments in the table the evidence on which we based our score. **Table 3** demonstrates that this framework can be used to compare two different alternatives of a public dashboard displaying the same data, as well as two completely different public dashboards. The total scores show that while our redesigned prototype of the HAI dashboard had a similar level of good faith implementation compared to the Rio traffic dashboard (scores 26 vs. 23, respectively), the original HAI dashboard had a very low level of good faith implementation compared to the other two (score − 20).

5. Discussion

As is consistent with the global trend, the state of MA implemented an OGD requirement to share HAI data with the public through posting a public dashboard on a web page. However, as residents of MA, the authors found that this dashboard

Dashboard function/ characteristic	Good faith	Bad faith
Access to underlying data	• Easy to download analytic dataset on which dashboard rides in *.csv format (e.g., through report export function) • Easy to access or download component datasets that went into the analytic dataset on which the dashboard rides	• Lack of downloading functions, or downloading functions that provide only reports in non-tabular and non-data format • Lack of transparency on how the analytic dataset was developed • Lack of transparency about source datasets
Transparency of underlying data	• For each native variable used in the dashboard, its source dataset is specified, and a link is given if available. • For each calculated variable used in the dashboard, clear documentation is available. • It is clear which data reflect real measurements, and which reflect simulations, imputations, or predictions	• Source datasets may be specified, but little information about the use of their variables in the dashboard is provided • It is not made clear which dashboard variables are calculated, and how they are calculated is also not made clear • It is not clear which data reflect real measurements, and which data have been simulated, imputed, or otherwise fabricated
Data classification	Data are classified in ways that are intuitive to the consumer, and results are presented according to those classifications	• Data are classified in ways that either make the development work easier for the analyst, or serve to mask negative indicators • Data are not grouped into classifications consumers use, making it impossible to obtain summary statistics for these classification levels
Comparison functions	Dashboard allows for comparisons that provide useful consumer decision-support	Dashboard prevents comparisons that would provide consumer decision-support, or makes them very difficult to make using the dashboard functions
Navigation functions	• Navigation functions reflect how users conceive of accessing the entities in the dashboard • Specifically, map navigation reflects how users conceive of their geographic locale when searching for information • This allows consumers to intuitively ingest and assimilate information as they interact with the dashboard	• Navigation functions reflect how public officials want consumers to navigate the entities in the dashboard • This forces consumers to think differently about the topic, and disrupts their ability to ingest and assimilate information • Map functions force the consumer to conceive of their geographic locale in an unintuitive way, making map navigation confusing
Metrics presented	• Metrics are intuitive to consumers • Metrics are presented in such a way that they are intuitive to ingest and assimilate • Metrics that are presented to make comparisons between entities intuitive to support decision-making	• Metrics reflect jargon, and are unintuitive to consumers • Metrics require consumers to read documentation to understand • Metrics are presented in such a way as to be confusing, making them impossible to be used for decision-support

Table 2.
Proposed framework for evaluation of good faith and bad faith public dashboards.

Dashboard function/ characteristic	**MA Hosp**	**Alt. MA Hosp**	**Comment: MA Hosp vs. Alt. MA Hosp**	**Rio**	**Comment: MA Hosp vs. Rio**
ACCESS TO UNDERLYING DATA	−5	5	The original solution had no access to underlying data (except by way of PDF-style reports). Alternative solution posts data publicly for download.	5	Rio dashboard uses open data from City Hall with user-generated content collected through Waze
Transparency of Underlying Data	−5	3	It was difficult to identify the source of the data in the original solution. The alternative solution uses the same data, which is from NHSN. Because NHSN itself is somewhat opaque, the final solution lacks transparency.	0	Unclear from the article, but it appears that it is possible to audit Rio dashboard design if a member of a certain role (e.g., public data scientist). Not all tools used were open source.
Data Classification	0	3	In informal user testing, public users found the data classifications much more intuitive in the alternative compared to the original solution. However, formal user testing was not conducted.	5	Much effort was made to classify data in Rio dashboard to make it useful for the public to make route decisions.
Comparison Functions	−5	5	In informal user testing, users found the comparison function in the alternative solution useful for decision-making, and could not find a comparison function in the original solution.	5	Rio dashboard was designed to allow the public to make comparisons about potential traffic routes.
Navigation Functions	0	5	In informal user testing, users reported being able to easily navigate the data and dashboard in the alternative solution, but having extreme difficulty in navigating the original solution.	5	Rio dashboard for the public had a very simple, intuitive interface with images and only a few metrics critical to decision-making. This made it possible to easily navigate the dashboard display.
Metrics Presented	−5	5	In informal testing, users indicated that they did not understand the metrics presented on the original solution but found the color-coding of the alternative solution intuitive for decision-making.	3	The few metrics presented on the Rio dashboard were geared specifically to helping the public make route decisions based on traffic metrics. However, no formal user testing is presented.
Total	−20	26		23	

Note: MA Hosp = original hospital-acquired infection (HAI) dashboard from the Commonwealth of Massachusetts (MA), MA Hosp Alt. = alternative solution, Rio = Rio traffic dashboard [2], and NHSN = National Healthcare Safety Network.

Table 3.
Application of rating system.

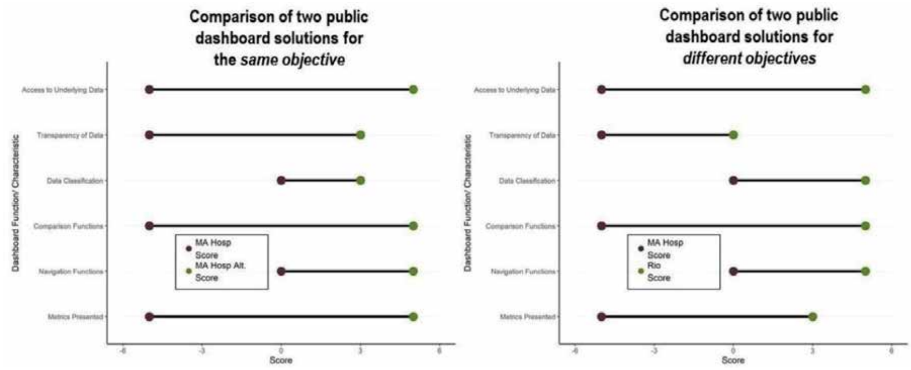


Figure 6.
Example of visualization of framework score comparison. Note: MA Hosp = original hospital-acquired infection (HAI) dashboard from the Commonwealth of Massachusetts (MA), MA Hosp Alt. = alternative solution, and Rio = Rio traffic dashboard [2].

did not serve our information needs, and essentially obscured the data it was supposed to present. To address this challenge, we not only redesigned the dashboard into a new prototype, but we also tested our proposed framework for evaluating the level of good faith in public dashboards by applying it. Using our proposed framework and rubric, we evaluated the original HAI dashboard, our redesigned prototype, and a public dashboard on another topic presented in the scientific literature on the level of good faith implementation. Through this exercise, we demonstrated that the proposed framework is reasonable to use when evaluating the level of good faith in a public dashboard.

The next step in the pursuit of holding governments accountable for meeting OGD standards in public dashboards is to improve upon this framework and rubric through rigorous research. As part of this research, entire groups of individuals could be asked to score dashboards on each of these characteristics, and the results could easily be summarized to allow an evidence-based comparison between dashboards. Results can be easily visualized in a dumbbell plot (using packages *ggplot2*, *ggalt*, and *tidyverse* [42–44]), which we have done with our individual scores, but could be done with summary scores (**Figure 6**).

As visualized in **Figure 5** and summed in **Table 3**, our scoring system suggested that the alternative HAI dashboard we developed was done in a level of good faith (score = 26) similar to that of the Rio traffic dashboard (score = 23), and that the original HAI dashboard appears to not have been done in good faith (score = −20), and may serve as a governmental attempt to hide or obscure uncomfortable data. This exercise shows that the framework and rubric developed can be used to compare the level of good faith in public dashboards, and to provide evidence-based recommendations on how governments can improve them so they meet both the spirit and the letter of OGD requirements.

6. Conclusion

In conclusion, in this chapter, we describe the challenge of holding governments accountable for developing public dashboards to meet OGD requirements in a way that also serves the public's information needs. To address this challenge, we propose a framework of six principles of good faith OGD by which public dashboards could be evaluated to ensure data shared by the government under OGD policies and laws are done so in good faith. We also demonstrate applying this framework to the use-case of a public dashboard intended for residents of MA in the US to use to compare and select hospitals based on their HAI rates. As a demonstration, we present our redesign of the dashboard, then use a rubric based on the framework to score and compare the original dashboard and our alternative in terms of levels of good faith OGD. We also demonstrate using the rubric on a published use-case in the literature. As our framework and rubric provide a reasonable starting point as a method for evaluating and comparing the level of good faith in public dashboards, we strongly recommend that future research into this topic consider our framework and rubric, and build upon it through gathering evidence in the field.

Author details

Monika M. Wahi[1]* and Natasha Dukach[2]

1 DethWench Professional Services (DPS), Boston, MA, USA

2 Boston University, Boston, MA, USA

*Address all correspondence to: dethwench@gmail.com

IntechOpen

References

[1] Ruijer E, Détienne F, Baker M, et al. The politics of open government data: Understanding organizational responses to pressure for more transparency. The American Review of Public Administration. 2020;**50**:260-274

[2] Matheus R, Janssen M, Maheshwari D. Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. Government Information Quarterly. 2020;**37**:101284

[3] Janssen M, van den Hoven J. Big and open linked data (BOLD) in government: A challenge to transparency and privacy? Government Information Quarterly. 2015;**32**:363-368

[4] Fareed N, Swoboda CM, Chen S, et al. U.S. COVID-19 state government public dashboards: An expert review. Applied Clinical Informatics. 2021; **12**:208-221

[5] Vila RA, Estevez E, Fillottrani PR. The design and use of dashboards for driving decision-making in the public sector. In: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance. New York: Association for Computing Machinery. pp. 382-388

[6] Joshi A, Amadi C, Katz B, et al. A human-centered platform for hiv infection reduction in New York: Development and usage analysis of the ending the epidemic (ETE) dashboard. JMIR Public Health Surveillance. 2017;**3**. DOI: 10.2196/publichealth.8312

[7] Guerrero LR, Richter Lagha R, Shim A, et al. Geriatric workforce development for the underserved: Using RCQI methodology to evaluate the training of IHSS caregivers. Journal of Applied Gerontology. 2020;**39**:770-777

[8] Mullins CM, Hall KC, Diffenderfer SK, et al. Development and implementation of APRN competency validation tools in four nurse-led clinics in rural east Tennessee. Journal of Doctoral Nursing Practice. 2019;**12**:189-195

[9] Giordanengo A, Årsand E, Woldaregay AZ, et al. Design and prestudy assessment of a dashboard for presenting self-collected health data of patients with diabetes to clinicians: iterative approach and qualitative case study. JMIR Diabetes. 2019;**4**. DOI: 10.2196/14002

[10] Jensen MC, Meckling WH. Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics. 1976;**3**:305-360

[11] Bugaric B. Openness and transparency in public administration: Challenges for public law. Wisconsin International Law Journal. 2004;**22**:483

[12] Martin R. Florida Scientist Says She Was Fired For Not Manipulating COVID-19 Data. 2020. Available from: https://www.npr.org/2020/06/29/884551391/florida-scientist-says-she-was-fired-for-not-manipulating-covid-19-data [Accessed: September 13, 2021]

[13] Lee CH, Yoon H-J. Medical big data: Promise and challenges. Kidney Research and Clinical Practice. 2017;**36**:3-11

[14] Wahi MM, Dukach N. Visualizing infection surveillance data for policymaking using open source dashboarding. Applied Clinical Informatics. 2019;**10**:534-542

[15] Yoo J, Jung KY, Kim T, et al. A real-time autonomous dashboard for the emergency department: 5-year case study. JMIR mHealth and uHealth. 2018;**6**:e10666

[16] Bernard J, Sessler D, Kohlhammer J, et al. Using dashboard networks to visualize multiple patient histories: A design study on post-operative prostate cancer. IEEE Transactions on Visualization and Computer Graphics. 2019;**25**:1615-1628

[17] Sedrakyan G, Mannens E, Verbert K. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. Journal of Computer Languages. 2019;**50**:19-38

[18] Thoma B, Bandi V, Carey R, et al. Developing a dashboard to meet competence committee needs: A design-based research project. Canadian Medical Education Journal. 2020;**11**:e16-e34

[19] Ahn J, Campos F, Hays M, et al. Designing in context: Reaching beyond usability in learning analytics dashboard design. Journal of Learning Analytics. 2019;**6**:70-85

[20] Mlaver E, Schnipper JL, Boxer RB, et al. User-centered collaborative design and development of an inpatient safety dashboard. Joint Commission Journal on Quality and Patient Safety. 2017;**43**:676-685

[21] Mass.gov. Healthcare Associated Infections Reports. Available from: https://www.mass.gov/lists/healthcare-associated-infections-reports [Accessed: March 18, 2021]

[22] Hassan KA, Fatima BK, Riffat M. Nosocomial infections: Epidemiology, prevention, control and surveillance. Asian Pacific Journal of Tropical Biomedicine. 2017;**7**:478-482

[23] Rhee C, Jones TM, Hamad Y, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. JAMA Network Open. 2019;**2**:e187571

[24] Bordeianou L, Cauley CE, Antonelli D, et al. Truth in reporting: how data capture methods obfuscate actual surgical site infection rates within a healthcare network system. Diseases of the Colon and Rectum. 2017; **60**:96-106

[25] Isikgoz Tasbakan M, Durusoy R, Pullukcu H, et al. Hospital-acquired urinary tract infection point prevalence in Turkey: Differences in risk factors among patient groups. Annals of Clinical Microbiology and Antimicrobials. 2013;**12**:31

[26] Neelakanta A, Sharma S, Kesani VP, et al. Impact of changes in the NHSN catheter-associated urinary tract infection (CAUTI) surveillance criteria on the frequency and epidemiology of CAUTI in intensive care units (ICUs). Infection Control and Hospital Epidemiology. 2015;**36**:346-349

[27] Conlon-Bingham GM, Aldeyab M, Scott M, et al. Effects of antibiotic cycling policy on incidence of healthcare-associated MRSA and clostridioides difficile infection in secondary healthcare settings. Emerging Infectious Diseases. 2019;**25**:52-62

[28] Homan P. Structural sexism and health in the United States: A new perspective on health inequality and the gender system. American Sociological Review. 2019;**84**:486-516

[29] Oehmke JF, Oehmke TB, Singh LN, et al. Dynamic panel estimate–based health surveillance of SARS-CoV-2 infection rates to inform public health policy: Model development and validation. Journal of Medical Internet Research. 2020;**22**:e20924

[30] de Garcia CL, de Abreu LC, JLS R, et al. Influence of burnout on patient safety: Systematic review and meta-analysis. Medicina. 2019;**55**:553

[31] Persson E. pdftables: Programmatic Conversion of PDF Tables. 2016. Available from: https://CRAN.R-project.org/package=pdftables [Accessed: March 20, 2021]

[32] Ooms J. pdftools: Text Extraction, Rendering and Converting of PDF Documents. 2020. Available from: https://cran.r-project.org/web/packages/pdftools/index.html [Accessed: March 20, 2021]

[33] PDFTables. Available from: https://pdftables.com/ [Accessed: March 20, 2021]

[34] Medicare.gov. Compare Care Near You. Available from: https://www.medicare.gov/care-compare/?providerType=Hospital&redirect=true [Accessed: March 18, 2021]

[35] Rural Information Hub. Rural health for Massachusetts. 2020. Available from: https://www.ruralhealthinfo.org/states/massachusetts [Accessed: March 20, 2021]

[36] US Census Bureau. QuickFacts. The United States Census Bureau. Available from: https://www.census.gov/programs-surveys/sis/resources/data-tools/quickfacts.html [Accessed: March 20, 2021]

[37] Frost J. Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models. Statistics By Jim Publishing; 2020

[38] Wahi M, Seebach P. Analyzing Health Data in R for SAS Users. London: CRC Press, Statistics by Jim Publishing; 2017. Available from: https://www.amazon.com/Regression-Analysis-Intuitive-Interpreting-Linear/dp/1735431184

[39] RStudio. Available from: https://www.rstudio.com/ [Accessed: March 22, 2019]

[40] Chang W, Cheng J, Allaire JJ, et al. shiny: Web Application Framework for R. 2018. Available from: https://CRAN.R-project.org/package=shiny [Accessed: March 12, 2019]

[41] Cheng J, Karambelkar B, Xie Y, et al. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. 2021. Available from: https://CRAN.R-project.org/package=leaflet [Accessed: September 15, 2021]

[42] Wickham H, Chang W, Henry L, et al. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. 2018. Available from: https://CRAN.R-project.org/package=ggplot2 [Accessed: April 3, 2019]

[43] Rudis B, Bolker B, Marwick B, et al. ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'. 2017. Available from: https://CRAN.R-project.org/package=ggalt [Accessed: September 15, 2021]

[44] Wickham H. tidyverse: Easily Install and Load the 'Tidyverse'. 2021. Available from: https://CRAN.R-project.org/package=tidyverse [Accessed: September 15, 2021]

Section 2

Case Studies of Open Data

Chapter 4

Intrusion Detection Based on Big Data Fuzzy Analytics

Farah Jemili and Hajer Bouras

Abstract

In today's world, Intrusion Detection System (IDS) is one of the significant tools used to the improvement of network security, by detecting attacks or abnormal data accesses. Most of existing IDS have many disadvantages such as high false alarm rates and low detection rates. For the IDS, dealing with distributed and massive data constitutes a challenge. Besides, dealing with imprecise data is another challenge. This paper proposes an Intrusion Detection System based on big data fuzzy analytics; Fuzzy C-Means (FCM) method is used to cluster and classify the pre-processed training dataset. The CTU-13 and the UNSW-NB15 are used as distributed and massive datasets to prove the feasibility of the method. The proposed system shows high performance in terms of accuracy, precision, detection rates, and false alarms.

Keywords: Intrusion detection, machine learning, Apache Spark, Big Data, CTU-13, UNSW-NB15, Feature selection, FCM clustering

1. Introduction

Recently, in computer networks the numbers of intrusions have grown extensively, and many new pirating tools and intrusive methods have appeared. To save the security of computer systems, several solutions have been identified like intrusion detection systems (IDS) which it is the mean solution to deal with suspicious activities in a network [1].

Using IDS tools, the presence of imperfect information greatly influences the response data under non-suitable as a medium for decision-making. Uncertainty is presented as imperfect data, the variability of the data that resides in the random nature of the information due to the heterogeneity of data sources, vagueness and incompleteness of data due to the lack of useful data [2]. Thus, fuzzy clustering as a robust artificial intelligent method has been successfully employed to reduce the amount of false alarm generated by the detection process and separate the overlap between normal and abnormal behavior in computer networks [3].

Hence, we use two intrusion detection datasets CTU-13 and UNSW-NB15 which contain varieties of intrusions, that we combine into one homogenous dataset and then we apply our ML model based on the Fuzzy C-Mean (FCM) clustering algorithm. We choose Microsoft Azure Blob Storage to load our datasets on.

This paper addresses the problem of generating application clusters from the network intrusion detection datasets. The Fuzzy C-Mean (FCM) clustering algorithms were chosen to be used in building an efficient network intrusion detection model. The paper is structured as follows: Section 2 provides related work of IDS using Big Data techniques, Section 3 introduces brief introduction about intrusion detection, Section 4 presents the used datasets, the proposed system and its components, Section 5 illustrates the evaluation metrics and results of the tested system, finally, Section 6 provides conclusions and further development of future work.

2. Related work

Several works of IDS using Big Data techniques exist. Jeong et al. [4] indicate that Hadoop can solve intrusion detection and big data issues by focusing specifically on anomalous IDSs. The experience of Lee et al. [5] with Hadoop technologies shows good feasibility as an intrusion detection instrument because they were able to reach up to 14 Gbps for a DDOS detector. M. Essid and F. Jemili [6] have combined and eliminated the redundancy of the alerts bases KDD99 and DARPA, they used Hadoop for data fusion. Besides, R. Fekih and F. Jemili [7] used Spark to merge and remove the redundancy of the three alerts bases KDD99, DARPA and MAWILAB. The main objective was to improve detection rates and decrease false negatives. Terzi et al. [8] created a new approach to unsupervised anomaly detection and used it with Apache Spark on Microsoft Azure (HDInsight21) to harness scalable processing power. The new approach was tested on CTU-13, a botnet traffic dataset, and achieved an accuracy rate of 96%. M.Hafsa and F.Jemili [9] created a new approach to intrusion detection. They used Apache Spark on Microsoft Azure (HDInsight21) to analyze and process data from the MAWILAB database. Their new approach achieved an accuracy rate of 99%. Ren et al. [10] created a new approach to unsupervised anomaly detection using the KDD'99 base to analyze and process the data, they achieved a low detection rate. Rustam and Zahras [11] compared two models, one supervised (the Support Vector Machine SVM model) and the other unsupervised (Fuzzy C-Means FCM) to analyze, process, and detect KDD'99 database intrusions. They found that SVM achieved an average accuracy rate of 94.43%, while FCM achieved an average accuracy rate of 95.09%. In this work, we propose an Apache Spark-based approach to detect intrusions. The goal of our system is to provide an efficient intrusion detection system using Big Data tools and fuzzy inference to treat uncertainties and provide better results.

3. Intrusion detection system (IDS)

Intrusion: An intrusion is any use of a computer system for purposes other than those intended, usually due to the acquisition of privileges illegitimately. The intruder is generally seen as a stranger to the computer system that has managed to gain control of it, but statistics show that the most common abuses come from internal people who already have access to the system [12].

Intrusion detection: Intrusion detection has always been a major concern in scientific papers [13, 14]. By security auditing mechanisms, it consists in analyzing the collected information in search of possible attacks.

Intrusion detection system (IDS): To detect and signal anomalous activities, an intrusion detection system (IDS) is used. To detect and signal anomalous activities, an intrusion detection system (IDS) is used. This system protects a system from malicious activities coming from known or unknown sources, this process is done automatically in order to protect confidentiality, integrity and availability of systems. Cannady et al. [15] states that an IDS has two detection approaches: an anomaly-based detection and signature-based detection. IDS are characterized by their surveillance domain which can monitor a corporate network, multiple machines or applications.

4. Data and tools

4.1 Data sets

4.1.1 CTU-13 data set

CTU-13 consists in a group of thirteen scenarios that each run a specific botnet performed in a real network environment. Each scenario includes a botnet pcap file, a tagged NetFlow file, a README file with the capture timeline, and the malware run file. The NetFlow (network flow) file is based on bidirectional flows that provide information about the communication between a source (a client) and a destination (a server). This dataset includes three types of traffic with a different distribution: Normal, botnet (or Malware), and background:

- **Normal traffic:** comes from normal hosts that are previously verified and very useful to check the actual performance of machine learning algorithms.

- **Malware or botnet traffic:** comes from malicious hosts or robots.

- **Background traffic:** is considered as unknown traffic that is needed to saturate the algorithms in order to check their speed performance and test if the algorithm merges with the other traffic [16].

The following **Figure 1** presents the distribution of experimental data for CTU-13 data set:

4.1.2 UNSW-NB15 data set

UNSW-NB 15 is a dataset that was created in an Australian Cyber Range Lab using an IXIA PerfectStorm tool to extract a hybrid of realistic modern natural activities and contemporary synthetic attack behaviors generated by network traffic. This dataset contains 49 features are categorized into five groups and which are explained in [17, 18].

The following **Table 1** represents the attack types which are classified into nine groups.

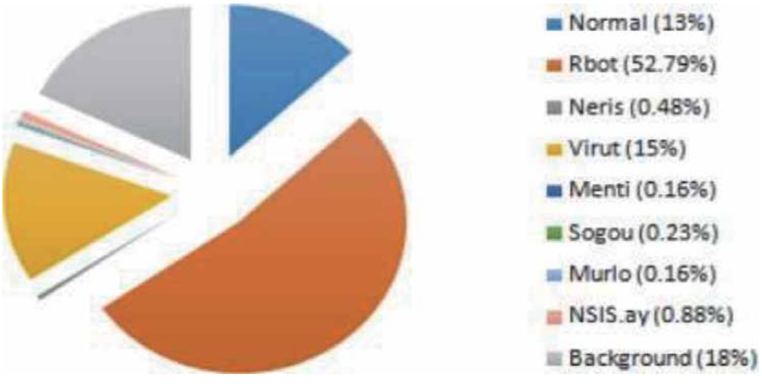


Figure 1.
Attack categories in CTU-13.

Types	Description
Fuzzers	The attacker attempts to cause a program or network suspended by feeding it the randomly generated data.
Analysis	It penetrates the web applications via ports (port scan), web scripts (HTML files), and emails (spam).
Backdoor	A technique in that a system security mechanism is bypassed to access a computer or its data.
DoS	An intrusion which attempts to make a network resource or a server unavailable to users, generally by temporarily suspending the services of a host connected to the Internet.
Exploit	It takes advantage of a glitch, bug, or vulnerability to be caused by an unintentional behavior on a network or a host.
Generic	A technique establishes against every block-cipher using a hash function to collision without configuration of the block-cipher.
Reconnaissance	It gathers information about a computer network to evade its security controls.
Shellcode	The attacker penetrates a slight piece of code starting from a shell to check the compromised machine.
Worm	The attacker replicates itself in order to advance on other computers. Frequently, it uses a computer network to spread itself, relying on the security failures on the target computer to access it.

Table 1.
UNSW-NB15 attack types.

4.2 Apache spark

Apache Spark, is powerful hybrid, scalable and fast distributed data processing engine most active open source project in big data. It was developed at UC Berkeley in 2009. It became one of the top projects in Apache in 2010 [19]. Spark provides APIs in Scala, Java, Python and R languages. To get a good hold on huge data, it must be fast enough by processing massive data at once. Therefore, it is necessary that Spark is available on several clusters rather than on a single machine. The result of the treatment provided by Spark is not written to the disk but kept in memory. This all-in-memory ability is a high-performance computing technique for advanced analytics, making Spark 100 times faster than Hadoop (**Figure 2**) [20].

Spark also has an ecosystem of libraries that can be used for Machine Learning, interactive queries. Which can have important implications for productivity. The project has been progressively enriched to provide a complete ecosystem today which is shown in **Figure 3**.

4.3 Microsoft azure

Microsoft Azure, formally known as Windows Azure, is a cloud computing platform for building, deploying and managing services and applications anywhere with the help of a global network of managed data centers located in 54 regions around the world [21]. Microsoft's HDInsight is a managed Hadoop service in Azure Cloud that uses the Hortonworks Data Platform (HDP). HDInsight clusters can be customized easily by adding additional packages and can scale up in case of high demand by allocating more processing power [22]. By the Azure Active Directory, The data is protected and persists even after the cluster is deleted.

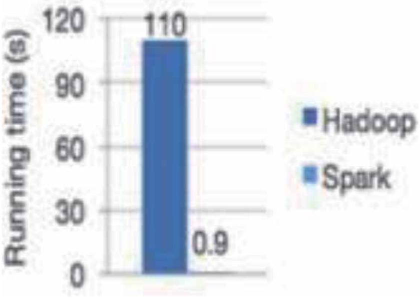


Figure 2.
Speed comparison chart between spark Hadoop.

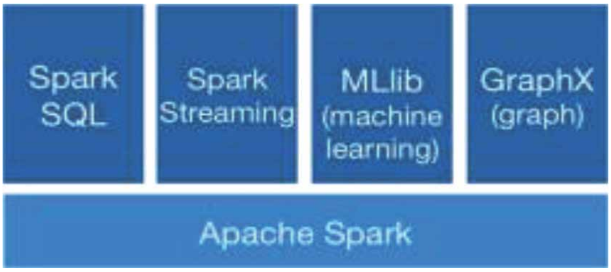


Figure 3.
Apache spark ecosystem.

4.4 Fuzzy C-means clustering (FCM)

The FCM algorithm is one of the most widely used fuzzy clustering algorithms [23] which attempts to partition a finite collection of elements into a collection of c fuzzy clusters with respect to some given criterion. This algorithm is based on minimization of the following objective function:

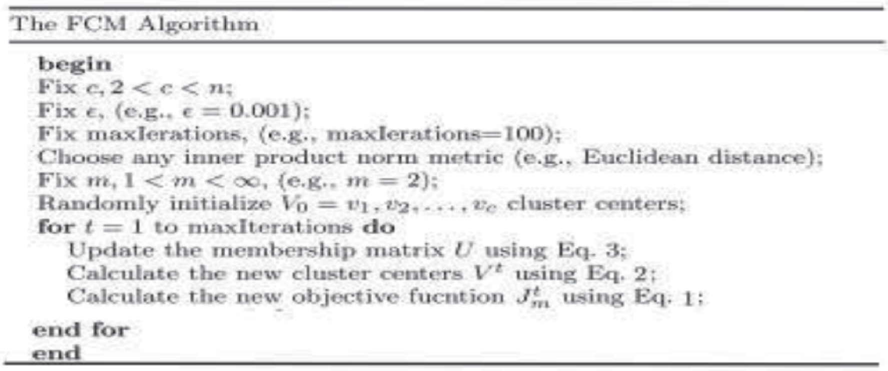
$$Jm = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, 1 \le m < \infty \tag{1}$$

where:

- m: Any real number greater than 1.
- u_{ij}: The degree of membership of xi in the cluster j.
- x_i: The ith of d-dimensional measured data.
- c_j: The d-dimension center of the cluster.
- ||*||: Any norm expressing the similarity between any measured data and the center. The algorithm FCM is composed of the following steps:

Step1: U0, Initialize U = [uij] matrix.
Step2: At k-step, calculate the centers vectors C (k) = [cj] with U(k) [24]

$$cj = \frac{\sum_{i=1}^{N} xiu_{ij}^{m}}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{2}$$



```
The FCM Algorithm
begin
Fix $c, 2 < c < n$;
Fix $\epsilon$, (e.g., $\epsilon = 0.001$);
Fix maxIerations, (e.g., maxIerations=100);
Choose any inner product norm metric (e.g., Euclidean distance);
Fix $m, 1 < m < \infty$, (e.g., $m = 2$);
Randomly initialize $V_0 = v_1, v_2, \ldots, v_c$ cluster centers;
for $t = 1$ to maxIterations do
    Update the membership matrix $U$ using Eq. 3;
    Calculate the new cluster centers $V^t$ using Eq. 2;
    Calculate the new objective fucntion $J_m^t$ using Eq. 1;
end for
end
```

Figure 4.
Pseudo code of FCM algorithm-.

Step3: Update U (k), U(k + 1).

$$cj = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|xi-cj\|}{\|xi-ck\|} \right)^{\frac{2}{m-1}}} \tag{3}$$

Step4: If || U(k + 1) -U(k) || $< s$ then STOP, else return step 2.

Step5: The Fuzzy partitioning [25] is realized out through an iterative optimization of the objective function in Eq. (1), with the cluster centers cj by using Eqs. (2) and (3) and the update of membership uij.

Step6: This iteration will stop when:

$$max_{\,ij} \left\| u_{ij}^{k+1} - u_{ij}^{k} \right\| < \epsilon \tag{4}$$

Where:

- ε: Termination criterion between 0 and 1.
- k: The iteration steps.

A pseudo code of the algorithm FCM is presented as follows (**Figure 4**).

5. Proposed method

The idea of our distributed architecture comes down to a process adaptation of in data fusion approach. This architecture allows us to facilitate data analysis with a powerful Spark big data tool (see **Figure 5**).

5.1 Converting incoming file

We will be using Jupyter Notebook with Apache Spark and the Python API (PySpark). In this stage, we will read CSV files and converting them to Apache Parquet format into Microsoft Azure Blob Storage. Apache Spark supports multiple operations on data, it bids the ability to convert data to another format in just one line of code. Developed by Twitter and Cloudera, Apache Parquet is an open-source columnar file format optimized for query performance and minimizing I/O, offering very efficient

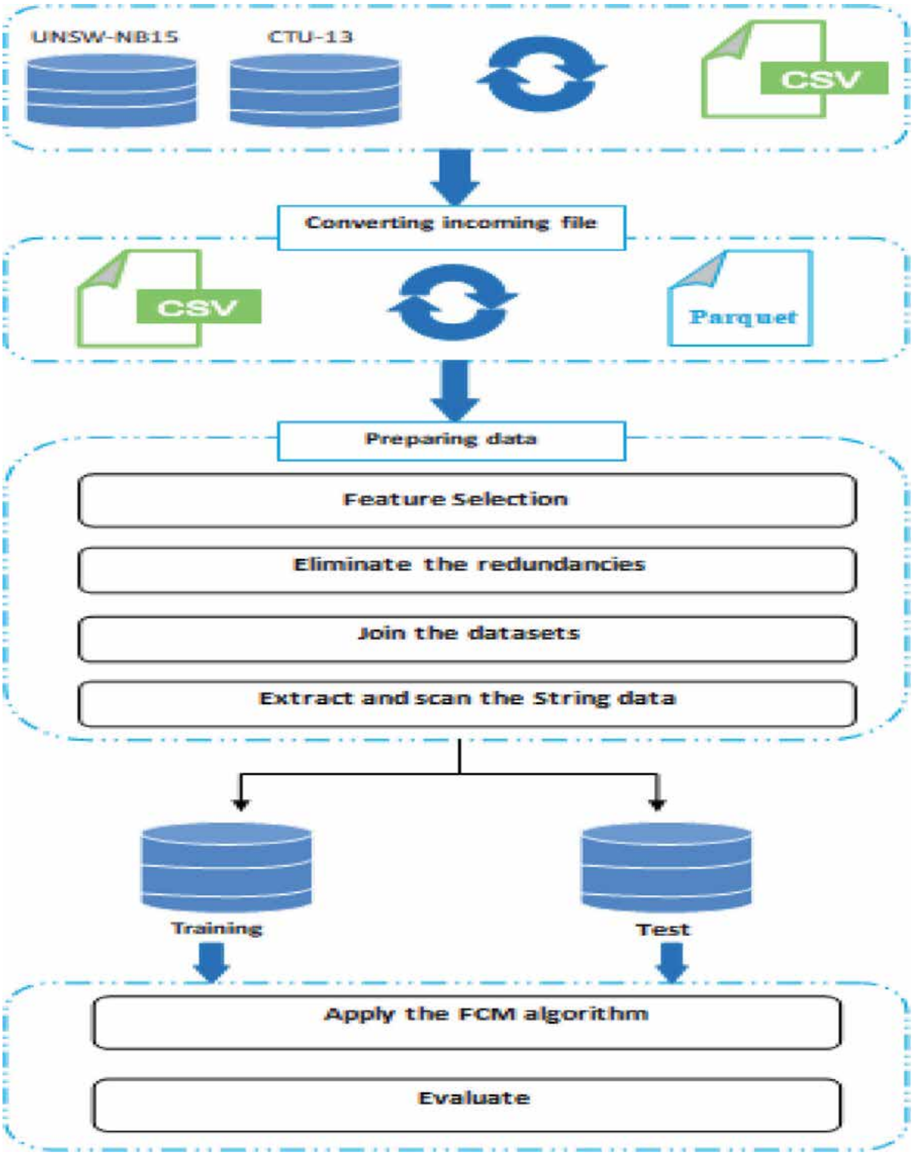


Figure 5.
Diagram of proposed approach.

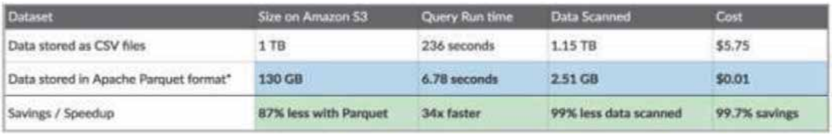
compression and encoding schemes [26]. **Figure 6** shows the efficiency of using the Parket format. This format minimizes storage costs and data processing time.

The following **Table 2** indicates the old and new size of each datasets after converting to Apache Parquet, We notice that by converting CSV to Parquet the storage costs are minimized.

5.2 Preparing data

5.2.1 Feature selection

The feature selection phase selects relevant attributes required for decision making. A pre-processing phase converts the flow records in a specific format which is acceptable to an anomaly detection algorithm [27].



Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

Figure 6.
Apache parquet advantages.

DataSets	Average Size (CSV)	Average Size (Parquet)	speedup
CTU-13	2600.96MO	555MO	x4.69
UNSW-NB15	559MO	202MO	x2.77

Table 2.
Average file size before and after converting.

with the CTU-13 dataset, We did not utilize the feature selection algorithm for this dataset, we instead selected columns that were pertinent and delete unnecessary features(empty columns). After the removing, we get with a total of 13 columns.

Using UNSW-NB15 dataset, we processed the data selection problem. We apply a combination fusion of Random Forest Algorithm with Decision Tree Classifier. V. Kanimozhi [28] decides that the combined fusion of these two algorithms provides 98.3% has listed the best four features are as sbytes, sttl, sload, ct_dst_src_ltm and the **Figure 7** labels the graphical representation of Feature Importances and the top four features.

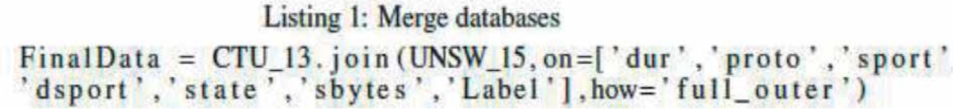
The goal of eliminating no-useful attributes is bring about a better performance by the system with a better accuracy.

5.2.2 Eliminate the redundancies

This eliminate redundancies task involves removing duplicates (removing all the repeated records) which helps with attack detection as it makes the system less biased by the existence of more frequent records.This tactic makes computation faster as it must deal with less data [29].

5.2.3 Join the datasets

Before the merge of the bases, some common columns have different names from one database to another (for example “Label” in CTU-13 named “attack_cat” in UNSW-NB15), in this state, we will rename these attributes then we will merge our bases. Since, Apache Spark offers the ability to join our databases in just one line of code. The following Listing shows the query used:



Listing 1: Merge databases

```
FinalData = CTU_13.join(UNSW_15,on=['dur','proto','sport'
 'dsport','state','sbytes','Label'],how='full_outer')
```

5.2.4 Extracting and scanning string data

Using the Apache Spark Machine Learning library, we create a Machine Learning pipeline. A pipeline is a sequence of stages where each stage is either an Estimator or a Transformer.

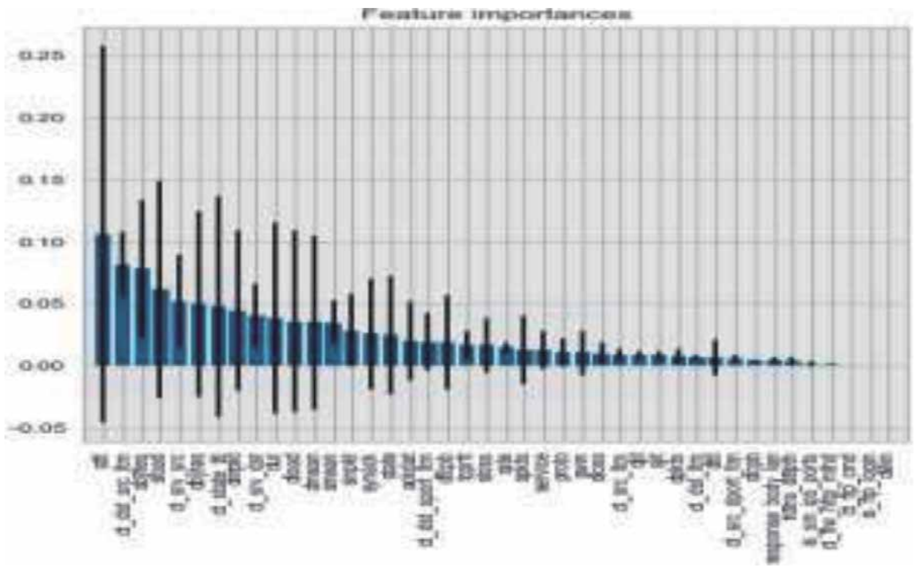


Figure 7.
Feature importance of UNSW-NB15 dataset.

In our final base, some attributes are of types string (Like: Label, sport, proto, ...), in this step we will convert all attributes of type string to attribute of type integer by using the transformer "StringIndexer" which encodes a string column of labels to a column of label indices. These indices are ordered by label frequencies, the most frequent label gets index 0.

StringIndexer classifies attacks automatically in class, it assigns the same index for attacks of the same category.

5.3 FCM application

In our experimental work and as we said above we will use Microsoft Azure as a cloud environment to upload and analyze the dataset with FCM algorithm. We use the training dataset to form and evaluate our model. The test dataset is then used to make predictions. We choose to train our Model with FCM algorithm. The first stage of the FCM algorithm is to initialize the input variable, the input vector includes the dataset features, the number of cluster is 2 (**1 = intrusion and 0 = normal**), and the center of cluster is calculated by taking the means of all feature in the final dataset.The use of fuzzy C-means clustering algorithm to classify data will generate a number of clusters, each cluster contains part of the data records [30]. The characteristics are different between normal and intrusion data records, so they should be in different clusters as shown in the following **Figure 8** which presents the data records clustering.

5.4 Performance metrics

Apache Spark Machine Learning provides a suite of metrics to evaluate the performance of Machine Learn- ing models [31]. To measure the performance in our work the metrics used are as present in below **Table 3** (Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives).

After apply the FCM to our final dataset(After merging our intrusion detection datasets) the result is shown in the following **Table 4**.

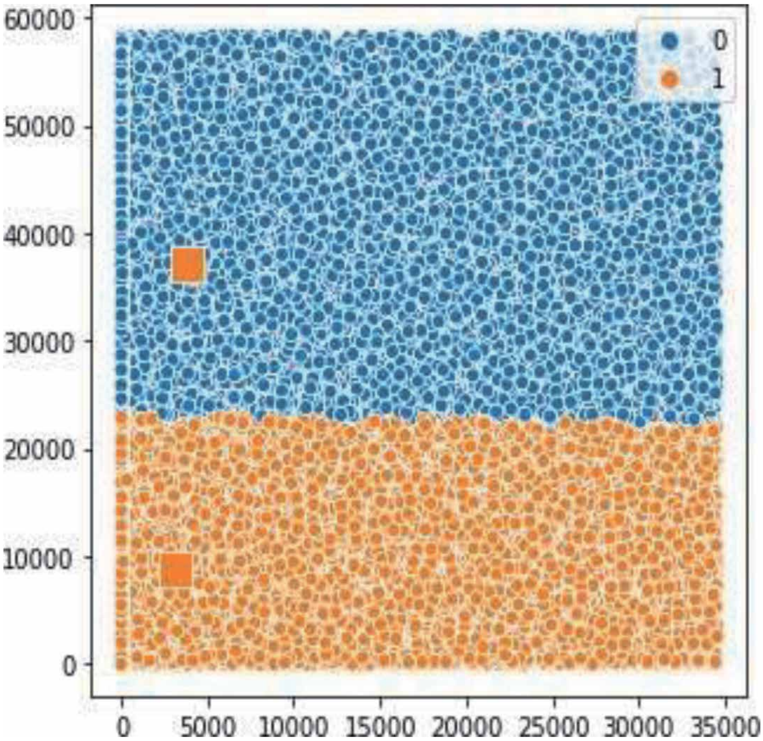


Figure 8.
The data records clustering.

Measure	Description	Formula
Accuracy	Accuracy measures performance across all labels	Accuracy = TP + TN/ TP + FP + FN + TN
Precision	The ratio of correctly predicted positive observations to the total predicted positive observations.	Precision = TP/TP + FP
Recall	The ratio of correctly predicted positive observations to the all observations in actual class	Recall = TP/TP + FN
F-measure	The weighted average of Precision and Recall	F1 = 2*(Recall * Precision) / (Recall + Precision)

Table 3.
Evaluation metrics.

As shown in **Table 4** the total input data is 845 721 records, 243 899 records as normal and 601 822 records as intrusion. After applying FCM algorithm, the result is 231 704 record for normal and 589 785 records for intrusion. Then we calculated the normal and intrusion classification rate by the following equation:

$$Classification\ rate = \frac{Number\ of\ classified\ patterns}{Total\ number\ of\ patterns} * 100 \tag{5}$$

The simulation results show that the classification rate is 96.4% by the FCM algorithm which means that the false positive rate(returns the rate of instances which are falsely classified) is 0.02%.

	Input data	Output data	Classification rate
Normal	243 899	231 704	94.9%
Intrusion	601 822	589 785	97.9%

Table 4.
Evaluation metrics.

6. Discussion

It is possible to obtain a very precise system (accuracy of 99%) but not very efficient with a recall of 10%. In our work, with an accuracy of 97.2% and a recall of 96.4%, we can say that our system is efficient. The use of the fuzzy algorithm in this experiment gave a good result. The advantage of our system is the fuzzy representation that is increasingly used to deal with missing and inaccurate data problems which is the disadvantage of most classification algorithms.

7. Conclusions and future work

In this paper, we achieved a successful distributed IDS. Using the FCM algorithm allows to effectively train and analyze our model after merging datasets. Proposing a distributed system and showing the power of Spark to combine and handle large and heterogeneous structures of training datasets present the main merits in our work.

In future work, we will perform our dataset analysis with another Big Data framework expected to reach faster results. In addition, we will develop our approach with other classifiers to get better results.

Author details

Farah Jemili[1]* and Hajer Bouras[2]

1 ISITCom, Mars Research Laboratory, University of Sousse, Hammam Sousse, Tunisia

2 ISITCom, University of Sousse, Hammam Sousse, Tunisia

*Address all correspondence to: jmili_farah@yahoo.fr

References

[1] Hafsa, M., Jemili, F. (2018). "Comparative Study between Big Data Analysis Techniques in Intrusion Detection. Big Data and Cognitive Computing", 2018.

[2] https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html.

[3] D. Song, M.I. Heywood, A.N. Zincir-Heywood, "Training Genetic Programming on Half a Million Patterns: An Example from Anomaly Detection," IEEE Transactions on Evolutionary Computation, 2005.

[4] H. W. L. J. Y. I. Jeong H, «Anomaly teletraffic intrusion detection systems on hadoopbased platforms: A survey of some problems and solutions,» 15th international conference on. IEEE, pp. 766-770.

[5] L. Y. Lee Y, «Toward scalable internet traffic measurement and analysis with hadoop,» ACM SIGCOMM Comput Commun Rev, vol. 43(1), pp. 5-13.

[6] F. J. Mondher Essid, «Combining intrusion detection datasets using MapReduce,» In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 10/2016.

[7] Ben Fekih, R., & Jemili, F. Distributed Architecture of an Intrusion Detection System Based on Cloud Computing and Big Data Techniques, 2018.

[8] D. S. Terzi, R. Terzi and S. Sagiroglu, "Big data analytics for network anomaly detection from netflow data," in International Conference on Computer Science and Engineering (UBMK), Antalya, 2017.

[9] Hafsa, M., Jemili, F. (2018). "Comparative Study between Big Data Analysis Techniques in Intrusion Detection. Big Data and Cognitive Computing", 2018.

[10] Ren, W., Cao, J., Wu, X. (2009). "Application of network intrusion detection based on Fuzzy C-means clustering algorithm", 3rd International Symposium on Intelligent Information Technology Application, IITA 2009.

[11] Rustam, Z., & Ariantari, N. P. A. A. (2018). Comparison between support vector machine and fuzzy Kernel C-Means as classifiers for intrusion detection system using chi-square feature selection, 2018.

[12] Moustafa, N., Slay, J. (2015). "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference", 2015.

[13] Sebastian Garcia, Martin Grill, Jan Stiborek, and Alejandro Zunino. "An empirical comparison of botnet detection methods", 2014.

[14] Moustafa, N., Slay, J. (2016). "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Information Security Journal", 2016.

[15] Dataricks. About Databricks. Available online: https://databricks.com/spark/about (accessed on 6 May 2018).

[16] Gaied, I., Jemili, F., Korbaa, O. (2016). "Intrusion detection based on Neuro-Fuzzy classification. Proceedings of IEEEACS International Conference on Computer Systems and Applications", AICCSA, 2016.

[17] Massimiliano, A.; Erbacher, R.F.; Jajodia, S.; Persia, M.C.F.; Picariello, A.; Sperli, G.; Subrahmanian, S.V.

Recognizing unexplained behavior in network traffic. Netw. Sci. Cybersecur. 2013.

[18] B. C.* Rhodes, J. A. Mahaffey and D. J. Cannady, "Multiple Self-Organizing Maps for Intrusion Detection," in National Information Systems Security Conference, Baltimore, 2000.

[19] Microsoft. Azure Regions. Available online: https://azure.microsoft.com/en-us/global-infrastructure/ regions/ (accessed on 5 May 2018).

[20] Neenu Daniel & Ritty Jacob, Intrusion Detection Techniques in Big Data: A Review, April 2017.

[21] Ren, W., Cao, J., Wu, X. (2009). "Application of network intrusion detection based on Fuzzy C-means clustering algorithm", 3rd International Symposium on Intelligent Information Technology Application, IITA 2009.

[22] Mllib Evaluation Metrics. Available online: https://spark.apache.org/docs/ 2.1.0/mllib-evaluation-metrics. html (accessed on 3 June 2018).

[23] Jawhar, M. M. T., Mehrotra, M. (2010)." Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network. International Journal of Computer Science and Security", 2010.

[24] P. Biondi, «Architecture expérimentale pour la détection d'intrusions dans un système informatique», 2001.

[25] Ar, L.; Levent, E.; Vipin, K.; Aysel, O.; Jaideep, S. A comparative study of anomaly detection schemes in network intrusion detection. In Proceedings of the SIAM Conference on Applications of Dynamical. Systems, 2003.

[26] Premasundari, M., Yamini, C. (2019). a Violent Crime Analysis Using Fuzzy C-Means Clustering Approach, 6956(April), 2019.

[27] K. Taha and P.D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", IEEE Transactions on Information Forensics and Security, 2016.

[28] Apache Parquet vs. CSV Files—DZone Database. Available online: https://dzone.com/articles/how-to-bea-hero- with-powerful-parquet-google-and (accessed on 6 February 2018).

[29] Umer, M. F., Sher, M., Bi, Y. (2017). "Flow-based intrusion detection: Techniques and challenges. Computers and Security", 2017.

[30] Kanimozhi, V., Jacob, P. (2019). "UNSW-NB15 dataset feature selection and network intrusion detection using deep learning. International Journal of Recent Technology and Engineering", 2019.

[31] Verma, R.; Kantarcioglu, M.; Marchette, D.; Leiss, E.; Solorio, "Security Analytics: Essential Data Analytics Knowledge for Cybersecurity Professionals and Students", IEEE Secur. Priv. 2015.

Chapter 5

Artificial Intelligence and IoT: Past, Present and Future

Kannadhasan Suriyan and Nagarajan Ramalingam

Abstract

Artificial intelligence (AI) approaches have recently made major impacts in the healthcare field, igniting a heated discussion over whether AI physicians would eventually replace human doctors. Human doctors are unlikely to be replaced by machines anytime soon, but AI may assist physicians make better clinical decisions or even replace human judgment in certain areas of healthcare (e.g., radiology). The increased availability of healthcare data and the rapid development of big data analysis tools have made recent productive applications of AI in healthcare possible. When driven by appropriate clinical queries, powerful AI systems may find clinically valuable information hidden in enormous volumes of data, which can help clinical decision making. The internet of things (IoT) is a network of many interconnected things that may communicate with one another across a computer network. We may get information from this global network by connecting sensors to it. Thanks to the computer network, we can obtain this information from anywhere on the planet. The internet of things (IoT) enables physical objects to connect to the internet and create systems using various technologies such as near-field communication (NFC) and wireless sensor networks (WSN).

Keywords: WSN, artificial intelligence, deep learning, IoT and health sector

1. Introduction

Wireless sensor networks (WSNs) have been shown in a range of applications during the past several years. A WSN is a collection of wireless devices that are typically small, battery-powered, and self-contained (also known as nodes). These devices include on-board computers, communication, and sensing capabilities, enabling them to monitor and transfer data on physical or environmental factors like temperature, sound, and pressure through a unidirectional or bidirectional network. The nodes are made up of a low-power CPU with limited processing, a memory device with limited storage capacity, a radio transceiver with low-power internal/external antenna, a low-data rate and limited range, sensors (scalar, cameras, microphones), and a power supply (batteries and solar cells). In most cases, each device is powered by a battery. Examining each of these gadgets individually may seem to be pointless.

WSNs play an important role in military applications. The increasing deployment of WSNs drives sensor network research. WSNs, on the other hand, maybe used for environmental monitoring, habitat monitoring, classroom/home monitoring, structural monitoring, and health monitoring, among other things. Based on its characteristics, each application has its own design concept and execution to meet its own demands. The qualities of WSN, together with technology improvements, give the greatest benefits to healthcare. A sensor network designed to identify human health indicators is known as a body sensor network (BSN). Because BSN nodes are directly attached to the human body, considerable vigilance is required. For many days, several healthcare applications need the BSN to collect patient data indefinitely without user intervention. Such applications must take into consideration the energy constraints of sensor networks [1–15].

The challenges of WBS networks healthcare is a need for everyone's quality of life in today's environment. The population of developed countries is growing at a pace that is proportional to the government's budget. Healthcare systems will face challenges as a result of this. One of the most difficult challenges is making healthcare more accessible to elderly people who live alone. In general, health monitoring is done on a check-in basis, with the patient remembering their symptoms; the doctor performs tests and develops a diagnosis, then monitors the patient's progress throughout therapy. Wireless sensor network applications in healthcare provide in-home assistance, smart nursing homes, clinical trials, and research advancement. Let's take a look at some of the challenges and basic features of BSNs before we get into the medical uses of this technology. In healthcare applications, low power, limited computing, security and interference, material restrictions, resilience, continuous operation, and regulatory requirements with elderly people are all issues.

Modern modelling approaches such as fuzzy logic (FL) and artificial neural networks (ANN) are frequently employed in hydrological modelling for a number of applications. The main benefit of these techniques is that they are not constrained by restrictive assumptions such as linearity, normality, or homoscedasticity and that they provide promising and acceptable alternatives to classical stochastic hydrological modelling in time series analysis, such as the auto regressive moving average exogenous (ARMAX) model (autoregressive moving average with exogenous inputs). When applied to hydrological systems, however, traditional stochastic models have many drawbacks, the most notable of which are short-time dependence and the normality assumption, as previously mentioned. Hydrological processes are well recognised for defying these assumptions. ANNs have been recognised as a tool for modelling difficult nonlinear systems and are widely used for hydrological prediction. Their applications range from forecasting hourly and daily river stages to further FL modelling applications in: rainfall-runoff groundwater; and time series modelling [6–10]. Fuzzy neural networks (FNN) are a unique approach for river flow prediction that blends FL and ANNs.

Because they can estimate any continuous function to any degree of accuracy, the Mamdani and tidal sequence (TS) systems are referred to as universal approximators. The smaller the error tolerance, the more fuzzy rules are necessary. In practise, fuzzy models can always yield nonlinear modelling solutions when the required number of fuzzy sets and rules are provided. In comparison to the TS approximator, the Mamdani approximator has the benefit of being able to use both numerical and verbal data produced from human knowledge and experience.

When nontrapezoidal/nontriangular input fuzzy sets are used, TS fuzzy systems may be more cost-effective than Mamdani fuzzy systems in terms of input fuzzy sets and fuzzy rules. They discovered that TS and Mamdani fuzzy systems have comparable minimal system configurations when trapezoidal or triangular input fuzzy sets are used. The performance of Mamdani (linguistic) and TS (clustering-based) fuzzy models was examined in the spatial interpolation of mechanical features of rocks. In terms of prediction performance, the clustering-based TS fuzzy modelling technique beats the Mamdani model, according to their results. The main purpose of this study is to develop a hybrid model for streamflow forecasting that incorporates both a genetic algorithm and fuzzy logic. Genetic algorithms and neural networks (NNs) were used to train the Mamdani and Takagi-Sugeno fuzzy logic modelling systems, respectively. According to the comparison, the Mamdani approach beats standard methods in terms of avoiding restrictive assumptions, insight into the modelling structure, and modelling accuracy.

Health is always a major concern as the human race improves in terms of technology. The current coronavirus outbreak, which has hurt China's economy to some extent, exemplifies how healthcare has become more vital. In areas where the pandemic has spread, it is always preferable to monitor these people using remote health monitoring equipment. As a consequence, the current solution is a health monitoring system based on the IoT. Remote patient monitoring provides for patient monitoring outside of typical clinical settings (for example, at home), increasing accessibility to human services offices while cutting expenses. This project's primary purpose is to design and build a smart patient-health monitoring system that uses sensors to monitor patient health and the internet to alert loved ones if there are any issues. The purpose of developing monitoring systems is to reduce healthcare costs by reducing the number of needed inspections. In an IOT-based framework, different consumers may be able to see sensitive aspects of the patient's blooming. Because the information should be double-checked by visiting a website or URL, this is the case. In GSM-based patient observation, the rising parameters are communicated utilising GSM through SMS techniques.

In most rural areas, the medical facility would not be within walking distance of the residents. As a consequence, the majority of people must attend doctor's visits, stay in hospitals, and undergo diagnostic testing procedures. Each of our bodies uses temperature and pulse recognition to determine our overall health. The sensors are linked to a microcontroller that monitors the state and, as a result, is interfaced to an liquid crystal display (LCD) as well as a remote connection that may send alarms. If the framework detects any unusual changes in heart rate or body temperature, it warns the client through IoT and also shows subtle features of the patient's pulse and temperature on the web in real-time. An IoT-based tolerant wellness monitoring framework efficiently leverages the web to monitor quiet wellbeing metrics and save time in this manner. There is a significant ability to disregard any form of minor health concern, as shown by changes in important components such as body temperature, pulse rate, and so on in the early stages.

When a person's health condition has developed to the point that his or her life is in peril, they seek medical assistance, perhaps wasting money. This is crucial to consider, especially if an epidemic spreads to a place where doctors are unavailable. Giving patients a smart sensor that can be monitored from afar to avoid the spread of sickness would be a realistic solution that might save many lives. Sensors monitor physiological signs, which are transformed into electrical impulses when a patient enters the healing centre. The basic electrical flag is then updated to an advanced flag (computerised data), which is then stored in RFID. To transfer computerised data to a local server, the Zigbee Protocol is employed. For this framework, Zigbee is a good choice. In this location, there are the most cell hubs. It's better for gadgets that are smaller and use less energy. A nearby server sends information to the therapeutic server through WLAN.

When the data is transmitted to the therapeutic server, it checks to see whether the patient already has a medical record, then adds the new information to that record and sends it to the specialist. If the patient has not had any prior treatment records, the server creates a new ID and stores the data in its database. The IoT is becoming more widely recognised as a feasible solution for distant value tracking, notably in the field of health monitoring. It permits the secure cloud storage of individual health parameter data, the decrease of hospital visits for routine checks, and, most critically, the remote monitoring and diagnosis of sickness by any doctor. In this research, an IoT-based health monitoring system was developed. Body temperature, pulse rate, and room humidity and temperature were all measured by sensors and shown on an LCD. The sensor data is then wirelessly sent to a medical server. These data are then delivered to a smartphone with an IoT platform that belongs to an authorised person. Based on the findings, the doctor diagnoses the condition and the patient's current state of health.

The advantages of AI have been extensively researched in the medical literature. Using complicated algorithms, AI can 'learn' characteristics from a large quantity of healthcare data and then apply the results to clinical practise. It might potentially include learning and self-correcting capabilities to improve accuracy as input changes. AI systems that give up-to-date medical information from journals, textbooks, and clinical practises may support physicians in providing proper patient care. In addition, an AI system might help to reduce diagnostic and therapeutic errors, which are inevitable in human clinical practise. Furthermore, an AI system extracts important data from a large patient population to assist in the generation of real-time health risk warnings and prediction findings [11–15].

In this chapter, we look at the current level of AI in healthcare and predict its future. First, we will go over four crucial factors from a medical researcher's perspective: (1) Justifications for AI use in healthcare, and (2) The sorts of data that AI systems have examined AI devices may be classified into two classes, according to the previous description. The first category includes machine learning (ML) approaches that analyse structured data such as imaging, genomics, and EP data. ML algorithms are used in medical applications to cluster people' traits or forecast the probability of sickness consequences. The second category includes natural language processing (NLP) tools, which extract information from unstructured data such as clinical notes and medical journals to supplement and enrich organised medical data. Texts are converted into machine-readable structured data, which may then be analysed using ML algorithms.

2. Artificial intelligence in health sector

Lung- and heart-related ailments are at the top of the list of health-related problems/complications. Wireless technology, which is a relatively new concept, may be used to track one's health. Wireless health monitoring systems make use of wearable sensors, portable remote health systems, wireless communications, and expert systems, among other technologies. Life is valuable, even a single life is valuable, but people are dying due to the lack of health facilities, sickness awareness, and sufficient access to healthcare systems. In all conditions, the IoT assists in the identification of diseases and the treatment of patients. In IoT healthcare systems, there are wireless systems in which different applications and sensors are linked to patients, information is gathered, and the information is communicated to a doctor or specialist through an expert system. Medical devices for the Internet of Things (MD-IoT) are connected to the Internet and use sensors, actuators, and other communication devices to monitor patient health. The expert system uses these devices to transfer patient data and information to a secure cloud-based platform, where it is stored and analysed.

Telemedicine is the practise of caring for a clinician and a patient while they are not physically present with each other. 'The delivery of healthcare services at a distance' is how telemedicine is defined. Telemedicine provides a variety of benefits, but it also has many disadvantages. Providers, payers, and politicians all recognise the difficulty of navigating some grey zones. While the sector will rapidly develop over the next decade, it will also provide practical and technological challenges. IoT is the most trustworthy and cost-effective alternative in certain circumstances, and the connection between different devices and interactive communication systems also need further formal examination. By communicating information to healthcare teams such as doctors, nurses, and specialists, IoT (Internet of Things) and mobile technologies make it easier to monitor a patient's health. Professionals would benefit from using the store and forward method to save and collect patient data that could be accessed at any time.

A smart healthcare system is a piece of technology that enables patients to be treated while also improving their overall quality of life. The smart health concept incorporates the e-health concept, which emphasises a number of technologies such as electronic record management, smart home services, and intelligent and medically connected items. Sensors, smart devices, and expert systems all help to create a smart healthcare system. Healthcare facilities are a big concern in today's globe, especially in developing countries where rural areas lack access to high-quality hospitals and medical experts. Artificial intelligence has benefited health in the same way it has benefited other aspects of life. The IoT is expanding its capabilities in many areas, including smart cities and smart healthcare. The IoT is now being used in healthcare for remote monitoring and real-time health systems. Controlling and preventing catastrophic events, such as the one that happened in 2020 when the coronavirus disease (COVID-19) ravaged the world, maybe done via IoT technologies without imposing severe restrictions on people and enterprises. COVID-19, unlike SARS in 2003, causes respiratory symptoms and seems to be more contagious. One way to restrict viral transmission until a vaccine is developed is to keep a close eye on physical (or social) distance. Improved surveillance, healthcare, and transportation networks will make it less probable for contagious diseases to spread. An IoT system combined with artificial intelligence (AI) may give the following advantages when considering a pandemic: (1) utilising surveillance and image recognition technologies to enhance public security, (2) using drones for supply, transportation, or disinfection, and (3) leveraging AI-powered apps and platforms to monitor and limit people's access to public places.

In healthcare, an IoT system is often made up of a number of sensors that are all connected to a computer and allow real-time monitoring of the environment or patients. AI-assisted sensors might be employed in the case of a pandemic to help predict whether or not people are sick based on symptoms like body temperature, coughing patterns, and blood oxygen levels. The ability to monitor people's locations is another useful function. During an outbreak of severe disease, tracking the distance between people may provide vital information. Using technologies like Bluetooth, we can get a good estimate of how much distance people maintain when walking in public places. This information might be used to target people who are not physically separated by a specified distance, such as 2 m, to stop the virus from spreading further. To prevent the abuse of personal information, security and data management must be addressed throughout the development of such platforms. Following a pandemic, governments may try to use these platforms and data for long-term monitoring to control and monitor people's behaviour.

One of the problems with traditional medical diagnosis is its inaccuracy and imprecision, which has resulted in the deaths of thousands of people. The development of various algorithms, models, and technologies to ensure accuracy and precision has considerably reduced the number of people who die every day in hospitals, and fuzzy logic, a branch of artificial intelligence, is one of these technologies. Medical diagnostic processes are carried out with the use of computer-assisted technologies, which are growing more common by the day. These systems are based on AI and are designed to diagnose as well as recommend treatments based on symptoms. Many decision support systems (DSSs) have been developed in the medical field, such as Aaphelp, Internist I, Mycin, Emycin, Casnet/Glaucoma, Pip, Dxplain, Quick Medical Reference, Isabel, Refiner Series System, and PMA, to assist medical practitioners in their decisions for diagnosis and treatment of various diseases.

The medical diagnostic System (MDS) is used to diagnose various ailments in an expert system like this. Fuzzy logic was chosen as the AI tool in the recommended system since it is one of the most efficient qualitative computational approaches. Fuzzy logic has been proved to be one of the most effective techniques to offer clarity to the medical field. Medical applications include CADIAG, MILORD, DOCTORMOON, TxDENT, MedFrame/CADIAG-IV, FuzzyTempToxopert, and MDSS, to name a few.

3. Artificial intelligence in open data

Control systems, household appliances, decision-making systems, and the medical and automotive industries all use fuzzy logic-based automated systems. Some of the concepts used in fuzzy logic include fuzzification, defuzzification, membership function, rules, domains, linguistic variables, and so on. While Boolean algebra's set values are confined to 0 and 1 or False and True, fuzzy logic proposes that there are extra values between 0 and 1 or False and True, referred to as in-between values. To put it another way, on its set of 0 and 1, Boolean logic employs entirely inclusive and exclusive rules, while fuzzy logic employs wholly inclusive, exclusive, and 'in between values' rules. Both expert systems and fuzzy logic control systems are designed to tackle difficult and intricate jobs, but a fuzzy logic control system has the benefit of being able to cope with ambiguity. Language standards are employed to enhance decision-making in the face of uncertainty, emulating a human operator. This decision-making power saves time and reduces or eliminates the need for the human element in control models, which was previously required. A closer look into this cluster reveals that similar themes include the use of intelligent data analysis and related domains to anticipate outbreaks, simulate disease transmission, and screen for the virus on a broad scale. Epidemiology is the term used to describe all of this. Modelling and forecasting the spread of COVID-19 using AI and ML methods may help governments, health organizations, corporations, and people manage the pandemic. In this regard, NNs have also played a significant role. To forecast situations, multi-layer feed-forward NNs and convolution neural networks (CNNs) were utilized. Other well-known algorithms for predicting time series data, such as ARIMA (auto-regressive integrated moving average model) and support vector machine (SVM), have been studied. Several of these models have been used to estimate daily infections during various sorts of lockdowns, assisting government decision-making. Public policies have been effectively planned using ML approaches.

In the creation of vaccines, AI and intelligent data analysis have also proven critical. ML and AI are particularly useful for repetitive activities that need large-scale data processing, making them ideal for drug-development. Deep learning has shown to be a very useful technique for predicting the qualities and uses of pharmaceutical compounds that might trigger a body's immune response to an illness. Because this analytical method often needs long periods of testing and a considerable expense, automating this contact would be quite beneficial. Scientists have developed algorithms that anticipate which immunogenic regions should be included in a vaccine, allowing the immune system to learn and prepare for specific antigens. Antigens already found in pathogens that may be related to antigens for a new infection may also be recognised by AI, speeding up the process even further.

AI is assisting in the development of vaccines by simplifying the comprehension of viral protein structures and assisting clinical professionals in sifting through hundreds of relevant study findings at a faster rate than would be achievable otherwise. The ability to understand the structure of a virus may aid in the creation of effective vaccination.

AI and soft computing (SC) play a crucial role in healthcare medical diagnostics. Doctors nowadays are unable to advance without the help of technological advancement. This digital advancement will be incomplete if AI and SC are not included. AI is a technique for constructing intelligent machines. The SC is a collection of computer algorithms for seeing and learning real-world information, which allows computers to create AI. As a consequence, the computer can perform as well as a person if the philosophy of human labour can be expressed using AI and SC technologies. In the healthcare sector, this technological development is being used for long-term medical diagnosis. AI is defined by Alan Turing, the discipline's inventor, as 'the science and engineering of building machines, especially sophisticated computer programmes.' Artificial intelligence systems are computer programmes that can mimic human cognitive processes.

In the early phases of AI, philosophy, potential, demonstrations, dreams, and imagination all played a part. In response to a variety of conflicting needs, possibilities, and interests, the field of IA developed. In a range of fields, including healthcare, AI combined with analytics (AIA) is becoming increasingly commonly employed. Medicine was one of the most successful applications of analytics, and it is now a prospective AI application sector. As early as the mid-twentieth century, clinical applications were designed and provided to physicians to assist them in their practise. Among the applications are clinical decision support systems, automated surgery, patient monitoring and assistance, healthcare administration, and others. The current methodologies are mostly focused on knowledge discovery via data and ML, ontologies and semantics, and reasoning, as we will see in the next sections. We will look at how AI has advanced in healthcare over the last 5 years in this piece.

Data mining, ontologies, semantic reasoning, and ontology-extended clinical recommendations, clinical decision support systems, smart homes, and medical big data will be the focus of the examination. The multiple artificial intelligence features of our study were not chosen at random. Indeed, we have noticed that they have developed a strong interest in medicine in recent years. Data mining methods are used in learning and prediction, as well as picture and speech processing, and anything involving emotion and sentiment. Because of their ability to reason, as well as its usage as a way of learning, sharing, reuse, and integration, ontologies have gained momentum in medicine. Clinical decision support systems that assist improve the quality of treatment in clinical practise draw on both disciplines. They are also used in smart homes to help those with cognitive impairments with daily tasks. Big data in medicine is becoming increasingly common, and its application in analytics is unavoidable.

Electricity engineers formerly concentrated their efforts on the production and transmission levels, with the distribution system receiving less attention. Engineers have only recently been provided with the tools necessary to cope with the computational burden of distribution systems to undertake realistic modelling and simulation. The majority of primary distribution systems are built up in a radial configuration, with one end providing each load point. The radial type system is the simplest and most often used for effective coordination of their protective systems. Fuzzy set theory has been developed and used in a range of engineering and non-engineering domains where the evaluation of actions and observations is 'fuzzy' in the sense that no clear boundaries exist. The fuzzy set theory provides for the inaccurate representation of evaluations and observations, which may then be utilised to describe and solve issues.

The use of fuzzy set theory to distribution system analysis may aid professional judgement and prior knowledge in distribution system planning, design, and operations. Future computer technology will be considerably more advanced than our greatest imaginations, and far more advanced than anything we can envision right now. The IoT is one of the most cutting-edge technologies, with IoT-enabled things all around us. With the help of RFID (radio frequency identification) and sensors, it will create its own world in which everything will be managed and transmitted over the Internet. The devices will create their own environment. The enormous amount of data created will be recorded, analysed, and presented in a timely, seamless, and understandable way. Cloud computing will provide us with virtual infrastructure for visualisation platforms and utility computing, enabling us to integrate device monitoring, storage, client delivery, analytics tools, and visualisation in one place. Cloud computing, which will provide an end-to-end solution, will allow users and businesses to access applications on-demand from anywhere. One of the most important IoT applications is in the field of healthcare. We designed a health monitoring device using current low-cost sensors to monitor and maintain human health parameters such as heart rate, temperature, and air quality. The approach of fuzzy logic was used. In 1965, Lotfi Zadeh presented the concept of fuzzy logic for the first time.

Fuzzy logic is a kind of multivalued logic with truth values ranging from 0 to 1. Fuzzy logic deals with the concept of partial truth, in which the truth value varies from completely false to completely true. The fuzzy logic technique includes fuzzification, inference, and defuzzification. The sensors capture crisp input data, which is then converted via membership functions into a fuzzy input set, linguistic words, and linguistic variables. The rules are used to make inferences. The system will work on the same principles as the IF-THEN system. The membership function is used to convert the fuzzy output to crisp output.

Vital signs are the four most important markers that reveal the condition of the body's vital functions. These measurements are used to assess a person's general physical well-being, detect probable diseases, and monitor healing progress. The fuzzy inference system is a computer framework that makes choices based on fuzzy set theory, fuzzy if-then logic, and fuzzy reasoning. Over the last decade, fuzzy set theory has advanced in many directions, with applications in taxonomy, topology, linguistics, automata theory, logic, control theory, game theory, information theory, psychology, pattern recognition, medicine, law, decision analysis, system theory, and information retrieval, to name a few. A fuzzy inference requires three parts: a membership function generation circuit that calculates the goodness of fit between an input value and the membership function of an antecedent part, a minimum value operation circuit that finds an inference result for each rule, and a maximum value operation circuit that integrates a plurality of inference results. When these components are combined into a system, the system can do inference. Each externally supplied input value, this membership function generating circuit creates one membership function value. The decision-making logic of the fuzzy inference machine is crucial, and it may be the system's most adaptive component. The fuzzification interface corresponds to our sensory organs (e.g., eye, ear), the de-fuzzification interface to our action organs (e.g., arms, feet, etc.), the fuzzy rule base to our memory, and the fuzzy inference machine to our thought process when a fuzzy system is compared to a human controller. It is called a fuzzy expert system when an expert system uses fuzzy data to reason. It is important to know what makes up a fuzzy expert system. The fuzzy expert system consists of a fuzzy knowledge base (based on fuzzy rules), an interference engine, a working memory subsystem, an explanation subsystem, natural language interference, and knowledge acquisition.

4. Conclusion

The number of individuals visiting hospitals has grown in recent decades as a result of population expansion and changing lifestyles throughout the world. As a result, the medicare healthcare system is overburdened. On the other hand, for the old, crippled, underprivileged, or those who live far away, visiting the hospital is quite tough. As a consequence, their health may worsen to the point where they die. The Internet has now become a necessary component of our daily life. Education, finance, business, industry, entertainment, social networking, retail, and e-commerce are just a few of the uses of the internet. The IoT is the web's next big thing (IoT). As a consequence, remote healthcare solutions were created to meet the medicare health system's above-mentioned difficult challenges. The patient's vital signs are monitored by electronic sensors and communicated to the hospital server through the Internet, allowing the doctor to examine, diagnose, and prescribe the required medicine to treat the patient without the patient needing to visit the hospital.

Author details

Kannadhasan Suriyan[1]* and Nagarajan Ramalingam[2]

1 Department of Electronics and Communication Engineering, Cheran College of Engineering, Karur, Tamilnadu, India

2 Department of Electrical and Electronics Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

*Address all correspondence to: kannadhasan.ece@gmail.com

References

[1] Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, et al. A view of cloud computing. Communications of the ACM. 2010; **53**(4):50-58

[2] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems. 2009; **25**(6):599-616

[3] Garcia Lopez P, Montresor A, Epema D, Datta A, Higashino T, Iamnitchi A, et al. Edge-centric computing: Vision and challenges. SIGCOMM Computer Communication Review. 2015;**45**(5):37-42

[4] Satyanarayanan M, Simoens P, Xiao Y, Pillai P, Chen Z, Ha K, et al. Edge analytics in the internet of things. IEEE Pervasive Computing. 2015;**14**(2):24-31

[5] Luan TH, Gao L, Li Z, Xiang Y, Wei G, Sun L. Fog computing: Focusing on mobile users at the edge. Networking and Internet Architecture. p. 11. Available from: https://arxiv.org/abs/1502.01815v3

[6] Bonomi F, Milito R, Zhu J, Addepalli S. Fog computing and its role in the internet of things. In: Proceedings of the Workshop on Mobile Cloud Computing. Vol. 2012. 2012. pp. 13-16

[7] Lancaster SS. A Fuzzy Logic Controller for the Application of Skin Pressure. Omaha, NE: Creighton University; 2004

[8] Mahfouf M, Abbod MF, Linkens DA. A Survey of Fuzzy Logic Monitoring and Control Utilization in Medicine. Sheffield, UK: Department of Automatic control and System Engineering, University of Sheffield; 2000

[9] Yue H, Yue G, Yi G. Application study in decision support with fuzzy cognitive map. International Journal of Computers. 2007;**12**(12)

[10] Lee CS, Nagy PG, Weaver SJ, et al. Cognitive and system factors contributing to diagnostic errors in radiology. American Journal of Roentgenology. 2013;**201**:611-617

[11] Neill DB. Using artificial intelligence to improve hospital inpatient care. IEEE Intelligent Systems. 2013;**28**:92-95

[12] Administration UFaD. Guidance for Industry: Electronic Source Data in Clinical Investigations. 2013. Available from: https://www.fda.gov/downloads/drugs/guidances/ucm328691.pdf [Accessed: June 1, 2017]

[13] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology. 2016;**278**:563-577

[14] Li CY, Liang GY, Yao WZ, et al. Integrated analysis of long noncoding RNA competing interactions reveals the potential role in progression of human gastric Cancer. International Journal of Oncology. 2016;**48**:1965-1976

[15] Shin H, Kim KH, Song C, et al. Electrodiagnosis support system for localizing neural injury in an upper limb. Journal of the American Medical Informatics Association. 2010; **17**:345-347

Edited by Vijayalakshmi Kakulapati

Open data is freely usable, reusable, or redistributable by anybody, provided there are safeguards in place that protect the data's integrity and transparency. This book describes how data retrieved from public open data repositories can improve the learning qualities of digital networking, particularly performance and reliability. Chapters address such topics as knowledge extraction, Open Government Data (OGD), public dashboards, intrusion detection, and artificial intelligence in healthcare.

Published in London, UK

IntechOpen

ISBN 978-1-83968-317-6

9 781839 683176