Harmonization and Categorization of Metrics and Criteria for Evaluation of Recommender Systems in Healthcare From Dual Perspectives

Adekunle Oluseyi Afolabi, University of Eastern Finland, School of Computing, Kuopio, Finland

https://orcid.org/0000-0003-4446-589X

Pekka Toivanen, University of Eastern Finland, School of Computing, Kuopio, Finland

ABSTRACT

Researchers' choice of metrics and criteria in evaluating recommender systems depends on what the researcher feels is popular among other researchers, or sometimes based on the objective of the research. There is no harmonized set of criteria and metrics that can be referenced when evaluating recommender systems in healthcare. In this article, a set of metrics and criteria are harmonized and categorized as a guide for evaluating recommender systems. By means of an online survey, the opinions of forty-four experienced researchers and other stakeholders from eight countries and four continents were sought on the relevance of identified metrics and criteria. Analysis of the results show speed and timeliness are at the top. Topping the list of criteria is the provision of information that will guide users to useful decisions. The result is presented from two logical perspectives. Four categories are then identified as a useful guide for evaluating recommender systems.

KEYWORDS

Categorization, Criteria, Evaluation, Harmonization, Healthcare, Metrics, Recommender Systems

INTRODUCTION

Recommender systems have played an important role in people's search for articles of interest, thus making searching an easy and enjoyable exercise. But more importantly, giving personalized recommendations is the beauty of any recommender system. Effectiveness of a recommender system in this regard essentially depends on the quality of recommendations obtained from it. This no doubt, and explains why many authors have placed high degree of importance on determining the quality of recommendations obtained. This can be judged using several metrics and criteria. In addition to using metrics and criteria, effectiveness can also be measured from users and system perspectives (Cremonesi et al., 2013). It is reasonable that, for an effective evaluation of recommenders systems to take place, this must be done from both perspectives. Evaluation of a recommender system is incomplete without evaluating user experiences in using a certain recommender system, no matter how

effective the system might seem from system's perspective. This perspective will involve considering evaluation metrics. There have been attempts to produce an evaluation framework that combines system-centric and user-centric evaluation methods (Kavu et al., 2017). This seems to be a way of not leaving out any part in trying to get a clear picture of effectiveness or quality of a recommender system. From these perspectives, harmonized metrics and criteria for carrying out such an evaluation is desirable, especially if they can be categorized. It has been affirmed that performing an evaluation before deploying a recommender system is very essential (Fazeli et al., 2017). This therefore requires a set of metrics and criteria that will provide an agreeable result among relevant stakeholders. While our work focuses on healthcare, there is general lack of uniformity in the metrics for evaluation of recommender systems because many of them are in use (Del Olmo & Gaudioso, 2008). A number of these metrics have been used by individuals based on their perceived appropriateness to their work. Therefore, harmonizing and categorizing them will provide a uniform platform for evaluating recommender systems regardless of the objectives of an individual researcher or stakeholder. Valdez et al. (2016) listed evaluation of a recommender system as one of the important steps. It can then be deduced that thinking about developing a recommendation system should go along with thinking about making it meet evaluation criteria. This obviously is better done as part of requirements and design of such a recommender system.

In justifying their choice of metrics and proving their appropriateness, many authors have chosen to evaluate their work using two or three metrics. Accuracy has been the most popular metric for evaluating recommender systems and this sometimes is based on the algorithm used (Del Olmo & Gaudioso, 2008; Sokolova & Lapalme, 2009; Rebouças Filho et al., 2017; Rodrigues et al., 2018). Related closely to this are Precision, recall and sort priority which have also been identified as the common evaluation metrics for evaluation of recommender systems. (Zhong & Li, 2016; Moreira et al., 2018) However, some have argued that there is the need to look beyond accuracy (Vargas & Castells, 2011; Wu et al., 2012; He et al., 2016). It has also been reported that quality of recommendation is an important metric that deserves an attention rather than just the predictive accuracy of algorithms (Ge et al., 2008). While accuracy has been given wider publicity as a metric for determining the effectiveness of a recommender system, some researchers have opined that it is not a good measure of quality perceived by the users (Cremonesi et al., 2011) but other metrics such as serendipity and coverage (Ge et al., 2008) along with confidence (Duan et al., 2011) have more roles to play in satisfying users recommendations desired other than the accuracy. While it may be true to have surprise recommendations that fit the needs of a particular user (serendipity), delivering health recommendations as a surprise to the user should be done with a lot of caution. In an ongoing project, Recommendations Sharing Community for Aged and Chronically Ill People (ReSCAP), concern is more about addressing the specific needs of the individuals within this community in order to reduce time spent in searching, and because of the chronic nature of the ailments of the individuals involved. In this type of project, timeliness as an important factor to be considered can enhance accuracy (Zhang et al., 2017).

In this work, Criteria is defined as a set of standards by which the effectiveness and quality of a recommender system is determined while Metrics are a set of measurable attributes that can be used to determine performance of a recommender system. Evaluation is the process of judging the quality, importance, or value of a recommender system. (Afolabi and Toivanen, 2018). The main objective of this work is to harmonize and categorize metrics and criteria for evaluating recommender systems and not how they are measured. Several authors have dealt with this other objectives of this work includes determination of relevant metrics and criteria for evaluation of recommender systems in healthcare, determination of the degree of importance of each metric or criterion to the stakeholders, and ascertaining the relationship between degree of important of the metrics and criteria and the institution where stakeholders carry out their research activities, After analyzing the result of an online survey, speed, timeliness, accuracy, and usefulness are on the top of metrics on the degree of importance to stakeholders. For criteria however, providing information that will guide users to

useful decisions, Personalization- ability to infer the needs of each person and then satisfy those needs, and taking into cognizance the progressions of some ailments and providing recommendations of immediate and future occurrences based on this follow in that order on degree of importance as indicated by the result of the analysis. This is the first attempt at harmonizing and categorizing metrics and criteria for evaluation of recommender systems known to the authors. Advantages of harmonizing and categorizing metrics and criteria for evaluation of recommender systems in healthcare are the ability to provide guidance for researchers during evaluation, help developers during development processes, provide a leap towards a universally acceptable checklist for evaluation, and ensure flexibility in the use of these metrics and criteria. With the presentation of two different perspectives and four different priority sets, the flexibility of the checklist is enhanced.

The remainder of this paper is organized as follows. Following this section is the background to the study, detailing the current and previous contributions to the subject. Research design follows after which the result of the study is presented. The discussion is thereafter presented and finally the conclusion and future work.

BACKGROUND

Recommender systems have come to play a crucial role in the life of users and other stakeholders. As a result, many sectors of economy have implemented recommender systems to provide effective services, although its use is popularly known in e-commerce. Effectiveness of these recommender systems are judged by evaluating them against known metrics and criteria. Unfortunately, however, according to our previous work, some of the recommender systems in healthcare have not been evaluated by researchers and developers. (Afolabi et al., 2015, Afolabi & Toivanen, 2018)) Those that have evaluated some of these recommender systems have done so using metrics that are popular, that have been used by famous researchers (Zhong & Li, 2016), or that appeal to them. This they have done using different metrics and criteria for evaluation without any collectively acceptable guide or standard. While it may appear from researchers' point of view that this method has no deficiency or that it is easy to use, it is not totally healthy for the field of recommender systems development and for users' community. Therefore, producing harmonized sets of metrics and criteria for valuating effectiveness of recommender systems is a reasonable investment in this research domain (Afolabi et al., 2015). Realizing this problem, this article presents a new viable method of using metrics and criteria in evaluating recommender systems. One major drawback that may come from using individual approach in selecting metrics and criteria for evaluation is the difficulty that will arise from measuring and comparing performance of recommender systems uniformly across regional and international boundaries, as is common in scientific communities. Thus, there is need to have a harmonized set of metrics and criteria for evaluation of recommender systems from the perspectives of different involving stakeholders. This work is essentially based on producing harmonized metrics and criteria for evaluation of recommender systems from the standpoint of researchers and other relevant stakeholders and categorizing them. The relevance of metrics to effectiveness of systems used in healthcare can be seen from how they have been considered in evaluating decision supports systems in addition to image processing systems such as Lungs, tissues segmentation, identification and classification of various diseases (Rebouças Filho et al., 2016; Reboucas Filho, 2017), stroke diagnosis (Rebouças Filho et al., 2017), Parkinson diseasedisgnosis (de Souza et al., 2018).

In recommender systems evaluation, one metric stands out and has been widely used. That metric is accuracy used both for evaluating recommender systems and other health-related systems such as decision support and diagnostic systems (Sokolova & Lapalme, 2009; Rebouças Filho et al., 2017; Rodrigues et al., 2018). Many researchers working in this field and other health related fields have argued in favour other metrics besides accuracy but also, Specificity among others. (Ramalho et al., 2014; Cavalcanti Neto et al., 2016; He et al., 2016). It is important to state that although decision support systems are different from recommender system on the issue of personalization

of recommendations, drawing examples from such other systems, as has been done here, helps to see the importance researchers have placed on evaluation metrics under consideration in this work.

Another metric of relevance in recommender systems evaluation is timeliness. Timeliness is the degree to which the information contained in each source is continuously updated (Aciar et al., 2007). Several researchers generally agree that timeliness is an important metric in evaluating the quality of recommenders systems and that some activities requires timeliness (O'Mahony and Smyth, 2009; Li and Kao, 2009; Wolfe and Zhang, 2010; Fong et al., 2011; Zhang et al., 2018; Xu et al., 2018), although it may not play a significant role in some instances (Rusell and Yoon, 2008); it has been seen to be as one of essential attributes that determines effectiveness of a recommender system (Aciar et al., 2007) and one of the limiting attributes for online information processing (Zaman et al., 2015). Some have specifically considered timeliness as one of attributes that determine the quality of recommendations (Ciaramella et al., 2010; Lv et al., 2015; Xu et al., 2018)

Furthermore, others have emphasized visibility and relevance based on the premise that for an item to be chosen it must be seen, and relevant items are likely to be chosen (Vargas &. Castells, 2011). In the categorization done in this work, relevance has been defined as relating to usefulness. Relevance metric measures whether the recommendation is useful to the user in performing the particular analytic task (Vartak et al., 2016). Concerning the relevance of an item produced by a recommender system, there is general agreement that this attribute should be considered (Anelli et al., 2017), and that it is an important attribute that is determined by other metrics (Frolov and Oseledets, 2017). Items that are highly rated are considered relevant (Bellogín et al., 2011). Relevance metrics evaluates how far items suggested by a recommender system are relevant for the user and this could be improved with diversification (Shafna and Rajendra, 2017). Others have indicated the importance of relevance in their works (Wu et al., 2016; Kamishima and Akaho. 2017). However, it has also been shown that order of recommendations and popularity of an item is determined by relevance (Vartak et al., 2016; Canamares and Castells, 2018).

Another important metric is adaptability. Adaptability has been defined as changes in recommendations as a result of changes in user's profile (McNee et al., 2006; Said et al., 2012). This operates in an environment where items collections, trends in interest on items change rapidly (Shani and Gunawardana, 2011). Therefore, adaptability is one of the key attributes to be considered (Wu et al., 2012) because it has the potential to make recommender systems perform better and is considered a recommender system personality (McNee et al., 2006; Torrent-Fontbona, 2018). Some recommender systems have been developed to explore this unique attribute (Nathanson et al., 2007; Yang et al., 2018). In other contexts, recommender systems must be adaptable so that they can deliver highly relevant information for patients (Wiesner and Pfeifer, 2014). Other researchers have advocated other metrics such as flexibility, scalability, and modifiability, changeability, efficiensy, and engagement. (Ross et al., 2008; Adams, 2015; He et al., 2016). Some of the metrics here mentioned are related in attribute or meausre to the metrics and criteria extracted for use in this work. In a relevant work, certain criteria have been combined with a framework for evaluation of recommender systems. The resulting model has some 60 questions (a simplified version has 15 questions) used to assess the quality of a recommender system (Pu et al., 2011). The framework is essentially predictive is evaluating user's motivation for choosing a recommender system. Now that relevance of evaluation metrics and criteria for recommender systems in the literature has been examined, attention will be directed to the research design in the following section.

RESEARCH DESIGN

The design of this work was divided into three stages. The first stage is identifying all relevant stakeholders. This includes researchers who have focused on and made important contributions in the field of recommender systems with specific reference to healthcare. Others are physicians, patients, and caregivers. The second stage is identifying the metrics and criteria that they used for evaluation

of their recommender systems and collating them. Since the earlier work of the authors indicates that many works reviewed on recommender systems in healthcare were not evaluated, the searchlight was taken to the domain of recommender systems in general. As a result of this some metrics and criteria were identified. These have been categorized as illustrated in Table 1. In this categorization, main metrics are indicated while other properties, qualities or metrics that are related to the main metrics are included. For example, nearness is related to accuracy which is the main metric. This was arrived at after several sessions of brainstorming. In the third stage, using the metrics and criteria that have been identified (as shown in Tables 2 and 3), an online survey was conducted to determine how much the stakeholders agree with each of the metrics and criteria on their importance in the evaluation of recommender systems. Key researchers who have made remarkable contributions along with other relevant stakeholders from hospital, research institute, and the industry were invited to respond to the online survey. Reminders were sent out at frequent intervals for some three months. The degree of importance was measured using the following categorization; least important, slightly important, somewhat important, very important, and extremely important. This is ranked from 1 as the least important to 5 as extremely important. The decision to use online survey was borne out of the need to make accessibility to the survey by respondents and its administration easier, without geographical boundaries hindering the process. Although questions were initially sent out to respondents, but it was found out that the best way to go would be online as suggested by one of the respondents.

Identifying Interest Groups and Stakeholders

As mentioned earlier, the first task was to identify the stakeholders (also called the interest group). This is done by identifying what they do and where they work. The interest group and the stakeholders have been categorized into the following:

- Researchers – Academic, industry
- Implementers – Industry, Government, Hospital
- Users–Hospital, Government, Hospice, Home

Table 1. Categorization of metrics for evaluation of recommender systems

S/N	Metric	Related Metrics/ Qualities/ Properties	Reference (Authors)
1	Speed		Pre-survey study/Authors' experience
2	Accuracy	Nearness, (precision, recall, fallout- Methods) relevance	Breese et al., 1998; Herlocker et al., 1999; Shardanand & Maes, 1995; Wiesner &Pfeifer, 2014; Ragone et al., 2017
3	Relevance	Precision	Zhong, & Li, 2016
4	Usefulness	Quality, closeness, utility, diversity, relevance	Del Olmo & Gaudioso, 2008; Ge at al., 2010
5	Specificity		Pre-survey study/Authors' experience Ramalho et al., 2014
6	Reliability	Confidence, trust	Duan et al., 2011 Herlocker, et al., 1999; Swearingen & Sinha, 2001
7	Adaptability	Scalability, adaptivity	Shani & Gunawardana, 2011; McNee et al., 2006
8	Robustness	Coverage, stability, novelty, serendipity	Ge et al., 2010; O'Mahony et al., 2004; Shani & Gunawardana, 2011
9	Timeliness	Context compatibility	Zhang et al., 2017

Table 2. Table of metrics

S/N	Metric
1	Speed
2	Accuracy
3	Relevance
4	Usefulness
5	Specificity
6	Reliability
7	Adaptability
8	Robustness
9	Timeliness

Table 3. Table describing evaluation criteria and variables describing them

| Variable	Criterion
V10	Personalization- ability to infer the needs of each person and then satisfy those needs
V11	Should be able to perform as many recommendations as possible per second for millions of users and items simultaneously
V12	Should have learning capabilities from user's preferences and current trends
V13	should be able to guide users in a personalized and useful way towards meeting their aims
V14	should take into cognizance the progressions of some ailments and provide recommendations of immediate and future occurrences based on this
V15	Should have the ability to give recommendations real-time
V16	Should provide information that will guide users to useful decisions
V17	Should have the ability to educate
V18	Should be user-based evaluation of the system as a whole
V19	Interesting to the user
V20	Should give predictions that match user's interests

This categorization is not a very rigid one. There could be overlap in the functions of individuals or institutions in these categories. This categorization helped in to reaching out to other stakeholders who are not in the mainstream of research and academic. This overlap of functions is depicted in Figure 1. This implies that academic stakeholders can function as a researcher, implementer, and user depending on what function the individual assumes at any given time.

Important Issues to Be Resolved

At the initial stage, some issues that might seem difficult to resolve and might hinder the progress of the project were considered and reviewed. Therefore, during brainstorming stage, a mock interaction between us and the intended respondents was created.

Figure 1. Interrelationship among stakeholders
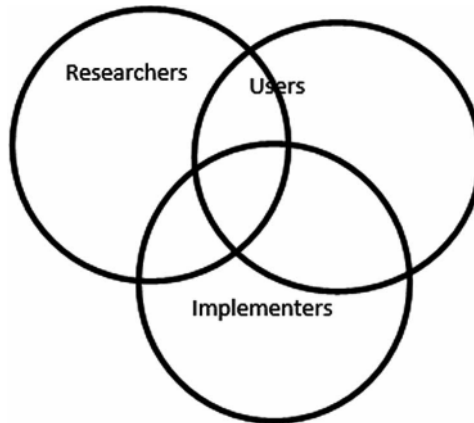
Setting the Objectives

In order to be well guided, there was the need to set the objectives for the research. The following are the objectives set:

1. To determine the relevant metrics for evaluation of recommender systems in healthcare;
2. To determine the degree of importance of each metric to the stakeholders;
3. To determine the relevant criteria for evaluation of recommender systems in healthcare;
4. To determine the degree of importance of each criterion to the stakeholders;
5. To determine the relationship between degree of important of the metrics and the establishment (job) of stakeholders;
6. To determine the relationship between degree of important of the criteria and the establishment of stakeholders (what he does, where he works);
7. To produce and categorize harmonized metrics for evaluation of recommender systems in healthcare;
8. To produce and categorize harmonized criteria for evaluation of recommender systems in healthcare.

RESULT

By means of an online survey, opinions of forty-four experienced researchers and other stakeholders from eight countries and four continents were sought on the relevance of identified metrics and criteria. At the expiration of the survey, the data collected was subjected to analysis in order to decipher the thinking of various stakeholders who responded and come up with some form of harmonized set of metrics and criteria for evaluation of recommenders systems in healthcare. The response was considered based on each of the objectives of this work, indicated as follows.

Determining the Relevant Metrics for Evaluation of Recommender Systems in Healthcare and Their Degree of Importance

From Table 4, it can be seen that, 25 respondents representing 56.8% feel that speed is extremely important while 24 respondents representing 54.5% choose timeliness as extremely important. Accuracy and timeliness have 19 and 23 individuals respectively who consider them extremely important. These represent 43.2% and 52.3%, respectively. The mean for these top four metrics:

Table 4. Frequency and mean table for metrics

Metrics	Least Important	Slightly Important	Somewhat Important	Very Important	Extremely Important	Mean	SD
Speed	1(2.3)	1(2.3)	2(4.5)	15(34.1)	25 (56.8)	4.41	0.87
Timeliness	1(2.3)	1(2.3)	5(11.4)	13(29.5)	24(54.5)	4.32	0.93
Accuracy	1(2.3)	6(13.6)	1(2.3)	17(38.6)	19(43.2)	4.23	0.87
Usefulness	1(2.3)	1(2.3)	8(18.2)	11(25.0)	23(52.3)	4.23	0.99
Relevance	1(2.3)	1(2.3)	8(18.2)	25 (56.8)	9(20.5)	3.98	0.71
Adaptability	1(2.3)	5(11.4)	8(18.2)	13(29.5)	15(34.1)	3.86	1.12
Specificity	1(2.3)	2(4.5)	16(36.4)	17(38.6)	7(15.9)	3.63	0.90
Reliability	1(2.3)	4(9.1)	15(34.1)	17(38.6)	6(13.6)	3.53	0.93
Robustness	2(4.5)	8(18.2)	12 (27.3)	13(29.5)	9 (20.5)	3.43	1.15

speed, timeliness, accuracy, and usefulness are 4.41, 4.32, 4.23, and 4.23 in that order. Adaptability also has 15 respondents which represents 34.1% choosing the metric as extremely important while relevance, specificity, reliability, and robustness have been chosen to be very important with 25, 13, 17, and 17 respondents respectively representing 56.8%, 29.5%, 38.6%, and 38.6% in that order. Their means are relevance (3.98), specificity (3.63), reliability (3.53), and robustness (3.43) while the mean for adaptability is 3.86.

Determining the Relationship Between Degree of Importance of the Metrics and Criteria and the Establishment (job) of Stakeholders

Table 5 shows the relationships between these variables and the choice of rank of the metrics across different continents where the respondents work as follows. For very important rank a total of 63.6% for Africa, 4.5% for Asia, 15.9% for North America, and 236.4% for Europe were gotten. As for the

Table 5. Cross tabulation of metrics against the country where stakeholders work

Degree of Importance			Continent				Total
			Africa	Asia	North America	Europe	
Least Important		Count	0	0	0	9	9
		% of Total	0.0%	0.0%	0.0%	20.5%	20.5%
Slightly Important		Count	3	1	2	17	23
		% of Total	6.8%	2.3%	4.5%	38.6%	52.3%
Somewhat Important		Count	6	4	7	63	80
		% of Total	13.6%	9.1%	15.9%	143.2%	181.8%
Very Important		Count	28	2	7	104	141
		% of Total	63.6%	4.5%	15.9%	236.4%	320.5%
Extremely Important		Count	26	2	20	89	137
		% of Total	59.1%	4.5%	45.5%	202.3%	311.4%
Total		Count	7	1	4	32	44
		% of Total	15.9%	2.3%	9.1%	72.7%	100.0%

extremely important rank, 59.1% for Africa, 4.5% for Asia, 45,5% for North America and 202.3% for Europe. Table 5 clearly indicates the result for the kind of institutions where they work and the bearing this may have on their choice of rank for the degree of relevance of these metrics. For very important 20.5% comes from research institute, 256.8% from University, 25% from hospital, and 18.2% from the industry. The result for extremely important shows 20.5%, for research institute, 247.7% from the university, 34.1% from the hospital, and 9.1% from the industry. The total percentage for very important rank is 320.5% while 311.4% is for the extremely important.

The scatter graphs in Figures 2a – 2c, show the relationship between each of the metrics and the institutions where respondents work. The essence of this is to ascertain if the choice of respondents on the metrics has anything to do with their places of work. For instance, is it possible that respondents working in the hospital will favor robustness over other metrics or the patients will favor timeliness over other metrics? The scatter dots simply show that no positive relationship exists between the choice of metrics by the respondents and the institution where they practice. Therefore, it may be concluded that the choice of metrics by the stakeholders does not have any relationship with the institutions where they practice but likely depends on individual. However, as the graphs have shown, there exists no such relationship. Although, this result seems interesting but further research will be conducted that will be institution-based which will help confirm the current finding.

Determining the Relevant Criteria for Evaluation of Recommender Systems in Healthcare and Their Degree of Importance

The relevance of the highlighted criteria in the evaluation of recommender systems in healthcare can be seen from Table 6. At the top of the table of mean is the recommender system's ability to provide information that will guide users to useful decisions with the mean of 4.43 and a count of 29 (65.9%) respondents feeling it is extremely imptable 7

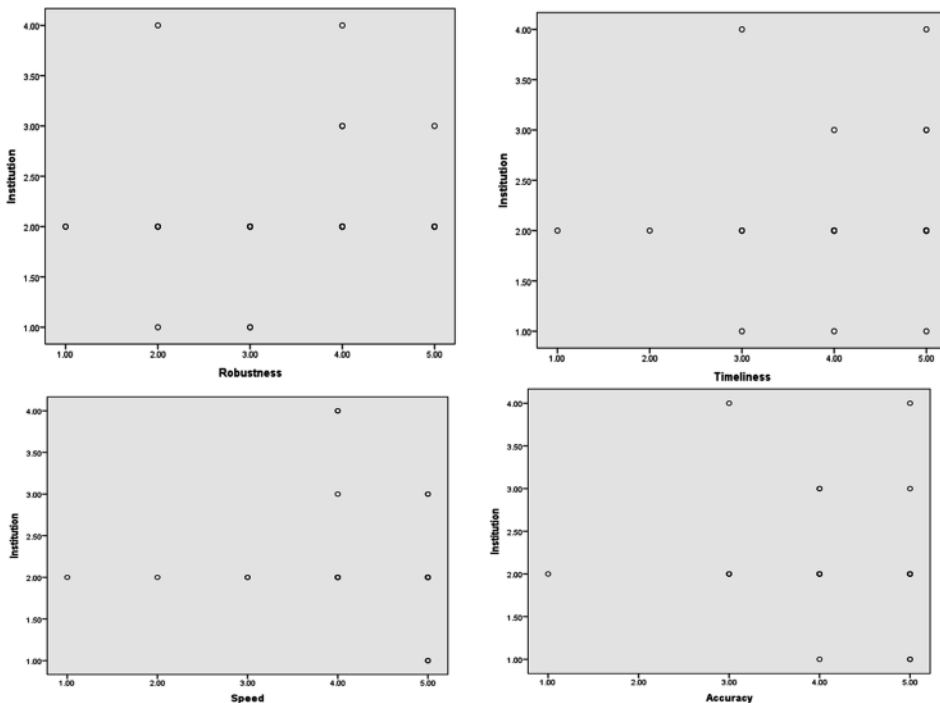Figure 2a. The scatter line graph of institutions against metrics (the dots show no agreement)

Figure 2b. The scatter line graph of institutions against metrics (the dots shows no agreement)
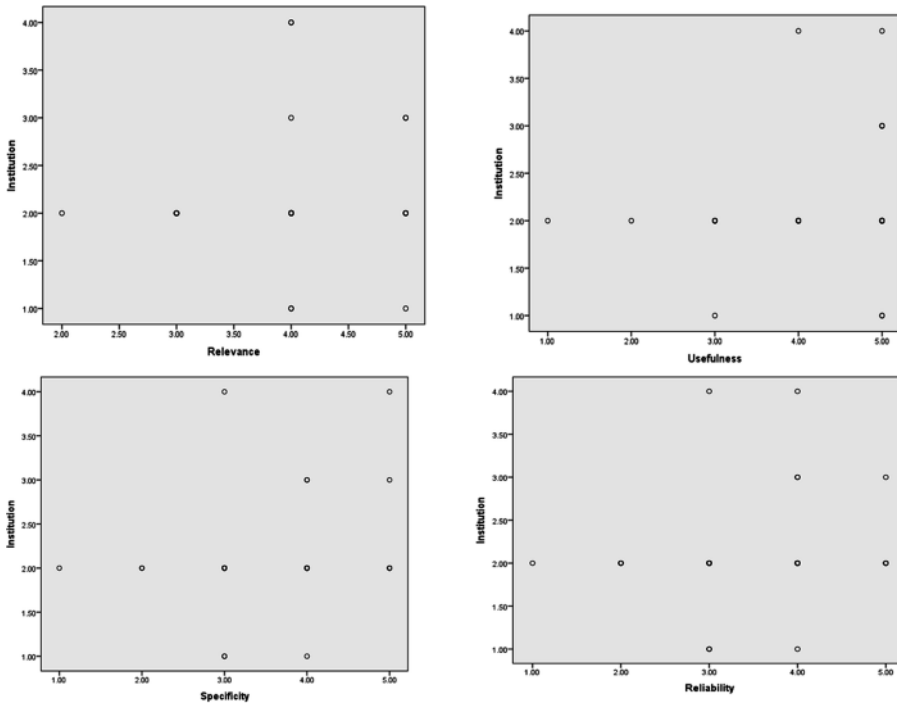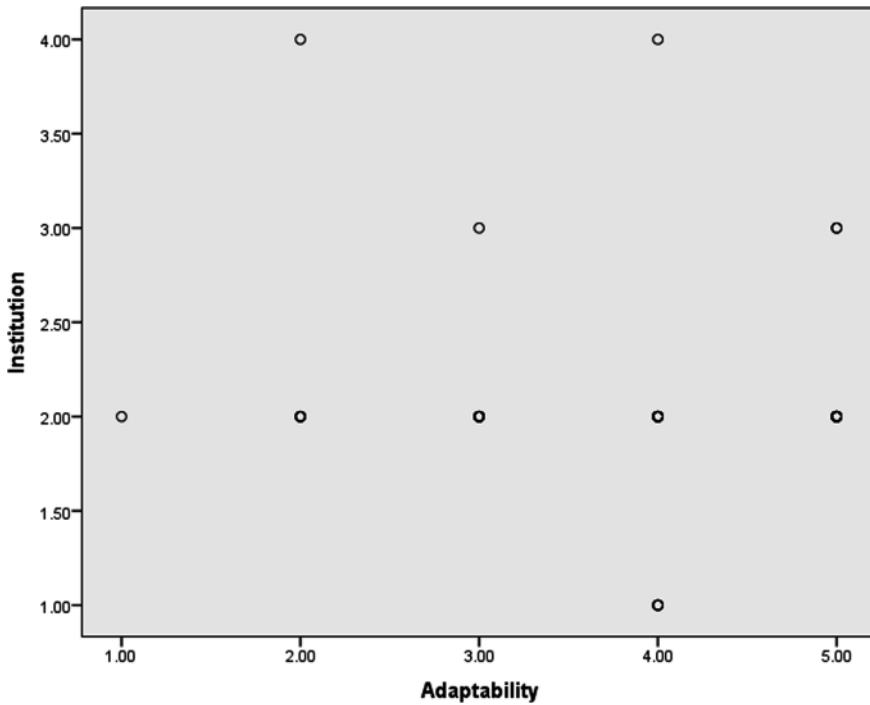

Figure 2c. The scatter line graph of institutions against metrics (the dots show no agreement)

Table 6. Frequency and mean table for criteria

Metrics	Least Important	Slightly Important	Somewhat Important	Very Important	Extremely Important	Mean	SD
V16	1(2.3)	1(2.3)	5(11.4)	8(18.2)	29(65.9)	4.43	0.95
V10	1(2.3)	1(2.3)	9(20.5)	17(38.6)	16(36.4)	4.05	0.94
V14	1(2.3)	2(4.5)	10(22.7)	17(38.6)	14(31.8)	4.00	0.87
V13	1(2.3)	2(4.5)	8(18.2)	19(43.2)	13(29.5)	3.95	0.95
V12	2(4.5)	1(2.3)	6(13.6)	27(61.4)	9(20.5)	3.93	0.87
V20	1(2.3)	3(6.8)	11(25.0)	16(36.4)	13(29.5)	3.91	0.92
V17	1(2.3)	3(6.8)	8(18.2)	19(43.2)	12(27.3)	3.88	0.98
V19	2(4.5)	5(11.4)	7(15.9)	14(31.8)	14(31.8)	3.79	1.18
V15	3(6.8)	5(11.4)	8(18.2)	13(29.5)	15(34.1)	3.73	1.25
V18	4(9.1)	17(38.6)	12(27.3)	10(22.7)	10(22.7)	3.65	0.95
V11	2(4.5)	7(15.9)	18(40.9)	10(22.7)	6(13.6)	3.26	1.05

ortant. Three other criteria on the top of the mean table are Personalization- ability to infer the needs of each person and then satisfy those needs (4.05), ability to take into cognizance the progressions of some ailments and provide recommendations of immediate and future occurrences based on this (4.00), ability to be able to guide users in a personalized and useful way towards meeting their aims (3.95). on the other hand, the ability to provide information that will guide users to useful decisions, and the ability to give recommendations real-time have extremely important as degree of importance to be highest, 29 (65.9%), and 15 (34.1%) respectively. However, 14 (31.8%) of the respondents choose both very important and extremely important for the ability of a recommender system to give recommendations that are interesting to the user.

DISCUSSION

While waiting for responses of the invited stakeholders a lot of anticipation in their diversity of opinions and the number of responses that would be obtained was built. As a result, when finally confronted with some feedback issues, it was not completely surprising. The following feedback issues were recorded from the online survey conducted and despite sending out repeated reminders to the invited stakeholders

Some of the invited respondents did not respond due to the following reasons:

1. Time to respond;
2. No interest any longer in the field;
3. Left academic and no longer active in the field (a couple of these specifically wrote back to mention this);
4. No interest in the research.

Determining the Relevant Metrics for Evaluation of Recommender Systems in Healthcare and Their Degree of Importance

The result of the analysis of the responses of the stakeholders shows the first four metrics that are considered relevant for the evaluation of recommender systems are speed, timeliness, accuracy and usefulness. When information is needed the users consider the time it takes to fetch the information

Table 7. Continental response cross tabulation for evaluation criteria

Degree of Importance			Continent				Total
			Africa	Asia	North America	Europe	
Least Important		Count	1	1	0	11	13
		% of Total	2.3%	2.3%	0.0%	25.0%	29.5%
Slightly Important		Count	3	0	5	25	33
		% of Total	6.8%	0.0%	11.4%	56.8%	75.0%
Somewhat Important		Count	9	4	6	88	107
		% of Total	20.5%	9.1%	13.6%	200.0%	243.2%
Very Important		Count	29	3	13	127	172
		% of Total	65.9%	6.8%	29.5%	288.6%	390.9%
Extremely Important		Count	34	3	19	95	151
		% of Total	77.3%	6.8%	43.2%	215.9%	343.2%
Total		Count	7	1	4	32	44
		% of Total	15.9%	2.3%	9.1%	72.7%	100.0%

Table 8. Institutional response cross tabulation for evaluation criteria

Degree of Importance			Institution				Total
			Research	University	Hospital	Industry	
Least Important		Count	1	12	0	0	13
		% of Total	2.3%	27.3%	0.0%	0.0%	29.5%
Slightly Important		Count	0	30	1	2	33
		% of Total	0.0%	68.2%	2.3%	4.5%	75.0%
Somewhat Important		Count	10	82	6	9	107
		% of Total	22.7%	186.4%	13.6%	20.5%	243.2%
Very Important		Count	14	136	13	9	172
		% of Total	31.8%	309.1%	29.5%	20.5%	390.9%
Extremely Important		Count	8	128	13	2	151
		% of Total	18.2%	290.9%	29.5%	4.5%	343.2%
Total		Count	3	36	3	2	44
		% of Total	6.8%	81.8%	6.8%	4.5%	100.0%

because time is an essential commodity. This may account for the reason for according time a high priority. Having timeliness in the top four priority list is not surprising however because everyone is concerned about getting information in a timely manner. In healthcare, this is important because getting information in a timely manner may save situation from deteriorating and may even result in saving lives. Accuracy no doubt should be important in healthcare because for any recommendations to achieve their desired objectives, then the accuracy cannot be compromised. In other domains, accuracy may not mean so much, as many researchers have opined, however, this is not the case in

healthcare where lives are at stake. Therefore, it is not surprising to see usefulness making the first four in the list of relevant metrics for evaluating recommender systems. In the candid opinion of the authors, any recommendations that do not prove useful to the user have no place in healthcare. One of the main reasons for seeking recommendations in healthcare is to find it useful. While it can be said that determining this may be a little difficult, since the determining whether a recommendation is useful or not depends largely on the judgment of the user, this metric is at the heart of rating a recommender system's effectiveness.

The result however shows other metrics not finding a place in the top list and this is appalling to the authors because of their perceived relevance in healthcare. One instance of this is reliability. This metric shows up as second to the last on the mean table. In the pre-survey opinion conducted we had thought that for any recommendations to be acceptable to the users, especially in healthcare, the source of such a recommendation must be reliable. Looking at this metric from this perspective, it seems that not user will be inclined to accept a recommendation whose source cannot be judged to be reliable. Although this result is not dismissing reliability as relevant, places it in the position that is contrary to what the authors had thought.

Determining the Relevant Criteria for Evaluation of Recommender Systems in Healthcare and Their Degree of Importance

Criteria for evaluating recommender systems are very important and are somehow related to the metrics in some ways. The result of the analysis however, has sprung surprises though. The top four criteria favored by respondents are providing information that will guide users to useful decisions, personalization-ability to infer the needs of each person and then satisfy those needs, taking into cognizance the progressions of some ailments and providing recommendations of immediate and future occurrences based on this, and able to guide users in a personalized and useful way towards meeting their aims. Most often, users dig out information when they are faced with decision making. As a result of their limited skill or know-how on the subject matter, they are open to help from experts in the field. This thus informs their decision to seek recommendations. Such information should not fail to guide them to useful decisions then.

It is not surprising however, that high priority is accorded personalization. Personalization has been the marked difference between the function of a recommender system and an ordinary predictive system. The result of the survey has just confirmed this. A recommender system should be able to infer the needs of the user and then make provisions for satisfying them. This therefore suggests the undisputable ability of a recommender system to make personalized recommendations to a particular user.

Of vital interest is the indication from the survey that suggests the need for a recommender system to be able to give predictive recommendations. This feature will find a great usefulness in healthcare, especially for chronic and progressive illnesses. When a recommender system is able to provide predictive recommendations, by means of a powerful artificial intelligence process, illnesses that are progressive in nature can be managed with less stress. Many individuals are faced with a situation in which they may not know what to expect as their illnesses progress, and thus may not be prepared for the challenges that come along with such. The ability of a recommender system to make recommendations that will address these challenges is important.

Determining the Relationship Between Degree of Important of the Metrics and Criteria and the Institutions Where Stakeholders Work

Although the result from the analysis of the response does indicate that continent where the respondents carry out their research and the institution where they work have no effect on their responses, it is obvious from the result that most stakeholders who had responded come from Europe. This may be due to the fact that most of the researchers whose work we had reviewed were from Europe, although we invited many researchers and other stakeholders from other parts of the world. Besides, other factors

that we have highlighted above may be responsible for this. The results indicated in Tables 5 and 7 clearly show no significant departure from each other. As shown, a great number of respondents are of the opinion that the metrics are very relevant and indicated by the total percentage of those who chose very important and extremely important.

Harmonizing the Metrics

In harmonizing the metrics, logical reasoning and rule of thumbs have been adopted. As a result, the following have been deduced from the result of the survey. Table 9 indicates the order of priority by the majority of the respondents and by the mean from the analysis.

From the table, it can be seen that there is a thin line between somewhat important and very important for the following metrics, Specificity (somewhat important- 16, very important - 17), Robustness (somewhat important- 12, very important - 13), and Reliability (somewhat important- 15, very important - 17), For accuracy however, there is a little gap between very important (17 or 38.6%) and extremely important (19 or 43.2%). This little gap between the two top degrees of importance may be an indication of some researchers' shift of attention from accuracy to other metrics as discussed in the background section.

Harmonizing the Criteria

As indicated under harmonizing metrics, no rigid rules have been followed in harmonizing the criteria as well. However, since much emphasis has not been placed on criteria in the past and in the literature, the same logical reasoning and rule of thumbs have been applied in deducing the criteria from the analysis of the result of the survey. Table 10 gives the summary of this.

The little gap can be observed in the following criteria in the degree of importance. Personalization, between very important (17 or 38.6%) and extremely important (16 0r 36.4%), and the ability to give recommendations real-time, between very important (13 or 29.5%) and extremely important (15 or 34.1%) have little gaps between the two top degrees of importance. However, for Interesting to the user, there is a tie between very important and extremely important with each having 14 or 31.8%.

Layman's Logical Perspective Versus Statistics Logical Perspective

The interpretations of the analysis of the result, though not rigid, are carefully illustrated in Figures 3 and 4. Interpretations have been presented from two distinct perspectives namely the Layman's logical interpretation and the statistics logical interpretation. In strongly held opinion by the authors

Table 9. Order of importance of the metrics

Metrics	Degree of Importance (Highest Count)	Count (%)	Mean	SD
Speed	Extremely important	25 (56.8)	4.41	0.87
Timeliness	Extremely important	24(54.5)	4.32	0.93
Accuracy	Extremely Important	19(43.2)	4.23	0.87
Usefulness	Extremely Important	23(52.3)	4.23	0.99
Relevance	Very Important	25 (56.8)	3.98	0.71
Adaptability	Extremely Important	15(34.1)	3.86	1.12
Specificity	Very Important	17(38.6)	3.63	0.90
Reliability	Very Important	17(38.6)	3.53	0.93
Robustness	Very Important	13(29.5)	3.43	1.15

Table 10. Order of importance of criteria

Criterion	Degree of Importance (Highest Count	Count (%)	Mean	SD
Should provide information that will guide users to useful decisions	Extremely Important	29(65.9)	4.43	0.95
Personalization- ability to infer the needs of each person and then satisfy those needs	Very Important	17(38.6))	4.05	0.94
Should take into cognizance the progressions of some ailments and provide recommendations of immediate and future occurrences based on this	Very Important	17(38.6)	4.00	0.87
should be able to guide users in a personalized and useful way towards meeting their aims	Very Important	19(43.2	3.95	0.95
should have learning capabilities from user's preferences and current trends	Very Important	27(61.4)	3.93	0.87
Should give predictions that match user's interests	Very Important	16(36.4))	3.91	0.92
Should have the ability to educate	Very Important	19(43.2)	3.88	0.98
Interesting to the user	Extremely Important / Very Important	14(31.8)	3.79	1.18
Should have the ability to give recommendations real-time.	Extremely Important	15(34.1)	3.73	1.25
Should be user-based evaluation of the system as a whole	Slightly Important	17(38.6)	3.65	0.95
Should be able to perform as many recommendations as possible per second for millions of users and items simultaneously	Somewhat important	18(40.9)	3.26	1.05

the two perspectives are logical. Simple and logical explanation for the perspective illustrated with Figure 3 is that, during evaluation the metrics in the tray labeled extremely important are to be given the highest priority before any other metrics in the remaining trays. On the other hand, it could also mean that during evaluation, the metrics in the tray labeled extremely important are to be used first, thereafter the metrics in the tray labeled very important. This order should be followed as indicated by the big arrow.

It may also be possible to use the statistical perspective to determine the evaluation metrics. This simply and logically implies that order of importance of the metrics is as indicated by the arrow. This stepwise order shows weight of each of the metrics in the evaluation process. One area that will need further light is how much of these metrics should be used before a recommender system can be said to have met the required standard. More research efforts and some form of standardization may be required. Another one may be considering factors that may make it mandatory to alter the order presented above. These are interesting subjects for research that can be explored in the nearest future. Figure 4 shows the evaluation order based on statistical means.

The approach adopted in presenting and explaining the result of analysis of metrics has also been adopted for presenting and explaining that of the criteria, Figure 5 gives a simple illustration of the order of criteria in a Layman's perspective that is logical and easy to understand. Whether to include the two criteria at the bottom of the staircase may depend on other factors such as the purpose of the recommender systems and the institution.

Categorization of Metrics and Criteria

Following the analysis of and logically reflecting on the result from the survey, there was the need to categorize the metrics and criteria for use in evaluation of recommender systems

Figure 3. The evaluation order based on the degree of importance (Layman's Logical interpretation)
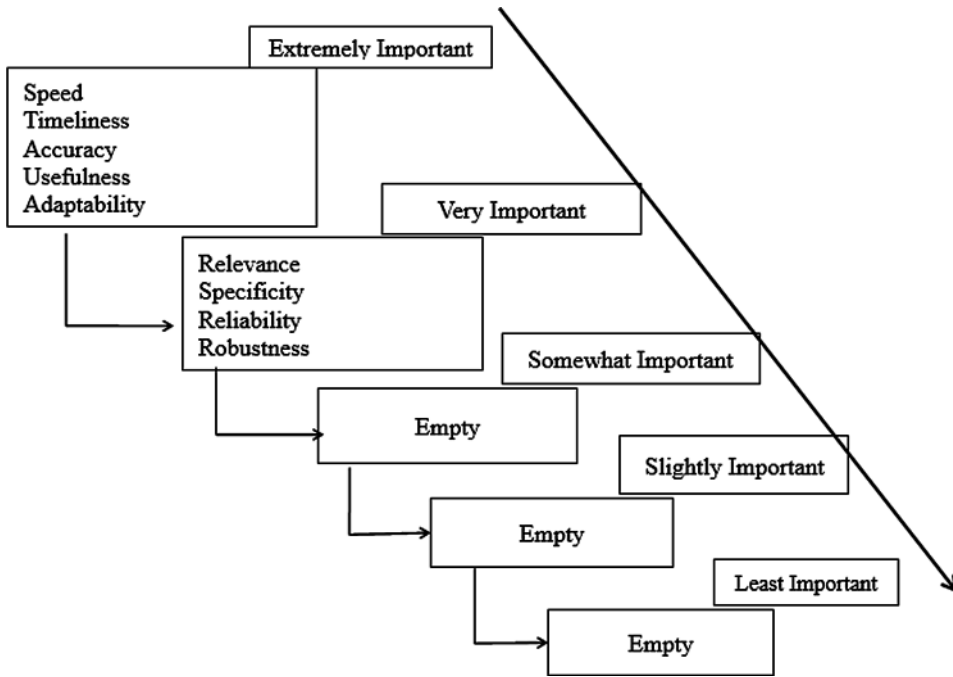

Figure 4. The evaluation order based on statistical means (Statistics Logical interpretation)
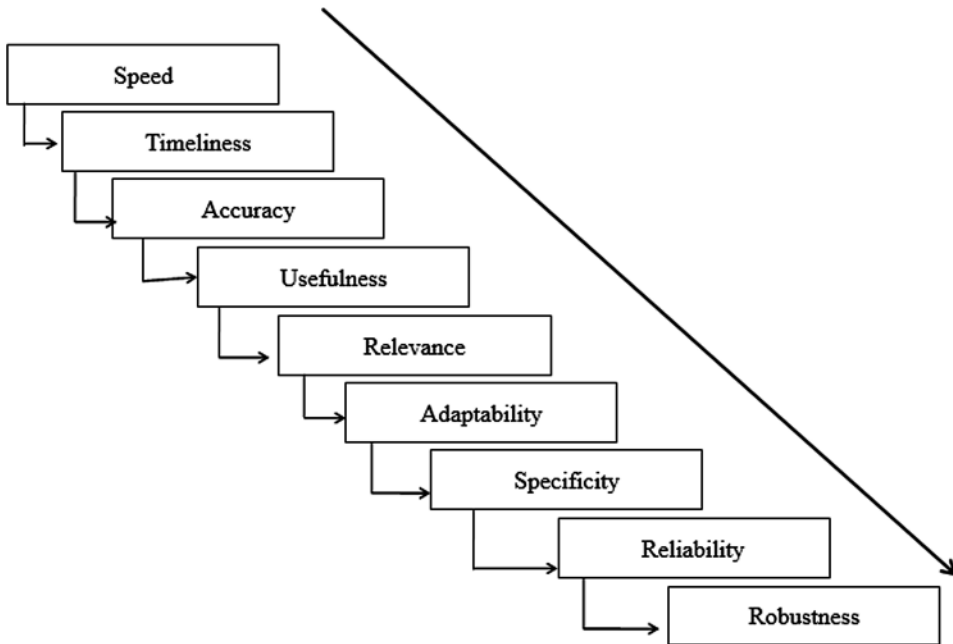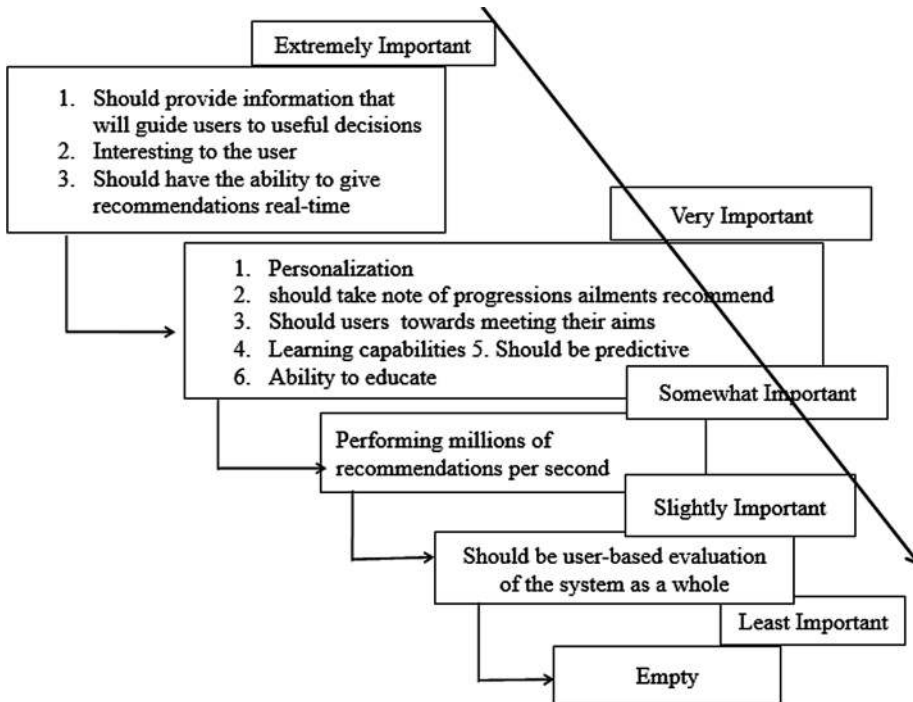
Figure 5. The evaluation order based on the degree of importance of criteria (Layman's interpretation)

in healthcare. Since this is the first attempt at doing this, no rigid rules are provided for this purpose, however, it is believed that this will prove useful in evaluation process and will lay a solid foundation for this process.

Looking at the metrics and criteria critically from the statistics logical perspective, the presentation and interpretation for use becomes clear and straightforward. However, an analytical examination of the Layman's logical perspective takes us a step further. Using the stepwise degree of importance depicted by the Layman's logical perspective, the metrics and criteria can be categorized into four different groups, namely high priority, low priority, weak priority, and conditional high priority. Candidly, we are persuaded of the logical soundness of this categorization. Each of the categorization is briefly described as follows. It should be noted that, by this categorization, the authors are not recommending that some of these metrics and criteria be used in the evaluation because they have been grouped as high priority while others not be used because of low or weak priority. For comprehensive evaluation of any recommender systems it may be necessary to use all metrics and criteria.

High Priority Set

High priority set includes those metrics and criteria that are extremely important to the stakeholders. These seem to essentially determine the effectiveness of a recommender system, and hence its acceptability by the users. From Figure 3, examples of metrics in this category are speed, timeliness, accuracy, usefulness and adaptability. Examples of criteria from Figure 5 are provision of information that will guide users to useful decisions, ability to give recommendations that will be Interesting to the user, and the ability to give recommendations real-time. Metrics and criteria that are so grouped are expected to be given high priority during evaluation.

Low Priority Set

As the name suggests, metrics and criteria grouped under low priority are given secondary priority during evaluation. They are to be considered after considering those belonging to the high priority set. This does not indicate that they are less important but that they fall behind those in the class of high priority. Examples of low priority set from Figure 3 are relevance, specificity, reliability, and robustness while the examples from Figure 5 are Personalization, ability to take note of progressions ailments and recommend, ability to help users towards meeting their aims, ability to have learning capabilities, ability to predictive, and the ability to educate.

Weak Priority Set

Weak priority set contains the metrics and criteria that fall below the low priority category. These are usually found in the somewhat, slightly, and least important ranks in the degree of importance ranking. This however, does not imply that they cannot and must not be used for evaluation. The prerogative of use depends of researchers and the efforts at carrying out comprehensive evaluation. There are no examples that can be used to illustrate this set from Figure 3 because the trays for those ranks are empty. However, in Figure 5, ability to perform millions of recommendations per second, and being user-based evaluation of the system as a whole fall under this set.

Conditional High Priority Set

Metrics and criteria under conditional priority set may not follow any of the above ranks in the degree of importance ranking. Other factors may determine the use of these metrics and criteria in the evaluation of recommender systems. Hence any metric or criterion can be pulled from any of the categories and be used. Such factors as the objective of the recommender systems, the target users, and the functional priority among others can be taken into consideration in determining the metric or criterion that falls into this category. A typical example from Figure 3 is reliability which has been grouped under low priority but may become a conditional high priority in a healthcare recommender system where the reliability of recommendations is extremely important to the users.

Advantages of Our Method

The following advantages can be deduced from the method of harmonization and categorization of metrics and criteria in the evaluation of recommender systems presented in this work.

Provision of Guidance in Evaluation

Before now, there had not been any known checklist of metrics and criteria for evaluation of recommender systems. As stated earlier, researchers' choice of metrics for evaluation of recommender systems are determined by their own research objectives, the kind of recommender system, and sometimes the view of known researchers in the field. With this work, it will be easy for researchers and other stakeholders to have a guide for evaluation of their work.

Helpful for Developers in System Development

Developers will not have to labor too much to determine what their clients will be looking for in any recommender system. Therefore, in their various recommender system development projects, the expected qualities of the system can complement their efforts at obtaining effective requirements for the system development.

Toward a Universal Guide for Evaluation

As of now and to the best of our knowledge, there has not been any universally accepted checklist of metrics and criteria for evaluation of recommender systems in healthcare. Researchers have used a

specific metric or a couple of metrics for evaluation based on what they feel should be evaluated. In this stance, this work has provided a suitable leap in this direction.

Flexibility on the Use of Any of the Perspectives as a Guide

In the authors' careful efforts to provide a guide for the use of metrics and criteria for evaluation of recommender systems, two perspectives namely Layman's Logical Perspective and the Statistic Logic Perspective have been proposed. Any of these perspectives can be used in the evaluation processes. This allows researcher some flexibility to choose the perspective that suits his objective. The categorization process also allows this flexibility.

Clinically Beneficial to Patients and Physicians

Since the list of metrics and criteria presented can be used to evaluate recommender systems before they are used, physicians and other caregivers will find it easy making recommendations to their patients on the choice of recommender systems for effective performance. In addition, the physicians and other caregivers themselves will find this decision easier to make. This is clinically beneficial to both the patient and care provider.

CONCLUSION

The relationship between the acceptability of a recommender system and its performance cannot be separated. Therefore, for a recommender system to pass the test of acceptability and find usefulness among teeming users whose lives and works depend on this system its performance must be well above average. Determining this sometimes may pose some challenges because of the diversity of opinions of researchers and other stakeholders working in this domain. Over the years many metrics have been used to evaluate the effectiveness of recommender systems and these are often based on the sole judgment of a researcher or stakeholder. As a result, there have been a lot of divergent opinions as to what sorts of metric should be used for evaluation of recommender systems.

In view of this, the subject was explored with a clear objective of harmonizing metrics and criteria for evaluating recommender systems in healthcare. Although our primary domain of application of this work is health, we strongly hold that our result can be applied to evaluation of recommender systems generally. The metrics and criteria that we identified were used in an online survey to elicit opinions of stakeholders on their degree of importance in using them for evaluation. The result of the survey was thereafter analyzed.

The result indicates that speed, timeliness, accuracy, and usefulness are at the top of metrics on the degree of importance to stakeholders. Topping the list of criteria however are, providing information that will guide users to useful decisions, Personalization- ability to infer the needs of each person and then satisfy those needs, and taking into cognizance the progressions of some ailments and providing recommendations of immediate and future occurrences based on this.

While this is the first attempt at harmonizing and categorizing metrics and criteria for evaluation of recommender systems known to the authors, the advantages of harmonizing and categorizing metrics and criteria for evaluation of recommender systems in healthcare have been presented. These are the ability to provide guidance for researchers during evaluation, help developers during development processes, provide a leap towards a universally acceptable checklist for evaluation, ensure flexibility in the use of these metrics and criteria, and clinically beneficial to patients and their physicians. With the presentation of two different perspectives and four different priority sets, the flexibility of the checklist is enhanced.

This research was confronted with certain limitations. Although the survey questions are easy to answer and not time consuming, many of the researchers, perhaps because of their busy schedules have limited time to respond despite repeated reminders sent. Another limitation is that not all desired researchers and stakeholders could respond. Some of the contacted researchers are no longer

REFERENCES

Adams, K. M. (2015). Adaptability, Flexibility, Modifiability and Scalability, and Robustness. In *Nonfunctional Requirements in Systems Analysis and Design* (pp. 169–182). Cham: Springer. doi:10.1007/978-3-319-18344-2_9

Afolabi, A. O., & Toivanen, P. (2018). Recommender Systems in Healthcare: Towards Practical implementation of Real-Time Recommendations to Meet the Needs of Modern Caregiving. In Handbook of Research on Emerging Perspectives on Healthcare Information Systems and Informatics (pp. 323-346). Hershey, PA: IGI Global.

Afolabi, A. O., Toivanen, P., Haataja, K., & Mykkänen, J. (2015). Systematic Literature Review on Empirical Results and Practical Implementations of Healthcare Recommender Systems: Lessons Learned and a Novel Proposal. *International Journal of Healthcare Information Systems and Informatics*, *10*(4), 1–21. doi:10.4018/IJHISI.2015100101

Anelli, V. W., Di Noia, T., Lops, P., & Di Sciascio, E. (2017, August). Feature Factorization for top-n Recommendation: from item rating to features relevance. In *Proceedings of the 1st Workshop on Intelligent Recommender Systems by Knowledge Transfer & Learning co-located with ACM Conference on Recommender Systems (RecSys 2017)*, Como, Italy (pp. 16-21).

Bellogin, A., Castells, P., & Cantador, I. (2011, October). Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 333-336). ACM. doi:10.1145/2043932.2043996

Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 43-52). Morgan Kaufmann Publishers Inc.

Cañamares, R., & Castells, P. (2018). Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems.

Ciaramella, A., Cimino, M. G., Lazzerini, B., & Marcelloni, F. (2010, November). Using context history to personalize a resource recommender via a genetic algorithm. In *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 965-970). IEEE. doi:10.1109/ISDA.2010.5687064

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011, September). Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (pp. 152-168). Springer. doi:10.1007/978-3-642-23765-2_11

Cremonesi, P., Garzotto, F., & Turrin, R. (2013, September). User-centric vs. system-centric evaluation of recommender systems. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (pp. 334-351). Springer. doi:10.1007/978-3-642-40477-1_21

de Souza, J. W., Alves, S. S., Rebouças, E. D. S., Almeida, J. S., & Rebouças Filho, P. P. (2018). A New approach to diagnose Parkinson's disease using a structural cooccurrence matrix for a similarity analysis. *Computational Intelligence and Neuroscience*. PMID:29853835

Del Olmo, F. H., & Gaudioso, E. (2008). Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, *35*(3), 790–804. doi:10.1016/j.eswa.2007.07.047

Duan, L., Street, W. N., & Xu, E. (2011). Healthcare information systems: Data mining methods in the creation of a clinical recommender system. *Enterprise Information Systems*, *5*(2), 169–181. doi:10.1080/17517575.2010.541287

Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Van der Vegt, W., & Sloep, P. B. (2017). User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg. *IEEE Transactions on Learning Technologies*.

Fong, A. C. M., Zhou, B., Hui, S. C., Hong, G. Y., & Do, T. A. (2011). Web content recommender system based on consumer behavior modeling. *IEEE Transactions on Consumer Electronics*, *57*(2), 962–969. doi:10.1109/TCE.2011.5955246

Frolov, E., & Oseledets, I. (2017). Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *7*(3), e1201. doi:10.1002/widm.1201

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010, September). Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 257-260). ACM. doi:10.1145/1864708.1864761

He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, *56*, 9–27. doi:10.1016/j.eswa.2016.02.013

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999, August). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 230-237). ACM. doi:10.1145/312624.312682

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250). ACM.

Kamishima, T., & Akaho, S. (2017). Considerations on Recommendation Independence for a Find-Good-Items Task. In *Proceedings of the Workshop on Responsible Recommendation*, Como, Italy

Kavu, T. D., Dube, K., Raeth, P. G., & Hapanyengwi, G. T. (2017). A Characterisation and Framework for User-Centric Factors in Evaluation Methods for Recommender Systems. *International Journal of ICT Research in Africa and the Middle East*, *6*(1), 1–16. doi:10.4018/IJICTRAME.2017010101

Li, Y. M., & Kao, C. P. (2009). TREPPS: A trust-based recommender system for peer production services. *Expert Systems with Applications*, *36*(2), 3263–3277. doi:10.1016/j.eswa.2008.01.078

Lv, G., Hu, C., & Chen, S. (2016). Research on recommender system based on ontology and genetic algorithm. *Neurocomputing*, *187*, 92–97. doi:10.1016/j.neucom.2015.09.113

Marinho, L. B., de Souza, A. H. Junior, & Rebouças Filho, P. P. (2016, December). A new approach to Human Activity Recognition using Machine Learning techniques. In *Proceedings of the International Conference on Intelligent Systems Design and Applications* (pp. 529-538). Cham: Springer.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006, April). Making recommendations better: an analytic model for human-recommender interaction. In CHI'06 extended abstracts on Human factors in computing systems (pp. 1103-1108). ACM. doi:10.1145/1125451.1125660

Moreira, M. W., Rodrigues, J. J., Kumar, N., Al-Muhtadi, J., & Korotaev, V. (2018). Nature-inspired algorithm for training multilayer perceptron networks in e-health environments for high-risk pregnancy care. *Journal of Medical Systems*, *42*(3), 51. doi:10.1007/s10916-017-0887-0 PMID:29392487

Nathanson, T., Bitton, E., & Goldberg, K. (2007, October). Eigentaste 5.0: constant-time adaptability in a recommender system using item clustering. In *Proceedings of the 2007 ACM conference on Recommender systems* (pp. 149-152). ACM. doi:10.1145/1297231.1297258

Neto, E. C., Cortez, P. C., Cavalcante, T. S., Rodrigues, V. E., Reboucas Filho, P. P., & Holanda, M. A. (2016). 3D lung fissure segmentation in TC images based in textures. *IEEE Latin America Transactions*, *14*(1), 254–258. doi:10.1109/TLA.2016.7430087

O'Mahony, M., Hurley, N., Kushmerick, N., & Silvestre, G. (2004). Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology*, *4*(4), 344–377. doi:10.1145/1031114.1031116

O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-Based Systems*, *23*(4), 323–329. doi:10.1016/j.knosys.2009.11.004

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 157-164). ACM. doi:10.1145/2043932.2043962

Ragone, A., Tomeo, P., Magarelli, C., Di Noia, T., Palmonari, M., Maurino, A., & Di Sciascio, E. (2017, April). Schema-summarization in linked-data-based feature selection for recommender systems. In *Proceedings of the Symposium on Applied Computing* (pp. 330-335). ACM. doi:10.1145/3019612.3019837

Ramalho, G. L. B., Rebouças Filho, P. P., Medeiros, F. N. S. D., & Cortez, P. C. (2014). Lung disease detection using feature extraction and extreme learning machine. *Revista Brasileira de Engenharia Biomédica*, *30*(3), 207–214. doi:10.1590/rbeb.2014.019

Rebouças Filho, P. P., Cortez, P. C., da Silva Barros, A. C., Albuquerque, V. H. C., & Tavares, J. M. R. (2017). Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images. *Medical Image Analysis*, *35*, 503–516. doi:10.1016/j.media.2016.09.002 PMID:27614793

Rebouças Filho, P. P., da Silva Barros, A. C., Ramalho, G. L., Pereira, C. R., Papa, J. P., de Albuquerque, V. H. C., & Tavares, J. M. R. (2017). Automated recognition of lung diseases in CT images based on the optimum-path forest classifier. *Neural Computing & Applications*.

Rebouças Filho, P. P., Rebouças, E. D. S., Marinho, L. B., Sarmento, R. M., Tavares, J. M. R., & de Albuquerque, V. H. C. (2017). Analysis of human tissue densities: A new approach to extract features from medical images. *Pattern Recognition Letters*, *94*, 211–218. doi:10.1016/j.patrec.2017.02.005

Rebouças Filho, P. P., Sarmento, R. M., Holanda, G. B., & de Alencar Lima, D. (2017). New approach to detect and classify stroke in skull CT images via analysis of brain tissue densities. *Computer Methods and Programs in Biomedicine*, *148*, 27–43. doi:10.1016/j.cmpb.2017.06.011 PMID:28774437

Rodrigues, M. B., Da Nóbrega, R. V. M., Alves, S. S. A., Rebouças Filho, P. P., Duarte, J. B. F., Sangaiah, A. K., & De Albuquerque, V. H. C. (2018). Health of Things Algorithms for Malignancy Level Classification of Lung Nodules. *IEEE Access*, *6*, 18592–18601. doi:10.1109/ACCESS.2018.2817614

Ross, A. M., Rhodes, D. H., & Hastings, D. E. (2008). Adaptability, Flexibility, Modifiability and Scalability, and Robustness. *Systems Engineering*, *11*(3), 242–262. doi:10.1002/sys.20098

Russell, S., & Yoon, V. (2008). Applications of wavelet data reduction in a recommender system. *Expert Systems with Applications*, *34*(4), 2316–2325. doi:10.1016/j.eswa.2007.03.009

Said, A., Tikk, D., Stumpf, K., Shi, Y., Larson, M., & Cremonesi, P. (2012, September). Recommender Systems Evaluation: A 3D Benchmark. In RUE@ RecSys (pp. 21-23).

Shafna, S., & Rajendran, V. V. (2017, June). Fuzzy ontology based recommender system with diversification mechanism. In *Proceedings of the 2017 International Conference on Intelligent Computing and Control (I2C2)* (pp. 1-6). IEEE. doi:10.1109/I2C2.2017.8321849

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257–297). Boston, MA: Springer. doi:10.1007/978-0-387-85820-3_8

Shardanand, U., & Maes, P. (1995, May). Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 210-217). ACM Press/Addison-Wesley Publishing Co. doi:10.1145/223904.223931

Swearingen, K., & Sinha, R. (2001, September). Beyond algorithms: An HCI perspective on recommender systems. *ACM SIGIR 2001 Workshop on Recommender Systems, 13*(5-6), 1-11.

Torrent-Fontbona, F. (2018). Adaptive basal insulin recommender system based on Kalman filter for type 1 diabetes. *Expert Systems with Applications*, *101*, 1–7. doi:10.1016/j.eswa.2018.02.015

Valdez, A. C., Ziefle, M., Verbert, K., Felfernig, A., & Holzinger, A. (2016). Recommender systems for health informatics: State-of-the-art and future perspectives. In *Machine Learning for Health Informatics* (pp. 391–414). Cham: Springer. doi:10.1007/978-3-319-50478-0_20

Vanesa Aciar, S., Serarols-Tarres, C., & Royo-Vela, M. (2007). Increasing effectiveness in e-commerce: Recommendations applying intelligent agents. *International Journal of Business and Systems Research*, *1*(1), 81–97. doi:10.1504/IJBSR.2007.014774

Vargas, S., & Castells, P. (2011, October). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 109-116). ACM. doi:10.1145/2043932.2043955

Vartak, M., Huang, S., Siddiqui, T., Madden, S., & Parameswaran, A. (2017). Towards visualization recommendation systems. *SIGMOD Record*, *45*(4), 34–39. doi:10.1145/3092931.3092937

Wiesner, M., & Pfeifer, D. (2014). Health recommender systems: Concepts, requirements, technical basics and challenges. *International Journal of Environmental Research and Public Health*, *11*(3), 2580–2607. doi:10.3390/ijerph110302580 PMID:24595212

Wolfe, S. R., & Zhang, Y. (2010, June). Interaction and personalization of criteria in recommender systems. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization* (pp. 183-194). Springer. doi:10.1007/978-3-642-13470-8_18

Wu, L., Liu, Q., Chen, E., Yuan, N. J., Guo, G., & Xie, X. (2016). Relevance meets coverage: A unified framework to generate diversified recommendations. *ACM Transactions on Intelligent Systems and Technology*, *7*(3), 39. doi:10.1145/2700496

Wu, W., He, L., & Yang, J. (2012, August). Evaluating recommender systems. In *Proceedings of the 2012 seventh international conference on Digital information management (ICDIM)* (pp. 56-61). IEEE. doi:10.1109/ICDIM.2012.6360092

Xu, Y., Yang, Q., & Chu, D. (2018). Exploring timeliness for accurate recommendation in location-based social networks. *Mathematical Foundations of Computing*, *1*(1), 11–48. doi:10.3934/mfc.2018002

Yang, L., Bagdasaryan, E., Gruenstein, J., Hsieh, C. K., & Estrin, D. (2018). OpenRec: A Modular Framework for Extensible and Adaptable Recommendation Algorithms. In *Proceedings of WSDM'18*, Marina Del Rey, CA, February 5–9. doi:10.1145/3159652.3159681

Zaman, F., Hogan, G., Van Der Meer, S., Keeney, J., Robitzsch, S., & Muntean, G. M. (2015). A recommender system architecture for predictive telecom network management. *IEEE Communications Magazine*, *53*(1), 286–293. doi:10.1109/MCOM.2015.7010547

Zhang, F., Liu, Q., & Zeng, A. (2017). Timeliness in recommender systems. *Expert Systems with Applications*, *85*, 270–278. doi:10.1016/j.eswa.2017.05.038

Zhang, Z., Chen, C., Zhou, J., & Li, X. (2018, May). An Industrial-Scale System for Heterogeneous Information Card Ranking in Alipay. In *Proceedings of the International Conference on Database Systems for Advanced Applications* (pp. 713-724). Springer, Cham. doi:10.1007/978-3-319-91458-9_44

Zhong, Z., & Li, Y. (2016, November). A Recommender System for Healthcare Based on Human-Centric Modeling. In *Proceedings of the 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE)* (pp. 282-286). IEEE. doi:10.1109/ICEBE.2016.055

Adekunle Oluseyi Afolabi received Bachelor's degree in Computer Science with Economics, the M.Sc. degree in Computer Science from Obafemi Awolowo University, Ile-Ife, Nigeria, and PhD degree in Computer Science from the University of Eastern Finland. His research interests are Digital and Connected Health, Software Engineering, Recommender Systems in Healthcare, Information Systems, Innovative Systems designs and digital storytelling in tourism.

Pekka Toivanen is a professor of Computer Science. He is the head of Computational Intelligence Laboratory at the University of Eastern Finland, Kuopio.