Towards Interdisciplinary Research: A Bibliometric View of Information Communication Technology for Development in Different Disciplines

Mehwish Kayani
Punjab University College of Information Technology (PUCIT)
University of Punjab
Lahore, Pakistan
mehwish.kayani@pucit.edu.pk

Saeed-Ul Hassan
Department of Computer Science
Information Technology University (ITU), Punjab
Lahore, Pakistan
saeed-ul-hassan@itu.edu.pk

Naif Radi Aljohani
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia
nraljohani@ksa.edu.sa

Darren Dancey
Department of Computing & Maths
Manchester Metropolitan University
Manchester, United Kingdom
d.dancey@mmu.ac.uk

Leo Liu
Department of Operations, Technology Events and Hospitality Management
Manchester Metropolitan University
Manchester, United Kingdom
l.liu@mmu.ac.uk

Raheel Nawaz
Department of Operations, Technology Events and Hospitality Management
Manchester Metropolitan University
Manchester, United Kingdom
r.nawaz@mmu.ac.uk

***Abstract*—We present a novel bibliometric view to create a taxonomy of the interdisciplinary field of Information Communication Technology for Development (ICTD or ICT4D), using scientific documents published in the fields related to information communication technologies that were indexed in the Scopus database from 2001 to 2015. Our research approach utilizes a set of relevant terms and venues (journals and conferences) to procure publications on ICTD topics, and further to analyze them statistically to identify emerging or emerged sub-disciplines. The sub-disciplines prominently include Economic Growth/E-Government/Political Economy, Cloud Enterprises/Entrepreneurship/ICT Innovation, and Smart Energy/Smart Grid. We identify the relevance and uniqueness of all ICTD topics, generated by the algorithm, by employing well-known indices of the coherence coefficient and Kullback–Leibler divergence, indicating the qualifications of a given topic as an established sub-discipline of ICTD. This paper attempts to discuss the interdisciplinary nature of the research field ICTD and recommends prospective directions for new sub-disciplines based on the text analysis.**

Keywords— Bibliometrics, Entrepreneurship, ICT Innovation, ICTD, ICT4D, Inter-disciplinary Research, Political Economy, Smart Energy, Semantic Analysis, Text Analysis, Topic Modeling.

I. INTRODUCTION

The bibliography databases such as Scopus, Web of Science or Google Scholar consist of large knowledge bases with immense amounts of information that is widely used to understand semantically scientific structures and impact [1-5]. Tapping on the advancements of bibliography databases, the research and development funding agencies along with Science & Technology (S&T) policy makers seek comprehensive and well-structured information to assess research activities systematically across a wide range of interrelated and overlapping scientific and technological disciplines [6]. For instance, to identify interdisciplinary developments; to obtain insight into the interaction between S&T; compile an inventory of institutional contributors in a particular research area; identify top researchers in a particular research area; and to identify centers of excellence [7].

The dataset for the types of analyses discussed above requires the documents extracted under a specialized research domain/discipline. The factor to which the extracted documents from a specific research domain/discipline are relevant can affect the quality of analyses - thus to procure these documents, a defined dataset is required. Scopus, among some bibliographic databases, works with journal mapping for the storage of documents. In the All Science Journal Classification (ASJC), research disciplines are mapped to journals, proceedings or books in the database; hence bibliographic information belonging to that particular discipline can be procured. This provides a systematic arrangement of documents belonging to disciplines and even sub-disciplines. However, scientific research belonging to interdisciplinary research areas is often difficult to categorize because the work belongs to more than one discipline. This may result in false categorization of documents which may exclude them from being a part of result set. This gives rise to the need to recognize these interdisciplinary research areas so that research may be categorized and placed in relevant journals [8].

There has been much research work and debate on the concept of interdisciplinary but no general indicator which is useful for S&T purposes has been accepted. The basic definition of interdisciplinary is approved by many research councils and science administrators but there is no agreement upon how to use this information for procuring scientific documents. Hence, there should be some way of representing an interdisciplinary research area. We believe, since a summary of a document can be understood through the keywords, so one potential effective approach could be to use a set of relevant keywords to procure documents belonging to any specific research area [9]. Learned from literature review, there is research on the extraction of a relevant set of keywords, but so far there is no automated tool available which can do the need full. One approach could be to use expert knowledge but that too can cause ambiguity at times, and it would be expensive to hire an expert for each field/area. Another approach could be to extract keywords manually (by humans) from resources like Wikipedia, which nevertheless requires excessive amount of time as one has to become an expert him/herself. Moreover, it is a tedious task to pull keywords manually and can lead to ambiguity. For example,

a word ICT is recognized as 'Information Communication Technology' and can have entirely different meaning as 'International Conference on Thermodynamics' in other fields.

In this paper we analyze the extracted keywords for the field of Information Communication Technology for Development (ICTD or ICT4D); their relevance to the field; and the accuracy with which they fit topic/bucket that they fall into, so we can identify emerging as well as emerged sub-disciplines in the field of ICTD. Specifically, we have made the following contributions: a) we semantically analyze the field of ICTD using bibliometric techniques for emerging as well as emerged domain in this discipline; b) we explore the techniques/methodologies to produce a set of keywords in defining a certain research area; c) we procure scientific publications from the bibliometric database Scopus using expert and author defined keywords for an interdisciplinary research area of ICT; and d) we analyze the extracted keywords for their relevancy to the emerging research area.

The rest of the paper has been organized as follows. Section 2 introduces the related work, and the methodology and experimental process is illustrated in Section 3. Section 4 presents the results. Finally, Section 5 concludes the paper with a discussion of the future work.

II. RELATED WORK

Several researchers have defined the terms interdisciplinary, multi-disciplinary and trans-disciplinary, and there is a debate on the definitions of these terms. Our literature review presents the related work for the identification procedures of discipline, in which we study the keyword based identification procedure using various algorithms, including a review on relevant clustering and classification mechanisms. We also reviewed the existing work on presenting ICTD field using bibliography databases.

A. Review on procuring publications for an inter-disciplinary field

Some approaches use ISI journal categories, funding organization's research area codes, or researcher's departmental education or affiliation to identify publications belonging to inter-disciplinarily fields [10]. These approaches do not specifically analyze the content of the work instead use the researcher's information or acquire information from the researcher's proposals and publications. Other approaches which are in practice, regarding the content of the publication include the identification of a research area with author-defined keywords, and these keywords are clustered into frequent item-sets, then based on these frequent item-sets, related documents are procured from the bibliographic databases.

Document clustering is an essential technique to discover topics from the text. It is used to group similar publications into relevant sets. It consists of two phases: firstly, feature extraction forms a keyword network which maps each document or record to a point in high-dimensional space; and secondly, a clustering algorithm automatically groups the points into a hierarchy of clusters [11]. The two significant categories of clustering algorithms are hierarchical and partitioning methods. A linear time approach for document clustering which offers a number of algorithm choices at every phase is used for extracting features and making a hierarchy of clusters, and then it also checks the quality of the clusters using F-measure [12]. The parameters that have been evaluated are clustering time and vector length; the two techniques i.e. feature extraction using term frequency-inverse document frequency (TF-ID) and then truncating feature vector improves the performance, somehow. The tradeoff in them is the cluster quality.

Another very prevalent method for document clustering is agglomerative hierarchical clustering [13]. This family of algorithm follows a same technique which is to merge the most similar pairs after computing the similarity of all cluster's pairs. There exist different similarity measuring schemes applied by different agglomerative algorithms. Steinbach et al. [14] show that the Unweighted Pair Group Method (UPGMA) with Arithmetic mean is the most accurate one in this category. The partitioning clustering algorithms category contain k-means and its variant algorithms. The efficiency and accuracy of basic k-means and agglomerative approaches is outperformed by one of the variants bisecting k-means. This algorithm selects a cluster to split, then employs the basic k-means to form two sub-clusters and reports until the preferred number of clusters is reached.

Kaur et al. [15] talks about introducing frequent item sets clustering as a new criterion, which can also be applied to document clustering. In this algorithm document is treated as a transaction. However, a hierarchy is not created. The Hierarchical Frequent Term-based Clustering (HFTC) proposed by Beil et al. [16] tries to address the special requirements in document clustering using the concept of frequent item sets. The HFTC works in a greedy manner in which it selects the next frequent item sets (demonstrating the next cluster) to minimize the overlapping of the documents that consist of both the item set and training item sets. The clustering result rest on the order of selecting item sets, which in turn depends on the greedy heuristic used.

Fung et al. [11] use the approach in which documents are assigned to the best clusters within all the available clusters. The technique works in a manner, that an item would have very less discriminating power if it is shared across different documents and vice versa. In their algorithm each document is described by a vector of weighted frequencies. Another approach is keyword plus which extracts keywords from individual documents based on titles of the cited articles in the references - which automatically creates a list of queries established on the log of queries that were previously submitted to the search engine using association rule, improves the searching capacity [17].

B. Review on procuring publication dataset in the field of ICTD

In In recent years, Information and Communication Technologies are penetrating human lives progressively, yet there is need to understand that what benefits and in which areas they are providing. The term ICTD refers to the

opportunities of Information and Communication Technology as an agent of development, a computer science perspective gives us the focus of implementation.

Heeks [18] was the first to work on setting up the ICTD journal ranking table to make it easier for researchers to target particular journals for their ICTD publications. The author identified the top 16 journals with titles that associates to ICTs or a division part or synonym, with locus to development (as in "international development") of either developing countries or regions consisting of developing countries. Following to the work of Heeks [18], Gomez et al. [19] conducted a content analysis of 948 ICTD papers from designated peer-reviewed journals and conferences published between 2000 and 2010 as mentioned in Heeks impacting ranking table. From results we can see that business and empowerment as the main domains of ICTD work, appeared to be dominating while less focus on ICT in general and information systems, as the most usable technology objects of analysis, with an increasing trend toward mobile phones. Similarly, Patra et al. [20] also examined the history and evolution of ICTD and the sub domains in ICTD that are getting maturity with time on the basis of a detailed literature review of the many projects in ICTD over the previous decade, and by conducting a survey of 50 researchers and practitioners in the ICTD domain. They determined that the field of ICTD is getting maturity, concentrating mostly on the domains of agriculture, education, (technical and social sciences) communication, governance, healthcare, design, a business of ICTD, and general ICTD.

Heeks [21] argues that there is much work done in the field of ICTD, with many emerging disciplines within the domain of development. As the ICTs enter into human lives, it gets crucial to study the influence of these technologies on the development of Human life. Therefore, the fundamental definition needs to be conceptualized theoretically about the role of ICTs in the perspective of development [22]. There is as yet no generalized taxonomy for the discipline of ICTD. However, the closest work is found in the field of ICT for governance and policy modelling as the governments across the world provide more sustenance to the open data initiative and social media channels to engage citizens. Therefore, researches tend to move towards future internet and suggestions or opinions of the crowd, like the trend analysis. In this context, ICT for governance and policy modelling has recently materialized. Hence, taxonomy is defined for the research areas and sub areas that challenge the domain in order to deal with its diversity and complexity [23].

The usage of ICTs by NGOs, small and medium-sized enterprises has revealed enable growth, mainly through sustained technology and training intervention. Mainly, the research in ICTD is at the organizational level [24]. It is challenging to confirm quantitatively about the overall perception of increased interdisciplinary fields in sciences. The interconnections in sciences can be seen through the journals of respective sciences, but with the increase in inter, multi and trans-disciplinary fields, the interconnections are not seen through journals. Silva et al. [25] study the interconnections of science journals and fields and conclude quantitatively that science fields are evolving into interdisciplinary fields. The unit of interdisciplinary (i.e. entropy) correlates strongly with the in-strength of journals and with the impact factor [26].

In general, approaches to creating taxonomy of interdisciplinary fields use ISI journal categories, funding organization's research area codes, or the researcher's department or affiliation. However, these methods do not take account of information from the content of the publication. The approaches in use for keyword extraction are tedious and require individual experts for every research area. The methods for document clustering generally consist of two phases: first, feature extraction, which forms a keyword network that maps each document or record to a point in a high-dimensional space; second, a clustering algorithm that automatically groups the points into a hierarchy of clusters. However, there is a trade-off in the cluster quality. There is a family of clustering algorithms with improved clustering techniques known as Latent Dirichlet Allocation (LDA) [27].

In summary, our literature review identifies a clear gap where no comprehensive exercise has been conducted to present a taxonomy of the prominent interdisciplinary field of ICTD.

III. METHODOLOGY

We experimented using the techniques of clustering and topic modelling. We have used the topic modelling algorithms to obtain the semantic meaning of frequently occurring keywords in the scientific publications related to ICTD. We used two features of the publication, title of the scientific publication and author-defined keywords. These features were then used to construct the topic models using the MALLET machine [28].

A. Data Set Collection

We procured 30,238 scientific publications using expertly annotated seed keywords, along with all the publications from the journals described by Heeks [17], from the Scopus database from 2001 to 2015. The seed keywords are as follows: "information and communication technology for development", "ict4d", "inform16ation communication technology for development", "information communication technologies for development", "ictd", "m4d", "mobiles for development", "mobile for development", "it4d", "information technology for development", "information communication technology" and "ict".

As the seed keywords contained "ICT" which covers a broader domain, moreover in the early years, there were many other meanings of the abbreviation "ICT". Thus, we had to clean the data for irrelevant results. The following are the many other meanings of "ICT" that we found in our dataset: Intermolecular charge transfer (ICT), intramolecular charge transfer (ICT), internal charge transfer (ICT), international conference on thermoelectrics (ICT), Indication and contraindication for intracoronary thrombolysis (ICT), intensified conventional insulin therapy (ICT), Immunoreactive Calcitonin (ICT), integer cosine chipset (ICT), image convertor tubes (ICT), Intensified Conventional Insulin Therapy (ICT), intercerebral tumors (ICT), Intracutaneous test (ICT), intercell transformer (ICT), Induction chemotherapy (ICT), industrial ct (ICT), immune complex transfer (ICT), ideal chemical theory (ICT), and icaritin (ICT).

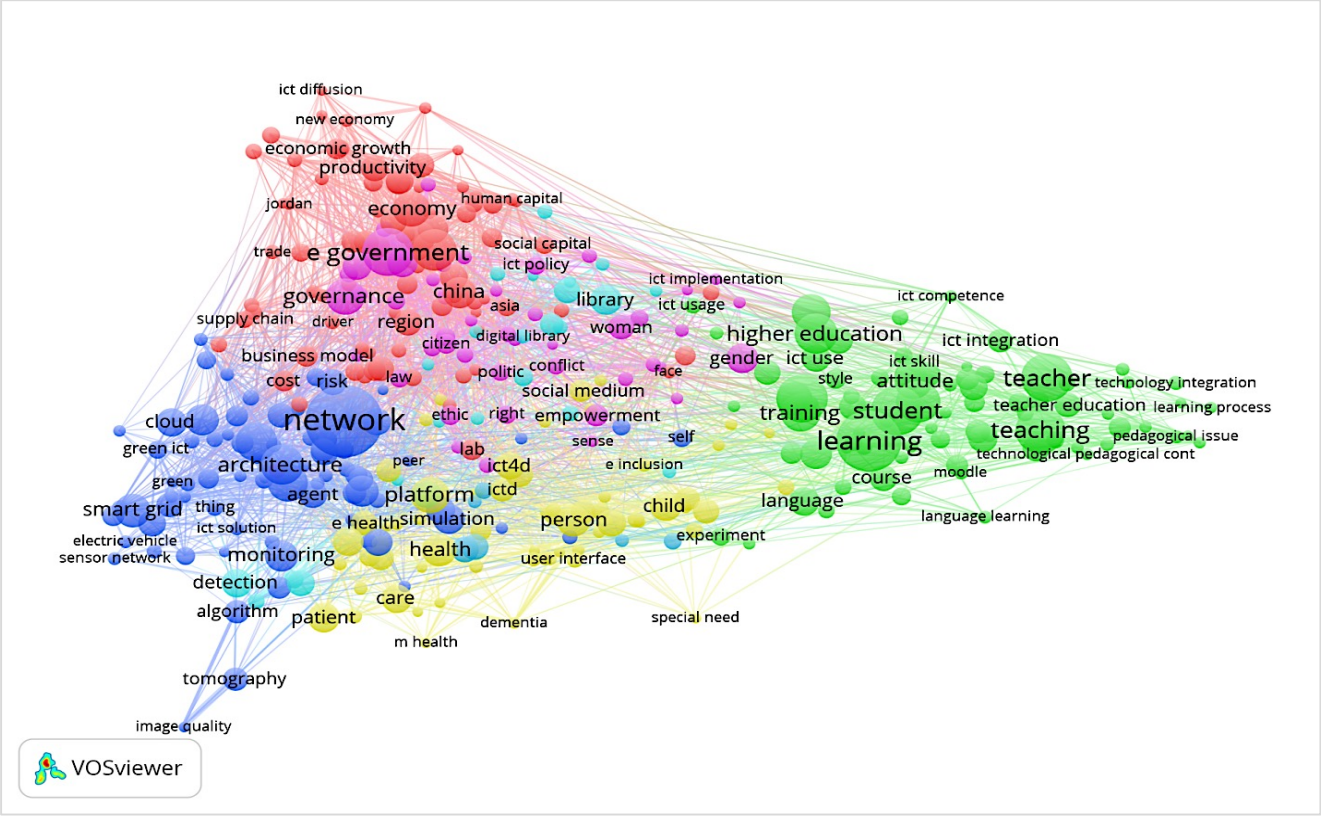

Fig. 1. Visualization of ICT publications from 2001-2015 using VOSviewer

B. Data Processing

We took the following steps to process the data: a) we download publications from Scopus in chunks of 2000 publications at a time; b) all the documents are populated into MS-Excel; c) converted the file to text file as per the need of the MALLET machine [27]; d) merged documents were then cleaned to eliminate all the irrelevant words; and finally e) all stop words were also removed. After removing the irrelevant publications, the remaining data was consisting of 25,560 publications from 2001 to 2015. Figure 1 visually portrays the publication corpus, using VOSviewer.com - a bibliometric mapping program for visualizing—a bibliography database.

After removing the irrelevant publications, the remaining data was consisting of 25,560 publications from 2001 to 2015. Figure 1 visually portrays the publication corpus, using VOSviewer.com - a bibliometric mapping program for visualizing—a bibliography database.

C. Approach

Our model identifies inter-disciplinary sub-areas within the field of ICTD using a topic modelling technique. We deployed the MALLET machine for training the data and then constructing Latent Dirichlet Allocation (LDA) topic models [28], using the scalable implementation of Gibbs sampling. Further, we used the hyper-parameter optimization technique in the topic models, which allows the model to fit the data better by allowing certain topics to be more prominent. We achieved 14 optimized topics. For intra-cluster and inter-cluster similarly, we computed Topic-word probability and Document-Topic probability to examine the relevance of topics formed. Table 1 shows the topic-word occurrence of 10 words selected from across the topics. The topic-word probability is calculated as in Eq. 1.

$$N_k = \sum_d N_{d,k} \,, \tag{1}$$

Here, d represents a document, k represents topic, so that N_k is the number of the topics. It processes the number of word tokens currently assigned to the topic. We rounded off the probabilities to counts for better understanding of the word count occurring in the topic. In addition, we also calculated the probability of a document (publication) in a given topic. We use this probability to sums the frequency of topic over all documents, and normalize to get distribution, as shown in Eq. 2.

$$P(d|k) = \frac{N_{d,k}}{\sum_{d`} N_{d`,k}} \,, \tag{2}$$

Here d is a document, and k represents a topic. Table 2 shows the document-topic probability of five randomly selected documents.

Finally, the topics were tested by the coherence coefficient, a measure of co-occurrence of the words in a topic, as computed using Eq. 3.

$$\sum_i \sum_{j<i} \log \frac{D(w_j w_i) + \beta}{D(w_i)} \,, \tag{3}$$

Here, w represents a word $D(w)$ is the number of documents that consists of at least one token of type w, and $D(w_i, w_j)$ is the number of documents that contain at least one w_i and one w_j.

Table I: Topic-word occurrence of 10 randomly selected words across topics

Words	Topic	Count	Topic	Count	Topic	Count	Topic	Count	Topic	Count
Learning	1	4688	9	519	5	30	–	–	–	–
Communication	15	3826	3	380	18	373	2	110	4	107
Education	1	2103	9	1516	15	111	2	100	3	18
Technologies	15	1894	3	203	18	168	1	108	4	83
Digital	2	1606	9	157	19	151	1	78	14	50
Government	14	1328	2	21	–	–	–	–	–	
Knowledge	5	1308	0	214	9	174	15	122	16	101
Health	17	1247	13	128	3	70	15	25	–	–
Smart	12	1005	13	77	16	1	–	–	–	–
Energy	12	968	10	30	–	–	–	–	–	–

Table II: Document-topic probability of five randomly selected documents

Document	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0	0.011162	0.01032	0.004749	0.017302	0.03257	0.021924
1	0.008356	0.007726	0.003555	0.012954	0.024384	0.016413
2	0.004764	0.004405	0.002027	0.007385	0.443788	0.009357
3	0.002562	0.002369	0.00109	0.158071	0.007475	0.005032
4	0.004167	0.003853	0.001773	0.006459	0.012159	0.008185

Further, we tested the topics by employing Kullback-Leibler (KL) Divergence, a useful measure to differentiate between two probability distributions, as shown in Eq. 4.

$$\sum_i P(w|k) log \frac{P(w|k)}{\frac{1}{|v|}}, \quad (4)$$

Here 'w' represents a given word. The distribution P(w|k) is the usual topic-word distribution proportional to the number of tokens of type 'w' that are assigned to the topic and V is the total vocabulary of words

IV. REUSLTS AND DISCUSSIONS

In this section, we discuss the results produced using LDA topic modelling and MALLET machine [26]. The topics are visualized as word clouds. While eight are specific to ICT, six of the identified topics focus exclusively on ICTD (see Fig. 2).

A. Discussion on Topics related to ICT

Fig. 2 shows the topics related to ICT related themes. Topic 0 features chiefly the use of technology for economic and industrial growth. This is an important dimension of ICT as it deals with economic growth, increase productivity and trade. Topic 1 shows the general adoption of ICTs in enterprises for commerce, trade and industry diffusion. Topic 2 discusses research themes related to social presence and the use of ICTs in specific online cultures due to the emergence of social media networks. Topic 3 discusses the role of ICTs in providing enterprise services over the cloud. Topic 4 is about the use of ICTs in the management of knowledge and transformation of knowledge to the right people, specifically the management of collaborative knowledge (also important under ICTD activities). Topic 5 discusses the role of open source web and software in development. The publications in these clusters highlight the importance of open source tools in the promotion of tools for decision making. Topic 6 has a theme of smart high-speed communication channels, such as fibre-optics, as the backbone of ICT. Topic 7 focuses on related incidents and the measures applied to ensure confidentiality, and this is the integrity of the data generated or stored by an ICT enabled system.

B. Discussions on Topics related to ICTD

The Fig. 3 shows the topics related to ICTD with the domain of ICT. Topic 8 discusses the role of technology in learning; and practices of ICT in virtual learning. It also presents the role of ICTs in education and its dispersion. It more broadly covers learning and study methods. Topic 9 covers the theme of libraries and is about access to digital information using ICTs. In addition, it discusses the issues of economic and social inequality in accessing or usage of ICTs – called digital divide. Topic 10 contains research on developments using mobile devices and the dispersion of cyber learning in rural areas of Africa. We also observe the theme of tele-learning in this topic, concerning with technology-based learning.

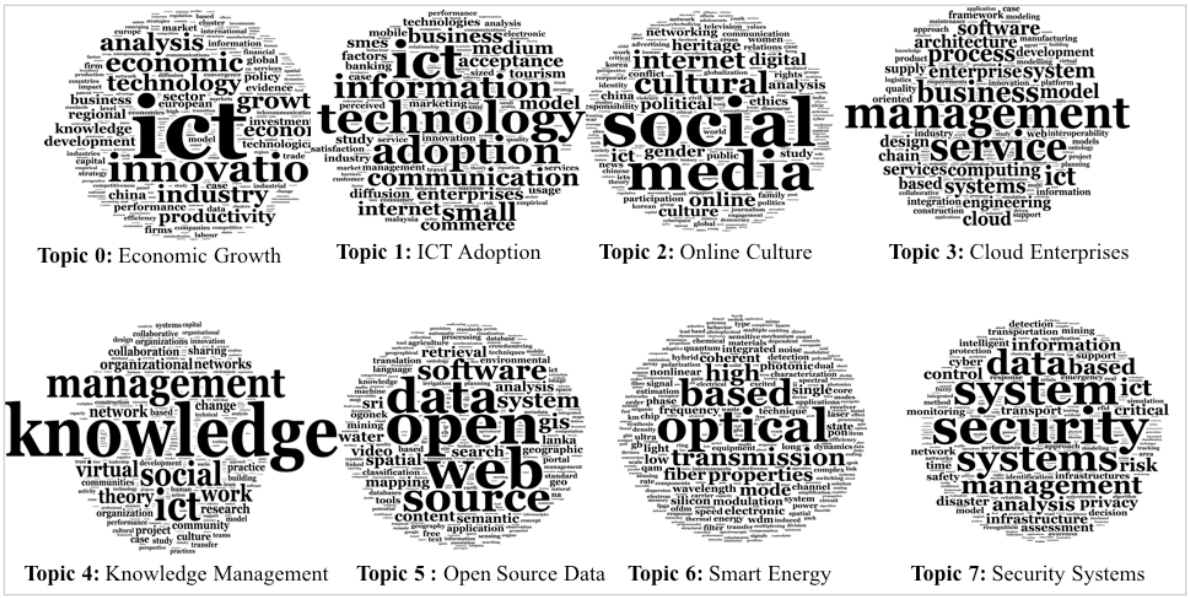

Figure 2: Visualization of ICT related topics [from 0 to 7] using LDA modeling

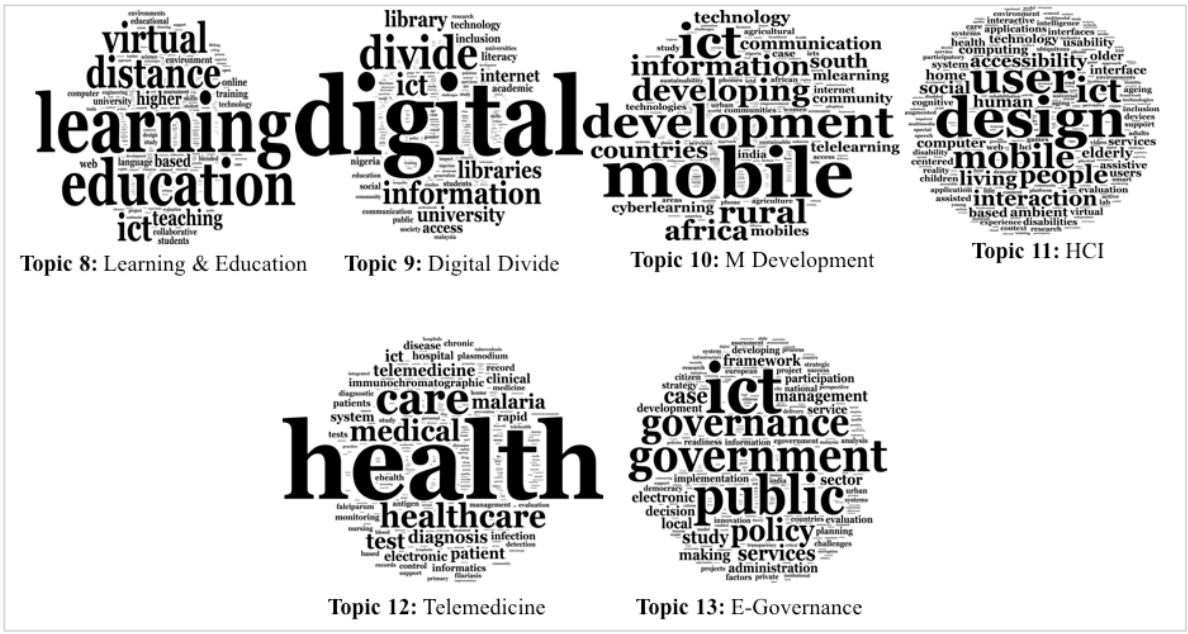

Figure 3: Visualization of ICTD-related Topics [8 to 13] using LDA modelling

Topic 11 has the theme of ICTs enhancing user experience and establishing designs that are usable by humans who have a disability and/or special needs. It basically shows that ICTD works with empathy for users in the domain of development. Topic 12 confers the role of ICTs in the development of healthcare diagnosis. It also highlights the use of ICT-enabled systems to control epidemics such as malaria. The less cohesiveness among the topic keywords shows its relevance with ICTD as compared to the field of ICT. Telemedicine is one wing in which healthcare is taken care of, and it relates to ICT systems use for Development purposes.

Finally, Topic 13 covers the role of ICT in bringing automation to government procedures and reducing costs by deploying ICT enabled large scale systems.

C. Discussion on Topic specficity and uniqueness

This sub section discusses the specificity and uniqueness of a given topic among the topics created by the LDA model. We deployed the measure of Coherence Coefficient and Kullback-Leibler Divergence to analyze the specificity and uniqueness respectively.

Figure 4 shows the coherence coefficient of all 14 topics. Interestingly, we find that all the topics related to general ICT appear to be less coherent than those related to core ICTD. It shows that the keywords constructing ICTD themes strongly co-occur. Since these values are log probabilities and they are negative, larger negative values indicate that the words do not co-occur while scores closer to zero mean that a topic is more specific and that words within topics do co-occur together. The least coherent is Topic 5, with the top keywords of- “openweb, data source, software, system, gist, spatial search, content, sri, water, retrieval, video, mapping, analysis, semantic, lanka, geographic, ogonek”.

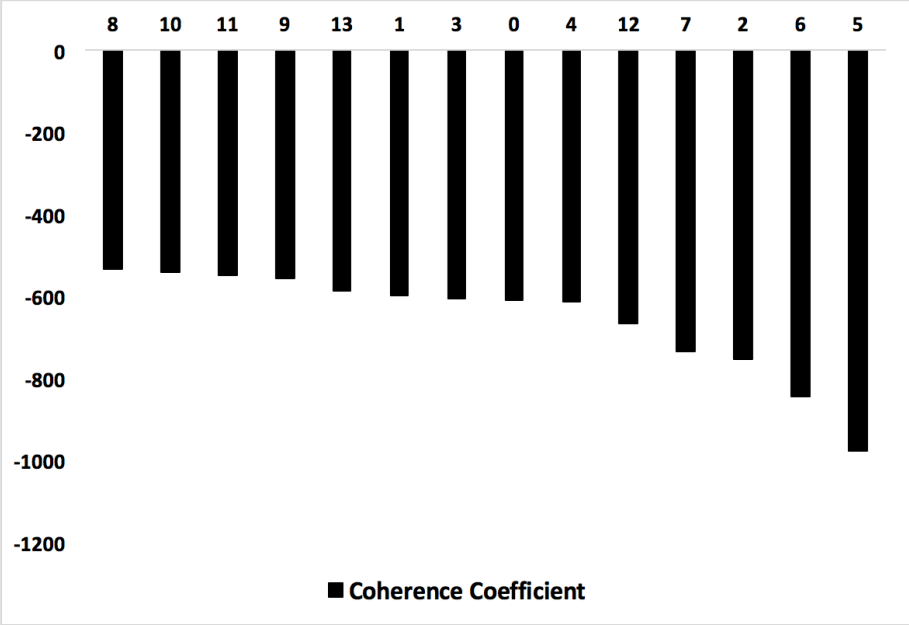

Figure 4: Coherence coefficient of Topics [0-13]

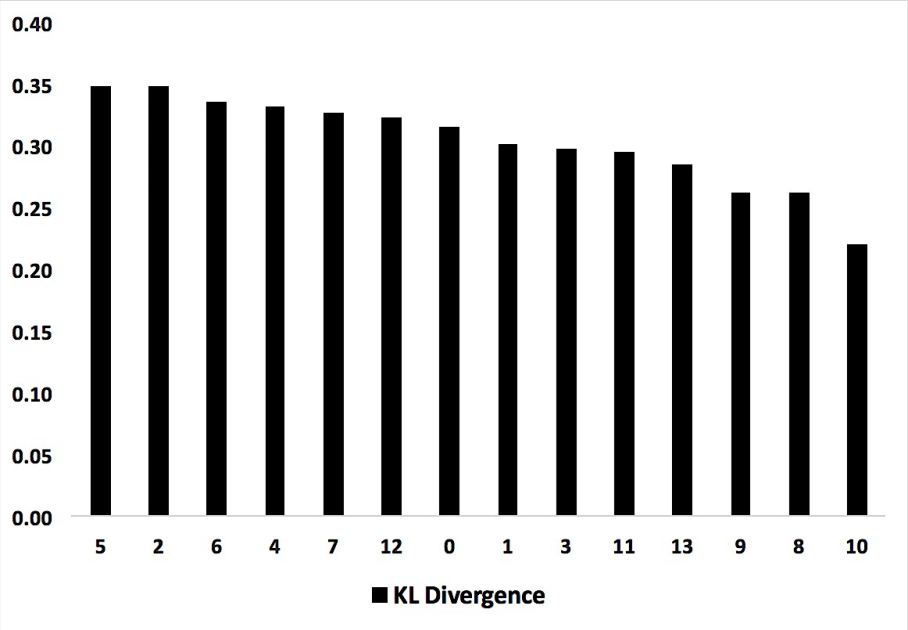

Figure 5. KL Divergence measure of topics [0-13]

We also observe that Topic 12, which covers the use of ICT for healthcare, shows relatively fewer coherent characteristics. One reason could be that the concept of tele-medicine/healthcare does not necessarily fall into any single ICTD domain, since many publications in this topic relate to medical information systems. The less cohesiveness of topic keywords shows the mix of ICT enabled systems use for healthcare purposes along with providing support for 'development' in medical/health care. Finally, Figure 5 shows the KL Divergence of all 14 topics. The value of divergence increases towards 1, which means a more divergent topic. We find that the topics related to ICTD show relatively less divergence than topics related to ICT. This shows that topics related to ICT have many keywords common to all topics. In contrast, topics related to ICTD are specific and defined, though ICTD emerges an interdisciplinary field within ICT.

V. CONCLUDING REMARKS

In this paper, we have presented a novel bibliometric view to create a taxonomy of the inter-disciplinary field of ICTD by deploying LDA modelling technique using scientific literature downloaded from the Scopus database. Further, to measure the specificity and uniqueness of the topics generated, we deployed the coherent coefficient and KL divergence measures. Our results indicate that topics related to ICTD are more coherent and less divergent than the overall topics generated. We conclude that the topic themes resulting from our research work can be used as a baseline to form taxonomies for the discipline of ICTD. Our literature review has shown that, to the best of our knowledge, there is no defined taxonomy for the domain of ICTD as yet, and thus the model presented is the first to show the themes comprehensively within ICTD. This work

could be enhanced for use as a baseline to present a temporal analysis that can show the growth and evolution of sub-fields within ICTD or related disciplines [29]. Last but the not the least, in addition to the employed supervised clustering techniques, Natural Language Processing techniques [30] along with Machine Learning based models [31] are recommended to detect emerging topics in ICTD.

REFERENCES

[1] R. Nawaz, P. Thompson, and S. Ananiadou, "Identification of Manner in Bio-Events", In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 3505-3510, 2012.

[2] S. Ananiadou, P. Thompson, R. Nawaz, "Enhancing Search: Events and Their Discourse Context. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing. Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg, 2013.

[3] M. Shardlow, R. Batista-Navarro, P. Thompson, R. Nawaz, J. McNaught and S. Ananiadou, "Identification of research hypotheses and new knowledge from scientific literature", BMC Medical Informatics and Decision Making, vol. 18, no. 1, 2018.

[4] P. Thompson, R. Nawaz, I. Korkontzelos, W. Black, J. McNaught and S. Ananiadou, News search using discourse analytics. In *2013 Digital Heritage International Congress (DigitalHeritage)*, Vol. 1, pp. 597-604. IEEE. 2013.

[5] P. Thompson, R. Nawaz, J. McNaught and S. Ananiadou, Enriching news events with meta-knowledge information. Language Resources and Evaluation, Vol. 51, No. 2, pp.409-438, 2017.

[6] G. Laudel, and G. Origgi, "Introduction to a special issue on the assessment of interdisciplinary research", Research Evaluation, Vol. 15 No. 1, pp. 2-4, 2006.

[7] D. Chapman, and C. L. Chien, "Higher Education in Asia: Expanding Out, Expanding Up", Montreal: UNESCO Institute for Statistics, 2014.

[8] S. U. Hassan, P. Haddawy, and J. Zhu, "A bibliometric study of the world's research activity in sustainable development and its sub-areas using scientific literature", Scientometrics, vol. 99, no.2, pp. 549-579, 2014.

[9] S. U. Hassan, P. Haddawy, P. Kuinkel, and S. Sedhai, "A bibliometric study of research activity in sustainable development", In Proceedings of 13th Conference of the International Society for Scientometrics and Informetrics (ISSI2011) Vol. 2, pp. 996-998, 2011.

[10] A. Duvvuru, S. Radhakrishnan, D. More, S. Kamarthi, and S. Sultornsanee, "Analyzing structural & temporal characteristics of keyword system in academic research articles", Procedia Computer Science, vol. 20, pp. 439-445, 2013.

[11] B. C. Fung, K. Wang, and M. Ester, (2003) "Hierarchical document clustering using frequent itemsets", In Proceedings of the 2003 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp. 59-70, 2003.

[12] B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto, and N. Ziviani, "Concept-based interactive query expansion", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 696-703, October 2005 .

[13] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections", In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 318-329, 1992.

[14] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques", In KDD workshop on text mining, Vol. 400, No. 1, pp. 525-526, 2000.

[15] R. Kaur, and A. Kaur, "Text Document Clustering and Classification using K-Means Algorithm and Neural Networks", Indian Journal of Science and Technology, Vol. 9 No. 40, 2016.

[16] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering", In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 436-442, 2002. ACM

[17] E. Garfield, and I. H. Sher, "Brief Communication Keywords Plus Algorithmic Derivative Indexing", Journal of the American Society for Information Science, pp. 1986-1998, Vol. 44 No. 5, 298, 1993.

[18] R. Heeks, "An ICT4D journal ranking table", Information Technologies & International Development, Vol. 6 No. 4, pp. 71-75, 2010.

[19] R. Gomez, L. F. Baron, and B. Fiore-Silfvast, "The changing field of ICTD: content analysis of research published in selected journals and conferences, 2000—2010". In Proceedings of the Fifth International Conference on Information and Communication Technologies and Development, pp. 65-74, 2012. ACM

[20] R. Patra, J. Pal, and S. Nedevschi, "ICTD state of the union: Where have we reached and where are we headed", In Information and Communication Technologies and Development (ICTD), pp. 357-366, 2009. IEEE.

[21] R. Heeks, "Deriving an ICT4D research agenda: a commentary on'Information and communication technologies for development (ICT4D): solutions seeking problems?", Journal of Information Technology, Vol. 27 No. 4, 339, 2012.

[22] M. K. Sein, and G. Harindranath, "Conceptualizing the ICT artifact: Toward understanding the role of ICT in national development", The Information Society, Vol. 20 No.1, pp. 15-24, 2004.

[23] M. Caminati and A. Stabile, "The pattern of knowledge flows between technology fields", Metroeconomica, Vol. 61 No. 2, pp. 364-397, 2010.

[24] P. J. Rousseeuw, and L. Kaufman, "Finding Groups in Data", Wiley Online Library, 1990.

[25] F. N. Silva, F. A. Rodrigues, O. N. Oliveira, and L. D. F. Costa, "Quantifying the interdisciplinarity of scientific journals and fields", Journal of Informetrics, Vol. 7 No. 2, pp. 469-477, 2013.

[26] L. Leydesdorff, L. Bornmann, and P. Zhou, "Construction of a pragmatic base line for journal classifications and maps based on aggregated journal-journal citation relations", Journal of Informetrics, Vol. 10 No. 4, pp. 902-918, 2016.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", Journal of machine Learning research, 3(Jan), pp. 993-1022, 2003.

[28] M. A. Mallet, "A machine learning for language toolkit". 2002.

[29] S.U. Hassan, and P. Haddawy, "Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy". Scientometrics, Vol. 103 No. 1, pp. 33-46, 2015.

[30] R. T. Batista-Navarro, G. Kontonatsios, C. Mihăilă, P. Thompson, R. Rak, R. Nawaz, I. Korkontzelos, and S. Ananiadou, 2013, March. "Facilitating the analysis of discourse phenomena in an interoperable NLP platform", In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 559-571, March 2013. Springer, Berlin, Heidelberg.

[31] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron", In 2017 Intelligent Systems Conference (IntelliSys), pp. 722-728, September 2017. IEEE