Systematic method for finding emergence research areas as data quality

Babak Sohrabi⁎, Ahmad Khalilijafarabad

⁎ Department of Information Technology Management, Faculty of Management, University of Tehran, Jalal Al Ahmad Street, Nasr Bridge, P.O. Box: 14155-6311, Tehran, Iran

ABSTRACT

The analysis of the transformation and changes in scientific disciplines has always been a critical path for policymakers and researchers. The current study examines the changes in the research areas of data and information quality (DIQ). The aim of this study was to detect different types of changes occurring in the scientific areas including birth, death, growth, decline, merge, and splitting. A model has been developed for this data mining. To test the model, all DIQ articles published in online scientific citation indexing service or Web of Science (WOS) between 1970 and 2016 were extracted and analyzed using the given model. The study is related to the Big Data as well as the integration methods in Big Data which is the most important area in DIQ. It is demonstrated that the first and second emerging research areas are sub-disciplines of entity resolution and record linkage. Accordingly, linkage and privacy are the first emerging research area and the entity resolution using ontology is the second in DIQ. This is followed by the social media issues and genetic related DIQ issues.

1. Introduction

The management of data and information quality (DIQ) dates back to the early days of information system studies, but at this era, it is believed that DIQ is reaching to its threshold of significant growth and becoming an essential research area (Shankaranarayanan and Blake, 2017).

Good quality data is essential for making a right decision. In other words, bad data can ruin the performance of a good design and affect the implementation of the information systems. It can deceive the decision makers influencing their decisions (Mecella et al., 2002). This research area is common between management, statistics, and computer science researchers (Batini and Scannapieca, 2006; Sadiq, 2013). The DIQ started by mathematicians and statisticians in late 1960 to find the duplicate values among the datasets and continued in 1990 by computer scientists for improving the quality of data in databases (Batini and Scannapieca, 2006). However, the first academic attempt starts with the Department of Computer Science, MIT that shows the DIQ is much more than the accuracy of the data. In their seminal work, they classified the DIQ dimensions in four classes of Intrinsic, Accessibility, Contextual, and Representational (Wang and Strong, 1996). Afterwards, a lot of studies proposed the definitions of DIQ management. For instance, DIQ management is defined as using the data in the right place, at the time, and in the format for completing the operation, providing the service for the customer and making the decision or implementing the strategy (Pyzdek and Keller, 2014). DIQ management also defines the degree of matching of data measured by the information system and its value in real-world (Orr, 1998). However, the most common definitions focus on the fitness for the use of data (Juran and Godfrey, 1999; Sadiq, 2013; Strong et al., 1997).

Due to the different technological advancements, DIQ management is becoming a critical issue for various industries and academia (Neely and Cook, 2008). In the last two decades, the nature of DIQ management has changed very fast, and it is going to become a strategic issue in organizations for different reasons (Sadiq et al., 2011). This research area is changing from content to contextual issues (Shankaranarayanan and Blake, 2017). There are various reason for this change, such as the emergence of Big Data technology (Cai and Zhu, 2015; Saha and Srivastava, 2014), social information systems (Tian et al., 2016; Tilly et al., 2015), and semantic web (Fürber, 2016), crowd technology (Lukyanenko and Parsons, 2015). Furthermore, transitions of organizational structures from a hierarchical structure to peer-to-peer structures that increase the volume, domains, and complexity of data (Batini et al., 2009).

Although the DIQ management is being used in different disciplines, there are not enough studies about the emergence research area of DIQ management. Most of the studies are focused on the definition of DIQ domain (Zhang et al., 2013), taxonomy of problems (Khalilijafarabad...
et al., 2016; Oliveira et al., 2005), and definition of DIQ as a scientific discipline (Blake, 2010; Blake and Shankaranarayanan, 2012; Helfert and Ge, 2006). But finding the emerging research areas can guide the researchers and people in business to focus on the right problems that will be mainstreams in the future. Due to their interest in the trending research areas, researchers often recourse to both qualitative and quantitative approaches. In the qualitative approach, they try to find an attractive research area based on the questionnaire (Kajikawa et al., 2008). However, considering the growth of research areas and multidisciplinarily approach of the DIQ, it is difficult to find the trending sub-disciplines. In fact, various studies have focused on the sub-disciplines of DIQ showing its multidisciplinary approach (Batini and Scannapieca, 2006).

Since scientific disciplines evolve, some researchers keep trying to come up with theories that show the mechanisms of the scientific changes. In this regard, the Kuhn science revolution theory states that the scientific progress is highly related to revolutionary ideas (Kuhn and Hawkins, 1963). The theory also states that although minor developments cause most of the scientific progress, some ideas can change the face and direction of science altogether. DNA invention can be taken as a supporting example (Madlock-Brown, 2014). Interestingly, these advancements are not limited to this example alone, but one can find the same in different branches and levels of science (Madlock-Brown, 2014).

It is also noteworthy to discuss the differences between the detection models, related to past and future. In other words, the detection models try to develop designs to find the phenomena in their early stages, but the prediction models try to come up with models to find out the future behavior of the phenomena (Kajikawa et al., 2008).

This study uses data mining models to find the emerging research areas of the DIQ. These emergence research areas are novel, growing fast and have high recency. Thus, all DIQ-related studies published between 1970 and 2016 in Web of Science (WOS) journals in the area of information system and computer science have been extracted and analyzed using the proposed method.

Rest of the paper is organized as follows. The first section is a literature review, which covers related concepts and the methods of finding scientific frontier. The second part is the research method and the steps taken in this study. The third part is the results and discussion, and the last section is conclusion and suggestions.

2. Literature Review

A lot of scientists came up with different theories about the dynamics of science and hypothesized about the reasons for emergence and changes in the scientific areas. One of the most famous theories is Gibbons Mode 2 theory of scientific changes. Mode 2 refers to application-oriented and trans-disciplinary, and Mode 1 refers to academic-oriented and disciplinary knowledge productions (Gibbons et al., 1994). Or the Helix model that shows the relationships between the industry, academics, and governments for knowledge production (Etzkowitz and Leydesdorff, 2014).

There is a body of knowledge that focuses on finding the emerging research areas and fore sighting of expertise. There are different definitions of emerging research areas (Day and Schoemaker, 2000; Rotolo et al., 2015). These studies used different indices for finding the emerging research areas. For example, the growth rate is defined as the emergence score (Kleinberg, 2002; Van Raan, 2000), or the rate of entrance for researchers and journals (Guo et al., 2011), or the novelty (Tu and Seng, 2012). But most of the studies used newness, radical novelty and growth to some extents (Crutchfield, 1994; Guo et al., 2011; Small et al., 2014; Tu and Seng, 2012).

Besides that, various studies draw the map of science or show the emerging research areas. One of the earliest attempts to draw and analyze the science map is the John Bernard’s study in 1939 that shows the social aspect of science (Bernal, 1939). After few years, in 1970, some researchers focused on detecting the emerging research areas began with Price index (Price, 1970). Two years later, some other indexes such as the half-life citation were developed (Garfield, 1972).

But the most commonly used maps are shaped by Henry Small seminal works of reference mapping (Small, 1973). After that, the time window and scientific string were proposed which helped the researchers to investigate the changes in scientific areas (Small, 2006). With these concepts, they can follow the evolutions of the scientific areas. In a study, this concept was used, and the average age of the clusters in combination with the size of the cluster was calculated to find attractive research areas (Takeda and Kajikawa, 2008). Another study used the citation network and clustering methods to find the sub-disciplines in combination with the lifecycle concept to find emerging areas (Kajikawa et al., 2008). In a study, the percentage of top-cited transfer from older to the new community was calculated to show the evolution and changes of research areas (Qian et al., 2008). Following this approach a research used citation network and followed the highly cited studies to show the citation structure of emerging research areas (Small and Upham, 2009). One year later in 2010, the same group used the co-citation to find the small clusters that are highly cited and claim that these clusters show the research fronts (Upham and Small, 2009).

Another study compared four types of citation networks and showed that using different citation models are not that much different (Boyack and Klavans, 2010). After that, another group of researchers demonstrated that just bibliographic coupling networks shows the changes faster than other types of networks and invent a mixed index of co-citation and bibliographic-coupling to find the emerging research areas (Small et al., 2014).

There are some other studies that used different text mining or data mining approaches. One of the most critical studies defined the emergence as a burst and suggested the Bayesian model to find the changes in the usage of words (Kleinberg, 2002). This approach also followed by some researchers. For example, some modification of Kleinberg model is suggested that can find the bursts more efficiently (Madlock-Brown, 2014) or a data mining tool called ‘DETECTS’ proposed finding bursting technologies and cluster them (Dennis et al., 2016).

Some studies used text mining methods to find the keywords or topics of the articles and analyzed the evolutions. (Glenisson et al., 2005) used a simple clustering method based on the full text of articles and bibliometrics to map the knowledge in the area of Scientometrics. But this study used only 85 articles. After that in 2006, a simple methodology of the fore sighting by text mining is proposed (de Miranda Santo et al., 2006). There is also an article that uses formal concept analysis to find the evolution of research areas. The problem with such studies is their need to expert opinions for defining the context (Lee et al., 2011). One of the best papers that use the evolution of networks of texts used the community size for changes in scientific areas but ignore some other indices (Chen et al., 2012). In some newer studies, they used unsupervised methods such as the Latent Semantic Analysis or Latent Dirichlet Analysis (X. Wang et al., 2013; Westgate et al., 2015).

In a more comprehensive study, researchers used the text mining and network analysis method to analyze the evolution of emerging research areas. But their approach suffers from complexity because they calculate all the similarities between all articles and generate the almost complete graph and then cut the edges (Furukawa et al., 2015). A study was focused on the forecasting of emerging trends by predicting keyword distribution through regression models (Asooja et al., 2016).

Another approach for analyzing the emerging research areas is using the network analysis. These methods try to analyze the topology of the graph of science map and find the relationship between the structure of graph and emergence of research areas. Betweenness centralities, path analysis, and other methods were used to analyze the research areas (Prabhakaran et al., 2015).

As mentioned before, the literature review could not find many models showing different behaviors of scientific domains such as birth,
death, growth, decline, merge, and split. There are also some pitfalls such as complexity of methods, focusing on one index of emergence. A model showing such behaviors becomes an acute necessity for researchers studying the lifetime of disciplines or tracking scientific areas. Hence, the present study tries to develop a model that can detect almost all macro behavior of scientific areas. This model is tested on the DIQ area to validate as well as find out the most critical sub-disciplines of the same.

3. Research method

This study uses the demonstrated approach to find the top trends and emerging research areas of DIQ. The following diagram (Fig. 1) clarifies the five main steps of the proposed research methodology to find the emergence research areas. In the following parts, a detailed explanation of each step will be presented.

3.1. Data gathering

For gathering the needed data for this study the abstract, title, and the publication year of the papers that are related to DIQ from the web of Science (WOS) and published between 1970 and 2016 are gathered. Although it is shown that using full-text data of the papers lead to more accurate results (Rezaeian et al., 2017) although because of the lack of accessibility to data, abstract and title is used. To cover related sub-disciplines, researchers, chose keywords extracted from taxonomy papers (Khalilijafarabad et al., 2016; Sadiq et al., 2011). For data gathering, the keywords “data quality,” “information quality,” “content quality,” “data consistency,” and “linkage” were selected. But the results were limited to areas dealing with computer science and information system. Altogether, this collection contained about 9000 research articles whose year-wise frequency and the trend could be visualized through the following graph (Fig. 2):

3.2. Science map generation

Developing a science map is the first step of a study aimed at finding scientific trends. This map shows the structure and interconnection between elements, i.e., the articles used for this study (Morris and Van der Veer Martens, 2008). In this study with the interval of two years, accumulative science map of DIQ was generated to analyze the evolution of this scientific area. This step has two parts, network generation and sub-discipline exploration of the DIQ research areas.

3.2.1. Network generation

In the present study, the network has been generated based on the abstract and title of the accumulated research articles. For this purpose, we produced a system using Text mining methods where nodes are the papers and edges are a weighted similarity between the content of the articles.

We used different methods such as stop word omitting, stemming, TF_IDF, and similarity calculation. First, we removed stop words like “the, a, is” that occurred frequently but were not important. Then we used the stemming method to find the root of words, for example, the words “works, work, working, worked” has the same root “work.” These words have the same meaning. Then we used weighting methods to omit the effect of less informative words and give more weight to more informative words. It is important to note that the most frequent words were not essentially the informative words (Huang, 2008). There were different weighting methods (Reed et al., 2006). In this study, we used the TF_IDF method. The basic form of TF_IDF was calculated as shown in Eq. (1):

$$W_{ij} = \log(f_j) \times \log\left(\frac{N}{n_j}\right)$$

In the above equation, $f_j$ refers to the occurrences of term $j$ in document $i$. $N$ is the total number of documents and $n_j$ is the number of documents that contain term $j$ at least once. The first part of this equation is TF and the second part is IDF.

After calculating the TF_IDF, we used the Cosine similarity measure to find the similarities of each pair of documents. The next step was to prune the edges with low similarity. Fig. 3 shows the histogram by bin numbers calculated from Rice Rule (Lane, 2003). We then used the trial and error method to find the best threshold for filtering and applied it to filter out the edges with a similarity of 0.3 or less. This method covered < 10,000 similarities.

The result of this part is the network that each node shows the paper, and each edge shows the similarity between the articles.

3.2.2. Sub-discipline exploration

Finding the sub-discipline of DIQ from the generated networks is known as community detection or graph partitioning problem in systems. There are a lot of different approaches and algorithms for finding the communities in the graphs. Detailed information about such algorithms is discussed in a comprehensive paper by Fortunato (Fortunato, 2010).

One of the fastest and accurate algorithms based on modularity and developed for the general case of weighted graphs was introduced by

![Fig. 2. Number of publications per year.](image-url)
3.4. Top trends extraction

In Fig. 4, the conceptual demonstration of threads that will be generated is displayed starting from the oldest timestamp to the newest one. The suicide code of this step is shown in (Fig. 5):

As it has mentioned in the literature review, the most frequent features for finding the emerging research areas are newness and growth. We also added the indices of importance.

The newness is the reverse of path length that will decrease as the age of the trend increase. The next important aspect was to analyze the growth of the threads which was done using the Eq. (3). This equation subtracts the community size in period t from the community size in t-1.

\[
\sum_{t=1}^{n} \left( f(t) - f(t-1) \right)
\]

(3)

These trends were analyzed to determine the most attractive research areas that have been named as emergence areas in this study. We calculated the average size of the communities by the Eq. (4).

\[
\sum_{t=1}^{n} \left( f(t) + f(t-1) \right) / n
\]

(4)

The emergence score is then calculated by the multiplication newness to growth rate. It means that the communities with maximum growth and minimum age are the emerging research areas.

3.5. Analysis

After generating the threads, their analytical and qualitative analyses were conducted to reveal the changes in the DIQ research areas. This investigation aimed finding the most important research areas and also the emerging research areas. In this step, we also used the topic modeling methods to find the focus of research communities. In this respect, the Latent Semantic Indexing (LSI) method was used in this study.

LSI is a mathematical-statistical method for extracting the concepts from the texts in the entirely unsupervised way. This technique needs no knowledge base, lexicon, or any other dictionaries (Tonta and Darvish, 2010). More information about LSI can be found in previous studies (Evangelopoulos, 2013; Landauer, 2006).

After extracting and analyzing the specified areas, we used the DIQ experts’ opinion to validate the findings especially the causes leading to such changes. We used four experts to check and validate the results.

4. Experiment and results

In order to do the experiments we used the Python 2.7 and also Gephi for visualization parts. As mentioned before, we used the cumulative network threads in the course of our study. It is worth mentioning that a cumulative network is generated and communities are extracted every 2 years. Table 1 shows the number of communities for each period:

<table>
<thead>
<tr>
<th>Year</th>
<th>Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>100</td>
</tr>
<tr>
<td>1981</td>
<td>105</td>
</tr>
<tr>
<td>1982</td>
<td>110</td>
</tr>
<tr>
<td>1983</td>
<td>115</td>
</tr>
<tr>
<td>1984</td>
<td>120</td>
</tr>
<tr>
<td>1985</td>
<td>125</td>
</tr>
<tr>
<td>1986</td>
<td>130</td>
</tr>
<tr>
<td>1987</td>
<td>135</td>
</tr>
<tr>
<td>1988</td>
<td>140</td>
</tr>
<tr>
<td>1989</td>
<td>145</td>
</tr>
<tr>
<td>1990</td>
<td>150</td>
</tr>
<tr>
<td>1991</td>
<td>155</td>
</tr>
<tr>
<td>1992</td>
<td>160</td>
</tr>
<tr>
<td>1993</td>
<td>165</td>
</tr>
<tr>
<td>1994</td>
<td>170</td>
</tr>
<tr>
<td>1995</td>
<td>175</td>
</tr>
<tr>
<td>1996</td>
<td>180</td>
</tr>
<tr>
<td>1997</td>
<td>185</td>
</tr>
<tr>
<td>1998</td>
<td>190</td>
</tr>
<tr>
<td>1999</td>
<td>195</td>
</tr>
<tr>
<td>2000</td>
<td>200</td>
</tr>
<tr>
<td>2001</td>
<td>205</td>
</tr>
<tr>
<td>2002</td>
<td>210</td>
</tr>
<tr>
<td>2003</td>
<td>215</td>
</tr>
<tr>
<td>2004</td>
<td>220</td>
</tr>
<tr>
<td>2005</td>
<td>225</td>
</tr>
<tr>
<td>2006</td>
<td>230</td>
</tr>
<tr>
<td>2007</td>
<td>235</td>
</tr>
<tr>
<td>2008</td>
<td>240</td>
</tr>
<tr>
<td>2009</td>
<td>245</td>
</tr>
<tr>
<td>2010</td>
<td>250</td>
</tr>
<tr>
<td>2011</td>
<td>255</td>
</tr>
<tr>
<td>2012</td>
<td>260</td>
</tr>
<tr>
<td>2013</td>
<td>265</td>
</tr>
<tr>
<td>2014</td>
<td>270</td>
</tr>
<tr>
<td>2015</td>
<td>275</td>
</tr>
<tr>
<td>2016</td>
<td>280</td>
</tr>
<tr>
<td>2017</td>
<td>285</td>
</tr>
<tr>
<td>2018</td>
<td>290</td>
</tr>
</tbody>
</table>

During the first phase, i.e., 1970–1989 when the number of published papers was not enough, there was no significant change. The visualized version of this thread is as follows. As is shown in the following figure, the result is a complex graph.

Fig. 6 shows the threads and their changes over the years. The proposed model can display different behaviors of scientific areas. For instance, one can find various types of behaviors in Table 2 that has been demonstrated using the proposed model:

To analyze the most important thread, we chose the top 100 communities and extracted the network of these communities.

According to the demonstrated network (Fig. 7), 22 threads were extracted where each thread had more than one node. But one of the threads has been omitted because it was dead before 2017. One can find...
all the behaviors mentioned above in these threads. Finding the thread topics as well as the most interesting threads were the last important tasks of the researchers. Table 3 shows results of the topic modeling:

To find the most important threads, researchers could also use some statistical studies. One of the most critical factors in the analysis was to know the extent of a community. Table 4 shows the related results:

As the table shows, THR12 is the thread with the most enormous community size as well as maximum growth among others. It is interesting to know that this thread type splits into two different areas. According to the result of the topic modeling, this thread was related to the perceived trust and satisfaction of web users that were also entirely related to data quality of the research area. The second most important area is thread 1 with the average size of 88.7 that is about the cleaning and integration of Big data. As it has discussed in previous parts, emerging research areas are not necessarily the hot topics, but the topics which have capability to grow will become hot topics in the future. According to the mentioned method, emerging research areas are the ones with high growth and shortage indicating that the area is
somewhat new to researchers and hence, it is in its early stages. To find the emerging research areas, we divided the growth on the age of the clusters (Table 5).

The table indicates that the privacy-preserving record linkage is the first emerging research area in the field of DIQ for the current year. It is worth mentioning that the record linkage is about combining multiple databases with disjunctive or additional information on the same person. Due to privacy of problems, most of the data were encrypted and hence; the privacy-preserving record linkage algorithms tried to find a way to link the data in such situations. Some researchers have given information about these issues and its usage (Vatsalan et al., 2013). It is also found that the concept of entity resolution using ontology is the second emergence research area in DIQ management. Although these two topics are somehow the same, their focus and methods are different (Brizan and Tansel, 2006).

It is also found that the issues of social media and mobile are the third emergence research area. This area is also emerging because of the growth of social media use in organizations. As a consequence, the content quality and the issues such as believability of data come up (Agarwal and Yiliyasi, 2010). In the course of the current study, it was also found that the data quality in genes and protein was the fourth most emerging research area. In other words, this focused on the dimensions and meaning of the DIQ at genetic or biomedicine studies.

5. Conclusion and suggestions

This paper covered two critical research areas of DIQ and science mapping. To find out the research trends, a scientific mapping technique based on the combination of text mining and network analysis approach was developed. We also used the concept of science thread to find the attractive research areas. The proposed methodology enabled us to show an overall behavior of the scientific field. Also, the DIQ research area was studied and analyzed using the proposed model. Accordingly, the topic related to Big Data and the integration methods in Big Data era was considered the most crucial issue in DIQ. It has also been shown that the topic associated with DIQ in manufacturing was the second significant extracted thread. But these threads were not much new and growing.

To find the emerging research areas, we used the division of growth overage. According to this score, the first two emergence areas are related to entity resolution. It is found that linkage and privacy are the primary emerging research areas of DIQ. This was followed by the DIQ issues of entity resolution using ontologies, social media and also genetic field problems. Although there is no research about the emerging research areas of DIQ management, the results of this study are entirely consistent with the previous studies of DIQ evolutions. For example, according to the research in 2016, the DIQ changes are entirely

Fig. 7. Threads of DIQ from top 100 communities.
find some more metrics for ranking the emerging areas. It can be helpful in correlating some other features like a citation, top-cited articles and authors on the growth of the proposed research area. In other words, there would be some correlation between the citation patterns or the author’s behaviors in shaping a new scientific discipline or emergence of modern science for future study. One of the problems with the proposed method is that there is no penalty function for the nodes to leave their previous communities and it will decrease the accuracy of the model to move in different timestamps. It will be critical to propose the models that penalty the nodes for leaving the community which were in the previous timestamps. It means that there will be just some nodes that create the new communities and also accept the penalty function.

References


Dernis, H., Squicciarini, M., Pinho, R., 2016. Detecting the emergence of technologies and some nodes that create the new communities and also accept the penalty function.

Table 3

Table 4

Table 5