The Trend of Researches in Digital Archives

Fang-Ming Hsu\textsuperscript{a}, Chun-Min Lin\textsuperscript{a,b} and Chiu-Tsu Fan\textsuperscript{c}

\textsuperscript{a}Department of Information Management, National Dong Hwa University, Taiwan, R.O.C.
\textsuperscript{b}Center of General Education, Taiwan Hospitality and Tourism College, Taiwan, R.O.C.
\textsuperscript{c}Department of Business Administration, Tungnan University, Taiwan, R.O.C.

\textit{fmhsu@mail.ndhu.edu.tw}

Abstract. Archives are extremely valuable parts of cultural heritage. Through various information technologies (IT), tremendous amount of digital archives (DA) are created and preserved. These archives are the critical for providing evidence in everlasting memory of human society. The management of digital archives becomes a fast growing field throughout last decade and introduces abundant articles in academic research area. However, the trend of researches in digital archives remains obscure in the academic community. To map the trend of DA research, this study uses Google Scholar as the source of articles for DA researches. 100 articles per year regarding DA research from 2004 to 2014 were collected as the base for analysis. In this study, text mining techniques, such as co-word and cluster analysis, have been deployed to investigate the trend of DA literature. The title, keywords and abstract of articles are used to count the frequency of noun terms. The noun terms in articles are isolated and their frequencies are accumulated. Besides the term frequency (TF), the term frequency–inverse document frequency (TFIDF) is also used to reflect how important the term is to an article in the collection. Only the nouns with high TFIDF and TF frequency are left. In co-word analysis, this study used the concept of association rule method to compute the count of two terms appearing simultaneously in articles. The frequency of two terms which appear simultaneously in articles represents their closeness. Therefore, we compute the amount of articles that each two terms exist as the co-word matrix. Finally, the co-word matrix of terms is aggregated into clusters by Ward’s method. Seven clusters of selected articles are elaborated as follows: Data and Information, Platform and Repository, IT and Lone-term Preservation, Information Interchange, Archival Science, Digital Archives Projects, and Electronic Records Management (ERM) and E-government. Concerning the distribution of terms and clusters, research articles in DA are more focused on technology than management issues. The findings of this study can help identify the research direction of DA research, provide a valuable basis for researchers to access DA literature, and act as an exemplary model for future research.

Keywords: Digital Archives, Co-word Analysis, Text Mining

INTRODUCTION

Archival records are important form of explicit knowledge in organizations. Since the late 1990s, many government agencies have launched projects with a particular
emphasis on using IT to provide electronic information and services to citizens and businesses [1]. This digital information cannot be considered trustworthy and is easily lost in a self-perpetuating and expensive cycle of obsolescence and incompatibility. Archives should be improved not only in terms of the completeness of planning, but also in terms of its specificity and consistency of services [2]. In the era of information technology (IT), digital archives are critical to knowledge management especially in digital society. Government agencies often organize their services and operations into programs that may be changed in response to a host of factors, including IT implementations. After the promotion of electronic government, tremendous records are created and preserved through records and archives management system. Through various IT and media, past activities of humanity are preserved as critical memory after filing. A digital archive is a repository that stores one or more collections of digital information objects with the intention of providing long-term access to the information. Agencies should carefully preserve digital archives to protect their authenticity, reliability, integrity and accessibility for long-term usage.

The long-term maintenance of the collective human memory in its original form is not sustainable. All physical carriers are subject to degradation and the information stored on such carriers is bound to vanish. Only a re-mediation of the original documents can prevent precious knowledge from being permanently lost [3]. Many digital archives (DA) programs regarding human activities have been developed in last decades. Several trends are emerging in the management of digital archives, including a shift from paper-based storage to computer-based systems, from managing information to supporting its access and retrieval, and from cost-reduction to continued process improvement. These trends all highlight the need for the management of digital archives in the operation of digitizing artifacts.

Digital archives are created by information systems (IS). Traditionally, the management of digital archives include all aspects of archival science – as more traditionally understood through the life cycle model, as well as all aspects of the creation, preservation, use, and disposition of archives though IT. The concept of DA research is similarly broadly construed and also includes research on archival and recordkeeping topics being undertaken by researchers. Many issues are blooming in this field such as authenticity, digital signature, migration, encapsulation, digital certification and social network. For better preservation and management of archives, agencies start to cooperate with IS vendors to meet the regulation of rules. These alliance activities have dependencies in the sense that the DA operations influence the others, and these dependencies need to be managed.

The development of archival research consciousness has unprecedented growth in the academy and in practice, as well as in scholarly awareness that the construct of the archives, and provides a rich locus for research and theorizing [4]. This tendency raises questions regarding what are important when their activities are not well known, what does IT consist of, and what is the focus of this management exercise. Research in DA management is unique insofar as it takes place within a multidisciplinary environment encompassing history, management, computer science, and archival science [5]. Due to various historical differences in organising, documenting and managing information across cultural institutions, cross-domain resource discovery in the cultural heritage sector remains problematic. One of the most critical problems for research archives is
their definition, which is also part of the recognition of their relevance. Except from the small group of specialists involved in documenting, recording, managing and preserving the digital records and heritage of science, the reasons and the solutions for these increasing complexities and the effort to transform the traditional frameworks and tools into efficient and updated proposals have not been fully investigated [6].

Although DA is becoming increasingly common in this information age, our understanding of their operation and management does not reflect their expanding role in organizations. The understanding regarding the intellectual structure of DA about technology, management and social behavior has been limited. The objective of this study is to investigate the trend and intellectual structure of DA among milieu factors of practice, concept, science and technology.

LITERATURE REVIEW

The literature on information technology, systems development, records management, and archives provides the theoretical basis for the DA management. Since the 1990s, digital archives have advanced tremendously – not only in terms of the scope of technology it encompasses, but also in the development of its knowledge.

Related Research on the Management of Digital Archives

A number of prior studies have reflected on archives management as an academic discipline. Assessment of DA as a field, from both an historical and a forward-looking perspective, served as one important theme for archive management. Archival literature could be classified into 10 categories: arrangement and description; history, organization, and activities of repositories; management of current records; general literature; preservation, restoration, and storage; application of photographic processes; appraisal and disposition; training and professional development; special physical types of records and manuscripts; and historical editing and documentary publication [7]. The typology of research fields in DA included: 1. Developing broader education and practical training in the DA management, 2. Employer needs in the DA managements, 3. Archival perspectives on the DA management, 4. DA management programs in organizations, 5. Archivists’ attitudes towards technology, 6. Archival profession and technology. A certain amount of effort should be put into records to assure the authenticity, integrity and accessibility of the records and archives [9], especially under the concept of records continuum.

Digital libraries are complex information systems and therefore demand formal foundations lest development efforts diverge and interoperability suffers. In this article, we propose the fundamental abstractions of streams, structures, spaces, scenarios, and societies, which allow us to define digital libraries rigorously and usefully. Streams are sequences of arbitrary items used to describe both static and dynamic (e.g., video) content. Structures can be viewed as labeled directed graphs, which impose organization. Spaces are sets with operations on those sets that obey certain constraints. Scenarios consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement. Societies are sets of entities and activities and the relationships among them. Together these abstractions provide a formal foundation to define, relate, and unify concepts - among others, of digital objects, metadata, collections,
and services - required to formalize and elucidate “digital libraries.” The applicability, versatility, and unifying power of the model are demonstrated through its use in three distinct applications: building and interpretation of a digital library taxonomy, informal and formal analysis of case studies of digital libraries, and utilization as a formal basis for a description language of digital libraries.

Academic disciplines typically seek to articulate the intellectual structures upon which they can cultivate their futures [10]. The typology of research fields in archival science included: 1. The object and aim of archival science, 2. Archives and society, 3. The history of archives and of archival science, 4. Archival functions, 5. The management of archival programs and services, 6. Technology, 7. Types of media and archives: electronic records, 8. Archival environments, 9. Specific issues related to archives [5]. The articles of archival science could generate seven subject categories: digital libraries and digital archiving technologies, online resources and finding aids, archives and archivists, legal and political issues, electronic records and technical issues, records and information management, and e-mail and information professionals. Finally, these seven subject categories were merged into three sectors: digital library, archives and management [11]. That study describes dynamic change in the 2001-2004 research themes from traditional single-subject areas to emerging, complex subject areas. Digital archives can be conceptualized as a package of standards. It builds on existing technical standards (e.g., with respect to operating systems, databases, and network standards). It embeds procedural and performance standards as well as numerous classification schemes and terminologies [12].

Prior Research on Intellectual Structure and Bibliometric Method

To understand the issues covered by DA along with their possible solutions, governments may find specific and objective references from numerous research topics in academic literature. The way to locate research trends through the classification of a large number of academic articles is to comprehend the intellectual structure of this subject area and its evolvement over time. The discipline of bibliometrics provides tools for the study of intellectual structure of research subjects. The method of co-word analysis, a powerful tool of bibliometrics, is used in this study to recognize the intellectual structure of DA.

Bibliometric methods have been successfully applied to examine the intellectual structure of several disciplines. Bibliometrics is a research method commonly used in library and information science. It uses quantitative analysis and statistics to depict patterns of publication within a given field or body of literature. Researchers may use bibliometrics to determine the influence of an article, for example, to describe the relationship between the given article and the other articles. Co-word analysis reflects many papers that have cited any particular pair of terms and it is explained as a measure of similarity of content of the two terms. Co-word analysis has been accepted as a reasonable way to map the relationships among concepts, ideas, and problems. In co-word analysis, it is assumed that keywords extracted from papers could represent a specific research direction, research topic or subject of a field. If two keywords co-occur within one paper, the two research topics they represent are related. Higher co-word frequency means stronger correlation in keywords pairs, which can further suggest that
two keywords are related to a specific research topic. Co-word analysis has the potential of effectively revealing patterns and trends in a specific discipline [13].

Co-word analysis is a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., words or noun phrases) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these texts. Indexes based on the co-occurrence frequency of items, such as an inclusion index and a proximity index, are used to measure the strength of relationships between items. Based on these indexes, items are clustered into groups and displayed in network maps. This study applied co-word analysis and cluster analysis to gain insights to the research paradigms of DA research field.

RESEARCH METHODOLOGY

The purpose of this study is to explore and map the trend and intellectual structure of DA studies. With bibliometric analysis, this study had five phases, each of which required different approaches to examine the evolution of the DA studies. In this study, Google Scholar (GS) is used as the database for DA articles. 100 research articles with DA highly cited were collected each year within 2004 and 2014. Words and terms used in the titles and abstracts of these literature articles were extracted and their numbers of frequency calculated. The co-occurrence of words and terms is used as the grouping measure. According to the literature of co-word analysis, such an approach can detect the distance among words, which is the similarity of themes, and hence create clusters among words. Further analysis of each cluster of words can help locate the topic they represented, namely the intellectual structure of this study. After identifying the intellectual structure of DA, the correlation between the two clusters is further examined to find the similarities and differences and to results in the intellectual structure of DS in integration.

This study utilized the keyword search strategy which involved search for the keyword Digital Archives in GS databases to identify the potential DA research articles. The data used in this study includes journals, publication titles, article names, publication dates, and keywords. Besides the field of traditional DA management, some fields, especially the medical field, also focused on digitizing information and brought about some research literature. After the digitalization of medical matters, some issues need to be addressed urgently, such as the legal issues of digitized medical records and the security of information transmission [14]. To exclude medical-related literature, four high-frequency medical-related terms were located from the literature, i.e. ‘health’, ‘medical’, ‘nursing’ and ‘treatment.’ These four terms were excluded in the re-queries of articles in GS.

When conducting text analysis, few issues were encountered, which included the selection of language, computer codes, duplication of information, special characters, stemming, abbreviations and conflation, stop-words and compound words, etc. 100 DA articles each year were obtained from 2004 to 2014 annually with a total of 1,100 articles. Besides, since English is the major language for data analysis, the articles not presented in English are eliminated. Furthermore, the articles duplicated from different sources or without abstract are also eliminated. After this condensation step, 886 articles are left. Then the text contents of an article are extracted. Title and keywords reflect the key
concept of an article. The title, abstract, and keywords are combined together and hereinafter referred to as an ‘article’ in this study.

This study extracts key terms to cluster the articles. The next step is to perform the data collection and analysis of terms in the data set. The most frequently used term in total time period were identified as the core terms in the field and further examined with co-word analysis. The co-word count for each pair of terms is retrieved through 2-items set of association rule method. The co-word count represents the similarity of each pair of terms. A series of operations are conducted, through which it is possible to identify the intellectual structure of DA studies.

Cluster analysis is also commonly used program to map the intellectual structure of studies and determine the common links between articles. The last step of cluster analysis is performed to group these articles according to the similarity of their research themes and focuses. By taking the co-citation matrix and grouping the articles using cluster analysis of the correlations between the entries, this study can determine which articles are grouped together as well as their common shared elements. The closeness of article points on these maps is algorithmically related to their similarity as perceived by citers.

In this study, the academic literature for DA was collected through GS by means of co-word analysis and formed the article bank. The words used in each article were analyzed into units of terms. Terms with representative value were then located through the analysis of occurring frequency. Finally, the relationship between and among terms were located by the use of co-word analysis and cluster analysis to establish the intellectual structure of DA. The process of research steps in this study is illustrated in Figure 1.
RESEARCH RESULTS AND FINDINGS

Highly Frequent Terms

This study used Google Scholar to retrieve the research articles in DA area. The title, keywords and abstract of articles are used to count the frequency of noun terms. The noun terms in articles are isolated and their frequencies are accumulated. Besides the term frequency (TF), the term frequency–inverse document frequency (TFIDF) is also counted to reflect how important the term is to a article in the collection. Only the nouns with high TFIDF and TF frequency are left. After comparison, this study finds that these terms are included in the glossary or thesaurus proposed by Australia, Canada, China, Taiwan or USA, e.g. the Australian Governments’ Interactive Functions Thesaurus (AGIFT) proposed by National Archives Australia. The top 100 terms with high frequency in total articles are described in table 1. All terms are converted to singular form. Since the key term Digital Archives may appear at any place of a research, it may not appear in title, abstract and keyword. Therefore, not all articles include the term Digital Archives in their extracted text. The term digital archive occurs with highest frequency. The following terms are data, archive, information, system, record, and management, etc.
Furthermore, the amount of articles which include key terms is counted from 2004 to 2014 as depicted in Table 2. The terms data, information and system keep the important position. The term record is increasingly important. The terms archive and content are less important than before.
Table 2. The amount of articles which include top 20 terms from 2004 to 2014

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>23</td>
<td>38</td>
<td>40</td>
<td>18</td>
<td>34</td>
<td>37</td>
<td>36</td>
<td>15</td>
<td>18</td>
<td>39</td>
<td>32</td>
</tr>
<tr>
<td>archive</td>
<td>21</td>
<td>33</td>
<td>34</td>
<td>28</td>
<td>40</td>
<td>42</td>
<td>38</td>
<td>15</td>
<td>19</td>
<td>42</td>
<td>2</td>
</tr>
<tr>
<td>information</td>
<td>32</td>
<td>28</td>
<td>27</td>
<td>23</td>
<td>30</td>
<td>39</td>
<td>41</td>
<td>22</td>
<td>25</td>
<td>27</td>
<td>18</td>
</tr>
<tr>
<td>system</td>
<td>21</td>
<td>35</td>
<td>32</td>
<td>14</td>
<td>26</td>
<td>25</td>
<td>25</td>
<td>29</td>
<td>24</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>record</td>
<td>11</td>
<td>10</td>
<td>13</td>
<td>33</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>20</td>
<td>34</td>
<td>22</td>
<td>61</td>
</tr>
<tr>
<td>management</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>29</td>
<td>18</td>
<td>14</td>
<td>23</td>
<td>24</td>
<td>18</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>access</td>
<td>9</td>
<td>22</td>
<td>20</td>
<td>13</td>
<td>25</td>
<td>16</td>
<td>20</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>electronic record</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>53</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>40</td>
<td>37</td>
<td>11</td>
</tr>
<tr>
<td>content</td>
<td>11</td>
<td>22</td>
<td>27</td>
<td>8</td>
<td>29</td>
<td>18</td>
<td>19</td>
<td>4</td>
<td>9</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>application</td>
<td>8</td>
<td>13</td>
<td>20</td>
<td>14</td>
<td>24</td>
<td>18</td>
<td>28</td>
<td>7</td>
<td>6</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>document</td>
<td>13</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>13</td>
<td>13</td>
<td>19</td>
<td>11</td>
<td>12</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>technology</td>
<td>7</td>
<td>16</td>
<td>15</td>
<td>9</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>5</td>
<td>4</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>web</td>
<td>12</td>
<td>15</td>
<td>13</td>
<td>4</td>
<td>20</td>
<td>18</td>
<td>23</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>structure</td>
<td>8</td>
<td>13</td>
<td>16</td>
<td>3</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>preservation</td>
<td>11</td>
<td>8</td>
<td>11</td>
<td>17</td>
<td>11</td>
<td>16</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>standard</td>
<td>16</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>12</td>
<td>9</td>
<td>8</td>
<td>14</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>program</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>16</td>
<td>9</td>
<td>12</td>
<td>3</td>
<td>13</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>network</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>file</td>
<td>6</td>
<td>10</td>
<td>11</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>

Co-word Analysis

In co-word analysis, this study used the concept of association rule method to compute the amount of articles which two terms appear simultaneously. The frequency of two terms which appear simultaneously in an article represents their closeness in co-word matrix. Therefore, we compute the amount of articles that each two terms exist as the co-word (100 x 100) matrix. The relationship among top 20 frequent terms is depicted in Figure 2.
The size of node in Figure 2 represents its TFIDF and the width of edge represents its amount of articles that two terms exist simultaneously. In Figure 2, there are close relationship among archive, data, information and system in digital archives researches. The co-word amounts of two terms are counted for co-word matrix of selected terms. This co-word matrix is then used as the correlation input for hierarchical cluster analysis with Ward’s method to evaluate the relationship of terms. Using hierarchical clustering analysis, these 100 selected terms can be aggregated into seven clusters. Seven clusters of selected articles are described as follows:

**Cluster 1: Data and Information**

Terms such as data, record, archive, information, content, access, web and application are included in cluster 1 regarding traditional research regarding archival information. It is the critical part of DA researches.

**Cluster 2: Platform and Repository**

Terms such as information system, electronic records management system (ERMS), file, digitization, platform, storage and repository are included in cluster 2 regarding ERMS. A trusted digital repository (TDR) is a set of metrics that are used to certify that a given repository is an appropriate custodian of a collection of digital assets. More than an array of abstract measures, however, a TDR represents a stable and sustainable organization, a set of policies and procedures for sound management of the digital objects, and a robust and secure technical platform. Digitization and repository are important to DA research.

**Cluster 3: IT and Long-term Preservation**

Terms such as information technology, archivist, description, index, strategy, planning, long-term preservation, national archives, authenticity, ontology, integrity, maintenance, and international standard are included in cluster 3 regarding long-term preservation. Archival quality is closely associated through the preservation management of digital surrogates.

**Cluster 4: Information Interchange**

Terms such as interoperability, xml, code, copyright, video, and server are included in cluster 4 regarding information interchange. Information technologies support the transmission and sharing of DA.

**Cluster 5: Archival Science**

Terms such as digital archives management, archives management, information management, evaluation, classification, and migration are included in cluster 5 regarding digital archives management. The archiving of items is not a mere copying process of bits and bytes from object to another but rather a transformation of a digital object that is made to fit the requirements of provenance and authenticity. This transformational process goes beyond the traditional practices of collection, documentation, and preservation, leading not merely to a change of the context in which the object is embedded but to a change of the object itself. Digital archives management is the central part of DA researches.
Cluster 6: Digital Archives Projects

Terms such as Electronic Records Archives (ERA) project, national digital archives program, and Taiwan e-learning and digital archives project are included in cluster 6 regarding digital archives projects. Many projects in different countries are issued for building practical DA.

Cluster 7: Electronic Records Management and E-government

Terms such as e-government, agency, electronic records management, e-learning, electronic document, arrangement, appraisal, and cloud computing are included in cluster 7 regarding the electronic records management in electronic government. DA applications used in institutions should be restructured providing an integrated and centralized digital recordkeeping system in order for controlling all records of the organization in all media and form. The understanding of electronic records is its analysis of the attributes of a record based on concepts and principles that have evolved over centuries of detailed study of the documentary process.

CONCLUSIONS AND IMPLICATION

In this century, the volume of digital information is increasing at an extraordinary rate. Owing to the rapid developments of information technology, our intellectual capital of digital objects is increasingly at risk by the volatile character. The preservation of digital archives compromising its authenticity and long-term access are fundamental challenges. The archival bond is defined as “the interrelationships between a record and other records resulting from the same activity.” The International Council on Archives (ICA) Guide notes that “the rapid proliferation of text and data files” made inventorying and preserving digital archives difficult, and turned archivists’ attention to the question of developing policies and practices to ameliorate this decentralized and uncontrolled situation. With the growth of networking and the development of paperless transactions, archivists have become increasingly concerned about the long-term preservation of digital archives. A number of researches have been done on the topic regarding DA.

Text mining techniques are used in this study. From the trend of key terms between 2004 and 2014, data, information and system play important roles in DA articles. Record is an increasingly important term also. From co-word analysis, we can find the close relationship among archive, data, information and system that these four terms often occur simultaneously in an article. It reflects that a system is needed for suitably organizing data in archive to render information.

Seven clusters are concluded as Archival Information, Electronic Records Management Systems, Lone-term Preservation, Interoperability, Digital Archives Management, Digital Archives Projects, and E-government and ERM. The central research issue is the Digital Archives Management cluster. The E-government and ERM cluster is closer to the Digital Archives Management cluster than others. Therefore, the management issues are the focused points in DA researches. Although, there are still some researches focused on the technology/system field than on the management field. Further study could investigate the intellectual structure effect on the articles from other publication database, e.g. Web of Science, for more information.

ACKNOWLEDGMENTS

The authors would like to thank the financial support from the Ministry of Science and Technology, Taiwan (MOST 103-2410-H-259-022).
REFERENCES


