International Journal of Medical Informatics 101 (2017) 58–67

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Review article

Computer-aided diagnosis: A survey with bibliometric analysis

Ryohei Takahashi*, Yuya Kajikawa

Department of Innovation Science, School of Environment and Society, Tokyo Institute of Technology, 3-3-6 Shibaura, Minato-ku, Tokyo, Japan

ARTICLE INFO

Article history:
Received 27 August 2016
Received in revised form 28 January 2017
Accepted 4 February 2017

Keywords:
Computer-aided diagnosis
CAD
Citation network analysis
Bibliometric analysis

ABSTRACT

Computer-aided diagnosis (CAD) has been a promising area of research over the last two decades. However, CAD is a very complicated subject because it involves a number of medicine and engineering-related fields. To develop a research overview of CAD, we conducted a literature survey with bibliometric analysis, which we report here. Our study determined that CAD research has been classified and categorized according to disease type and imaging modality. This classification began with the CAD of mammograms and eventually progressed to that of brain disease. Furthermore, based on our results, we discuss future directions and opportunities for CAD research. First, in contrast to the typical hypothetical approach, the data-driven approach has shown promise. Second, the normalization of the test datasets and an evaluation method is necessary when adopting an algorithm and a system. Third, we discuss opportunities for the co-evolution of CAD research and imaging instruments—for example, the CAD of bones and pancreatic cancer. Fourth, the potential of synergy with CAD and clinical decision support systems is also discussed.

1. Introduction

In the last two decades, significant progress has been made in Information Technology. This includes the proposal of new imaging modalities, because of which developments in the computerized analysis of medical images are expected in the future. Diagnosis through computerized analysis, known as computer-aided diagnosis (CAD), was originally developed for breast cancer using mammography in the 1960s, and has since been extended to the diagnosis of lung cancer, colorectal cancer, and so on [1–3].

Using CAD, doctors, primarily radiologists, can use computer support as a "second opinion" and make a final decision more quickly and with greater confidence. There are two types of CAD research—"Detection" and "Diagnosis"—and it consists of two phases—a "propose phase" and an "evaluation phase" [4] (Fig. 1).

In CAD research, "Detection" implies a technology designed to reduce observational oversight in general by marking the regions of an image that have potential for specific abnormalities. "Diagnosis" implies a technology designed to assess a disease using image-based information. Thus, CAD is an important technology to reduce the burden on doctors and medical staff, and shorten the time required for the interpretation of medical images. Furthermore, CAD systems have increased the accuracy of diagnosis, which has led to their increased use over the years, such that CAD technology is now a major research subject in medicine [1,5–8]. CAD systems might have been separately developed in each department of medicine, such as imaging modalities. However, it is too difficult to comprehensively determine the research trends and the "big picture" in CAD.

In this article, we present an overview of recent developments in CAD to support future cross-sectional studies. In general, there are two main approaches to grasping a research overview: expert-based approach and computer-based approach [9]. The expert-based approach involves convincing, but is often subjective and admits of the possibility that some subjects are overlooked in the research. On the contrary, the computer-based approach presents objective results and provides an accurate overview of the medical scenario in question. Several studies have been conducted using the expert-based approach in CAD [4,6,7,10–15], but these researches have a little chance to overlook the important fact which author did not grasp. Hence, we used a computer-based approach called bibliometric analysis for objective discussion in this paper.

2. Methodology and data

2.1. Methodology

Bibliometrics is a field of research in library and information science (LIS) that features various methods to quantitatively analyze

* Corresponding author.
E-mail addresses: takahashi.r.ay@m.titech.ac.jp, ryohei.takahashi.1988@gmail.com (R. Takahashi).

http://dx.doi.org/10.1016/j.ijmedinf.2017.02.004
1386-5056/

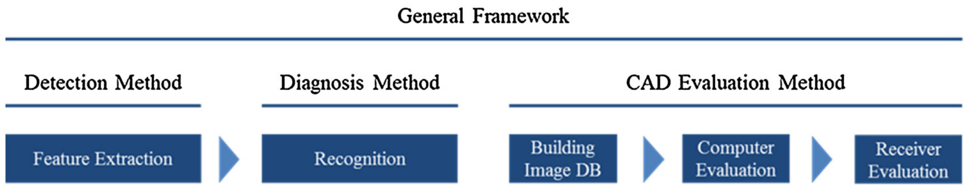

Fig. 1. General CAD framework.

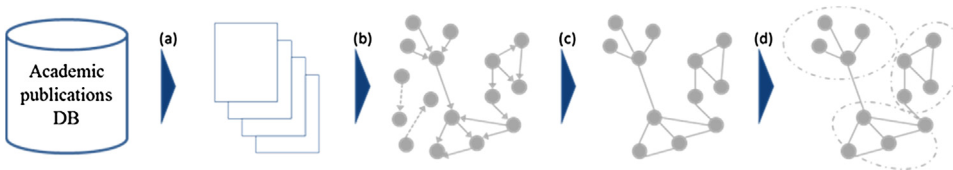

Fig. 2. Methodology overview.

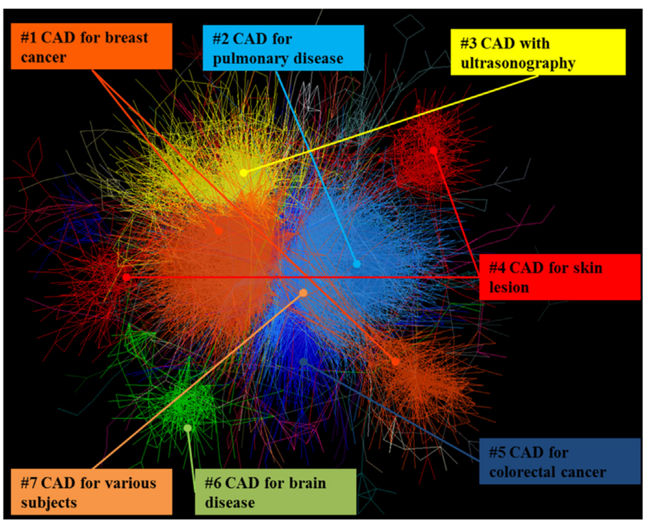

Fig. 3. CAD Academic research overview.

the bibliographic information of papers, patents, and so on. Bibliometric methodologies generally use Information Technology to process and analyze quantitative as well as qualitative data from bibliographic information and provide meaningful implications.

In this research, we selected citation network analysis, which is an effective bibliometrics methodology to identify an overview of an academic field. This technique analyzes the characteristics of a field with little chance of missing important research in each domain [16–19].

Fig. 2 shows an overview of our research methodology [16]. We first acquired relevant CAD research papers from each academic domain of interest using an academic publication database by using selected queries (Fig. 2a). We then constructed a citation network by regarding papers as nodes and direct citations as links (Fig. 2b). We did this because a previous study had indicated that direct citation is the best approach to detect emerging trends [20]. Following this, we eliminated irrelevant papers that were not linked to other papers in the largest graph component of the citation network to focus on the mainstream of research (Fig. 2c). Finally, we organized the network into clusters (Fig. 2d) using a topological clustering method known as Newman's algorithm [21]. In Newman's algorithm, clusters are divided into subsets in accordance with a rule that maximize modularity: Q. Then, Q is defined as follows:

$$Q = \sum_{s=1}^{M}\left[\frac{l_s}{l} - \left(\frac{d_s}{2l}\right)^2\right], \tag{1}$$

where Q is the independence of a cluster, M is the number of clusters, s is the cluster ID, l is the number of links in the entire network, l_s is the number of links inside module S, and d_s is the total number of links of nodes in S.

Newman's algorithm has been noted to build well-separated clusters in terms of research domains [16–19,22]. The network is visualized using a large graph layout (LGL) [23]. The LGL can help visualize large-scale networks containing thousands of nodes and millions of links within a reasonable computational time. For ease of recognition, intra-cluster links in the same network are expressed in the same color.

Following clustering, we analyzed the characteristics of each cluster, including the average publication year of papers, the number of citations, journal name, and the term frequency-inverse document frequency (tf-idf). Tf-idf is the best approach for discovering corresponding relationships between papers [18], and is defined as follows:

$$\text{Tf-idf} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log \frac{D}{\{d : d \ni t_i\}}, \tag{2}$$

where t_i is the given term, $n_{i,j}$ is the number of occurrences of term t_i, D is the total number of documents, and d is the number of documents containing term t_i.

In addition to simple keyword analysis; semantic similarities between clusters were measured to investigate semantic linkages of topics in each CAD field [24]. Semantic similarity was measured by cosine similarity [22]; defined as follows;

$$\text{CosineSimilarity}(t, s) = \frac{\bar{j_t} \bullet j_s}{\sqrt{\sum_i j_t^{(i)} \bullet j_s^{(i)}}}, \tag{3}$$

where t and s are clusters in each domain, and j_t and j_s are term vectors of clusters t and s, respectively. Cosine similarity increases when each cluster tends to share the same words more frequently, which implies the existence of common research topics among clusters

2.2. Data

We collected bibliographic data from academic publications related to CAD. Our data, including title, author, publication year,