Automation in Construction 59 (2015) 68–80

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis

Mehmet Yalcinkaya *, Vishal Singh [1]

Department of Civil and Structural Engineering, Aalto University, Rakentajanaukio 4A, 02150 Espoo, Finland

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 30 May 2015
Accepted 16 July 2015
Available online 4 September 2015

Keywords:
BIM
Latent Semantic Analysis
Research trends
Patterns

ABSTRACT

Building Information Modeling (BIM) has emerged as one of the key streams in construction and civil engineering research within the last decade. Given this interest in BIM and the rapidly increasing volume of BIM literature, it is important to understand and discern the core themes and trends emerging in BIM research, and its implications for broader research. The previously reported studies to identify the core of BIM research are typically subjective and qualitative, and hence, prone to bias and interpretation of a limited number of reviewed papers. There is a lack of comprehensive, quantified and systematic classification of the BIM literature. This research brings some clarity by synthesizing and labeling a large corpus of BIM research studies published from 2004 through 2014. Latent Semantic Analysis (LSA), a natural language processing technique was applied to the abstracts of 975 academic papers. This objective analysis reveals twelve principal research areas. Various specific research themes associated with each principal area have been identified. These principal research areas and research themes indicate the patterns and trends in BIM research.

1. Introduction

Building Information Modeling (BIM) has emerged as one of the key streams in construction and civil engineering and received a considerable amount of attention by researchers within the last decade, with a rapid increase in the number of related publications. The rapid increase in the volume of BIM literature poses a critical challenge in terms of identifying the research direction and trends, because despite the increasing volume of the literature, the core of BIM research and related themes remain unclear and poorly understood. Previous efforts to review the trends in academic BIM literature are typically qualitative, subjective and based on manual review. These reviews such as [1–6], provide an elaborate guidance, categorization and framework for implementation, adoption and technologies of BIM to the various participants of a project and to the researchers. Others such as [7–11] have discussed trends and future directions based on surveys, critical literature reviews, often with a specific frame of reference such as BIM for facilities management or so on.

The manual review, while insightful, is prone to be biased and limiting in terms of the number of articles that can be reviewed. In such manual reviews there is a potential tendency to read the more cited and influential papers, which while being a valid approach for review, might not be a comprehensive representation of the patterns in BIM literature. That is, it is important to note at the outset that we distinguish between critical review and patterns in literature. Instead of conducting a qualitative and subjective evaluation as to which themes are more important, we are interested in quantitative assessment based on textual analysis on how often the different themes have emerged in the BIM literature. The frequency and occurrence of themes can be assessed through semantic grouping of the keywords using techniques such as Latent Semantic Analysis (LSA), and such methods are well established as valid and useful methods to understand the patterns in literature.

LSA, as a text data mining and natural language processing (NLP) technique allows a systematic, computational and comprehensive analysis of a large corpus of literature for matching between query and document at topic level. In addition to the application areas of text data mining and NLP techniques across various disciplines, recent studies also demonstrate the viability of these techniques for the architecture, engineering and construction (AEC) industry. Similar text-based data mining and NLP techniques have been shown to be useful in the AEC industry for various aspects such as information retrieval, increasing the efficiency in decision making, predicting cost over-runs, reducing the human efforts, reducing errors in recognitions, and dispute resolution in construction accidents [12–21].

This research brings a structured representation of the patterns in BIM research by systematically applying LSA techniques to a large corpus of academic BIM articles published from 2004 through 2014. This systematic analysis of 975 paper abstracts reveals twelve principal BIM research areas. In addition, ninety specific research themes associated with the twelve principal areas have been identified and clustered.

* Corresponding author. Tel.: +358 40 321 8075.
E-mail addresses: mehmet.yalcinkaya@aalto.fi (M. Yalcinkaya), vishal.singh@aalto.fi (V. Singh).
[1] Tel.: +358 50 431 5667.

http://dx.doi.org/10.1016/j.autcon.2015.07.012
0926-5805/

These principal research areas and research themes comprise the comprehensive patterns of BIM research, which indicate the directions and trends in BIM research. The outcomes of analysis presented in this paper also provide opportunities for the researchers/professional to position their future BIM research and/or implementation strategies.

The paper is organized into three main sections. The first section describes the research data and the research methodology. This includes a detailed description of the data collection process, the LSA technique and its application to the collected data. The second section describes the results and findings from the analysis. Results are presented at various levels of details, ranging from twelve key principal research areas to ninety specific research themes. A cross analysis across the twelve principal research areas and ninety specific research themes is presented to identify their links. The third and final section concludes the paper with a brief discussion on the usefulness of the research findings, the limitations of the reported research, and the implications for future research.

2. Methodology and research data

The research data comprised of academic journals and conference papers collected through multiple academic databases and sources. Conference papers are in general less rigorously reviewed compared to the journal articles. However, they are included in the corpus, because instead of comparing or evaluating the strength or quality of studies, this paper aims to reflect the interest of research community on BIM topic. The following definition of BIM by the National Institute of Building Sciences buildingSMART alliance is considered as a fundamental reference to choose the relevant papers.

> "*A BIM is a digital representation of physical and functional characteristics of a facility. As such it serves as a shared knowledge resource for information about a facility forming a reliable basis for decisions during its lifecycle from inception onward*"
>
> [22]

There are three steps applied to collect and finalize the data. Initially the articles were identified using the search phrase "building information modeling" in their titles and/or abstracts and/or keywords. Instead of using "bim" by itself, which also represents other research topics from disciplines such as genetics or economics, phrases "bim" AND "building" and "bim" AND "information modeling" were used together to enlarge the corpus size. BIM-related textbooks, trade magazines, brochures and product/software related white papers were not considered, because the analysis focused on research publications. Since most research reports and theses result in academic articles, they were also not considered to prevent duplication in the corpus. Initially 980 paper abstracts were collected in the corpus. In the second step, "building information modeling" AND "bim" search phrases together in "full text" of the documents. In this step, the queried papers were manually screened to decide whether it is relevant to include in the corpus. If the searched keyword exists in an example, citation or acknowledgement, the related paper was not added to the corpus. At the end of this step, 89 more abstracts were added to the corpus. Table 1 summarizes the document collection process.

Table 1
Criteria for paper search and collection.

Round #	Search criteria	Number of papers collected
1	Search phrases/Within databases/ In Title OR Abstract OR Keywords	980
2	Search phrases/Within databases/ In full text	89
3	Elimination of duplications	94
Total		975

The 1069 abstracts (documents) collected through different sources were reviewed to eliminate duplication in the final step. After the final processing, 975 documents were saved in the research corpus. Tables 2 and 3 summarize the top sources of the included documents. Table 2 shows the distribution across the databases with the number of journals (J) and conferences (C), and Table 3 lists the top 20 contributing journals (J), showing the distribution of papers by year.

3. Data analysis

3.1. Introduction to Latent Semantic Analysis

The BIM literature dataset (corpus) is subjected to Latent Semantic Analysis (LSA), a data/text mining technique used to facilitate retrieving and querying large corpus of data [23–25]. As a mathematical and statistical method, LSA is used to identify the latent concepts within the textual data at the semantic level [26]. LSA employs a set of algorithms to convert unstructured text into structured data objects, and analyze these data objects to identify patterns for the discovery of knowledge [27]. The main idea behind LSA is to collect all the contexts belonging to the words in the corpus, and derive associated factors that represent related concepts. In this paper, a factor can be defined as a latent class representing multiple observed entities which have similar patterns that associated with the latent class. For example, in a 12 factor analysis, all the relevant entities that were found in the corpus relevant to the analysis are classified or represented through 12 latent classes, such that each of the entities in corpus is associated with or shows similar patterns to one of the 12 latent classes. Thus, each factor reserves and represents a certain amount of the overall observed entities, and the factors are organized in the order of how many entities they explain. A sample representation of factors and their associated entities are presented in Table A1 for information science discipline in [25]. A similar example is also presented in Table 4 for a five-factor solution to articulate the relationship between factors and corresponding high-loading terms. In any text, multiple words may share the same meaning and one word may have many synonyms in different contexts. LSA "loads" the words that share the same meaning to their associated concept and also "loads" one word to various latent semantics other than its main associated concept.

LSA differs from traditional factor analysis by applying Singular Value Decomposition (SVD) to reduce the dimension of the original data. This helps to present the collected 975 papers under the categories aggregated variously. Using SVD, LSA generates two sets of loadings in matrix format, one for the terms and one for the documents. Each factor in the matrices is associated with a high-loading term or a high-loading document, where each factor represents a research theme in the corpus. Higher term/document loading values for each factor indicate a greater possibility that the related term/document discloses certain theme. The researcher can alter the level of detail for identifying the research

Table 2
Number of papers collected across different databases.

Database Name / Years	2014–12		2011–09		2008–06		2005–04		TOTAL	
	J.	C.	J.	C.	J.	C.	J.	C.	J.	C.
ASCE	63	53	15	30	4	22	0	2	82	107
CuminCAD	4	47	2	39	4	30	0	35	10	151
CIB-Library	0	9	0	50	0	9	0	2	0	70
EBSCO	2	0	7	0	2	0	0	0	11	0
ELSEVIER-Science Direct	153	0	58	0	21	0	7	0	239	0
EMERALD	15	0	7	0	5	0	2	0	29	0
IEEE Xplore	0	39	0	31	0	12	0	0	0	82
Proquest	54	22	8	10	7	3	1	2	70	37
SpringerLink	9	0	5	0	0	0	2	0	16	0
Taylor & Francis	35	0	11	0	2	1	1	1	49	2
Wiley Online Library	9	0	2	0	7	0	1	1	19	1
TOTAL	344	170	115	160	52	78	14	43	525	450