ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

AI-Based Project Similarity Evaluation Model Using Project Scope Statements

Taewoo Ko[1]*, H. David Jeong[2], JeeHee Lee[3]

[1] *Department of Construction Science, Texas A&M University, 3137 TAMU, College Station, TX 77843, USA,* E-mail address: woowoo1127@tamu.edu
[2] *Department of Construction Science, Texas A&M University, 3137 TAMU, College Station, TX 77843, USA,* E-mail address: djeong@tamu.edu
[3] *Department of Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy. Las Vegas, NV 89154, USA,* E-mail address: jeehee.lee@unlv.edu

Abstract: Historical data from comparable projects can serve as benchmarking data for an ongoing project's planning during the project scoping phase. As project owners typically store substantial amounts of data generated throughout project life cycles in digitized databases, they can capture appropriate data to support various project planning activities by accessing digital databases. One of the most important work tasks in this process is identifying one or more past projects comparable to a new project. The uniqueness and complexity of construction projects along with unorganized data, impede the reliable identification of comparable past projects. A project scope document provides the preliminary overview of a project in terms of the extent of the project and project requirements. However, narratives and free-formatted descriptions of project scopes are a significant and time-consuming barrier if a human needs to review them and determine similar projects. This study proposes an Artificial Intelligence-driven model for analyzing project scope descriptions and evaluating project similarity using natural language processing (NLP) techniques. The proposed algorithm can intelligently a) extract major work activities from unstructured descriptions held in a database and b) quantify similarities by considering the semantic features of texts representing work activities. The proposed model enhances historical comparable project identification by systematically analyzing project scopes.

Key words: project similarity evaluation, project scope, natural language processing, BERT, cosine similarity

1. INTRODUCTION

Obtaining and using knowledge and insights from comparable historical projects is desirable for better and reliable project planning and management at the pre-construction phase. Project owners typically store substantial amounts of data generated throughout project life cycles in digitized databases

However, with a large amount of available and growing project data, one of the most time-consuming but important processes is to identify past projects that are comparable to an ongoing project [1]. These projects can serve as references through which project owners can obtain useful knowledge and information for a new project. For example, the identification of similar projects in the preconstruction phase may allow cost estimators to quickly a) organize work breakdown structures (WBSs), b) determine major work items, and c) estimate the costs of those items [2].

An early project scope document defines project requirements and provides an initial project in terms of the extent of the project and major project requirements [3]. The descriptive narratives in the project scope documents are useful to identify similar projects. However, currently, comprehending content and capturing vital information on project similarity basically rely heavily on human experiences and professional knowledge [4]. Such dependence may not only require a significant amount of time but also create biases when practitioners have insufficient experience and knowledge. To overcome these limitations of current practice, there is a need to develop an approach to identifying similar projects by analyzing project scopes.

Recent advances in NLP techniques have the potential to address this problem. As an artificial intelligence (AI) subfield, NLP techniques allow computers to comprehend the contexts of human language and extract significant information from unstructured statements [5]. Specifically, an NLP-based text similarity assessment can determine how close two words or phrases are to each other in terms of their context or meaning [6]. Finally, NLP can be the most appropriate technique for analyzing unstructured project scope descriptions and evaluating the similarity between historical and ongoing projects. This study proposes a model that can assess project scope similarities to improve the identification of comparable projects.

2. BACKGROUND

2.1. Project similarity evaluation

An accurate evaluation of project similarities is a significant work task prior to obtaining useful information from completed projects for a new project. Previous studies on improving the accuracy of identifying similar projects have applied statistical or computational techniques. Du and Bormann [7] proposed an algorithm for appropriate case retrieval. This research adopted case-based reasoning (CBR) as a projection process and applied global sensitivity analysis (GSA) to reflect quantitative relations between project features in the CBR process. Hyung et al. [8] used generic algorithms for CBR to compute optimized weights of project details, including general information, material features, or construction area factors, for an application to a new project.

Qiao et al. [9] proposed a new methodology to quantify project similarities based on budget line items. They assumed that if two projects had a large percentage of similar budget line items, they could be considered identical. Torkanafar and Azar [10] used the work breakdown structure (WBS), a hierarchical decomposition of total projects, for a more accurate representation of projects.

The review of previous studies revealed that detailed and reliable project data is a significant factor for systematic similarity computation. However, incomplete project plans and a lack of complete project information poses a significant challenge to using this data in the early preconstruction phase.

2.2. Text encoding and Bidirectional Encoder Representations from Transformers (BERT)

Advancement in NLP techniques enables a computer program to comprehend the semantics of text data, compute the closeness between pairs of two words, phrases, or sentences, and then determine the similarity of the two project scopes [11]. Text data must be converted into a numeric format to perform any data analytics and information extraction. NLP techniques have enhanced their ability to recognize semantic relationships between two words and measure the similarities of sentences by encoding them into vectors [12]. NLP-driven text encoding preserves the context and relationship between sentences.

As an advanced text encoder, BERT enables the recognition of textual data by leveraging adjacent words to establish context [13]. Recent advancements in machine learning and deep learning have resulted in developing state-of-the-art text embedding methods that represent words as real-valued vectors [14]. In the process of text embedding, the represented vectors adequately preserve context or the semantic relation between words. Most text encoders apply neural network methods, which process words sequentially for a represented vector generation [15][16]. However, these text encoders contain an inherent limitation: If the sentence is excessively long, the information from the initial words is gradually blurred during the embedding process, resulting in information loss of the previous texts in the represented vector [17].

BERT can provide an alternative to these challenges. Specifically, the attention mechanism in BERT can deal with entire words in a sentence at once instead of sequentially and then learn the context of the target words based on all the surrounding words in a sentence [18]. Many previous studies have confirmed that BERT, which uses an attention mechanism, can achieve a more accurate outcome than other encoders that use sequential or directional methods.

Another significant point is that BERT serves as a pre-trained model that can be efficiently fine-tuned for a domain-specific task. Specifically, BERT can offer a pre-trained language representation model to understand natural language based on a large general text corpus [19]. Such a pre-trained model substantially helps in the accuracy of enhanced word embedding through fine-tuning as compared to training from scratch [20]. Furthermore, the already-encoded representation vectors from the pre-trained BERT model allow fine-tuning for specific needs to occur, with fewer data required than for training from scratch.

This paper demonstrates that BERT encodes the required work activities included in project scopes into sequential vectors. The encoded activities are applied to quantify the similarities between the two activities, which serve as inputs for calculating project scope similarities scores.

3. PROJECT SCOPE-BASED SIMILARITY EVALUATION METHOD

Figure 25 illustrates the systematic process for project scope-based similarity evaluation. The proposed method consists of three main tasks:

1. Text pre-processing: It converts raw project-scope data into a well-organized structured format for computational techniques.
2. Activity-level similarity evaluation: It calculates the similarity between work activities described in project scopes and the number of common work activities between two comparison projects.
3. Scope-level similarity evaluation: It quantifies project scope similarities based on common work activities and calculates similarity scores ranging from 0 to 1.

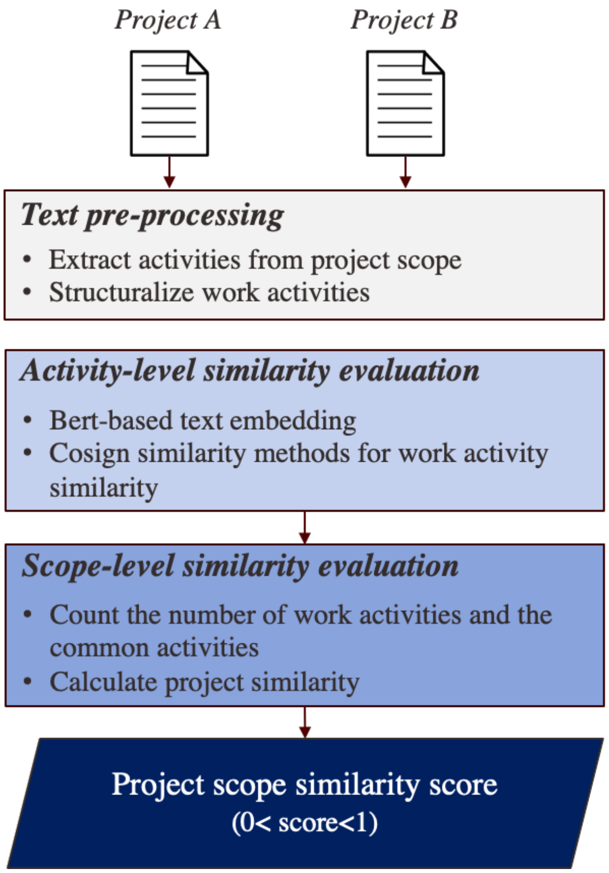

Project A
Project B
Text pre-processing
• Extract activities from project scope
• Structuralize work activities
Activity-level similarity evaluation
• Bert-based text embedding
• Cosign similarity methods for work activity similarity
Scope-level similarity evaluation
• Count the number of work activities and the common activities
• Calculate project similarity
Project scope similarity score
(0< score<1)

Figure 25. Project scope-based similarity evaluation process

3.1. Text pre-processing

Text pre-processing is performed to a) eliminate redundant texts and b) structuralize the description of the project scope for better performance of NLP techniques. A review of project scope documents reveals no specific structures or outlines for project scopes. Project owners typically develop their documents in free format or using bullet points. Thus, the unstructured raw texts and various representations of project scopes require text pre-processing or the structurization of project scopes.

Pre-processing involves extracting significant work activities and organizing them into a structured data set. This extraction begins with determining sequential syntactic patterns. The pre-defined patterns can serve as indicators for identifying phrases or clauses related to work activities. Specifically, this task analyzes the parts of speech (i.e., verb, adjective, noun, or preposition) of raw texts for project scopes. It then identifies text sequences whose parts of speech correspond to pre-defined syntactic patterns. The extracted work activities are stored in a structured data set for computational, technique-based similarity evaluation.

3.2. Activity-level similarity evaluation

This task aims to evaluate work activity similarities and count the number of common activities involved in the two projects. This task uses BERT to encode extracted work activities into real-valued vectors (see Figure 26). The cosine similarity, which computes the pairwise similarity between two chunks of text using the dot product of vectorized data, measures how similar two work activities are likely to be [21]. Cosine similarity represents the cosine of the angle between two vectors projected in a multidimensional space. The cosine similarity is calculated based on the formula below; where $\|A\|$ is the Euclidean norm of vector $A = (A_1, A_2, A_3, \ldots, A_n)$, defined as $\sqrt{A_1^2 + A_2^2 + \cdots + A_n^2}$.

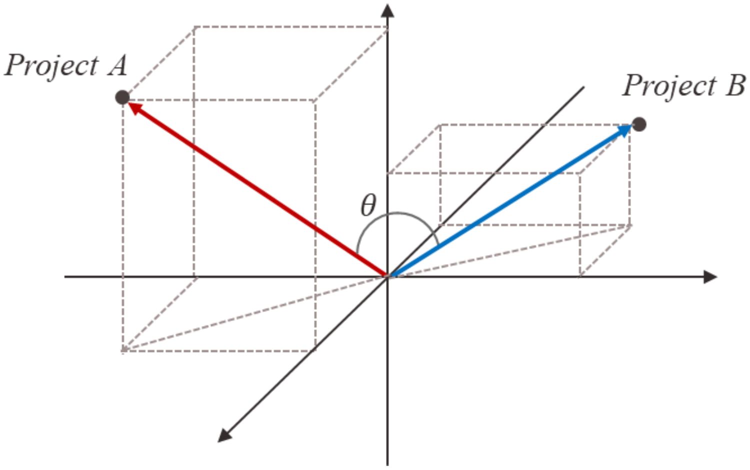


Figure 26. Project Vectorization and Cosine similarity

$$similarity(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

The similarity score of 0 indicates that the two vectors are 90° (orthogonal) and have no match with each other. The closer the similarity score is to 1, the smaller the angle and the greater the match between vectors. Whether the two compared work activities are identical is determined by binary classification based on a threshold score. The similarity score is more accurate than the threshold to determine if the two work activities are the same. On the other hand, if the similarity score is less than the threshold, the proposed process decides that the two are different. The proposed method compares various work activities individually and then counts the number of common work activities across the two projects.

3.3. Scope-level similarity evaluation

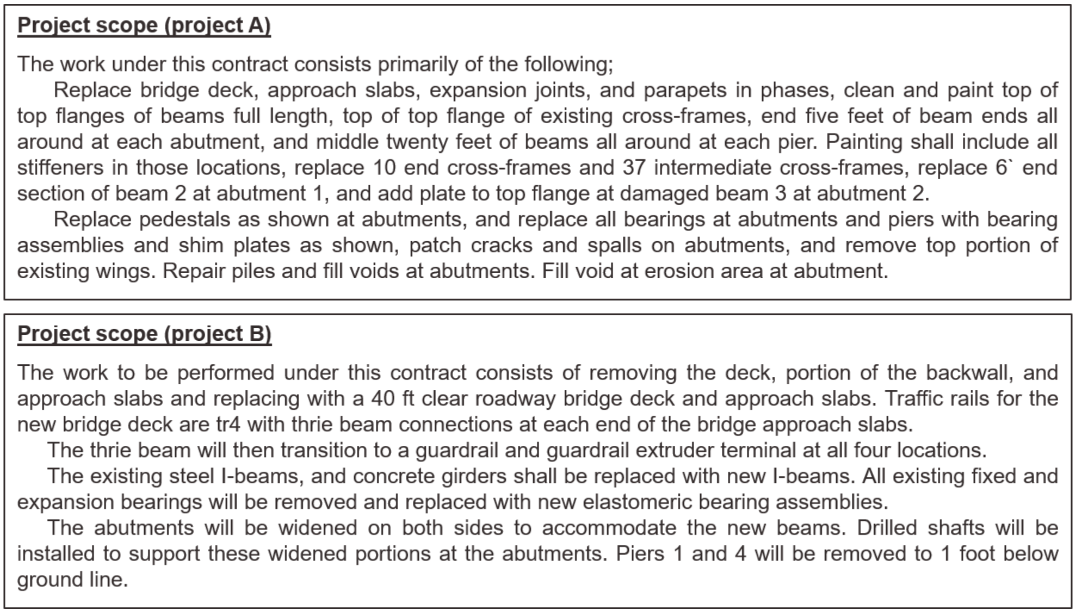
The objective of a scope-level similarity evaluation is to quantify the similarity by using activity-level similarity evaluation outcomes. The project similarity is measured as the ratio of common work activities to the number of activities in each of the two projects. Below is the formula for evaluating scope-level similarity:

$$similarity\ score\ based\ on\ project\ scope = \frac{2 * \gamma}{\alpha + \beta} \tag{2}$$

Where α and β refer to the total work activities of projects to be compared and γ indicates the number of common work activities.

4. CASE STUDY

This research uses a case study to demonstrate the entire process with actual data. The study assesses the accuracy and validity of the research outcomes and shows the practical issues and concerns for improved applicability. The scope-related documents used for the case study were gathered from a state department of transportation (DOT). Figure 27 shows the project scope description examples of two bridge rehabilitation projects.



Project scope (project A)

The work under this contract consists primarily of the following;

Replace bridge deck, approach slabs, expansion joints, and parapets in phases, clean and paint top of top flanges of beams full length, top of top flange of existing cross-frames, end five feet of beam ends all around at each abutment, and middle twenty feet of beams all around at each pier. Painting shall include all stiffeners in those locations, replace 10 end cross-frames and 37 intermediate cross-frames, replace 6` end section of beam 2 at abutment 1, and add plate to top flange at damaged beam 3 at abutment 2.

Replace pedestals as shown at abutments, and replace all bearings at abutments and piers with bearing assemblies and shim plates as shown, patch cracks and spalls on abutments, and remove top portion of existing wings. Repair piles and fill voids at abutments. Fill void at erosion area at abutment.

Project scope (project B)

The work to be performed under this contract consists of removing the deck, portion of the backwall, and approach slabs and replacing with a 40 ft clear roadway bridge deck and approach slabs. Traffic rails for the new bridge deck are tr4 with thrie beam connections at each end of the bridge approach slabs.

The thrie beam will then transition to a guardrail and guardrail extruder terminal at all four locations.

The existing steel I-beams, and concrete girders shall be replaced with new I-beams. All existing fixed and expansion bearings will be removed and replaced with new elastomeric bearing assemblies.

The abutments will be widened on both sides to accommodate the new beams. Drilled shafts will be installed to support these widened portions at the abutments. Piers 1 and 4 will be removed to 1 foot below ground line.

Figure 27. Two examples of project scope statements

Table 1. Similarity evaluation results of projects A and B

Attribute	Values
Number of work activities for project A	14
Number of work activities for project B	13
Number of common activities	9
Project similarity score	0.667

The proposed process analyzed the scope descriptions for the above two projects and extracted 14 work activities from project A and 13 activities from project B. Additionally, the activity-level similarity evaluation revealed that nine work activities are common to projects A and B. Then, the scope-level similarity evaluation calculated similarity scores using the results addressed by two prior tasks, which indicates that the similarity between projects A and B is 0.667.

Two discussion points arise from the above results. First, in the text pre-processing task, the quality of syntactic patterns is significant for enhancing the accuracy of work activity extraction. As project scopes are described in an unstructured manner, developing various syntactic patterns is required to eliminate redundant texts and minimize the omission of any work activity. Second, determining a threshold score is a major requirement for accurately identifying similar activities. A low threshold may cause an error in recognizing different work activities as the same.

Conversely, an excessively high threshold value may lead to errors in recognizing that the assumed similarities are different. These errors can cause an incorrect number of common work activities, which in turn leads to inaccuracies in determining project similarity. The case study used a threshold of 0.7. In order to increase the reliability of this study, additional work is required to analyze how project similarity is changed by applying various threshold values.

5. CONCLUSION AND FURTHER STUDY

This proposed method can help project owners identify comparable historical projects for more reliable planning. In the early pre-construction phase, with the lack of available data for comparable project identification, this method can analyze free-formatted project scope documents using cutting-edge NLP techniques. Systematic analysis can provide more reliable insights than the current practice of relying on simple project characteristics, including type, size, or location. Specifically, the scope analysis results can prioritize historical projects with the same characteristics as an ongoing project through quantified similarity scores. This prioritization contributes to effectively filtering out historical projects with identical characteristics but different scopes. Additionally, NLP techniques allow a computer to understand the context of project scope descriptions, which facilitates the efficiency of project scope recognition.

For improving logical coherence and applicability, additional tasks are needed. First, setting the appropriate threshold values for the same work activity identification should be performed. Because the threshold may vary for specific work activities, it is necessary to determine a particular threshold value that can minimize identification errors. It is also recommended to develop a visualized map of project scope similarities to provide project engineers with fast, clear, and easy-to-understand results.

REFERENCES

[1] M. Al Qady, A. Kandil, “Automatic clustering of construction project documents based on textual similarity”, Automation in construction, vol. 42, pp. 36-49, 2014.

[2] J. Ahn, M. Park, H.S. Lee, S.J. Ahn, S.H. ji, K. Song, “Covariance effect analysis of similarity measurement methods for early construction cost estimation using case-based reasoning”, Automation in Construction, vol. 81, pp. 254-266, 2017.

[3] M.K. Fageha, A.A. Aibinu, “Managing project scope definition to improve stakeholders’ participation and enhance project outcome”, Procedia-Social and Behavioral Sciences, vol. 74, pp. 154-164, 2013.

[4] N. Torkanfar, E.R. Azar, “Quantitative similarity assessment of construction projects using WBS-based metrics”, Advanced Engineering Informatics, vol. 46, 2020.

[5] R. Collobert, J. Weston, L. Bottou, M Karlen, K. Kavukcuoglu, P. Kuksa, “Natural language processing (almost) from scratch”, Journal of machine learning research, vol, 12, 2011.

[6] J. Wang, Y. Dong, “Measurement of text similarity: a survey”, Information, vol. 11, no. 9, p.421, 2020.

[7] J. Du, J. Bormann, “Improved similarity measure in case-based reasoning with global sensitivity analysis: An example of construction quantity estimating”, Journal of Computing in Civil engineering, vol. 28, no. 6, 2014.

[8] W.G. Hyung, S. Kim, J.K. Jo, “Improved similarity measure in case-based reasoning: A case study of construction cost estimation”, Engineering, Construction and Architectural Management, 2019.

[9] Y. Qiao, J.D. Fricker, S. Labi, “Quantifying the similarity between different project types based on their pay item compositions: application to bundling”, Journal of Construction Engineering and Management, vol. 145, no. 9, 2019.

[10] N. Torkanfar, E.R. Azar, “Quantitative similarity assessment of construction projects using WBS-based metrics”, Advanced Engineering Informatics, vol. 46, 2020.

[11] Y. Ji, J. Eisenstein, “Discriminative improvements to distributional sentence similarity”, Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 891-896, 2013.

[12] M. Farouk, “Measuring sentences similarity: a survey”, arXiv preprint arXiv:1910.03940, 2019.

[13] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.

[14] C. Liu, D. Yu, “BLCU-NLP at COIN-Shared Task1: Stagewise Fine-tuning BERT for Commonsense Inference in Everyday Narrations”, Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, 2019.

[15] J. Mueller, A. Thyagarajan, “Siamese recurrent architectures for learning sentence similarity”, In Proceedings of the AAAI conference on artificial intelligence, vol. 30, No. 1, 2016.

[16] J.W.G. Putra, T. Tokunaga, “Evaluating text coherence based on semantic similarity graph”, Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pp. 76-85, 2017.

[17] M. Mosbach, M. Andriushchenko, D. Klakow, “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines”, arXiv preprint arXiv:2006.04884, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need”, Advances in neural information processing systems, Vol. 30, 2017.

[19] J. Howard, S. Ruder, “Universal language model fine-tuning for text classification”, arXiv preprint arXiv:1801.06146, 2018.

[20] C. Sun, X. Qui, Y. Xu, X. Huang, “How to fine-tune bert for text classification?”, In China national conference on Chinese computational linguistics, pp. 194-206, 2019.

[21] A.R. Lahitani, A.E. Permanasari, N.A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment”, In 2016 4th International Conference on Cyber and IT Service Management, pp. 1-6, 2016.