Using Jaccard Similarity to Identify New Issues from AEC Project Team Meeting Minutes

**Hasan Gokberk Bayhan,[1] Yao Ma,[2] Joseph Thekinen, Ph.D.,[3]
Jiliang Tang, Ph.D.,[4] and Sinem Mollaoglu, Ph.D., A.M.ASCE[5]**

[1]Ph.D. Student, Construction Management Program, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824; e-mail: bayhanha@msu.edu
[2]Ph.D. Candidate, Department of Computer Science and Engineering, Michigan State Univ., East Lansing, MI 48824; e-mail: mayao4@msu.edu
[3]Research Associate, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824; e-mail: thekinen@msu.edu
[4]Assistant Professor, Department of Computer Science and Engineering, Michigan State Univ., East Lansing, MI 48824; e-mail: tangjili@msu.edu
[5]Professor, Construction Management Program, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824; e-mail: sinemm@msu.edu

ABSTRACT

Keeping track of issues and their documentation in Architecture, Engineering, and Construction (AEC) projects demand significant amounts of time, budget, and effort. While various types of documents and software aid coordination in AEC projects, project team meeting minutes, developed as a follow-up to periodic project team meetings, continue to be the most common and prominent type of documentation across project types for recording team communications, tasks, and assignments. Presently, identifying unique project issues and tracking their progress from meeting minutes is a manual process that is time-consuming and susceptible to error. Dynamic organizational structures to project teams, varying document formats from project to project and even within projects based on leading organizations during delivery, and changing milestones from different projects create challenges in automating this task. This study aims to automate the identification of project issues and track resolution timelines using project team meeting minute documents via the Jaccard Similarity method. In this study, over 50 AEC project team meeting minutes documents of varying formats from three different projects of various sizes were collected, automatically converted, and coded to train the Jaccard Similarity model for detecting new and continuing issues. Specifically, we treated individual entries in each meeting minute document as an issue data point on the date they first appeared and used key information from those entries as features. We modeled the task as a classification problem, labeling each item to either a new or a continued issue. Accuracy, precision, recall, and F1 parameters were tested, and the accuracy rates of 81.86% to 94.18% were obtained. The study provides the groundwork to automate the analysis of important information in project meeting minutes that include but are not limited to issue complexity, detection of bottlenecks, and analysis of expertise assignments for issue resolution.

INTRODUCTION

In AEC projects, periodic project team meetings that take place on weekly or biweekly intervals serve as an important communication channel to discuss project issues between responsible parties, including owner, designer, and contractor (Mincks and Johnson 2010). The meeting minutes document the agenda and topics discussed while tracking project progress and creating an execution plan for the unresolved issues (Javanmardi et al. 2020). Continuation of issues at periodic project team meeting minutes with or without resolution over time can serve as an indicator to assess process efficacy, issue complexity, and level of team integration or the lack thereof. Traditionally, issue identification from meeting minutes is handled manually by a human expert. Manual identification of issues from a high number of unstructured text-based meeting minutes documents is a tedious task.

Contractual clauses determine the responsibilities of the project partners as well as the communication processes and frequency. However, project team meeting minutes document the issues in a format that might not be fully structured. The dynamic structure of the construction process leads to changing formats even within projects that cause a lack of continuity. Keeping track of this documentation to present meaningful results could be dramatically catalyzed by the recent information technologies taking over manual efforts (Caldas et al. 2002). Even though utilizing cloud-based project management software providing standardized forms are proliferated among construction companies in recent years, the use is limited due to affordability issues. More than 90% of the industry’s companies in most developed countries belong to the group of Small Medium Enterprises with limited cash flow (Kumar et al. 2010). Moreover, in many cases, developing new frameworks for meeting minutes can improve the communication between parties.

This study aims to automate the identification of project issues and track resolution timelines using project team meeting minute documents via the Jaccard Similarity method. Four different meeting minute formats from three different projects were evaluated to train and test the model's validity. The objectives are to (1) automatically code the meeting minutes generated in the form of .pdf to .xls or .csv file in a hierarchical form; (2) identify the characteristic of the issue at hand by Jaccard similarity – whether it is a continuous or new task. This way, notwithstanding the format of the documents, the type of issues were identified directly from meeting minutes.

LITERATURE REVIEW

Meeting minute is the written representation of formal information exchange followed at the meetings. In the AEC industry, informal meeting characteristics are studied for inter-organizational teams considering the decision-making and goal-setting characteristics, scheduling, solving problems, and information sharing (Gorse and Emitt 2007, 2009, Wu et al. 2007, Javanmardi et al. 2020). Uses of project meeting observations and team meeting minute documents data include back charge claims (Kisi et al. 2020), constraint management and removal (Hamzeh et al. 2015, Wang 2016), and performance prediction via data mining during project delivery (Van Niekerk 2020). Kisi et al. (2020) also emphasized the importance of proper documentation and early notices via project meeting minute documentation.

A large percentage of the data in construction inter-organizational information systems is stored in text documents (Caldas et al. 2002). With the technological developments, especially in machine learning, a great deal of opportunity exists in performance tracking for all participants involved. Text recognition abilities do exist (Moon et al. 2021) but have not expanded to issue identification and resolution in AEC project management.

As a step forward in this direction, the Jaccard index (Jaccard 1902), as a superior and straightforward string-based similarity measuring index (Bag et al. 2019, Diana and Ulfa 2019) is a promising method to recognize text. In the construction industry, Abd Jamil and Fathi (2020) adopted Jaccard's coefficient in hierarchical cluster analysis for a more comprehensive data interpretation in the Building Information Modeling framework. It works on string chains and character organization using a term-based similarity procedure, ranging between 0 and 1, as the intersection is divided by the union of the objects. In other words, for text documents, the Jaccard similarity coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms (Huang et al. 2008).

DATASET and METHODOLOGY

Our dataset consists of meeting minutes archived from three different types of AEC projects with the following size and timelines:

- An ongoing construction project of 600+ million USD budget spanning over two years.
- A completed expansion project of 12+ million USD budget spanning over one year.
- A completed infrastructure project of 10+ million USD budget spanning over two years.

The coordination meetings are held weekly or biweekly, depending on the project needs and deadlines. The interdisciplinary meetings are held between different participants, including Owners or Owner’s representatives, Designers, Contractors, Design Assist Partners, and Subcontractors. At the end of each meeting, new issues since the last meeting and existing unresolved issues from previous meetings are documented in the meeting minutes. A meeting minute documents new and existing issues as bulleted points in a multi-level hierarchical structure. A multi-level hierarchical structure organizes the issues depending on responsible parties and the nature of the project.

The format of the meeting minute and its organization vary from one project to another. Even within a project, the format differs from one meeting type to another. As an example, some coordinators assign origin dates of individual issues discussed, and others skip this step. Independent of the format, we coded the issues in the meeting minutes in a consistent format following the rules, also could be seen from examples in Table 1: (a) Each line in our coding is a bullet point at the bottom of the hierarchy; and (b) we carry over the information from all upper-level headings.

We automated the process of converting text from meeting minute document format to our coding structure. Our Python code reads pdf files line by line using the PyDF2 package, filters text irrelevant to the meeting, and identifies headings and their hierarchy to code them in the structure shown in Table 1. It works on a variety of meeting minute formats commonly used in AEC meeting documentation. In Table 1, two different meeting minute types from two projects are coded, and the hierarchical structure is preserved.

Table 1. Issue Coding Examples Across Differently Formatted Meeting Minutes Documents using Python Code.

	Original Formats with Different Hierarchies	Hierarchy Converted into Sequence Logic Via Python Coding
Project 1 Issues	1 – SAFETY SITE-SAFETY o Abatement start on 2/2, air-monitoring in place o Abatement complete on 1-East, Demo to continue next week 2 – ADMINISTRATION BUDGET REVIEW o Need to schedule a Budget Review meeting with head	• SAFETY; SITE-SAFETY; Abatement start on 2/2, air-monitoring in place • SAFETY; SITE-SAFETY; Abatement complete on 1-East, Demo to continue next week • ADMINISTRATION; BUDGET REVIEW; Need to schedule a Budget Review meeting with head
Project 2 Issues	1) Overall Design Status a) Architect group i) Floor plans "approved" ii) Drawing status update a. Electrical - MDF room UPS is undetermined b. Electrical - clarify space for medical equipment b) DD and CD Deliverable coming end of this week 2) Design-Assist Responsibilities a) Plumbing i) Progress on underground drain b) Mechanical	• Overall Design Status; Architect group; Floor plan "approved" • Overall Design Status; Architect group; Drawing status update; Electrical - MDF room UPS is undetermined • Overall Design Status; Architect group; Drawing status update; Electrical - clarify space for medical equipment • Overall Design Status; DD and CD Deliverable coming end of this week • Design-Assist Responsibilities; Plumbing; Progress on underground drain • Design-Assist Responsibilities; Mechanical

After we converted all project meeting minute documents to lists of project issues as shown in Table 1, human coders coded labeled each issue as:

- **Continuous issue** if the issue is a continuation of an existing issue discussed in a previous meeting minute, or
- **New issue** if appeared for the first time.

To establish the dataset, five human coders (C1, C2, …, C5) coded four types of meeting minutes (M1A, M1B, M2, and M3) from these three projects (P1, P2, P3). The human coders are selected from the construction management profession to enhance the integrity of the coding process. For M1A, a total of 1440 row amounts of data are coded for 21 meeting minutes, where for P3, only 173 row amounts of data are coded for ten meeting minutes.

In the coded datasets, each issue is summarized by a description, which is in the form of several sentences. Hence, an issue can be represented as a sequence of words, i.e., $S = [w_1, \dots, w_{n_S}]$. We denoted a project P with n_P meetings as $\{M_1, M_2, \dots, M_{n_p}\}$, where M_i denote the i^{th} meeting minute. A meeting minute M with a total of n_M new and existing issues is represented as $\{s_1, s_2, \dots, s_{n_M}\}$. Comparing text similarity among the issues, we used Jaccard Similarity Coefficient. Figure 1 summarizes the Jaccard Model process.

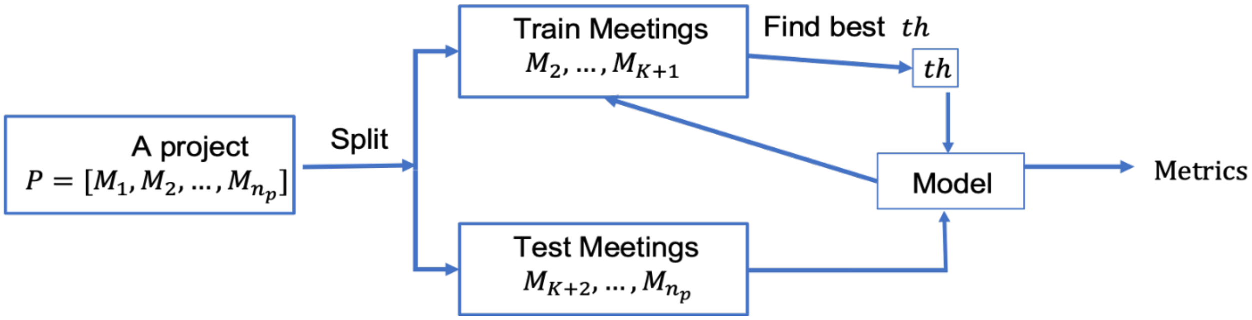

Figure 1. Jaccard Model Process

Given two sets $\mathrm{A} = \{\mathrm{a}_1, \ldots, \mathrm{a}_{\mathrm{n_A}}\}$ and $\mathrm{B} = \{\mathrm{b}_1, \ldots, \mathrm{b}_{\mathrm{n_B}}\}$, the Jaccard Coefficient is defined as follows:

$$\mathrm{JC(A, B)} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|},$$

Where $A \cap B$ is the intersection of the two sets, and $A \cup B$ is the union. We used the $|\cdot|$ to denote the cardinality of sets. We next introduced the proposed model, which is based on the Jaccard Similarity Coefficient. Specifically, given a new meeting minute M_k, for any issue s_j in M_k, we define a maximum similarity score as follows:

$$h(s_j) = \max\{\, \mathrm{JC}\,(\mathrm{s_o}, \mathrm{s_j}) | \forall\, \mathrm{s_o} \in \mathrm{M_i}, \mathrm{i} < \mathrm{k}\}.$$

This score measures the maximum similarity between the issue s_i and any exiting issue in previous meeting minutes. We used this score to perform the classification. If the score is larger than a certain threshold, the issue is similar to existing issues and hence labeled as a continuous issue; otherwise, it is a new issue.

To find a suitable threshold, we utilized the first K meeting minutes, also called “training”. The first meeting minutes of projects contained all new issues, so we ignored them in our training data. Hence, when we mention the first K meetings here, we refer to the second meeting as the $(K+1)st$ meeting. Then, we used the remaining meeting minutes, also called “test”, to test our model's performance with the selected threshold. We select the suitable threshold th from the range (0,1). For convenience, we set a step size of 0.05, i,e, we searched from the list [0.05, 0.1, … 0.90, 0.95]. Then, we implemented our model with all the possible thresholds in the list and chose the one with the largest accuracy as the best threshold, which is utilized as the final model for the test. For each of the projects, we increased K from 1 to 5, respectively, to find the better fit (Niwattanakul et al. 2013).

We adopted four different metrics, including accuracy, precision, recall, and F1 score, to measure the model performance. The definitions of these metrics are as follows:

$$Accuracy = \frac{\#correctly\ predicted\ issues}{\#total\ issues}$$

$$Precision = \frac{\#correctly\ predicted\ continuous\ issues}{\#total\ issues\ being\ predicted\ as\ continuous\ issues}$$

$$Recall = \frac{\#correctly\ predicted\ continuous\ issues}{\#total\ continuous\ issues}$$

$$F1 = 2 \cdot \frac{precision \times recall}{precision + recall}$$

RESULTS

The Jaccard similarity related performance results for different training levels and testing numbers are presented in the following two tables. The lower accuracy of the parameters at $K = 1$ is not given in a separate table. We tried larger numbers for K, such as 4 and 5 (not listed in the tables), and observed that the found thresholds are similar to those when we use $K = 3$. The project number, meeting minute type, and coder number (P_MM_C) is identified in the first column of the tables below.

Table 2. Training Results with Two Meeting Minute Documents

P_MM_C	# Train	#Test	Threshold	Accuracy	Precision	Recall	F1
P1_M1A_C1	2	18	0.35	0.8775	0.9557	0.8935	0.9236
P1_M1A_C2	2	16	0.3	0.8885	0.9448	0.9208	0.9326
P1_M1B_C2	2	10	0.7	0.7222	0.975	0.4432	0.6093
P1_M1B_C4	2	6	0.4	0.755	0.9315	0.5152	0.6634
P2_M2_C3	2	11	0.3	0.9453	0.9417	0.9898	0.9652
P3_M3_C5	2	7	0.4	0.8815	0.9588	0.8857	0.9208

Table 3. Training Results with Three Meeting Minutes

P_MM_C	# Train	#Test	Threshold	Accuracy	Precision	Recall	F1
P1_M1A_C1	3	17	0.35	0.8806	0.9557	0.8981	0.9260
P1_M1A_C2	3	15	0.3	0.8926	0.9451	0.9263	0.9356
P1_M1B_C2	3	9	0.2	0.8186	0.7529	0.9296	0.8319
P1_M1B_C4	3	5	0.25	0.8277	0.8690	0.7849	0.8249
P2_M2_C3	3	10	0.25	0.9418	0.9340	0.9926	0.9624
P3_M3_C5	3	6	0.4	0.8696	0.9647	0.8723	0.9162

In most cases, the proposed model works quite well, producing high performance concerning all metrics due to its consistency in iterating each word. On average, Jaccard similarity performed better when K (training) is increased and selected 3. Especially, in the M1B, the model works quite well when $K = 3$. After the data examination, it is found that the M1B meeting minutes are mainly containing "new" issues. The dataset structure follows this trend because of the higher number of parties involved in this type of meeting minute. Hence, there is not enough data (especially "continuous" issue) to find the best threshold th to differentiate between "new" issues and "continuous" issues. Apart from the table above, coders' accuracy is 94.2% and %79.1 respectively in M1A and M1B. Therefore, deciding on M1B items is also more challenging for the human coders could explain the relatively lower accuracy rates. The number of files coded in M1A and M1B is different for coders. However, different meeting minute codes and the slight differences in issue type decisions did not affect the parameters with a statistically meaningful trend. The comparative results verified the functionality of the threshold number adjustments.

CONCLUSION

Even though informal information may affect AEC, formal documents are the primary measure determining the resolution status and construction performances. The most reliable and common type of documentation in AEC projects is project meeting minutes. Evaluating unique issues and assessing their resolution status from the meeting minutes is a tedious and manually managed task. Our study achieved automatically converting and coding the meeting minutes to .xls or .csv file in a hierarchical form. Moreover, by utilizing a Jaccard similarity model, regardless of the coding and labeling format, our approach successfully tolerated the lack of sufficient labeled documents and handled the diverse styles of various meeting notes. The new and continuous types of issues are trained and tested in four differently formatted meeting minutes from three different scales and types of projects.

According to the results, when the training number is increased from 2 to 3, the model's accuracy is improved on average. The accuracy of the model is ranging from 81.86% to 94.18%. F1 values are ranging from 82.49% to 96.24%. One limitation in our study is that the meeting minutes' issues are accepted to resemble the actual project progress and updated accordingly. Our model and test results proved the efficiency of the Jaccard similarity model in categorizing and identifying the issues in text-based files. This approach could be utilized in many aspects of dynamic documents in the AEC, such as RFIs and submittals. This study provides the groundwork to automate the determination and estimation of issue complexity, bottlenecks, and expertise assignments analysis for issue resolutions. Future research will aim to test the model with different files and categories.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation through Grant No. 1928278. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the researchers and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Abd Jamil, A. H., and Fathi, M. S. (2020). "Enhancing BIM-Based Information Interoperability: Dispute Resolution from Legal and Contractual Perspectives". *Journal of Construction Engineering and Management*, 146(7), 05020007.

Bag, S., Kumar, S. K., and Tiwari, M. K. (2019). "An efficient recommendation generation using relevant Jaccard similarity". *Information Sciences*, 483, 53-64.

Caldas, C. H., Soibelman, L., and Han, J. (2002). "Automated classification of construction project documents". *Journal of Computing in Civil Engineering*, 16(4), 234-243.

Diana, N. E., and Ulfa, I. H. (2019, March). "Measuring Performance of N-Gram and Jaccard-Similarity Metrics in Document Plagiarism Application". *In Journal of Physics: Conference Series* (Vol. 1196, No. 1, p. 012069). IOP Publishing.

Gorse, C. A., and Emmitt, S. (2007). "Communication behaviour during management and design team meetings: a comparison of group interaction", *Construction management and economics*, 25(11), 1197-1213.

Gorse, C. A., and Emmitt, S. (2009). "Informal interaction in construction progress meetings". *Construction Management and Economics*, 27(10), 983-993.

Hamzeh, F. R., Zankoul, E., and Rouhana, C. (2015). "How can 'tasks made ready'during lookahead planning impact reliable workflow and project duration?". *Construction Management and Economics*, 33(4), 243-258.

Huang, A. (2008, April). "Similarity measures for text document clustering". *In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand (Vol. 4, pp. 9-56).

Jaccard, P. (1902). “Lois de distribution florale dans la zone alpine” *Bull Soc Vaudoise Sci Nat*, 38, 69-130.

Javanmardi, A., Abbasian-Hosseini, S. A., Liu, M., and Hsiang, S. M. (2020). "Improving Effectiveness of Constraints Removal in Construction Planning Meetings: Information-Theoretic Approach". *Journal of Construction Engineering and Management*, 146(4), 04020015.

Kisi, K. P., Kayastha, R., and Shrestha, P. P. (2020). "Back Charges in Construction Contract: Case Study of Airport Project". *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 12(3), 05020010.

Kumar, B., Cheng, J. C., and McGibbney, L. (2010, June). “Cloud computing and its implications for construction IT”. In *Computing in Civil and Building Engineering*, Proceedings of the International Conference (Vol. 30, p. 315).

Mincks, W. R., and Johnston, H. (2010). *Construction jobsite management*, Cengage Learning.

Moon, S., Lee, G., Chi, S., & Oh, H. (2021). “Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing”. *Journal of Construction Engineering and Management*, 147(1), 04020147.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013, March). “Using of Jaccard coefficient for keywords similarity”. *In Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 6, pp. 380-384).

van Niekerk, J. (2020). "A Study of the Data Mining of Meeting Minutes of Construction Projects", *Master's Thesis*, Faculty of Engineering at Stellenbosch University, NZ.

Wang, J., Shou, W., Wang, X., and Wu, P. (2016). "Developing and evaluating a framework of total constraint management for improving workflow in liquefied natural gas construction". *Construction Management and Economics*, 34(12), 859-874.

Wu, G., Liu, C., Zhao, X., and Zuo, J. (2017). "Investigating the relationship between communication-conflict interaction and project success among construction project teams." *International Journal of Project Management*, 35(8), 1466-1482.