Automation in Construction 64 (2016) 78–88

Contents lists available at ScienceDirect

Automation in Construction

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

Evaluating the strength of text classification categories for supporting construction field inspection

Nai-Wen Chi [a], Ken-Yu Lin [b,*], Nora El-Gohary [c], Shang-Hsien Hsieh [a]

[a] Dept. of Civil Engineering, National Taiwan Univ., No. 1, Roosevelt Rd., Sec. 4, Taipei City 10617, Taiwan
[b] Dept. of Construction Management, Univ. of Washington, 130L Architecture Hall, Box 351610, Seattle, WA 98195-1610, USA
[c] Dept. of Civil Engineering, Univ. of Illinois at Urbana-Champaign, 3129c Newmark Civil Engineering Laboratory, 205 N. Mathews Ave., Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 6 January 2016
Accepted 12 January 2016
Available online 28 January 2016

Keywords:
Construction safety
Field inspection
Text classification

ABSTRACT

Field inspection is a common approach to the prevention of on-site accidents in the construction industry, which aims to identify and correct violations before they result in accidents. While conducting a field inspection, quite often safety professionals need to consult applicable construction safety standards. By doing so, they can make informed judgments on the violations and reference applicable standards. Text classification (TC) can be used to classify safety standards based on the types and causes of violations. Safety professionals can therefore use violation types and causes as indices to quickly locate applicable standards. Defining TC categories (or labels) is the first important step in performing TC, because satisfactory results cannot be achieved without appropriate TC categories. Researchers often determine applicable TC categories based on the important topics within a knowledge domain. However, not all TC categories can yield satisfactory TC results because some of them are not associated with strong and specific keywords that can be identified by text classifiers. This paper proposes a methodology with two strength measures for evaluating the appropriateness of candidate TC categories. The measures were tested on two alternative sets of candidate categories that were drafted for supporting construction field inspections. The results showed that the measures could accurately predict the relative TC performance and the satisfaction levels (satisfactory or unsatisfactory) of TC categories. Beyond the construction domain, this research provides a generalized procedure for evaluating the strength of candidate TC categories.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Occupational safety of workers exposed to potentially severe hazards in the construction industry is an important concern [1]. Such occupational hazards are caused by various factors such as lack of skilled laborers [2]. To eliminate or reduce the number of fatalities, various safety approaches, including field inspection, are commonly applied throughout the different project phases [3–6]. Field inspection aims to identify and correct unsafe behaviors and environments in a timely manner. During field inspection, safety professionals audit construction sites to detect violations and take corrective actions. The traditional field inspection process has three main stages: recording, correcting, and reporting [7]. The quality of traditional safety reports generated through field inspections has been recognized as being inadequate [8] mainly because of the inefficiency and ineffectiveness of the process [4]. Several research efforts have attempted to improve the traditional field inspection workflow and to develop innovative tools that can be used on portable computing devices for digitalizing and automating the generation of inspection reports [4,7,9].

In the summer of 2013, the authors held a workshop to investigate how safety professionals use (or would like to use) information communication technologies for field inspections. Twelve construction safety professionals participated in the workshop. The results of the workshop showed that safety professionals wish to access applicable safety standards on their portable computing devices while conducting field inspections, especially when they encounter issues that may cause fatalities such as a fall. For safety inspectors to assess and correct violations, they need to retrieve applicable safety standards for two main reasons. First, they need to refer to the official safety standards to establish an observed issue as a violation, convince field workers that a violation has occurred, and/or make an unbiased judgment if there is a disagreement about a violation. Second, they need to recall and verify the details of the standards because many standards are associated with specific conditions and thus their details must be clarified. For example, the following U.S. Occupational Safety and Health Administration (OSHA) standard contains specific geometrical requirements:

* Corresponding author.
E-mail addresses: naiwenchi@caece.net (N.-W. Chi), kenyulin@uw.edu (K.-Y. Lin), gohary@illinois.edu (N. El-Gohary), shhsieh@ntu.edu.tw (S.-H. Hsieh).

http://dx.doi.org/10.1016/j.autcon.2016.01.001
0926-5805/© 2016 Elsevier B.V. All rights reserved.

"When used to control access to areas where overhand bricklaying and related work are taking place: The controlled access zone shall be defined by a control line erected not less than 10 ft (3.1 m) nor more than 15 ft (4.5 m) from the working edge." [10].

A number of research efforts have addressed automated safety compliance checking in the construction safety domain. For example, Melzner et al. [11] proposed a rule-based checking algorithm for supporting fall protection. Zhang et al. [12] proposed an ontology-based approach for supporting automated safety planning for job hazard analysis. Text classification (TC) offers another potential solution for supporting construction safety by retrieving applicable safety standards. TC is the task of classifying text units (such as documents, paragraphs, or sentences) according to one or more categories [13–16]. In a field inspection scenario, areas of compliance on the inspection checklist can be regarded as pre-defined categories. When an issue is observed during field inspection, safety professionals flag corresponding compliance categories on the checklist. Ideally, TC can be used during this stage to classify safety standards and identify which issues belong to the flagged categories based on their textual features (e.g., terminologies).

The notion of using TC to identify safety standards is novel, and the usefulness of TC, especially when dealing with an application area such as occupational safety, where compliance is extremely important, depends on its performance. While many existing research efforts use technical approaches to improve the performance of their TC applications, their performance is limited by the nature of the TC categories. Satisfactory TC results cannot be achieved without a set of well-defined and appropriate TC categories because the categories reflect the information needs of the users, similar to how search engine users cannot find the correct documents if they do not effectively translate their information needs into appropriate queries. A number of research studies in the Computer Science domain have investigated the impact of category definition in cases where there are a large number of categories that cause negative impacts on TC performance [17,18]. However, most of the existing TC research efforts implicitly assume that the quality of the categories is sufficient and does not affect the performance [19]. In the field inspection case, for example, it might seem reasonable for safety professionals to use the areas of compliance from the inspection checklists as the TC categories, but those categories might not be sufficiently distinctive for text classifiers to yield satisfactory TC performance. This is because the standards that belong to these categories might not contain strong text features (e.g., keywords) that can be identified and discriminated by text classifiers. To this end, it is clear that defining appropriate TC categories that are both practical and technically distinctive for obtaining satisfactory TC performance is important, especially in case of domain-specific applications that require highly specialized categorizations, such as field inspections. In the remainder of this paper, the term "strength" is used to represent a category's distinctiveness and "weak categories" are those with less distinctiveness strength.

In this paper, a methodology for evaluating the strength of potential TC categories (defined as "TC category strength" in the later sections) as a way to filter out weak categories is proposed. To the best of the authors' knowledge, existing category assessment approaches only focus on eliminating the negative effects caused by a large number of categories and there is still a lack of understanding on how to evaluate the distinctiveness of categories. This methodology aims to support the process of category selection during TC. To provide the framework for discussing the proposed methodology, Section 2 reviews machine learning (ML)-based TC techniques, how such techniques have been applied in the Architecture/Engineering/Construction (AEC) domain, and how category evaluation, while being important, has not been studied in the AEC domain. Section 3 describes how the proposed methodology for TC category strength assessment fits into the overall ML-based TC process, and Section 4 introduces the proposed methodology in more detail. Section 5 describes the testing of the methodology in a construction field inspection application. The test results are discussed in Section 6, and concluding remarks are provided in Section 7.

2. Machine learning-based text classification

ML techniques are often applied in TC applications to reduce human effort and can be divided into two primary types: supervised and unsupervised [20]. The main difference between the two types is that unsupervised ML-based TC does not require a training process for learning how to classify text into proper categories, whereas supervised ML-based TC needs a gold standard for training the classifier. A gold standard is a set of pre-labeled text units that can be applied to perform the training process. A text unit is labeled positive for a specific category if human experts or text classifiers determine that it is related to that category; otherwise, it is labeled negative for that category. For conducting the training, the gold standard is divided into two sets: (1) a training data set, which is used for training the classifier and (2) a testing data set, which is used for evaluating the performance of the classifier. Although supervised ML TC requires human effort to create the training and testing data sets, it allows researchers to define their own categories, which is necessary in applications that require pre-defined categories to represent specific information needs. Different algorithms have been used for supervised ML-based TC, such as Naïve Bayes [21,22], maximum entropy [23], k-nearest neighbors (k-NN) [24], and support vector machines (SVM) [25]. Brank et al. [26] verified that an SVM algorithm has the best TC performance compared with other classifiers. An SVM is also supported by many existing toolkits such as libsvm [27] and has been therefore selected as the theoretical TC foundation in this research.

3. Text classification process

TC techniques are widely applied in various research fields. There are also several TC-related research efforts in the AEC domain. For example, Caldas and Soibelman [28] integrated a mechanism for embedding hierarchical document classification into a construction management information system. Tsoumakas and Katakis [15] used SVM for classifying different types of construction documents and developed an unstructured document classification model for the AEC domain. Amor and Xu [29] adapted latent semantic indexing for matching AEC Web pages to codes automatically in a TC system. Mahfouz [30] developed different SVM automated classification models for unstructured construction documents such as correspondences, meeting minutes, claims, and differing site condition cases. Salama and El-Gohary [31] presented an ML-based TC algorithm to support automated compliance checking. Chi et al. [32] applied TC on the classification of construction fatality case reports to support job hazard analysis. Peng and El-Gohary developed a hierarchical ML-based TC algorithm [33] and proposed an ontology-based TC approach [34] for supporting automated environmental regulatory compliance checking.

To describe how the proposed TC category strength assessment methodology (explained in Section 4) can be conducted within a TC process, the four primary steps of ML-based TC, that are commonly undertaken, are first introduced: text preprocessing, feature selection, feature weighting, and performance evaluation.

3.1. Text preprocessing

In ML-based TC, each unique term in the document collection represents an individual feature that informs the classifier. However, one term may have different morphologies (e.g., different verb tenses or single/plural forms of a noun) and without appropriate preprocessing, these terms might be treated as different features. In addition, there are also stop words (e.g., in, on, and at) that are not meaningful for the purpose of classification. Stop words are low-content words that

Download English Version:

https://daneshyari.com/en/article/246227

Download Persian Version:

https://daneshyari.com/article/246227

Daneshyari.com