Expert Systems with Applications 41 (2014) 1153–1167

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Knowledge discovery in inspection reports of marine structures

Seung-kyung Lee [a], Bongseok Kim [a], Minhoe Huh [a], Jooseoung Park [a], Seokho Kang [a], Sungzoon Cho [a,*], Dongha Lee [b], Daehyung Lee [b]

[a] Department of Industrial Engineering, Seoul National University, 1 Gwanakro, Gwanakgu, 151-744 Seoul, Republic of Korea
[b] Central R&D Institute, Daewoo Shipbuilding & Marine Engineering Co., Ltd., Republic of Korea

ARTICLE INFO

Keywords:
Knowledge Discovery in Textual Databases
Text mining
Shipbuilding and marine engineering industry
Inspection process

ABSTRACT

Inspection reports, commonly called "punches" in the marine structuring domain, are written documents about defects or supplementations on marine structures. Analyzing the inspection reports improves the construction process for the structure and prevents additional "punches." This consequently reduces construction delays and supplementary costs. The free-form texts of the reports, however, hinder management from understanding the nature of defects. Therefore, we applied Knowledge Discovery in the Textual Databases (KDT) process to answer the questions, "what kinds of defects are reported while inspecting a marine structure, and which of them are closely related?" In particular, we propose a concept extraction and linkage approach as an "add-on" module for the Self-Organizing Map (SOM), a clustering algorithm for document organization. A purely data-driven graph is derived for defect-types, which gives it in an easy-to-understand form for domain experts and reduces the gap between data analysis and its practical use. Interpretation with domain experts showed that our KDT process is useful in understanding the nature of defects in the domain and systematically responding to some other related defects.

1. Introduction

Knowledge Discovery in Databases (KDD) is a non-trivial process of identifying novel and useful patterns in numerical data by utilizing data mining as the essential step (Choudhary, Harding, & Tiwari, 2009a; Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996; Menon, Loh, & Keerthi, 2005). This process has been successfully applied in a variety of industrial applications such as semiconductor manufacturing (Kang, Kim, joo Lee, Doh, & Cho, 2001), shipbuilding and marine engineering (Lee et al., 2013), etc. From Gartner[1] research reports, however, approximately 80% of all enterprise data is a form of unstructured text (Ur-Rahman & Harding, 2012; Yu, Wang, & Lai, 2005), e.g. post-project reviews (Choudhary, Oluikpe, Harding, & Carrillo, 2009b), voice of customer reports (Godbole & Roy, 2008), and problem reports (Malin, Millward, Gomez, & Throop, 2010). This leads to a strong demand for utilizing a massive amount of texts as knowledge for enterprises (Fan, Wallace, Rich, & Zhang, 2006; Gupta & Lehal, 2009; Yu et al., 2005).

In the marine structuring domain, there is also a need to analyze the unstructured texts. In the construction of marine structures such as oil rigs, floating production systems (FPS), and offshore power plants, an inspection is conducted to get more reliable structures. Approximately 100 inspectors manually input about 300 reports each per day. These reports are written about defects or supplementations in the marine structure, and are collected in a database. The analysis of inspection reports is crucial in improving the current construction process. This reduces both construction delays and supplementary costs from additional tasks.

The free-form text descriptions of the inspection reports, however, hinder management from understanding the "what" and "where" of defects. Answering these questions regarding defects in a marine structure requires engineers to read and summarize all the inspection reports. Thus, only manual analysis has been done on a small sample of reports with huge workloads (Godbole & Roy, 2008; Nasukawa & Nagano, 2001). For efficient search and summarization, these kind of reports are typically stored with a human-defined "category" field, i.e. defect-type in our domain (Malin et al., 2010). In spite of partial success in using this approach, the category setting does not help much as it has some practical limitations (Menon et al., 2005). In our domain, a defect-type for each report has been defined as one of about 150 possible types by inspectors. There are some managerial limitations: the difference between these defect-types is ambiguous, and it is

* Corresponding author. Tel.: +82 2 880 7025.
E-mail addresses: sklee83@snu.ac.kr (S.-k. Lee), kiolol2@gmail.com (B. Kim), dninb.kr@gmail.com (M. Huh), uni208@snu.ac.kr (J. Park), prokids@snu.ac.kr (S. Kang), zoon@snu.ac.kr (S. Cho), dongha@dsme.co.kr (D. Lee), dhlee8@dsme.co.kr (D. Lee).

[1] http://www.gartner.com.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.109

very common for reports to have the same and meaningless defect-type, e.g. "0001" code, as pointed out in other database study (Menon et al., 2005). Thus, some reports of similar contents were often assigned to different defect-types.

In this paper, we answer the questions: "what kinds of defects are reported during inspection, and which defects are closely related?" To achieve this purpose, we use the Knowledge Discovery in Textual Databases (KDT) process (Fan et al., 2006; Feldman & Dagan, 1995; Gupta & Lehal, 2009;Hearst, 1999; Ur-Rahman & Harding, 2012; Yu et al., 2005). In particular, we propose a concept extraction and linkage approach based on the Self-Organizing Map (SOM) (Kohonen, 1995) to derive a purely data-driven concept graph that outlines defects or supplementations in the marine structure. This approach consists of three steps. First, SOM is used to group (or summarize) the reports into representative ones, called codebooks to form a two-dimensional grid on a visual map. In this map, we demonstrated that local regions that cover some similar codebooks tend to form topical clusters of reports and may correspond to domain-specific concepts, i.e. defect-types, as pointed out by Lagus, Kaski, and Kohonen (2004). Second, hierarchical clustering is applied to the codebooks of the report map in order to define the local regions. Then, in a bottom-up manner for a codebook hierarchy, it is determined whether the local regions correspond to semantically coherent concepts, using the keyword extraction method (Azcarraga, Yap, Tan, & Chua, 2004) for WEBSOM, the information retrieval and organization application of SOM (Lagus, Honkela, Kaski, & Kohonen, 1999, 2004). The method does not force keywords if there are no keywords for a region. Otherwise, if some meaningful keywords for a region are extracted, the codebooks in the region are thought to be coherent and form a concept characterized in terms of those same keywords. Third, concepts are linked vertically in a top-down manner from a dummy node named "punch," with horizontal linkages among child concepts that share the same keywords. Consequently, a concept hierarchy or graph, which is different from the codebook hierarchy, is obtained and visualized some defect-types as concepts with their structural relations as linkages.

We expect that the snapshot of the easy-to-understand form provides information on the supplementation process, i.e. "which defects are found at which component or process?". This leads to an evidence-based decision making in the domain. For example, a decision regarding the replacement of a component in a marine structure may trigger defects in related components as side effects. We expect the snapshot for inspection reports in the marine structure helps domain experts to consider such side effects systematically before making decisions. Moreover, the accumulation of such knowledge for the supplementation process prevents similar defects from occurring, thus contributing to cost reduction.

In Section 2, we explain the KDT process and summarize industrial cases that apply the process. In Section 3, we outline our KDT process including the proposed concept extraction and linkage approach as a core step. In Section 4, we demonstrate the obtained results by applying our KDT process into the real inspection reports for a marine structure. In Section 5, we provide conclusions and future work.

2. Knowledge Discovery in Textual Databases and its applications

By linking a human's linguistic capability with a computer's computational one (Fan et al., 2006), various text processing techniques has been utilized for searching, organizing documents and currently discovering knowledge in documents, as shown in Fig. 1. As pointed out by Nasukawa and Nagano (2001), text mining in a broad sense is roughly classified into three methodologies according to their functional purposes. Fig. 1 summarizes the text mining methodologies, their functional purposes, and outputs. First, information retrieval is the methodology used to rank documents optimally for given keywords as a query so that relevant documents would be ranked above non-relevant ones (Zhai, 2008). This helps people narrow down the number of documents related to keywords that they are interested in. Second, document clustering, i.e. segmentation, and classification, i.e. categorization are the methodologies used to boil down documents into some smaller groups with similar content or pre-defined labels, i.e. topic (Fan et al., 2006; Hearst, 1999; Iiritano & Ruffolo, 2001; Nasukawa & Nagano, 2001; Sedding & Kazakov, 2004). These methodologies organize documents and provide an overview of them, which is useful in web browsing, e.g. Vivismo, a meta search engine based on commercial clustering interface (Carpineto, Osiński, Romano, & Weiss, 2009), or enterprise information management (Fan et al., 2006; Sedding & Kazakov, 2004).

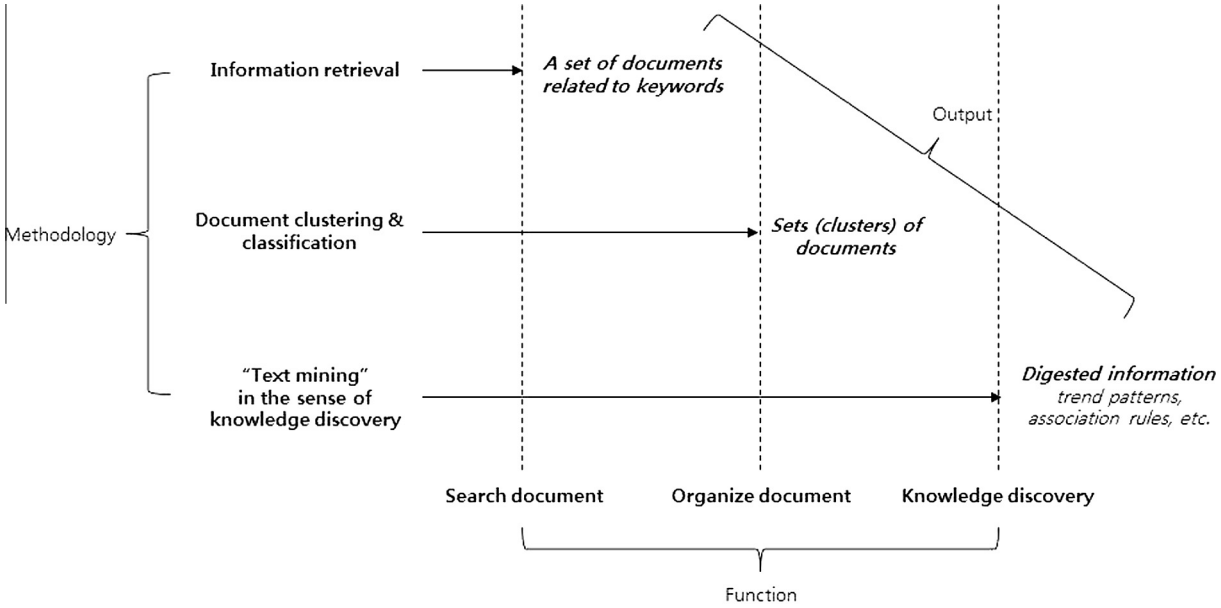

Fig. 1. The classification of some text mining methodologies, their functions and outputs in the Nasukawa and Nagano (2001) study.