Advanced Engineering Informatics 27 (2013) 615–624

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Data mining and knowledge discovery in materials science and engineering: A polymer nanocomposites case study ☆

O. AbuOmar [a,b], S. Nouranian [b], R. King [a,b,*], J.L. Bouvard [c], H. Toghiani [d], T.E. Lacy [e], C.U. Pittman Jr. [f]

[a] Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[b] Center for Advanced Vehicular Systems (CAVS), Mississippi State, MS 39762, USA
[c] Center for Material Forming (CEMEF), Mines ParisTech, 06904 Sophia Antipolis Cedex, France
[d] The Dave C. Swalm School of Chemical Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[e] Department of Aerospace Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[f] Department of Chemistry, Mississippi State University, Mississippi State, MS 39762, USA

ARTICLE INFO

Article history:
Received 10 December 2012
Received in revised form 23 May 2013
Accepted 13 August 2013
Available online 5 September 2013

Keywords:
Materials informatics
Data mining
Vapor-grown carbon nanofiber
Vinyl ester
Unsupervised learning

ABSTRACT

In this study, data mining and knowledge discovery techniques were employed to validate their efficacy in acquiring information about the viscoelastic properties of vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposites solely from data derived from a designed experimental study. Formulation and processing factors (VGCNF type, use of a dispersing agent, mixing method, and VGCNF weight fraction) and testing temperature were utilized as inputs and the storage modulus, loss modulus, and tan delta were selected as outputs. The data mining and knowledge discovery algorithms and techniques included self-organizing maps (SOMs) and clustering techniques. SOMs demonstrated that temperature had the most significant effect on the output responses followed by VGCNF weight fraction. SOMs also showed how to prepare different VGCNF/VE nanocomposites with the same storage and loss modulus responses. A clustering technique, i.e., fuzzy C-means algorithm, was also applied to discover certain patterns in nanocomposite behavior after using principal component analysis as a dimensionality reduction technique. Particularly, these techniques were able to separate the nanocomposite specimens into different clusters based on temperature and tan delta features as well as to place the neat VE specimens (i.e., specimens containing no VGCNFs) in separate clusters. Most importantly, the results from data mining are consistent with previous response surface characterizations of this nanocomposite system. This work highlights the significance and utility of data mining and knowledge discovery techniques in the context of materials informatics.

1. Introduction

Data mining is a field at the intersection of computer science and modern mathematical analysis [1–4]. It is used for discovering patterns in large datasets using predictive modeling techniques, where hidden data trends can be found [2]. The overall goal of the data mining process is to extract information from a large complex dataset and transform it into an understandable structure, thus enabling knowledge discovery. This transformation of massive amounts of structured and unstructured data into information and then into new knowledge using a myriad of data mining techniques is one of the great challenges facing the engineering community. The use of data mining techniques in the context of materials science and engineering is considered an important extension of materials informatics [5–8]. This interdisciplinary study integrates computer science, information science, and other domain areas to provide new understanding and to facilitate knowledge discovery. Materials informatics is a tool for material scientists to interpret vast amounts of experimental data through the use of machine learning approaches integrated with new visualization schemes, more human-like interactions with the data, and guidance by domain experts. It can also accelerate the research process and guide the development of new materials with select engineering properties. Material informatics is being fueled by the unprecedented growth in information technology and is driving the interest in the application of knowledge representation/discovery, data mining, machine learning, information retrieval, and semantic technology in the engineering disciplines.

There are several recent published applications utilizing material informatics and data mining. Hu et al. [9] used material

☆ Handled by C.-H. Chen.

* Corresponding author at: Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762, USA. Tel.: +1 662 325 2189.

E-mail address: rking@cavs.msstate.edu (R. King).

1474-0346/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2013.08.002

informatics to resolve the problem of materials science image data sharing. They presented an ontology-based approach that can be used to develop annotation for non-structured materials science data with the aid of semantic web technologies. Yassar et al. [10] developed a novel computational model based on dislocation structures to predict the flow stress properties of 6022 aluminum alloy using data mining techniques. An artificial neural network (ANN) model was used to back-calculate the *in situ* non-linear material parameters and flow stress for different dislocation microstructures [10]. Sabin et al. [11] evaluated an alternative statistical Gaussian process model, which infers a probability distribution over all of the training data and then interpolates to make predictions of microstructure evolution arising from static recrystallization in a non-uniform strain field. Strain, temperature, and annealing time were the inputs of the model and the mean logarithm of grain size was its output. Javadi and Rezania [12] provided a unified framework for modeling of complex materials, using evolutionary polynomial regression-based constitutive model (EPRCM), integrated in finite element (FE) analysis, so an intelligent finite element method (EPR–FEM) was developed based on the integration of the EPR-based constitutive relationships into the FE framework. In the developed methodology, the EPRCM was used as an alternative to the conventional constitutive models for the material. The results of the analyses were compared to those obtained from conventional FE analyses. The results indicated that EPRCMs are able to capture the material constitutive behavior with a high accuracy and can be successfully implemented in a FE model.

Brilakis et al. [13] presented an automated and content-based construction site image retrieval method based on the recognition of material clusters in each image. Under this method, the pixels of each image were grouped into meaningful clusters and were subsequently matched with a variety of pre-classified material samples. Hence, the existence of construction materials in each image was detected and later used for image retrieval purposes. This method has allowed engineers to meaningfully search for construction images based on their content. Sharif Ullah and Harib [14] presented an intelligent method to deal with materials selection problems, wherein the design configurations, working conditions, as well as the design-relevant information are not precisely known. The inputs for this method were: (1) a linguistic description of the material selection problems (expressing the required levels of material properties/attributes and their importance), and (2) the material property charts relevant to the linguistic description of the problem. The method was applied to select optimal materials for robotic links and it was found that composite materials were better than metallic materials for robotic links.

A class of advanced materials, nano-enhanced polymer composites [15], have recently emerged among the more traditional structural metals. Polymer nanocomposites have been used in a variety of light-weight high-performance automotive composite structural parts where improved specific properties and energy absorption characteristics are required [16]. Though polymer nanocomposites have recently been widely investigated [17,18], they have never been studied in the context of material informatics. Therefore, the purpose of this study is to apply data mining and knowledge discovery techniques, as a proof of concept, to a thermosetting vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite system. Nouranian et al. [19–21] and Torres et al. [22] developed a relatively large dataset for this material system suitable for data mining. This study seeks to use this dataset to demonstrate the usefulness of knowledge discovery and data mining techniques for nanocomposite material property characterization.

VGCNFs are commercially viable nanoreinforcements with superb mechanical properties [23]. VEs are thermosetting resins suitable for automotive structural composites due to their superior properties in comparison with unsaturated polyesters [20–22, 24,25]. Incorporating VGCNFs into VEs may provide improved mechanical properties relative to the neat matrix. These mechanical properties, however, are dependent on the degree of VGCNF nanodispersion in the matrix achieved during the mixing stage of the process. Examples of good and poor nanofiber dispersion in the matrix are given in Fig. 1, where two transmission electron micrographs of VGCNF/VE specimens are compared. Large nested groups of nanofibers (agglomerates) are a sign of poor VGCNF dispersion in the matrix, often resulting in inferior mechanical properties.

Data mining and knowledge discovery techniques can help discover and map patterns in the physical, mechanical, and system properties of VGCNF/VE nanocomposites, thereby aiding the nanocomposite design, fabrication, and characterization without the need to conduct expensive and time-consuming experiments.

In this study, several unsupervised knowledge discovery techniques were used to explore a large VGCNF/VE dataset [19]. The dataset consisted of 240 data points each corresponding to the combinations of five input design factors and three output responses, i.e., a total of eight "dimensions." The dimensions in data mining are the combination of both inputs and outputs of the developed model. The dimensions of the VGCNF/VE dataset are VGCNF type, use or absence of dispersing agent, mixing method, VGCNF weight fraction, temperature, storage modulus, loss modulus, and tan delta (ratio of loss to storage modulus), where the last three dimensions correspond to measured macroscale material properties. Kohonen maps [26,27], or self-organizing maps (SOMs), were applied to the dataset in order to conduct a sensitivity analysis of all of these factors and responses. In addition, principal component analysis (PCA) [28] was used to provide a two-dimensional (2-D) representation of nanocomposite data. This facilitated application of the fuzzy C-means (FCM) clustering algorithm [29,30] to characterize the physical/mechanical properties of VGCNF/VE nanocomposites.

2. Materials and methods

A brief summary of the statistical experimental design and testing procedures to generate the VGFCNF/VE dataset is given here. A more detailed discussion can be found in [19–21].

2.1. Statistical experimental design

The effect of five input design factors on the viscoelastic properties (storage and loss modulus) of VGCNF/VE nanocomposites were investigated using a general mixed-level full factorial experimental design [31]. These carefully selected factors, based on the state-of-the-art formulation and processing procedures, included: (1) VGCNF type (designated as A), (2) use of a dispersing agent (B), (3) mixing method (C), (4) VGCNF weight fraction in parts per hundred parts of resin (phr) (D), and (5) the temperature (E) used in dynamic mechanical analysis (DMA) testing. Experimental design factors and their associated levels are given in Table 1.

A total of $2 \times 2 \times 3 \times 5 \times 4 = 240$ "treatment combinations" (different combinations of the factor levels in Table 1) were randomized to eliminate bias in preparing the specimens. Each treatment combination resulted in three specimens prepared from the same material batch [20,21]. Each specimen was tested using a dynamic mechanical analyzer (single cantilever/flexure mode) to measure average storage modulus, loss modulus, and tan delta for each treatment combination. Storage and loss moduli are dynamic mechanical properties and indicative of the polymer nanocomposite's stiffness and energy dissipation capability, respectively.