Computers in Industry 61 (2010) 66–74

Contents lists available at ScienceDirect

Computers in Industry

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

Labeling design documents based on operators' consensus—A case study of robotic design

Sun Jie *, Lu Wen Feng, Loh Han Tong

Department of Mechanical Engineering, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 14 April 2008
Received in revised form 1 June 2009
Accepted 5 July 2009
Available online 3 August 2009

Keywords:
Product design
Knowledge retrieval
Text classification
Hypothesis test

ABSTRACT

Designers usually begin with a database to look for historical design solution, available experience and techniques through design documents, when initiating a new design. This database is a collection of labeled design documents under a few of predefined categories. However, little work has been done on labeling a relatively small number of design documents for information organization, so that most of design documents in this database can be automatically categorized.

This paper initiates a study on this topic and proposes a methodology in four steps: design document collection, documents labeling, finalization of documents labeling and categorization of design database. Our discussion in this paper focuses on the first three steps. The key of this method is to collect relatively small number of design documents for manual labeling operation, and unify the effective labeling results as the final labels in terms of labeling agreement analysis and text classification experiment. Then these labeled documents are utilized as training samples to construct classifiers, which can automatically give appropriate labels to each design document.

With this method, design documents are labeled in terms of the consensus of operators' understanding, and design information can be organized in a comprehensive and universally accessible way. A case study of labeling robotic design documents is used to demonstrate the proposed methodology. Experimental results show that this method can significantly benefit efficient design information search.

1. Introduction

When initiating a new design, designers usually begin with a database to look for historical design solution, past experience and available techniques through design documents. This database is a collection of labeled design documents under predefined categories, and covers the concerned aspects from product structure, functionality, manufacturing processing and materials, in which all have an effect on the final design quality. To achieve a desired product design, Lakshminarayanan et al. built a database with historical operating conditions and product quality, and presented a methodology to analyze it [1]. Ferguson suggested product designers to access a range of corporate databases [2], in particular, customer complaints, product material features, and R&D testing, with the use of data mining techniques. This investigation enabled information from later life cycle stages to be used by earlier ones, and made this information understandable and useable to other product functions. In order to effectively use past design knowledge, Wu built up a design knowledge database through an elaborate process of data collection, mining, integration, storing, management and maintenance [3]. It included design standards and requirements, product samples, experimental data, computational models, designers' experiences, and failure information. Product designers need material information relevant with product issues to adjust design solutions. To meet their needs, Kesteren investigated the current utilization of material properties database, and proposed some strategies for database developer to improve information presentation [4]. All these databases are becoming electronically accessible and growing at an explosive rate.

The applications of text mining technique to support product design increase dramatically in the last five years. It provides an effective tool to process and analyze huge amount of design documents and to eventually gain valuable insight into product design solutions. In most of these applications, a labeled database is essential. However, manually labeling these databases is of high cost [5]. The alternative is to manually label the relatively small number of design documents, and use them as training samples to build a set of text classifiers with artificial intelligence techniques. These classifiers can categorize the whole design document database automatically and yield appropriate labels to each document.

Several research papers have been focused on generation of labeled datasets. Adami et al. proposed a semi-automatic process

* Corresponding author.
E-mail address: mpesunj@nus.edu.sg (S. Jie).

0166-3615/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.compind.2009.07.002

to obtain a hierarchical document classifier [6]. For unlabeled documents, they used bootstrapping process to make a hypothesis of categorization, and organize them according to the given taxonomy with revised Self-Organizing Map. Godbole et al. presented a document annotation method based on active learning algorithm, which collected user opinion on feature representations as well as whole-document labels to minimize the user's efforts [7]. With the existing hierarchical directory structures, Davidov et al. described a system to automatically acquire labeled datasets for text categorization from websites [8]. The generation process did not consider the content of the datasets, and its difficulty was controlled by parameters. Lewis reported to label Reuters-produced stories in three stages [9]: autocoding, manual editing, and manual correction. Stories first passed through a rule-based text categorization system, and the output was automatically checked for compliance with the minimum code policy. If so, the story was sent to a holding queue; otherwise, to a human editor for labeling. Finally, the stories were reviewed to correct labeling mistakes, and loaded into the database. However, this method is not suitable to build a new database. Zhang et al. proposed an automatic method to collect labeling documents as training samples and build hierarchical taxonomies [10]. In this method, the category was initially defined by some keywords, the web search engine was then used to construct a small set of labeled documents, and a topic tracking algorithm with keyword-based content normalization was applied to enlarge the training corpus on the basis of the seed documents. However, this method is sensitive to the selection of key words.

Document categorization was an attractive technique for information organization, and the classification accuracy was believed to be as good as human performance [11]. However, the classification accuracy of text mining methods depends on the number of available training samples, and their quality. Hence, how to prepare a relatively small number of labeled design document database with low cost and high-quality for training purpose is a big challenge. This motivates us to establish a somewhat general methodology to tackle this problem.

In this paper, the proposed methodology is illustrated in Section 2. An example on collecting robotic design database is reported in Section 3. Section 4 describes the evaluation process of operators' labeling performance, so that the consensus can be extracted from effective labeling results. The finalization of documents labeling is introduced in Section 5, and a summary is given in Section 6.

2. Methodology

Many real-world categorization problems can be inherently quite imprecise, and as such open to interpretation by individuals that apply them. Further, human labeling is inevitably a subjective process due to limited or diverse understanding in this filed. There can be considerable variation in inter-indexer agreement for data sets [12,13], therefore, the labeling results may not be totally valid and could not be directly used to form the final document labels.

Clearly, it would be of great benefit if some quantitative measure of labeling agreement could be applied. The ideal approach is to compare with the outcome against some benchmark standards; however, it is not practical in real cases. Hence, an evaluation process is needed to reveal each operator's labeling performance and filter prejudiced understanding before finalizing the document labels.

Fig. 1 shows the flow chart of design document database building process. As indicated, the product design document database is created through four phases: design document collection, documents labeling, finalization of documents labeling, and categorization of product design database.

In a design document database, past design knowledge and useful information from new techniques are collected in terms of categories related to product design. To collect a small set of design documents for manual labeling, a proper sampling scheme is chosen and then a series of labeling policies are determined according to the property of input data source. Before manually labeling operation, a joint discussion with one expert in this domain is essential to promote operators' understanding, since people dedicates to a certain domain may not be fully knowledgeable with all the sub-knowledge branches. In order to ensure the quality of documents labeling, performance evaluation is carried out to identify operators' biased understanding. This document labeling process is repeated until the effective labeling results are achieved. To unify labeling results, the minimum number of agreeable operators is estimated by labeling agreement analysis, and tested by text classification experiment. These two steps are carried out in the phase of finalization of documents labeling. Eventually, the collected documents are labeled in terms of the consensus of operators' understanding. Using these documents as training samples, a set of classifiers can be constructed which are capable to categorize the whole product design database into the predefined categories. Hence, the rest documents in this product design database could be automatically labeled. Our discussion here is mainly focused on manual labeling operation, performance evaluation and subsequent finalization.

2.1. Manual labeling operation

In manual labeling, human operators are asked to identify whether a design document belongs to one or more categories based on their give their personal understanding. Evidently, the manual labeling process may never be perfect—it was difficult to produce perfectly consistent annotations, particularly under

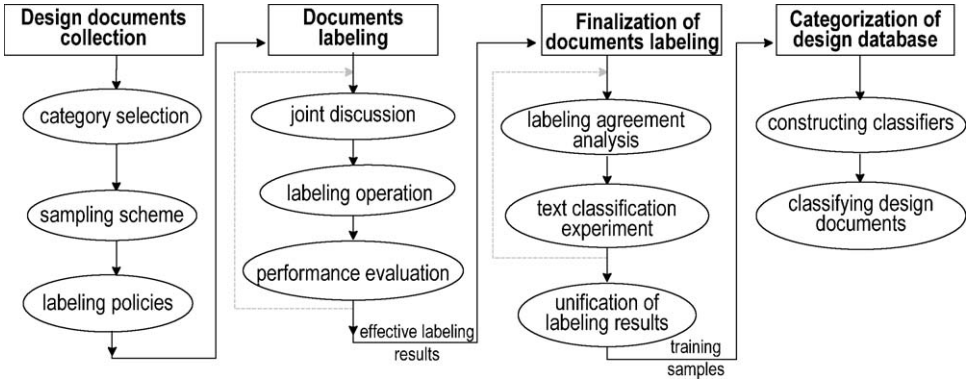

Fig. 1. Design document database buildup scheme in product design.