Domain driven data mining in human resource management: A review of current research

Stefan Strohmeier *, Franca Piazza

Chair of Management Information Systems, Saarland University, 66123 Saarbruecken, Germany

ARTICLE INFO	ABSTRACT
	An increasing number of publications concerning data mining in the subject of human resource management (HRM) indicate the presence of a prospering new research field. The current paper reviews this research on HR data mining to systematically uncover recent advancements and suggest areas for future work. Based on the approach of domain driven data mining, an initial framework with significant domain-specific requirements is elaborated. Relevant research contributions are identified and reviewed against the background of this framework. The review reveals that HRM constitutes a noteworthy new domain of data mining research that is dominated by method- and technology-oriented work. However, specific domain requirements, such as evaluating the domain success or complying with legal standards, are frequently not recognized or considered in current research. Therefore, the systematic consideration of domain-specific requirements is demonstrated here to have significant implications for future research on data mining in HRM.

1. Introduction

Data mining refers to the non-trivial process of identifying novel, potentially useful and valid patterns in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). There is a broad set of data mining application domains and corresponding research domains, including the management domain with well-established sub-domains, such as customer management (e.g. Ngai, Xiu, & Chau, 2009), manufacturing management (e.g. Choudhary, Harding, & Tiwari, 2008) or financial management (e.g. Ngai, Hu, Wong, Chen, & Sun, 2011). Recently, these sub-domains appear to be complemented by human resources management (HRM). In recent years, a quickly growing number of research contributions aim at supporting the practical adoption of HRM data mining. Research contributions refer to various HRM activities and processes, such as selecting employees (Aiolli, de Filippo, & Sperduti, 2009) and predicting employee turnover (Aviad & Roy, 2011) in the function of *staffing*, ascertaining competencies of employees (Zhu, Goncalves, Uren, Motta, & Pacheco, 2005) and career planning (Lockamy & Service, 2011) in the function of *development*, planning HR costs (Juan, 2009) and predicting the acceptance of severance pay (Ramesh, 2001) in the function of *compensation*, and predicting (Thissen-Roe, 2005) and evaluating (Zhao, 2008a,b) employee performance in the function of *performance management*. To support these func-

tions, the entire spectrum of data mining methods – decision trees (Sivaram & Ramar, 2010) cluster analysis (Karahoca, Karahoca, & Kaya, 2008), association analysis (Zhang & Deng, 2011), support vector machines (Li, Xu, & Meng, 2009) and neural nets (Ning, 2010) – is employed, while methodical advancements and innovations are presented (Goonawardene, Subashini, Boralessa, & Premaratne, 2010). In summary, browsing the literature indicates a prospering new field of data mining research that provides ample insights in how to generate advanced information and decision support within the HR domain.

This paper is the first comprehensive review of this new field of research, and it purposes to uncover recent advancements and suggest remaining tasks for future research. Given the growing doubts concerning the ability of conventional data mining research to meet the requirements of practice and the corresponding calls for more relevant data mining research (Adejuwon & Mosavi, 2010; Cao, 2010; Cao & Zhang, 2007; Puuronen & Pechenizkiy, 2010; Weiss, 2009), the concept of *domain driven data mining* (Adejuwon & Mosavi, 2010; Cao, 2010; Cao & Zhang, 2007; Wang & Wang, 2009) is employed as a reference framework for the review.

To provide a systematic review, the method of identifying relevant research contributions is described, and an initial framework of domain driven data mining research in HR is substantiated. Subsequently, this framework is employed to systematically review the discovered research contributions and to derive implications for future research.

* Corresponding author. Tel.: +49 68130264750; fax: +49 681 30264755.
 E-mail address: s.strohmeier@mis.uni-saarland.de (S. Strohmeier).

2. Method and framework

2.1. Method

To comprehensively identify relevant research contributions, a combination of keyword searches and forward and backward searches was employed. To perform the keyword search, a scholarly Internet search engine (scholar.google.com) and several scholarly online databases (ABI/Inform, Business Source Premier, Scopus, and Science Direct) were employed. Given that HRM data mining refers to an intersection of method and domain, respective pairs of search terms such as "data mining" and "HRM" were employed. Beyond synonyms such as "KDD" and "talent management", multiple HRM sub-domains such as "recruiting" and "compensation", method-categories such as "decision trees" and "cluster analysis", and single mining methods such as "C5.0"and "k-means" were used as search terms. Ultimately, there were over 400 pairs of search terms from the possible combinations. Although relevant to HRM, the area of e-learning (educational data mining) was excluded because there are already corresponding research reviews for this topic (Baker & Yacef, 2009; Romero & Ventura, 2010). Moreover, work on data mining in employee fraud (e.g. Luell, 2010) was excluded because fraud is usually classified within compliance rather than as an HR task. The search was further restricted to English publications. This keyword-based search was complemented by a backward search, which systematically searched the references sections of the selected research contributions to identify additional relevant publications, and by a forward search, which searched for additional articles that cited the previously found contributions. To comprehensively map the existing research and to mirror research activities over time, the period from 1991 to 2011 was used.

Employing this procedure initially revealed 172 publications. The abstracts of these research contributions were reviewed independently by two researchers who are familiar with both data mining and HRM. The research contributions were included only if both researchers agreed on their topical relevance (Cohen's $\kappa = 0.96$). Based on this procedure, 111 relevant research contributions were initially rated as relevant. However, because some work was published several times, 11 duplicate research contributions were subtracted, leading to 100 total research contributions. All relevant research contributions are marked with a consecutive number in square brackets and all duplicate research contributions are marked with [D] in the reference section. The selected research contributions were independently reviewed again by two researchers, while the results were stored in a database with structured and unstructured fields. The results of both researchers were compared in the final phase, and in cases of disagreement, the corresponding research contributions were reviewed again to ensure consistency and accuracy. The final data pool was used for this review with automatic and manual information retrieval.

2.2. Framework

To offer a framework with review criteria that evaluate the *relevance* of research, the approach of domain driven data mining (Cao, 2010; Cao & Zhang, 2007; Pechenizkiy, Puuronen, & Tsymbal, 2008; Puuronen & Pechenizkiy, 2010; Wang & Wang, 2009) was employed. Domain driven data mining emerged as a response to the lack of relevance of conventional method-oriented data mining research. With method-oriented research, it has been noticed that "... findings are not actionable and ... lack power in solving real-world problems" (Puuronen & Pechenizkiy, 2010) and that a "... big gap between academic outputs and business expectations ..." (Cao, 2010) has emerged. Simply offering advanced mining methods with exam-ples of mined patterns obviously underestimates the complexities of real-world data mining. Consequently, a readjustment of data mining research is suggested to shift from a *method driven approach* to a *domain driven approach*. The basic idea of domain driven data mining is offering utility and relevance by carefully considering relevant domain-specific requirements and constraints. To realize this objective, a *multi-dimensional research approach* (Cao, 2010; Puuronen & Pechenizkiy, 2010) is suggested to consider additional requirement dimensions beyond the hitherto considered method dimension. The domain driven approach is used as framework to ensure a *relevance-oriented review* of current research. Given that *none* of the research contributions claim to follow the domain driven approach, using a domain driven framework might appear to be unreasonable at first glance. However, even explicitly method-oriented data mining research (e.g. Aviad & Roy, 2011; Yang, Lin, Chen, & Shi, 2009; Zhang, Zhu, & Hua, 2009; Zhu et al., 2005) must consider relevance and practical applicability, otherwise such research is of no use. It is both permissible and necessary to evaluate HRM data mining research with regard to its practical relevance. However, due to the present early stage of HRM data mining research, there is no theoretical, conceptual and/or empirical work that specifies concrete requirements of the HR domain. Therefore, a basic model of relevant requirement dimensions and criteria has to be derived from the general work on domain driven data mining. This paper therefore will offer insights into the requirements and provide guiding examples for a future, relevance-oriented research.

Data mining research should focus on *relevant problems* to provide information and decision support that is relevant and useful for HR practice (Coppock, 2003; Puuronen & Pechenizkiy, 2010). Assuring the functional relevance of research is, however, by no means a trivial task. As a consequence of the absence of domain expertise in method- and technology-oriented research, over-trivialized and even imaginary problems are often addressed in data mining research (Cao, 2010; Weiss, 2009). Therefore, domain driven data mining research must consider a *functional dimension*. As significant criteria of this dimension, the following criteria are reviewed:

- functional domain (whether and which functional HR problem are treated);
- functional relevance (whether and how the relevance of the functional HR problem is justified); and
- functional success (whether and how the success of data mining in solving the functional HR problem is evaluated).

Moreover, as in any other domain, practical data mining in the HR domain relies on suitable *data mining algorithms*. Given the comprehensive research on this topic there is a broad spectrum of available methods (Puuronen & Pechenizkiy, 2010; Wang & Wang, 2009). However, it is necessary to check whether general prefabricated algorithms fit the domain requirements or whether there is a need for domain-specific, i.e., customized or newly developed, algorithms in HR that deal with specific domain problems. Within this *method dimension*, the following are used as review criteria:

- methodical category (whether and which mining methods are employed); and,
- methodical adjustment (whether these data mining methods are general or domain-specific customized/developed).

The success of practical data mining depends, crucially, on the availability of the right number and kind of data (Bole, Jakličb, Papac, & Žabkard, 2011; Wang & Wang, 2009). It is commonly thought that in any domain the amount of available data is ever