Chapter 48
Applications of Pattern Discovery Using Sequential Data Mining

Manish Gupta
University of Illinois at Urbana-Champaign, USA

Jiawei Han
University of Illinois at Urbana-Champaign, USA

ABSTRACT

Sequential pattern mining methods have been found to be applicable in a large number of domains. Sequential data is omnipresent. Sequential pattern mining methods have been used to analyze this data and identify patterns. Such patterns have been used to implement efficient systems that can recommend based on previously observed patterns, help in making predictions, improve usability of systems, detect events, and in general help in making strategic product decisions. In this chapter, we discuss the applications of sequential data mining in a variety of domains like healthcare, education, Web usage mining, text mining, bioinformatics, telecommunications, intrusion detection, et cetera. We conclude with a summary of the work.

HEALTHCARE

Patterns in healthcare domain include the common patterns in paths followed by patients in hospitals, patterns observed in symptoms of a particular disease, patterns in daily activity and health data. Works related to these applications are discussed in this sub-section.

DOI: 10.4018/978-1-4666-2455-9.ch048

Patterns in patient paths: The purpose of the French Diagnosis Related Group's information system is to describe hospital activity by focusing on hospital stays. (Nicolas, Herengt & Albuisson, 2004) propose usage of sequential pattern mining for patient path analysis across multiple healthcare institutions. The objective is to discover, to classify and to visualize frequent patterns among patient path. They view a patient path as a sequence of sets. Each set in the sequence is a hospitalization instance. Each element in a hospitalization can be any symbolic data gathered by the PMSI (medical data source). They used the SLPMiner system (Seno & Karypis, 2002) for mining the patient path database in order to find frequent sequential patterns among the patient path. They tested the model on the 2002 year of PMSI data at the Nancy University Hospital and also propose an interactive tool to perform inter-institutional patient path analysis.

Patterns in dyspepsia symptoms: Consider a domain expert, who is an epidemiologist and is interested in finding relationships between symptoms of dyspepsia within and across time points. This can be done by first mining patterns from symptom data and then using patterns to define association rules. Rules could look like ANOREX2=0 VOMIT2=0 NAUSEA3=0 ANOREX3=0 VOMIT3=0 ⇒ DYSPH2=0 where each symptom is represented as <symptom>N=V (time=N and value=V). ANOREX (anorexia), VOMIT (vomiting), DYSPH (dysphagia) and NAUSEA (nausea) are the different symptoms. However, a better way of handling this is to define subgroups as a set of symptoms at a single time point. (Lau, Ong, Mahidadia, Hoffmann, Westbrook, & Zrimec, 2003) solve the problem of identifying symptom patterns by implementing a framework for constraint based association rule mining across subgroups. Their framework, Apriori with Subgroup and Constraint (ASC), is built on top of the existing Apriori framework. They have identified four different types of phase-wise constraints for subgroups: constraint across subgroups, constraint on subgroup, constraint on pattern content and constraint on rule. A constraint across subgroups specifies the order of subgroups in which they are to be mined. A constraint on subgroup describes the intra-subgroup criteria of the association rules. It describes a minimum support for subgroups and a set of constraints for each subgroup. A constraint on pattern content outlines the inter-subgroup criteria on association rules. It describes the criteria on the relationships between subgroups. A constraint on rule outlines the composition of an association rule; it describes the attributes that form the antecedents and the consequents, and calculates the confidence of an association rule. It also specifies the minimum support for a rule and prunes away item-sets that do not meet this support at the end of each subgroup-merging step. A typical user constraint can look like [1,2,3][1, a=A1&n<=2][2, a=B1&n<=2][3, v=1][rule, (s1 s2) ⇒s3]. This can be interpreted as: looking at subgroups 1, 2 and 3, from subgroup 1, extract patterns that contain the attribute A1 (a=A1) and contain no more than 2 attributes (n<=2); from subgroup 2, extract patterns that contain the attribute B1 (a=B1) and contain no more than 2 attributes (n<=2); then from subgroup 3, extract patterns with at least one attribute that has a value of 1 (v=1). Attributes from subgroups 1 and 2 form the antecedents in a rule, and attributes from subgroup 3 form the consequents ([rule, (s1 s2) ⇒ s3]). Such constraints are easily incorporated into the Apriori process by pruning away more candidates based on these constraints.

They experimented on a dataset with records of 303 patients treated for dyspepsia. Each record represented a patient, the absence or presence of 10 dyspepsia symptoms at three time points (initial presentation to a general practitioner, 18 months after endoscopy screening, and 8–9 years after endoscopy) and the endoscopic diagnosis for the patient. Each of these symptoms can have one of the following three values: symptom present, symptom absent, missing (unknown). At each of the three time points, a symptom can take any of these three possible values. They show that their approach leads to interesting symptom pattern discovery.

Patterns in daily activity data: There are also works, which investigate techniques for using agent-based smart home technologies to provide at-home automated assistance and health monitoring. These systems first learn patterns from at-home health and activity data. Further, for any new test cases, they identify behaviors that do not conform to normal behavior and report them as predicted anomalous health problems.