การสร้างแบบจำลองข้อมูลในลักษณะของโมเดลที่มีความแม่นตรง เป็นจุดมุ่งหมายหลักของกระบวนการทำเหมืองข้อมูล การจะได้ผลผลิตสุดท้ายเป็นโมเดลที่แม่นตรงจะต้องใช้การค้นหาจากข้อมูลที่ได้รับการเก็บรวมรวมและผ่านกระบวนการเตรียมข้อมูลมาอย่างดี ข้อมูลที่ดีและมีคุณภาพจะเป็นปัจจัยสำคัญ ที่ช่วยให้โปรแกรมทำเหมืองข้อมูลสามารถค้นหาโมเดลที่ดีได้ภายในระยะเวลาอันสั้น แต่การเตรียมข้อมูลเป็นงานที่เสียเวลามากและมักจะต้องทำด้วยมือ หรือใช้โปรแกรมพื้นฐานอื่นๆ ช่วยเช่นโปรแกรมสเปรดชีตหรือแผ่นตารางทำการ งานวิจัยนี้จึงถูกเสนอขึ้นเพื่อพัฒนาเครื่องมืออัตโนมัติที่จะช่วยในกระบวนการเตรียมข้อมูล กระบวนการนี้จะประกอบด้วยสามขั้นตอนย่อยได้แก่ การแปลงรูปแบบข้อมูล การปรับปรุงข้อมูลให้สมบูรณ์ การคัดเลือกและลดขนาดข้อมูล ผู้วิจัยได้พัฒนาเทคนิคและเครื่องมือในการแปลงข้อมูลที่เป็นข้อความให้เป็นรูปแบบฮอร์นคลอส หรือข้อความแบบฮอร์นที่จะสามารถประมวลผลได้โดยโปรแกรมเชิงตรรกะ ในกรณีที่ข้อมูลมีบางค่าสูญหาย เครื่องมือที่พัฒนาขึ้นสามารถจัดการได้ด้วยสามแนวทางคือ การแทนที่ค่าสูญหายด้วยค่าคงที่บางค่า การแทนที่ค่าที่สูญหายด้วยค่าส่วนใหญ่ หรืออาจจะตัดทิ้งแอททริบิวต์ที่มีค่าสูญหายในสัดส่วนที่สูงมาก ด้านการลดขนาดข้อมูลงานวิจัยนี้ใช้เทคนิคการสุ่มเป็นพื้นฐานหลัก เครื่องมือที่พัฒนาขึ้นเพื่อการสุ่มข้อมูลประกอบด้วยเทคนิคการสุ่มแบบใส่ค่ากลับคืน การสุ่มแบบไม่ใส่ค่ากลับคืน และการสุ่มตามความหนาแน่นของกลุ่มข้อมูล จากการทดสอบเครื่องมือต่างๆที่พัฒนาขึ้นพบว่าสามารถลดเวลาการเตรียมข้อมูลลงได้มากและช่วยให้ผู้วิเคราะห์ข้อมูลทำความเข้าใจ รวมถึงทำการสำรวจลักษณะต่างๆ ภายในข้อมูลได้อย่างมีประสิทธิภาพ

225145

Modeling data accurately is the main focus of data mining process. Reliable model is the final product of the search process through the well-prepared and properly collected data records. High quality of the original data can greatly help the data mining tools to discover better models in less time. Data preparation is, however, an extremely time-consuming process and traditionally has been done manually or semi-automatically with some simple tools such as spreadsheet. This project is thus proposed to automate the tedious data preparation tasks. These tasks are comprised of data transformation, data cleaning, data selection and reduction. We develop techniques to transform textual data into Horn clauses that are ready to be processed by logic-based mining programs. All missing values in the data set are handled by several methods, i.e. replace with some constants, replace with majority values, or remove feature in which its values are highly missing. Data reduction is achieved via a sampling method. We implement three sampling methods: random sampling with replacement, random sampling without replacement, and density-biased sampling. The developed data preparation tools have been proved beneficial to data analysts as they can reduce the preparation time and gain more insight in their data.