# SUT Filter ระบบเตรียมข้อมูลเพื่อสนับสนุนการค้นหาความรู้*

# SUT Filter: A System for Data Preparation to Support Knowledge Discovery

กิตติศักดิ์ เกิดประสพ, นิตยา เกิดประสพ, สมหมาย เชิดชูชัยทิพย์, สุรเชษฐ์ ชุดพิมาย, ศิวพงศ์ รื่นเริง, อมรชัย ใจซื่อกุล

Kittisak Kerdprasop, Nittaya Kerdprasop, Sommai Cherdchuchaitip, Surachet Chutpimai, Siwapong RuenRueng, Amornchai Chaisuekul

School of Computer Engineering, Suranaree University of Technology, 111 University Ave., Muang District, Nakorn Ratchasima 30000, Thailand. e-mail address: kerdpras@ccs.sut.ac.th

**บทคัดย่อ:** งานวิจัยนี้เป็นการออกแบบและพัฒนาซอฟต์แวร์ที่ใช้ในการนำเข้าข้อมูลขนาดใหญ่ และปรับปรุงข้อมูลนั้นให้มีขนาดและคุณภาพที่เหมาะสมกับขั้นตอนการสังเคราะห์โมเดลในงานค้นหาความรู้จากข้อมูล  ข้อมูลที่นำเข้าสามารถมาจากหลายแหล่งและอยู่ในรูปแบบที่แตกต่างกัน  การปรับปรุงข้อมูลจะหมายรวมถึงการทำให้ข้อมูลสมบูรณ์ถูกต้อง  และลดแอททริบิวต์รวมทั้งจำนวนให้เหมาะสมกับงานค้นหาความรู้

**Abstract:** This paper describes an infrastructure for information loading and treatment to support the analysis phase of knowledge discovery. We designed and developed a system to specialize in the utilization of heterogeneous information sources and the preparation of those information to achieving the best knowledge discovery results. The preparation tasks include data format transformation, data cleaning and reduction. The final product of this cooperative multi-component system is the clean data set well-prepared for the knowledge mining task.

**Introduction:** Knowledge discovery in databases has recently received considerable attention due to the proliferation of large databases. Traditional data analysis methods that require humans to process data sets of huge size are completely inadequate. Fayyad, Piatetsky-Shapiro, and Smyth [2] describe knowledge discovery as the process of identifying valid, novel, potentially useful and understandable patterns in data. This complex process involves five phases: data selection, data preprocessing, data transformation, data analysis (mining), and interpretation and evaluation. While previously most attention is on the analysis phase involving the mining of patterns from data, the other phases are also currently admitted considerably important to the success of the process. Most mining algorithms presume that a cleansed and appropriately transformed data set is already available. This setting is not realistic in real world applications in which data are corrupted, noisy and format incompatible. It has been shown [1] that almost 75% of the discovery time has been spent on populating, cleansing, and transforming the data to a format appropriate for the mining algorithm.

The data preparation not only supports the mining phase, but also significantly influences the quality of the mined patterns. Our focus is thus on the preparation phases of the knowledge discovery process. We design and develop a data filtering system to access and prepare data for the mining algorithm. The proposed system supports user interactivity; i.e., the user is allowed to interactively direct the process towards an efficient result, or to supply background knowledge to some components for achieving the best performance.

**Methodology:** Data preparation tasks in the process of knowledge discovery comprise of the component to locate and access the relevant data set, the component to clean the data including fill-in the missing values, the component to reduce the data set to a sufficient size, and the component to transform the data format to match the requirements of the mining algorithm. Although the mining component is essential, the other components often require more time and effort to complete and can be the primary factors to the success or failure of the knowledge discovery process. Most knowledge discovery systems, e.g. WEKA [8] , assume a closed environment in that a database and a set of tools for cleaning, preparing, and analyzing data are combine within the same environment. Our data filtering system is designed to support the access of heterogeneous data sources and the preprocessing tasks that exist outside the mining environment. The proposed system may be viewed as a distributed infrastructure for the knowledge discovery task. The framework of a data filtering system is illustrated in Figure1.

* ได้รับทุนสนับสนุนการวิจัยจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ รหัสโครงการ NT-B-06-4C-19-524

Figure1 presents an overall perspective of the data filtering system. It illustrates the resources and all the components required to accomplish the data preparation tasks. The framework of a data filtering system divides the knowledge discovery process into three layers: locating and accessing, filtering, and mining layers. Our project has focused on the critical labor-intensive parts; i.e., locating, accessing, and filtering data.

*Locating and accessing layer.* At the bottom layer of the framework multiple, heterogeneous data sources are located in an enterprise environment. These data sources may be distributed across a network, such as the intranet or internet. Thus, the data filtering system must provide a component to dynamically locate the relevant data sources. Moreover, during the process of knowledge discovery if the underlying data sources are updated, there must be a monitoring component to notify and migrate the updated data to the upper layer.

*Filtering layer.* Once the relevant data are located and gathered, they will then be migrated to the next layer where all the necessary actions for preparing and filtering data are taken place. This filtering layer can be viewed as a data treatment phase in that the extracted data are carefully cleansed, the missing fields are treated, attributes are analyzed with an appropriate technique (e.g., symbolic attributes are analyzed on the basis of information theoretical measurement, whereas numeric attribute analysis is based on statistical measurements [5,6,7]) and only promising attributes towards the success of the mining process are selected. In order to support the mining algorithm to discover valuable knowledge, the sampling component [3,4] is another essential part. It is responsible for approximating the size of the data set. The justification of this size approximation role is that extracting knowledge from the massive data set is a computational and I/O intensive process which results in the difficulty of guaranteeing quick response time of the knowledge discovery process. Therefore, sampling is an effective technique to minimize the I/O traffic involved in the distributed environment in which the data filtering system and the mining system may be at the separate sites. Finally, the data samples are transformed to the format appropriate for each mining algorithm.

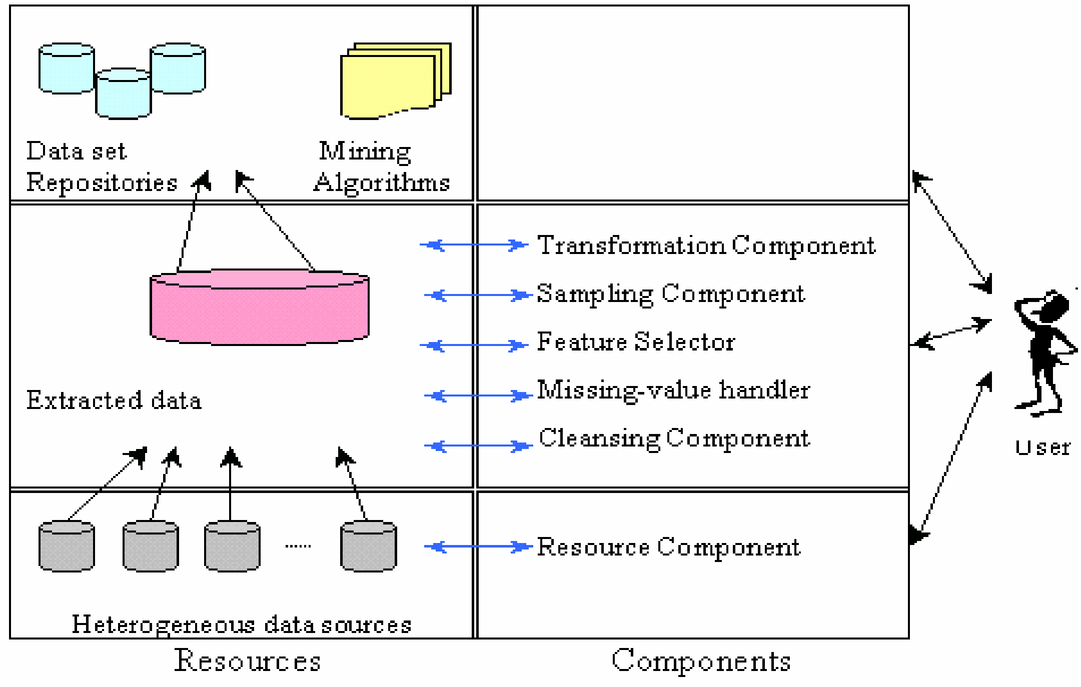


**Fig. 1.** Data filtering system framework

**Results and Discussion:** The components at the filtering layer are composed of:

- the cleansing component to get rid of noise in the data contents,
- the missing-value handler to seek for the appropriate method of filling in some missing parts in the data contents,
- the feature selector to efficiently evaluate and select the most promising features out of the available data set,
- the sampling component to obtain the representatives appropriate for a specific mining task, and
- the transformation component to turn the data to the right format.

Samples of major components are shown in Figures 2, 3, and 4.

**Conclusion:** The contribution of this paper is the design and implementation of a system to provide an integrated, flexible, and efficient pre-mining platform supported by various components. This platform provides mechanisms of data browsing and extracting, data arrange, data quality for knowledge discovery process. The interactions of modules in the system are controlled by one component acting as a supervisor of the whole system. Components are designed to be active and intelligent. They are able to react appropriately to unpredictable situations, evaluate and apply their own problem solving strategies. However, the current design has to be extensively tested on various application domains. Several areas of extensions are currently being investigated. The functionalities of filtering components can be extended to support new techniques of cleansing and adaptive sampling. Finally, we plan to extend the scope of this project to cover the mining and post-mining phases to obtain a complete distributed, agent-based knowledge discovery system.

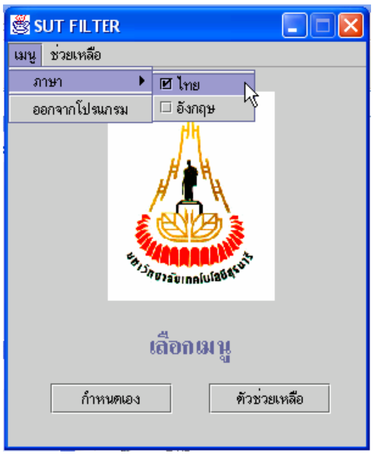
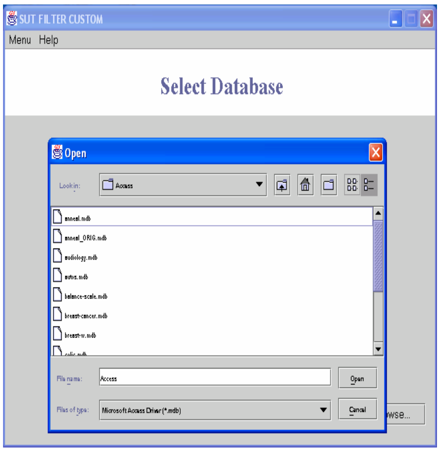


**Fig. 2.** Main menu of SUT Filter



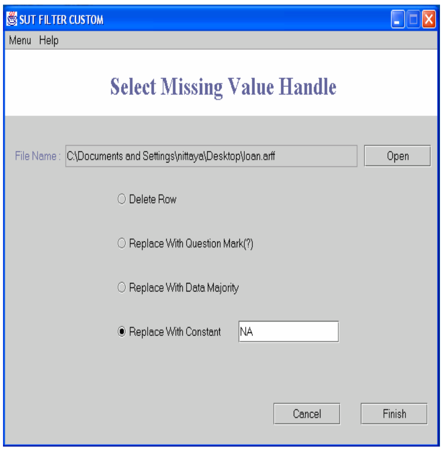
Look in: Access

anneal.mdb
anneal_ORIG.mdb
audiology.mdb
autos.mdb
balance-scale.mdb
breast-cancer.mdb
breast-w.mdb
colic.mdb

File name: Access

Files of type: Microsoft Access Driver (*.mdb)

**Fig. 3.** Data source selection and loading



File Name : C:\Documents and Settings\nittaya\Desktop\loan.arff

○ Delete Row

○ Replace With Question Mark(?)

○ Replace With Data Majority

● Replace With Constant NA

**Fig. 4.** Missing-value handling Component

**References:**

1. Engels, R., Theusinger, Ch.: Using a Data Metric for Preprocessing Advice for Data Mining Applications. In: *Proceedings of 13th European Conference On Artificial Intelligence* (1998) 430-434
2. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (eds.): *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA (1995) 1-34
3. Fink, A.: *How to Sample in Surveys, volume 7.* Sage Publications, Thousand Oaks, CA (2002)
4. Henry, G. T.: *Practical Sampling*. Sage Publications, Thousand Oaks, CA (1990)
5. Rubin, D.B.: Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, Vol. 91 (1996) 473-489
6. Schafer, J.L.: Multiple Imputation: A Primer. *Statistical Methods in Medical Research* (1999)
7. Shafer, J.L., Olsen, M.K.: Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multi variate Behavioral Research*, Vol. 33 (1998) 545-571
8. Witten, I., Frank, E.: *Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco (2000)