

The Chamois Reconfigurable Data-Mining Architecture

Won Kim*, Ki-Joon Chae, Dong-Sub Cho, Byoungju Choi, Anmo Jeong, Myung Kim, KiHo Lee, Meejeong Lee, Minsoo Lee, Sang-Ho Lee, Seung-Soo Park, Hwan-Seung Yong, Ho-Sook Kim, Jung-Won Lee, Wol-Young Lee

Department of Computer Science and Engineering, Ewha Women's University, Seoul, Korea

* also Cyber Database Solutions, Inc., Austin, Texas, U.S.A.

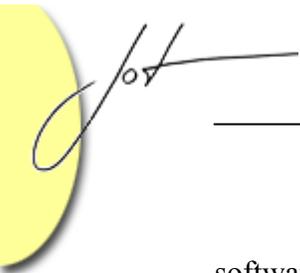
Abstract

The process of knowledge discovery in data (KDD) stored in computers in general requires iterations of three stages: data preparation, data mining, and results analysis. A variety of software tools are available for each of the stages. KDD environments, objectives of KDD, and types of data to be mined affect the choice of software tools in each stage. This article proposes a component-based architecture for an "end-to-end" integrated suite of KDD software tools that supports the entire KDD process. The architecture allows the configuring of an integrated tool suite with software tools appropriate for a given KDD environment and a given set of KDD objectives. The architecture is a part of the Chamois component-based knowledge-engineering framework under development at Ewha Women's University in Korea.

1 INTRODUCTION

Businesses and governments are increasingly adopting knowledge discovery in data (KDD) (often somewhat erroneously called data mining) technology to automatically discover unforeseen knowledge (in the form of patterns of data) from huge amounts of raw data at their disposal that can lead to business and governance advantages. The use of KDD technology in detecting frauds involving credit cards, phone cards, and insurance claims; and segmenting people and customers, etc. is already fairly well established.

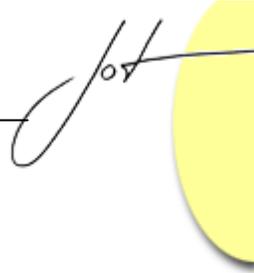
The process of discovering knowledge from raw data stored in computers is rather complex, involving iterations of three stages: data preparation, data mining, and results analysis. A variety of software tools are on the market to aid each stage of the KDD process, and most software tools offer different advantages and disadvantages. The KDD process today suffers from two major problems: weak interoperability among the



software tools, and inability to substitute one tool with another for the same “application”. By “application” we mean a major task in each stage of the KDD process, such as dirty data cleansing, data extraction from one or more data sources, data transformation, data loading, data clustering, data classification, data analysis, etc. It is important that the output of one stage of the KDD process be usable as input for the next stage without a high data transformation overhead. The need to transform data from one stage to the next can lead to a serious KDD administration overhead and a slowdown in the overall KDD process. An integrated software tool suite that takes the user through all three stages of the KDD process in a single workflow, without forcing “data transformation” stops along the way, is highly desirable. Further, the software tools from different vendors for the same “application” in each of the three stages offer different advantages and disadvantages, and even the same vendor offers a variety of tools for the same “application”. Most vendors of data-mining algorithms provide more than one data-mining algorithms, such as a decision-tree algorithm, a neural net algorithm, an association rules algorithm, etc., since no data-mining algorithm is best for all applications. As such the “best” tool for each stage depends on such factors as the types of data to be mined, the primary operations of the application, the set of data sources, the degree of dirty data, etc. Hence it is highly desirable for the user to be able to create an “optimal mix” of software tools that best suits the user’s KDD requirements. However, it is difficult to substitute one software tool for another in the same stage without a serious data transformation and software engineering overhead.

The Department of Computer Science and Engineering in Ewha (pronounced ee-hua) Women’s University of Seoul, Korea is currently developing a mammoth prototype component-based framework for supporting the development of enterprise business intelligence applications. The framework, named Chamois [Kim et al 2002], consists of commercial software and research prototype components built by the Chamois project team (hereafter to be referred to as “Chamois research modules”). The framework is currently operational, and several components have been integrated into the framework, with more to follow. One of the possibilities that the Chamois framework offers is precisely the integrated suite of tools that aid the three stages of the KDD process described above. The Chamois framework consists of components, and each component is made to communicate in a common Chamois API (Application Programming Interface). Each KDD software tool becomes merely a component in the Chamois framework.

In this article, we propose the Chamois component-based framework as a research vehicle for studying feasibility of building an “end-to-end KDD tool suite” that makes it possible to easily substitute software tools in each of the three stages of a KDD process as appropriate for different KDD requirements. We first examine the KDD process in somewhat greater depth, and then describe the Chamois project and architecture in some depth in order to illustrate how the Chamois framework can serve as a framework for supporting software tools in an “end-to-end” KDD process.

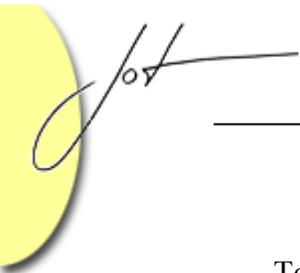


2 THE KDD PROCESS AND SOFTWARE TOOLS

Let us examine each of the three stages of the KDD process a little more closely. The data preparation stage basically prepares raw data to apply data-mining algorithms. This is the most time-consuming and administration-intensive stage in the KDD process. It often represents about 80 percent of the time and efforts in a KDD process. Data preparation includes in general the creation of a data mart (or data warehouse) in a relational database (RDB), and transformation of some of the tables in the data mart into a format suitable for processing by the data-mining algorithms. Data mart creation in turn involves data extraction, transformation, and loading (ETL). The data-mining stage involves the running of data-mining algorithms chosen for a given application. This stage requires following all the steps required by the algorithms (e.g., model building, model training, model validation), and the setting of all parameters required for the algorithms. The results analysis stage is necessary for analyzing the results of the data-mining algorithms. Often the results of running data-mining algorithms are not meaningful for a variety of reasons, such as errors in running the algorithms, the effect of dirty data on the algorithms, the results being non-actionable, etc. Based on the results analysis, the KDD process may be repeated either by running the data-mining stage or the data-preparation stage.

A variety of software tools are on the market for each of the three stages of the KDD process. For the data-preparation stage, database design tools and ETL tools are used for the creation of data marts (data warehouses). Database design tools are used to create the schema of the data mart (data warehouse), often in a star-join or snowflake structure of tables. RDB vendors and such tool vendors as Rational Software and Metrowerks offer database design tools. RDB vendors and ETL vendors such as Ascential Software, Informatica, Sagent Technology, Acta Technology, etc. offer ETL tools. These tools differ in the types of data sources they support, performance and scalability, and flexibility to transform data. In the data-preparation stage, data-quality software is also used. Ascential Software, First Logic, and Trillium Software offer data-cleansing tools for use in detecting and repairing data in non-standard form, missing data, mistyped data, etc. For the data-mining stage, there are lots of data-mining algorithms [Berry and Linoff 1977]. For the results analysis stage, query tools, data-mining visualization tools, and statistical analysis packages are used. Business Objects, Cognos, Brio Technology, Silicon Graphics, SAS Institute, SPSS, etc. offer results analysis tools.

Besides the software tools used to directly support the tasks in each stage of the KDD process, a variety of additional software tools form the “infrastructure” for the KDD process. Such infrastructure software tools include a metadata repository for managing the metadata (e.g., for ETL and for the data marts), an RDB for storing and managing data in data marts, and an OLAP (online analytical processing) server for aiding multidimensional data analysis.



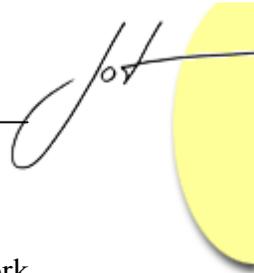
To be sure, most data-mining software vendors provide somewhat integrated tool suites. For example, the MineSet product from Silicon Graphics, the Enterprise Miner product from SAS Institute, the Darwin product from Oracle each include several data-mining algorithms, data transformation utilities, and data visualization utilities. These products are effective in addressing certain application requirements. Further, a data-mining algorithm from one vendor is often integrated with third-party query tools and RDBs. An “end-to-end” KDD process, however, can give rise to numerous workflows with different requirements for different parts of the workflow, and the current level of interoperability falls far short of allowing for the configuring of an optimal mix of tools from a variety of vendors. The following is a possible KDD workflow.

A KDD application may require data from a few different data sources. An ETL tool may be needed to extract, transform and load data from such data sources into a data mart. A dirty data analysis may be needed before the data mart is used, as the data-mining algorithm selected may be sensitive to dirty data. Then a data-cleansing tool may be used to cleanse the data mart: for example, missing data may be filled in, wrong data may be corrected, data formats may be standardized, etc. A data transformation tool may now be used to transform the cleansed data mart: for example, an integer-valued field may be changed to a categorical-valued field, a string-valued field may be changed to a numerical-valued field, records that contain field values that fall outside a certain range are deleted, etc. A data-mining algorithm selected is then run. The results of the algorithm are then analyzed using a data visualization tool. Depending on the data-mining algorithm selected, a possibly different data-mining algorithm is run to test for staleness of the data-mining model that was created earlier, and if the model is deemed stale, the original data-mining algorithm is re-run.

3 THE CHAMOIS PROJECT

Chamois is the code name for the Integrated Knowledge Engineering Architecture (IKEA) project which has been underway in the Department of Computer Science and Engineering at Ewha Women’s University since the fall of 1999. It is a 5-year, large-scale advanced research and development (R&D) project aimed at developing a unified framework (platform) for supporting development of knowledge engineering (i.e., business intelligence) applications. The Chamois project provides a single research focus for 11 of the faculty members and 80 graduate students in the Department. The goal and scope of the project are described in [Kim et al 1999]. (* The chamois is a goatlike antelope that lives in high mountains of Europe and southwestern Russia, and is known for its ability to jump very high after only a few short momentum-generating gallops. We named our project Chamois in our hope that the project will serve as a vehicle for sharply elevating the research capabilities of the Department and the University. *)

The framework is to provide core services such as data warehousing, online analytical processing (OLAP), data and text mining, XML document storage and query



processing, data quality management, QoS transmission of multimedia data, network security, information personalization, etc. Today, these core services exist in the form of independent software from a variety of different software vendors that are not designed to flow together. One objective of the Chamois project is to integrate most of these technologies (software products) into a single unified framework, so that knowledge engineering application developers may develop applications, such as customer relationship management, business analytics, executive decision support, electronic commerce, etc., with a much higher productivity. The framework is built using component-based design, integration, and testing technologies [Schmidt and Assmann 1998]. Another objective of the Chamois project is to make significant research advances in each of the core service technologies for the framework.

The Chamois framework includes a number of commercial software and Chamois research modules. Broadly, the framework consists of an “infrastructure” part and an “application part”. The infrastructure part consists of commercial software and prototype modules that provide all key knowledge engineering infrastructure technologies. The application part consists of applications that make use of the infrastructure part as sources of data and knowledge. The primary application is an electronic commerce system. However, there are also various simpler applications, such as a university registration data miner.

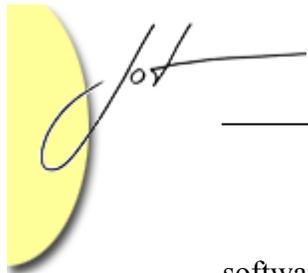
4 CHAMOIS ARCHITECTURE

The architecture of the Chamois framework is described in substantial detail in [Kim et al 2002]. To make this article self-contained, however, we repeat below a small part of the descriptions from [Kim et al 2002] that is of relevance to the subject of this article.

The Chamois framework is a scalable three-tier architecture that is similar to the EAI (Enterprise Application Integration [Ruh et al 2001]). It consists of the following major components: Common Communication Bus, Workflow Designer, Automation Server, Integration Manager, Transformation Server, Metadata Repository, Database Adapter, Commercial Software Packages, Chamois Software Modules, Chamois API, Commercial Software Adapters, Chamois Software Adapters, Chamois-Based Applications, Adapter SDK, and System Monitor.

The Common Communication Bus connects various commercial software with Chamois research modules. It also connects knowledge engineering applications developed using Chamois. As a common bus, the Chamois framework uses Microsoft’s Component Object Model (COM+) [Platt 1999] for an intranet environment and Web Service [Kirtland 2001] for an Internet environment.

Commercial Software Adapters are runtime libraries that dynamically convert proprietary APIs of commercial software to the Chamois common API. They provide a common way to access various proprietary APIs supported by commercial business



software that the Chamois team has adopted. They include the MaxScan high-performance data server, Informatica PowerMart ETL tool, the Accrue/Pilot Software MOLAP server, Microstrategy ROLAP server, IBM DB2, Microsoft SQL/Server, and Oracle RDB system, IBM Intelligent Miner text mining toolkit, DB Miner data-mining tool, etc.

Chamois research modules have their own proprietary APIs. They include JAVA/CORBA, Call Level Interface (CLI), and COM+. Chamois Software Adapters are runtime libraries that dynamically convert these proprietary APIs to the Chamois common API.

The Chamois framework provides two knowledge engineering services in support of a KDD process. One allows the construction of a knowledgebase by extracting, transforming, and loading operational data into the data warehouse, storage of XML documents in an RDB, generating OLAP summary tables for fast query processing, and training data-mining models, etc. The Workflow Designer and major Chamois components provide this service. Another service allows queries against the knowledgebase from Chamois-based applications, and presents and visualizes the results.

The users of the Chamois project may use the Workflow Designer to design automated processes between applications and some components of the Chamois framework. Chamois provides a graphical user interface (GUI) for the design of a workflow process that consists of data transfers among the components and executions of the components. This process is executed periodically via a scheduler, so that a knowledgebase is automatically maintained for fast response to client's queries. The Automation Server uses a developed workflow process to coordinate the invocation and notification of events, and message transmission among all parts of the Chamois framework.

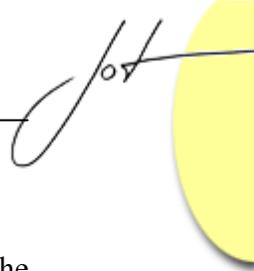
5 DATA MINING RESEARCH RESULTS INTEGRATED INTO CHAMOIS

Below we outline the Chamois research modules that have been integrated into the Chamois framework that complement commercial software tools in the data preparation and data-mining stages of the KDD process. Somewhat more detailed descriptions of these and other Chamois components are given in [Kim et al 2002].

Data Mining

Data-mining Model Designer

We have studied some 30 existing data-mining algorithms with the view to bring some organization to the many data-mining algorithms. In particular, we wanted to understand the strengths and weaknesses of the algorithms, the types of applications the algorithms



are suited for, and the ways in which the algorithms are invoked. We grouped the algorithms into 7 categories: association rules, clustering, neural networks, decision trees, genetic algorithms, memory-based reasoning, and Bayesian networks. We defined 10 criteria for assessing these algorithms: input data type, time complexity, scalability, explainability of the result, rate of the model going stale, ease of use, degree of human interaction required, training time, supervised/unsupervised, and applicability [Lee et al 2001a].

This effort has clearly exposed to us the need to provide Chamois users with a means to create a meta data-mining model for describing the essential characteristics of the data-mining algorithms. Chamois allows the user to create a meta data-mining model from both multidimensional and RDB sources by using the Mining Model Wizard provided with Analysis Services of Microsoft SQL server.

Spatial Data Mining

We have developed and integrated into Chamois a new spatial clustering algorithm named DBSCAN-W. Spatial data mining is a process to discover interesting relationships and characteristics that exist implicitly in a spatial database [Han and Kamber 2001]. We have implemented the algorithm by using an Informix Spatial Datablade module [Informix 1997]. DBSCAN-W is an extension of the existing density-based clustering algorithm DBSCAN [Ester et al 1996].

XML Mining

We have developed a new methodology for computing similarity between XML documents by taking account of XML semantics (i.e., meanings of the elements and nested structures of XML documents) [Lee et al 2001b]. We expect that many Web applications that process XML documents, such as grouping similar XML documents and searching for XML documents that match a sample XML document, will require the clustering and classifying of XML documents.

Data Preparation

Data Quality Monitor

We have developed a methodology for measuring the quality of data, and developed and integrated into the Chamois framework a Data Quality Measurement (DAQUM) prototype that reflects the methodology [Kim et al 2001].

An XML Server

We have developed an XML data server for storing and querying XML data in an RDB [Kim and Lee 2001]. We extract an RDB schema from XML documents. To optimize queries against XML documents stored in an RDB, we use a cost model that we have developed on the basis of an analysis of all possible types of queries against XML documents.

An OLAP Server

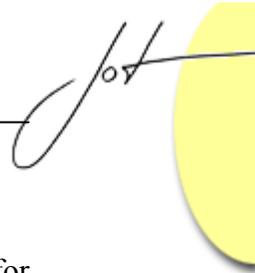
We are developing a high-performance scalable prototype OLAP server. OLAP is a process and methodology for a multidimensional data analysis from an enterprise data warehouse. As part of our efforts, we have extended existing techniques for speeding up OLAP cube generation, and devising efficient OLAP cube storage schemes for fast query processing [Kim and Lim 2001][Kim and Song 2002]. Chamois provides a GUI to allow users to easily control, access, and integrate OLAP components. It also provides a Dimension Wizard for the definition of dimensions and levels on a data mart, and a Cube Wizard for designing the structure of an OLAP cube. Metadata are stored in the Metadata Repository.

XML-based Metadata Converter

We have developed an XML-based translator that integrates the export and import functions to exchange metadata between a pair of RDBs [Lee and Lee 2000]. The translator translates source metadata into an XML equivalent, and converts it back to the metadata of a target database system. It preserves consistency of data exported and imported. The translator may be used for the exchange of metadata among all components of Chamois that involves metadata.

6 CONCLUDING REMARKS

In this article we described how the Chamois component-based framework may be used to support the implementation of an “end-to-end integrated suite of software tools” for supporting the entire process of automatic discovery of knowledge from data stored in computers. The Chamois framework, currently operational at Ewha Women’s University, is a framework that allows the integration of commercial business intelligence infrastructure software and prototype modules developed by the Chamois project team that supplement the commercial software. The KDD process involves iterations of three stages: data preparation, data mining, and results analysis. The Chamois framework includes various commercial software for use in each of these stages. It also includes Chamois research modules, including a couple of novel data-mining algorithms that have been reported in the literature; a few data preparation tools such as an XML server, an OLAP server, and a metadata converter. The Chamois framework also includes



infrastructure prototype software, such as a workflow designer, and some applications for validating the operation of Chamois, such as an electronic commerce system.

Using the facilities of the Chamois framework, we plan to conduct two lines of research into a software architecture for supporting the automatic knowledge discovery in data and data mining. One is feasibility of building an “end-to-end integrated suite of software tools” for supporting the entire KDD process. Another is feasibility of “tailoring such a tool suite” with dynamic substitution of software tools to suit the changing requirements of the applications and application environments.

REFERENCES

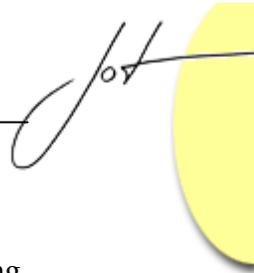
- [Berry and Linoff 1997] M. Berry and G. Linoff. “Data Mining Techniques for Marketing, Sales and Customer Support”. Wiley, 1997.
- [Ester et al 1996] M. Ester, H. Kriegel, J. Sander, and X. Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, In Proc. 3rd Int’l Conf. on Knowledge Discovery and Data Mining, pages 226-231, Montreal, Canada, Jun. 1996.
- [Han and Kamber 2001] J. Han and M. Kamber. “Data Mining: Concepts and Techniques”, pages 405-412, Morgan Kaufmann, 2001.
- [Informix 1997] Informix, “Informix Spatial Datablade Module: User’s Guide”, Informix Press, 1997.
- [Kim et al 1999] W. Kim, et al. “A Component-Based Knowledge Engineering Architecture”, Journal of Object-Oriented Programming, Sep. 1999.
- [Kim et al 2001] W. Kim, B. Choi, E. Hong, S. Kim, D. Lee. “A Taxonomy of Dirty Data”, Journal of Data Mining and Knowledge Discovery, the Kluwer Academic-Publishers, to appear in 2003.
- [Kim et al 2002] W. Kim, et al. “The Chamois Component-Based Knowledge Engineering Framework”, IEEE Computer, May 2002, pages 44-52, IEEE CS Press.
- [Kim and Lee 2001] J. Kim, W. Lee, K. Lee. “The Cost Model for XML Documents”, In Proc. of ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, Jun. 2001.
- [Kim and Lim 2001] M. Kim and Y. Lim. “A Z index based MOLAP cube storage scheme”, Ewha Institute of Science and Technology Research Report Series EIST-CSE-01002, 2001.

- [Kim and Song 2002] M. Kim and J. Song. “An Efficient ROLAP Cube Generation Scheme”, Journal of Korea Information Science Society, Vol. 29(2), April 2002.
- [Kirtland 2001] M. Kirtland. “Web Services Essentials”, Microsoft Developer Network, Jan. 2001.
- [Lee and Lee 2000] W. Lee and K. Lee. “An XML-based metadata interchange among the RDB systems”, In Proc. of IASTED Int’l Conf. on Internet and Multimedia Systems and Applications, Las Vegas, Nevada, Nov. 2000.
- [Lee et al 2001a] J. Lee, H. Kim, J. Choi, H. Kim, H. Yong, S. Lee, and S. Park. “Classification and Analysis of Data Mining Algorithms”, to appear in Journal of Korea Information Science Society, Vol. 28(3), Sept. 2001.
- [Lee et al 2001b] J. Lee, K. Lee, and W. Kim. “Preparations for Semantics-based XML Mining”, in Proc. 1st IEEE Int’l Conf. on Data Mining (ICDM 2001), San Jose, CA, Nov. 2001, IEEE CS Press.
- [Platt 1999] D. S. Platt. “Understanding COM+”, Microsoft Press, Jul. 1999.
- [Ruh et al 2001] W.A. Ruh et al. “Enterprise Application Integration”, John Wiley & Sons, 2001, pp. 12-38.
- [Schmidt and Assmann 1998] R. Schmidt and U. Assmann. “Concepts for Developing Component-Based System”, Int’l Workshop on Component-Based Software Engineering, Kyoto, Japan, Apr. 1998.

About the authors



Won Kim is President and CEO of Cyber Database Solutions (www.cyberdb.com) and MaxScan (www.maxscan.com) in Austin, Texas, USA. He is also Dean of Ewha Institute of Science and Technology, Ewha Women's University, Seoul, Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology (www.acm.org/toit), and Chair of ACM Special Interest Group on Knowledge Discovery and Data Mining (www.acm.org/sigkdd). He is the recipient of the ACM 2001 Distinguished Service Award.



Ki-Joon Chae, professor of computer science and engineering, (network security, active network management and security, network protocol design and performance evaluation), Ph.D. in electrical and computer engineering, North Carolina State University, kjchae@ewha.ac.kr



Dong-Sub Cho, professor of computer science and engineering, (computer systems architecture, web engineering, interactive game engines), Ph.D. in computer engineering, Seoul National University, dscho@ewha.ac.kr



Byoungju Choi, associate professor of computer science and engineering, (component-based software engineering, software testing, data and software quality), Ph.D. in computer science, Purdue University, bjchoi@ewha.ac.kr



Anno Jeong, President and CEO of Artist, Inc. and adjunct professor of computer science and engineering, (knowledge acquisition and management, component-based system architecture design, enterprise application integration), B.S. in physics education, Seoul National University, jam96@ibis.co.kr



Myung Kim, associate professor of computer science and engineering, (OLAP, high performance computing, parallel/distributed computing), Ph.D. in computer science, the University of California at Santa Barbara, mkim@ewha.ac.kr



KiHo Lee, professor of computer science and engineering, (programming languages, compilers, XML), Ph.D. in computer science, Seoul National University, khlee@ewha.ac.kr



Meejeong Lee, associate professor of computer science and engineering, (QoS networks, multicast communications, wireless mobile network protocols), Ph.D. in electrical and computer engineering, North Carolina State University, lmj@ewha.ac.kr



Minsoo Lee, assistant professor of computer science and engineering, (database systems, data warehousing, business intelligence, and Web technology), Ph.D. in computer and information science and engineering, University of Florida at Gainesville, mlee@ewha.ac.kr



Sang-Ho Lee, professor of computer science and engineering, (algorithm design, information security, bioinformatics), Ph.D. in computer science, Korea Advanced Institute of Science and Technology, shlee@ewha.ac.kr



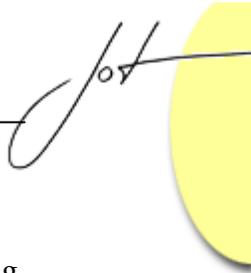
Seung-Soo Park, Dean of Engineering College, associate professor of computer science and engineering, (artificial intelligence, data mining and bioinformatics), Ph.D. in computer science, the University of Texas at Austin, spark@ewha.ac.kr



Hwan-Seung Yong, associate professor of computer science and engineering, (multimedia database systems, data mining and bioinformatics, Internet computing), Ph.D. in computer engineering, Seoul National University, hsyong@ewha.ac.kr



Ho-Sook Kim, Ph.D. candidate in computer science and engineering, (database systems, geographical information systems, and data mining), khosook@ewha.ac.kr



Jung-Won Lee, Ph.D. candidate in computer science and engineering, (data mining, XML, and programming languages), jungwony@ewha.ac.kr



Wol-Young Lee, Ph.D. candidate in computer science and engineering, (programming languages, XML, and database systems), wylee@ewha.ac.kr