A Prototyping Framework of the Documentation Retrieval System for Enhancing Software Development Quality

Wen-Kui Chang*, **Tzu-Po Wang**

Department of Computer Science and Information Engineering
Tunghai University, Taichang, Taiwan
*wkc@mail.thu.edu.tw

Abstract

This paper illustrates a prototyping framework of the documentation-standards retrieval system via the data mining approach for enhancing software development quality. We first present an approach for designing a retrieval algorithm based on data mining, with the three basic technologies of machine learning, statistics and database management, applied to this system to speed up the searching time and increase the fitness. This approach derives from the observation that data mining can discover unsuspected relationships among elements in large databases. This observation suggests that data mining can be used to elicit new knowledge about the design of a subject system and that it can be applied to large legacy systems for efficiency. Finally, software development quality will be improved at the same time when the project managers retrieving for the documentation standards.

Keywords : software quality assurance, data mining, distributed systems, database management

1. Introduction

Knowledge discovery research and development in database management is often called data mining, which emerged in the early 1990's. It aims at the discovery of useful information from large collections of data [1]. The data mining can be based on rules describing properties of the data, frequently occurring patterns, and clusters of the objects in the database. Current technology makes it fairly easy to collect data, but data analysis tends to be slow and expensive. There are several successful applications of data mining in [7].

Essentially, this paper briefly gives a documentation-standard retrieval system via the data mining approach for enhancing software development quality, with the three basic technologies of machine learning, statistics and database management, applied to this system to speed up the searching time and increase the fitness [9]. We start in Section 2 by briefly discussing the data mining approach, including the role of machine learning, statistics and database management in data mining. Section 3 discusses the system design concept and architecture in the standard retrieval system. In Section 4, the standard retrieval system for distributed software standards is introduced briefly. Finally, a demonstration is illustrated in Section 5.

2. Data mining approach

Generally data mining may be described in the following aspects:

2.1. Overview of data mining

With the proliferation of data warehouses, data mining tools are flooding the market [5, 10]. Their objective is to discover hidden gold in the data. The ultimate objective of data mining is knowledge discovery. Data mining methodology extracts hidden predictive information from large databases. That's where technology comes in: for true knowledge discovery a data mining tool should unearth hidden information automatically. By this definition, data mining is data-driven, not user-driven or verification-driven.

The tool processes every field in every record in the database until it sufficiently splits the common use from the non-common use and learns the main differences between them [1, 4]. Once the tool has learned the crucial attributes it can rank them in order of importance. A user can then exclude attributes that have little or no effect on targeting potential new applications.

2.2. Data mining techniques

Data mining combines methods from at least three areas: machine learning, statistics, and database management [6, 8].

The close links between machine learning, statistics and data mining are fairly obvious. All three areas aim at locating interesting regularities, patterns, or concepts from empirical data.

2.2.1. Machine learning

Machine learning methods form the core of data mining: decision tree learning or rule induction is one of the main components of several data mining algorithms.

2.2.2. Statistics

In data mining, statistics is also an important framework and technology for knowledge retrieving. Statistical analysis can be used to find the relationships existing in large bodies of raw data.

2.2.3. Database Management

Database management systems are systems especially developed for the storage and flexible retrieval of large masses of structured data. The detailed database schema is discussed in the following section.

3. Derived classification approach

According to the data mining methodology, the derived classification approach is illustrated as follows [9, 11, 12].

3.1. Classification by IEEE 1074

By the IEEE 1074 classification methodology, the raw data is first divided into larger groups, and each group process is then divided into smaller processes.

3.2.Classification by project schedule

Another method is classification by the project schedule. This includes the life cycle model, project management, pre-development, development, post-development and integration processes.

3.3. Classification by project category

Finally, a project category is provided to classify these documentation-standard documents. The details are listed in Table 1.

Table. 1 Project category and project schedule

Group	Process
Life Cycle Model	Selection of a Life Cycle Model
Project Management	Project Initiation Project Monitoring and Control Software Quality Management
Pre-development	Concept Exploration System Allocation
Developmen t	Requirements Design Implementation
Post-development	Installation Operation and Support Maintenance Retirement
Integration	Verification and Validation Software Configuration Manage-ment Documentation Development Training

Lastly, another group named "other" is added for the process "other". If the standard documents cannot be classified into any of the above groups, they will be classified into the "other" group.

3.4. Database schema

The database schema in the standard retrieval system is illustrated in the following. It is noted that the concerned schema is skeptically simplified since this prototyping framework was emphasized on the generic usage. Thus, the following database schema may not consider complete in the sense of database integrity.

4. Standard retrieval system

Empirically the prototyping standard retrieval system was implemented by the ASP (Active Server Page) and the database was built using Microsoft Access. The major database is located at the server, which runs Windows NT, and the other databases are located at different servers, which run different operating systems. The following table illustrates the distributed database and its operating system.

Table. 2 Distributed database and its operating system

Standards Database	Operating system
IEC and ISO	Windows 98
ANSI	Linux
IEEE	Windows 2000
NASA	Solaris

4.1. System functions

The prototyping standard retrieval system, including two kinds of users, common users and administrators, is represented by an UML use case diagram as illustrated in Fig. 2. The functions for common users include data browsing, data searching, statistical data and frequently asked questions (FAQ). The functions for the administrator include data inserting, data deleting, data modifying, statistical data collecting and information management. Among these functions, the information management is used for data mining. By means of the data mining results, some raw data will become useful information. For example, the most frequently referenced documents will be listed at the beginning of the result.

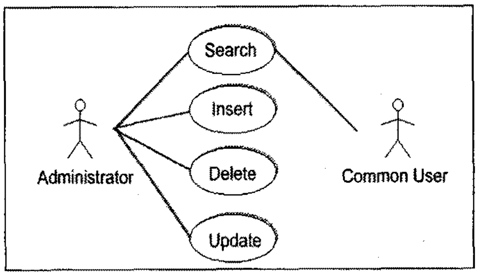


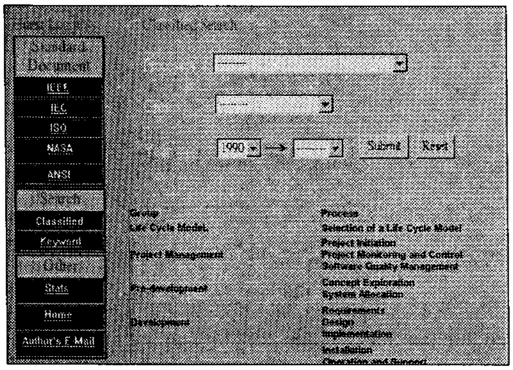
Fig. 1 UML use case diagram

4.2. System architecture

The standard retrieval system is divided into the client component and the server component, which is illustrated in the following [2, 3]:

4.2.1. Client component

The client interface is shown in Fig. 3. It includes two major users: common users and administrators. Common users use the search form to query the information they want. Two search forms are provided, including classified search and keyword search. In addition to the search function, the administrator maintains the database, including the *"insert"*, *"update"*, and *"delete"* functions. The hierarchical function view is illustrated in Fig. 4.



Group	Process
Life Cycle Model	Selection of a Life Cycle Model
Project Management	Project Initiation Project Monitoring and Control Software Quality Management
Pre-development	Concept Exploration System Allocation
Development	Requirements Design Implementation
	Installation Operation and Support

Fig. 2 Client interface

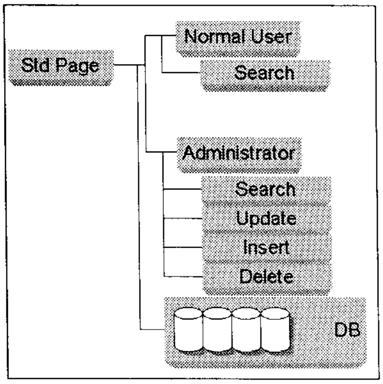


Fig. 3 Hierarchical function view

4.2.2. Server component

The server component consists of four sites, each of which is in a different operating system, as shown in Fig. 5. The standard document database is divided into IEEE, ISO&IEC, NASA, and ANSI. The IEEE standard document database is put on site S1, which uses Windows 2000. By the "my network places" function and "map network driver" function, we put the ISO and IEC standard document database on site S2, which uses Windows 98. The other two sites are not Microsoft family operating systems. However, Samba technology is added to communicate with Linux and Solaris environments. We put the NASA standard document database on site S3 that uses the Solaris 2 operating system, and the ANSI standard document database on site S4 that uses the Linux operating system.

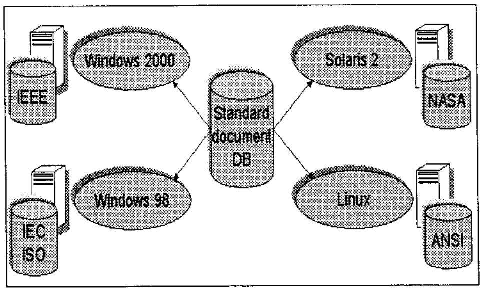


Fig. 4 The four parts of the distributed database

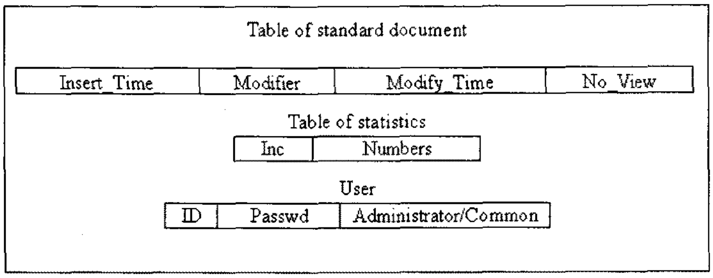


Table of standard document

Insert_Time	Modifier	Modify_Time	No_View

Table of statistics

Inc	Numbers

User

ID	Passwd	Administrator/Common

Fig. 5 Table schema in the prototyping standard retrieval system

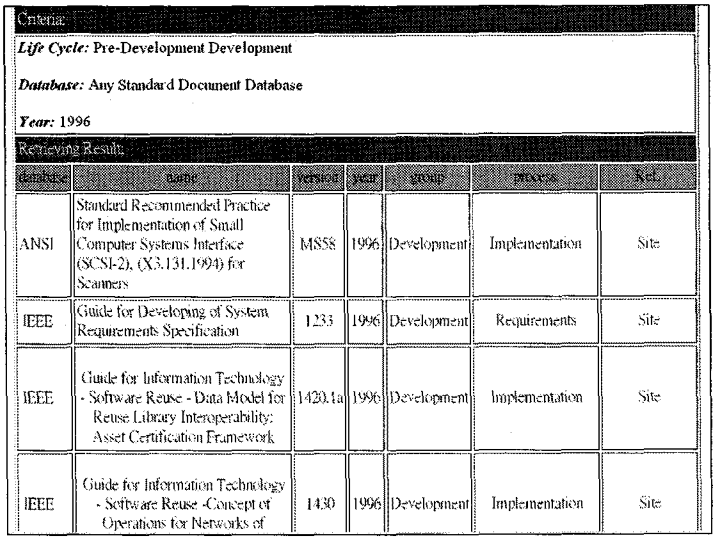
5. Demonstration example

5.1. Description

In this case, we desire to retrieve a standard document from a database that collected a group of standards established up to 1996. After assigning the specified values of the required fields, retrieved result is illustrated in Fig. 6. The most frequently retrieved standard documents are shown at the beginning of the result lists. The standard documents, which were produced at the same time, are also sorted in the same form.

5.2. Discussion of results

In this case, the three basic technologies are used successfully in the standard



Criteria:

Life Cycle: Pre-Development Development

Database: Any Standard Document Database

Year: 1996

Retrieving Result:

database	name	version	year	group	process	Ref.
ANSI	Standard Recommended Practice for Implementation of Small Computer Systems Interface (SCSI-2), (X3.131.1994) for Scanners	MS58	1996	Development	Implementation	Site
IEEE	Guide for Developing of System Requirements Specification	1233	1996	Development	Requirements	Site
IEEE	Guide for Information Technology - Software Reuse - Data Model for Reuse Library Interoperability: Asset Certification Framework	1420.1a	1996	Development	Implementation	Site
IEEE	Guide for Information Technology - Software Reuse -Concept of Operations for Networks of	1430	1996	Development	Implementation	Site

Fig. 6 The retrieval result

retrieval system. By machine learning, the system records the search times and results for each user for each search. Statistically, the system handles all the raw data on search times and results searching for relationships. Finally, according to the database schema design, each successive retrieval becomes more efficient.

As shown in the previous result, the relationship among these standard documents may be illustrated from the perspective of groups and project processes category via the standard retrieval system.

6. Conclusion

In literature, we found that the field of data mining is rich in research and application development opportunities. Current data mining approaches are enhanced through cutting edge research from statistics, machine learning, and visualization and database management.

We have presented our initial work using data mining techniques to design a standard retrieval system via the data mining approach. We propose a retrieval algorithm using the three technologies of data mining, which are machine learning, statistics and database management, applied to this system to speed up the search time and increase the fitness. Accordingly, software development quality can be improved by shorter of retrieval time and more meaningful results.

Our experience shows that data mining can be used to produce a logical decomposition of a software system. Data mining offers the advantage that it can identify data cohesive subsystems without any knowledge of the subject system. Moreover, data mining is capable of producing meaningful results regardless of the size of the database. These properties of data mining make this approach especially appropriate to the analysis of large undocumented software systems. Furthermore, data mining in a distributed environment is a natural extension of the research considering the decentralized and collaborative nature of the work and data source locations.

Acknowledgment

This research is partially supported by the National Science Council: Grant NSC: 90-2213-E-029-005: A mobile agent for testing distributed applications. The authors of this paper would also like to thank Mr. Chi-Chang Chen and Mr. Jay-Len Ho who implemented the prototyping documentation-standards retrieval system.

References

[1] Carlos Montes de oca and Doris L. Carver (1998), "Identification of Data Cohesive Subsystems Using Data Mining Techniques," Proceedings of the International Conference on Software Maintenance.

[2] Chang, Wen-Kui and Tzu-Po Wang (1999), "Performance Modeling of Client/Server Distributed Architectures Applications," the Eleventh Workshop on Object-Oriented Technology and Applications, pp.633-642.

[3] Chang, Wen-Kui and Tzu-Po Wang (2000), "Issue of quality assurance on e-business," The 5th Annual International Conference on Industrial Engineering Theory, Applications and Practice-IE.

[4] Chen, Ming-Syan, Jong Soo Park, Yu, P.S. (1996), "Data mining for path traversal patterns in a web environment," Proceedings of the 16th International Conference on Distributed Computing Systems (ICDCS).

[5] Chen, Ming-Syan, Jiawei Han and Philip S. Yu (1996), "Data Mining: An Overview from a Database Perspective," Vol. 8, No. 6, pp. 866-883.

[6] Elmasri and Navathe (2001), Fundam entals of Database Systems, 3rd version, Addison Wesley.

[7] Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (1996), "Advances in Knowledge Discovery and Data Mining," AAAI Press, Menlo Park, CA.

[8] Heikki Mannila (1996), "Data mining: machine learning, statistics, and databases," Proceedings of the 8th International Conference on Scientific and Statistical Database Management (SSDBM).

[9] Ian Sommerville (2001), Software Engineering, 6th edition, Addison Wesley.

[10] Lam, W.; Segre, A.M. (1997), "Distributed data mining of probabilistic knowledge," Proceedings of the 17th International Conference on Distributed Computing Systems (ICDCS).

[11] Lowry, D.D.; Lowry, M.R. (1995), "Legal issues in knowledge-based software engineering," Proceedings of the 10th Knowledge-Based Software Engineering Conference (KBSE).

[12] Singh, R. (1995), "Harmonization of software engineering and system engineering standards," Proceedings of the 2nd IEEE Software Engineering Standards Symposium.