Knowledge Management System for Failure Analysis in Hard Disk Using Case-based Reasoning

Parinya Wichawong and Prabhas Chongstitvatana
Software Engineering Program
Department of Computer Engineering, Faculty of Engineering
Chulalongkorn University, Bangkok, Thailand
Parinya.Wic@student.chula.ac.th, Prabhas.C@chula.ac.th

***Abstract*— Hard disk failure is a serious problem in term of product quality and credibility to customers. All hard disk drive companies need to be aware and address how to get rid of failure and prevent the repeat of the problem in their products. The quality of failure analysis process depends on the person who has most experience. It would not be so efficient if the company has no experienced person to perform the analysis. A knowledge management system can store the knowledge of experienced engineers. It can help new engineers to learn the craft. It would reduce a knowledge gap issues and bring up efficiency for failure solving process. This paper presents a design and implementation of knowledge management system for failure analysis in hard disk with case-based reasoning. The existing cases are stored and a new case can be compared to the existing one in order to retrieve the relevant existing knowledge to help the analysis. Once the new case is solved, it can be stored to aid the future cases. A prototype of the system has been implemented and the assessment of user satisfaction shows that it can improve the failure analysis process effectively.**

Keywords—knowledge management; case-based reasoning; failure analysis; root cause analysis; vector space model

I. INTRODUCTION

Failure is a serious problem for the hard disk drive companies. They need to fix the root cause of failure in order to prevent repeating of the same problem in the future. Failure analysis is a process to identify the cause of the failure. The failure analysis requires knowledge from person who has most experience. However there are knowledge gap between expert and new engineer. It would be a problem if the company has no experienced person for analyzing the failure. Reducing the knowledge gap issues and the failure analysis process efficiency will increased.

In order to store the accumulated experience of the senior engineers in failure analysis, knowledge management can be used. Knowledge management [1] is a process of identifying, capturing, evaluating, retrieving, sharing, and effectively using the knowledge in the organization. Using knowledge management will reduce a knowledge gap issues and bring up efficiency for failure analysis process.

The problem solving process in failure analysis of the hard disk can be modelled by case-based reasoning. Case-based reasoning (CBR) is a generic methodology for building the knowledge-based systems for solving the problem on specific tasks. CBR solves the new problems by adapting the successful solutions from the similar problems in the past, then store the solution as a new case.

This paper presents a knowledge management system for failure analysis of hard disk that employs case-based reasoning. The information retrieval techniques including vector space model and cosine similarity are applied. The eight disciplines problem solving is used to design the document template. The system design and development process are presented. Finally, the assessment of the prototype system is discussed.

Section II discusses related theory. Section III explains our methodology. Section IV describes the details of the proposed system. Section V shows the system assessment. Finally, the conclusion and future work are discussed in Section VI.

II. RELATED THEORIES

The related theories are presented in this part including knowledge management, problem-solving and case-based reasoning, information storage and retrieval, vector space model and cosine similarity, and eight disciplines problem solving.

A. Knowledge management

Knowledge is a fact or condition of knowing something with familiarity gained through experience or association. There are two kinds of knowledge namely explicit knowledge and tacit knowledge [1] [2]. Explicit knowledge is the knowledge that set out in tangible form. It is articulated knowledge which can be stored in certain media such as files, databases, documents, emails, or software codes. Tacit knowledge is the knowledge that is difficult to transfer or access. This knowledge is obtained by internal individual processes like experience or individual talents. Therefore, the tacit knowledge cannot be managed and taught in the same manner as explicit knowledge.

Knowledge management (KM) is a method that simplifies the process of sharing, distributing, creating, capturing and understanding of a company's knowledge. These assets may include databases, documents, procedures, and experience in individual workers. KM refers to multi-approach for achieving organizational objectives by making the best use of knowledge. Knowledge management typically focus on the organizational objectives such as improved performance, innovation, sharing

of lessons learned, integration, and continuous improvement of the organization.

In this work, we proposed the knowledge management system that store explicit knowledge in failure analysis process. The explicit knowledge is transformed from tacit knowledge by the engineers who perform the analysis. They will record their tacit knowledge into documents and store them into the system.

B. Problem-solving and case-based reasoning

Problem-solving is the process to resolve various difficulties. It consists of using generic or ad hoc methods for finding solutions [3]. When products or processes fail, corrective actions and preventive actions can be taken to fix the root cause and prevent failures in the future. The past experience is useful for enhancing the problem-solving process.

Case-based reasoning is a generic method for building the knowledge-based systems that is capable of solving problems on specific tasks. CBR solves a new problem by adapting the previous successful solutions of the similar problems [4]. CBR systems can acquire new knowledge as a case. Sometime the solutions of similar problems might be directly applicable to current problem. Usually the adaptation is required when problems are not exactly the same as the past cases. The adaptation is based upon the differences between current problem and similar problem in past. Once the solution to the new problem has been verified, the system will define and store it into the memory as a new case.

The CBR life cycle is shown in Fig. 1. There are four stages of the CBR. The first stage is the retrieve stage [5]. It starts from measuring the similarity of the current problem and the previous problems that stored in the memory called case-based with their solutions. Then retrieving one or more similar cases. Next, the reuse stage, in this stage the retrieved solution is adapted to solve the new problem. After that, the revision stage, revise the new proposed solution, including the information or knowledge gained from solving the new problem. Finally, the retain stage, store the new case into the system memory.

We proposed the knowledge management system which apply CBR concept to help users to solve their problems in failure analysis process by adapting the previous successful solutions of the similar failures.

C. Information storage and retrieval

Information retrieval (IR) accesses to information items in natural language such as documents, Web pages, structured and unstructured records [7]. An information retrieval system contains two parts. The first part is the indexing part for creating the term index to represent the individual document. The second part is the searching part for retrieving the information items that are relevant to user query.

IR process begins when a user enters a query into the system. After that, the user's query is matched against each

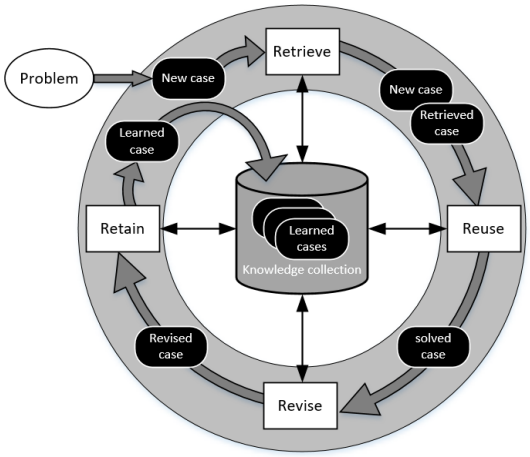

Fig. 1. Case-based reasoning life cycle. [6]

information items and similarity score is computed in this step. The information items are ranked by similarity scores. Finally, the relevant information is displayed to the user.

In this work, the information retrieval technique is applied in the retrieve stage and the retain stage. The indexing technique is used in the retain stage of CBR in order to create the index that represents the case. Searching technique is used in the retrieve stage of CBR in order to retrieve the case that relevant to the new failure.

D. Vector space model and cosine similarity

Vector space model [8] is an approach which is based on vector formed by each word contained in the document or the query. Document is a vector which has direction in n-dimensional space and magnitude that considers as the frequency of each words.

The relevance of a document to a query is based on the similarity between the document vector and the query vector. Cosine formula is used to measure the similarity by measuring the cosine of the angle between their corresponding term vectors [9]. The degree of similarity is higher if the cosine angle between two vectors is close to 1.

E. Eight disciplines problem solving

Eight disciplines problem solving (8D) [10] is a problem-solving method that is highly effective and is popular among manufacturers. 8D comprised of eight stages. First, Form a team, establish a team including members from many areas in the organization. Second, Describe the Problem, specify the problem properly. Third, Interim containment, implement actions to isolate the problem from customers. Fourth, Root cause analysis, identify root cause and explain why the problem has occurred. Fifth, Verify permanent corrective action, establish a long term solution and verify that it will actually solve the problem. Sixth, Implement a permanent corrective action, fix the problem at the root cause. Seventh, Prevent recurrence, define and implement the preventive action to prevent recurrence of current problem and similar problems. Finally, Team celebration, celebrate successful completion and recognizing both team and individual efforts.

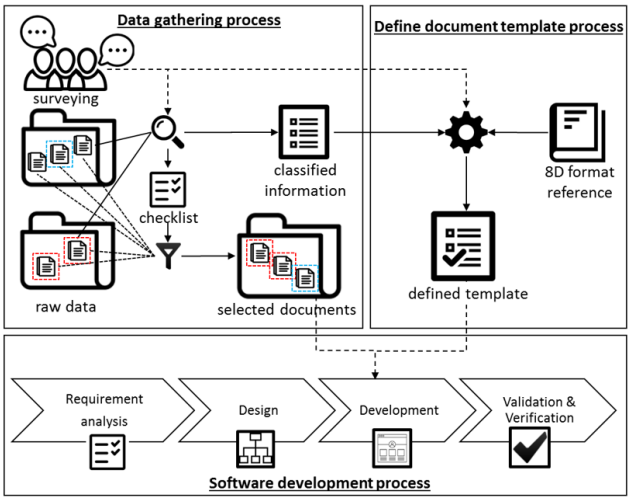

Fig. 2. System development methodology.

III. METHODOLOGY

There are three processes to develop the system starting from data gathering process, define document template process, and software development process as shown in Fig. 2. The explanation of data gathering process and define document template process is described in details, while the proposed system constructed by software development process is described in the next section.

A. Data gathering

There are two types of data. The first type of data is the existing raw data in the failure analysis process. This data are contained within the documents in the process such as presentations, emails, and weekly reports. The second type of data is user requirement data. This data are gathered for understanding the user's standpoint on each data. In this work, the data gathering process is separated into three stages.

First stage, the documents in the current failure analysis process are collected to classify the types of data that is contained in document content. Starting from identifying types of failure in current failure analysis process. Then, select one or two representative for representing the data of each group. Then, mapping the data type between each selected representative documents. In this step, the common data and the specific data that contained in the current process are collected. Then, analyze and define which data are required in the proposed system. Finally, verify the classified data against user requirement.

Second stage, select the documents as the initial data in the proposed system. Starting from analyzing the classified data from previous stage to define the checklist. Then, select the documents by using the checklist as defined in previous step. Finally, reformatting the selected documents. The format of the document is contained by the classified data from the previous stage.

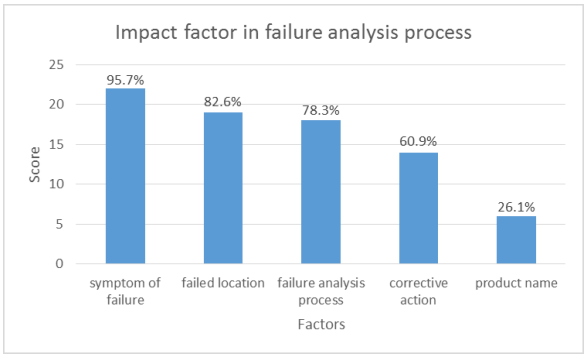

Fig. 3. Impact factors in failure analysis process.

Failure name:	Short description of the failure.		
Case ID:	this id is generated by system.	**Reference case:**	the case-based.
Product name:	The name of product.	**Firmware version:**	version of firmware.
Failed location:	Location of failure.		
Background: (symptom)	- what - where - when - why - how - how many		
Containment action:	Define containment actions to isolate the problem from the customer - short term solution		
Root cause analysis:	Identify all applicable causes that could explain why the problem has occurred. Also identify why the problem was not noticed at the time it occurred. All causes shall be verified or proved. - Working procedured to find the root cause of failure.		
Corrective action:	Define and implement the best corrective actions. - Long term solution - Fix the root cause		
Preventive action:	Define procedures to prevent recurrence of the failure and all similar failure.		

Fig. 4. Defined template of document for failure analysis.

Third stage, understand the user's standpoint on each classified data. Data gathering are executed using survey technique. The short questionnaire on topic of failure analysis is issued to the engineering team. The analysis of their responses are used to validate and improve the proposed system.

There are three outcomes generated by this process. First, the list of classified data. Second, the set of selected documents. Last, the set of impact factors of data in the current process (Fig. 3).

B. Define document template

Document template serves as a starting point for creating new document with pre-formatted. There are several advantages of using templates. First is the consistency of the document. Every documents are constructed by a template so they will have the same structure. Second, it reduces an error because using template will enforce users to fill in all critical elements in order to complete their task. Moreover, the good template will increase speed of filling in the information since users can simply change the desired information instead of developing a new document every time.

The process to develop the template are separated to three steps. First, mapping the classified data from previous process with 8D format. Then, analyze and design the document

template. In this step, the impact factors from previous process are applied. Next, define a template. Finally, verify the document template against system requirement. The defined template is shown in Fig. 4.

IV. Proposed System

The proposed system contained three domains as shown in Fig. 5. First, user domain presented the user workflow. The case-based reasoning concept is applied in this domain with supporting functions provided by the application domain. Next, application domain is a part of system that provide interface to the user. Finally, system domain is a core function of the system. There are two main functions including in the system domain: the indexing function and searching function.

A. User domain

The user activities consist of retrieve, reuse and revise, and retain. These activities follow the CBR concept (Fig. 6). The activity begins when a user has the new failure need to be analyzed. The user would search using the symptom of the new failure as a query. The system will suggest cases that are similar to the query. In this step, vector space model and cosine similarity are used.

Next, the user reviews each case. If the retrieved case is relevant to the new failure, the user can duplicate the relevant case and uses it as the reference of failure analysis procedure. In the other hand, if the retrieved case does not match the new failure, the user can create the new case document without a reference. Then, the new failure is analyzed by adapting the knowledge in the reference case. After the problem has been solved, the user saves the new case document into the system. The system will index and store it into the case-bases library and the index term table. Finally, the new case is stored in the system and is ready for searching in the future.

B. Application Domain

The application domain is an interface that is connected to the user domain. This domain is the interface between the user domain and the system domain. The system workflow is controlled by this domain. There are several functions in this domain to support the CBR concept such as retrieve, suggest the cases for reusing, serve an information for adapting the knowledge in the reference case to solve the new problem, create a new case document, and retain a new case into the system.

C. System Domain

The system domain is the main function responsible for computational tasks. There are four main sections in the system domain including storage, retrieval, Lucene.NET, and data layer as shown in Fig. 7. Storage is responsible for document processing in order to store the case into the knowledge base library and update the case information in the relational database. There are two modules in this section.

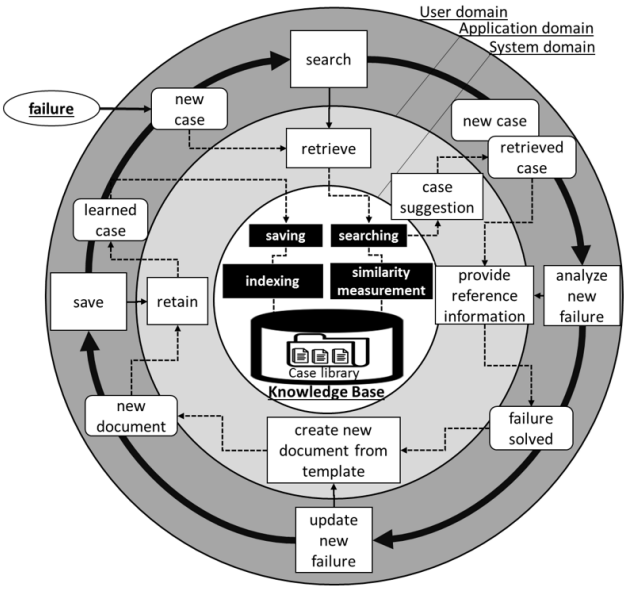

Fig. 5. System overview.

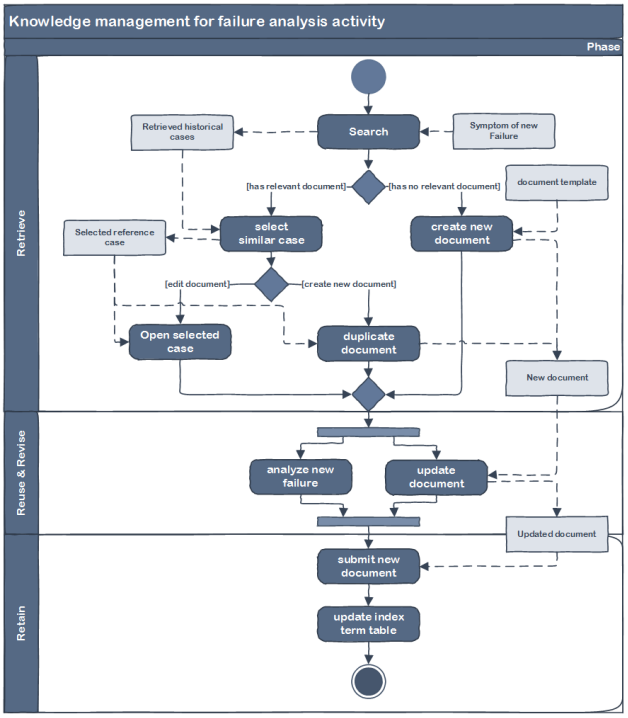

Fig. 6. Activity diagram to explain the user workflow.

The first module is the content extraction module. This module is the interface that receives the new case document from the application domain. It extracts the content according to the document template structure. The next module, the document management module, is responsible for creating the new case and save it into the knowledge base library then update the case status in the database. This module also selects the content to send to Lucene.NET for indexing.

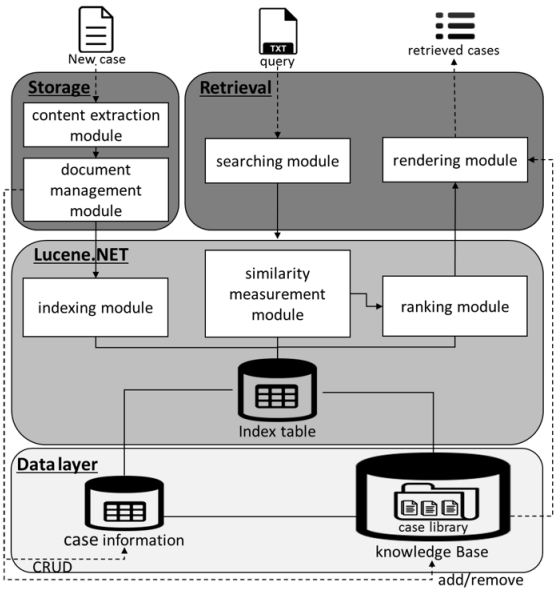

Fig. 7. System domain architecture.

Retrieval is responsible for searching the relevant cases and returning the set of relevant cases back to the application domain ordering by the similarity score. There are two modules in this part. The first module is the searching module. It is an interface for receiving the query from the application domain and pass it to the Lucene.NET. The second module is the rendering module. It receives the search result from Lucene.NET and generates the content of each case then sends the result back to the application domain.

Lunene.NET is C# library that provide an infrastructure for the information retrieval system. In this work, we separate the function of Lucene.NET to three modules. The first module is the indexing module. It is responsible for document processing. It creates the term index of the document and stores it into the index table. The second module is the similarity measurement module. This module is responsible for searching activity. It computes the similarity between query and every document in the index table using vector space and cosine similarity. In the third module, the search result would be ranked and ordered by the similarity score and pass them to the rendering module.

Data layer is responsible for managing the data. There are two types of data in the system. The first type is knowledge base. It is a library that stored the case documents. The documents stored in the knowledge base are in XML file format. The second type of data is the case information table. It is a relational database that stored the case information such as case id, owner, date create, date modify, status, and reference case.

D. System Implementation

A prototype has been developed to demonstrate the proposed knowledge management system for failure analysis. Several software tools are used including the Microsoft Visual Studio 2015, ASP.NET, MySQL database, and Lucene.net library. The knowledge management system is integrated with the case-based reasoning to provide more effective knowledge support for analyzing the failure. The interface of the knowledge management system includes the search engine and failure analysis environment with supporting knowledge. Once a worker save a new document after completed the failure analysis process, the system will create a new case and store it into the system. This stored cases will be a good reference knowledge for the future.

V. SYSTEM ASSESSMENT

The assessment consists of two parts: assessing user satisfaction and assessing the searching effectiveness.

A. User satisfaction assessment

The user satisfaction is assessed using the questionnaire. The subject for assessment is divided into four subjects. The first is learning organization. This subject measures the user expectation in term of knowledge exchangeability in the organization. The second is the improvement of the process. This subject is intended to assess the suitability of proposed system to failure analysis process. The third is the saving of time. This subject measures the user expectation for reducing the time to analyze the failure by using the proposed system. The fourth is the improvement of correctness. This subject measures the user expectation in term of accuracy and efficiency improvement by using the system.

The score of each subject is ranged between 1 to 5. 5 is a highest satisfaction score. There are two sections in the assessment document. The first section presents the prototype system including the workflow and example for using the system. In the second section, four questions are presented.

The questionnaire are distributed to 92 persons that work on failure analysis area in one hard disk drive company. Totally 23 from 92 responded. 65% of the responder has more than 5 years' experience on failure analysis, 23% has experience between 4 to 5 years, and 22% less than 3 years. The assessment result is shown in Table I. The learning organization shows highest score at 4.52, improve effectiveness 4.39, reduce time 4.39, and improve correctness 4.22.

From Table I the learning organization shows highest score at 4.52. The assessment score of every subject is in high level, so we can conclude this system is successful in term of user satisfaction.

B. Searching effectiveness assessment

This assessment measures whether the users are getting relevant documents at the top of ranking or not. The retrieval evaluation methods are precision at 5 (P@5) and precision at 10 (P@10). The idea is that the higher number of relevant documents at the top of ranking should have more positive score. The P@5 and P@10 measure the precision when 5 or 10 documents have been seen from the search result.

There are a several steps for this assessment process. First, ten queries are prepared for searching. Then, analyze each individual item from the search result.

TABLE I. USER ASSESSMENT RESULT

subject	Score (5 = best , 1 = worst)					
	5	4	3	2	1	average
learning organization	14	7	2	0	0	4.52
improve effectiveness	11	10	2	0	0	4.39
reduce time	10	12	1	0	0	4.39
improve correctness	7	14	2	0	0	4.22

TABLE II. CASE RETRIEVAL ASSESSMENT RESULT

	ranks	query number									
		1	2	3	4	5	6	7	8	9	10
search result (relevant = 1)	1	1	1	1	0	1	1	1	1	1	1
	2	1	1	0	1	1	1	1	1	1	1
	3	1	1	0	1	1	1	1	1	0	1
	4	1	0	1	0	0	0	1	0	0	1
	5	1	0	0	0	0	0	0	0	1	1
	6	1	0	0	1	1	0	0	1	0	0
	7	1	0	1	0	0	0	0	1	0	0
	8	1	0	0	1	0	0	0	0	0	0
	9	1	0	0	1	0	0	0	0	0	0
	10	1	0	0	0	0	0	0	0	0	0
	11	1	0	0	0	0	0	0	1	0	0
	12	0	0	0	1	0	0	0	0	0	0
	13	0	1	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0
P@5		1.0	0.6	0.4	0.4	0.6	0.6	0.8	0.6	0.6	1.0
P@10		1.0	0.3	0.3	0.5	0.4	0.3	0.4	0.5	0.3	0.5

Give 1 point if it is relevant to the query, while give 0 point otherwise. After completed all queries the P@5 and P@10 are calculated as shown in Table II.

From Table II the P@5 is 66%, and P@10 is 45%. That means the users are getting relevant document at the top 5 more than a half and get 45% relevant at the top 10 by average. We can conclude that the searching effectiveness of the proposed system is acceptable.

VI. CONCLUSION AND FUTURE WORK

This paper presents a knowledge management system for failure analysis in hard disk with case-based reasoning. Document template is designed by using the real data in failure analysis process. The prototype system is implemented. Based on assessment result it can be concluded that the system has high score in term of user satisfaction, and the searching effectiveness is acceptable. In summary, this system is successful. However, the proposed system presented in this paper has a limitation. The cases in this system are represented by using the text only. The actual picture of the cases is a useful information for helping user to analyze the failure but it is not included in the database. In the future, variety of data should be presented in the database.

REFERENCES

[1] F. O. Bjornson and T. Dingsoyr, "Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used," *Inf. Softw. Technol.*, vol. 50, no. 11, pp. 1055–1068, 2008.

[2] S. Vasanthapriyan, J. Tian, and J. Xiang, "A Survey on Knowledge Management in Software Engineering," *2015 IEEE Int. Conf. Softw. Qual. Reliab. Secur. - Companion*, pp. 237–244, 2015.

[3] D. R. Liu and C. K. Ke, "Knowledge support for problem-solving in a production process: A hybrid of knowledge discovery and case-based reasoning," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 147–161, 2007.

[4] Z. Teng, J. Chen, and H. Xia, "Study on case-based reasoning-inspired approaches to machine-learning," *Proc. - 2015 Int. Conf. Intell. Transp. Big Data Smart City, ICITBS 2015*, pp. 760–763, 2016.

[5] C. Jian, T. Zhe, and L. Zhenxing, "A Review and Analysis of Case-Based Reasoning Research," *2015 Int. Conf. Intell. Transp. Big Data Smart City*, pp. 51–55, 2015.

[6] K. Chantanaporn and C. Numthong, "Development of Failure Analysis Case-Based Expert System for Computer Network Equipment Products," 2012.

[7] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval: The Concepts and Technology behind Search," *Inf. Retr. Boston.*, vol. 82, p. 944, 2011.

[8] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[9] S. Zhu, J. Wu, H. Xiong, and G. Xia, "Scaling up top-K cosine similarity search," *Data Knowl. Eng.*, vol. 70, no. 1, pp. 60–83, 2011.

[10] "Eight Disciplines Problem Solving - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Eight_Disciplines_Problem_Solving.