Expert Systems with Applications 36 (2009) 4753–4764

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A document classification and retrieval system for R&D in semiconductor industry – A hybrid approach

Shui-Shun Lin *

Department of Business Administration, National Chin-Yi University of Technology, Taiwan, ROC

ARTICLE INFO

Keywords:
Document classification and retrieval
Vector space model
Document management system

ABSTRACT

In this paper, a hybrid methodology with a vector space model (VSM) and process-oriented attributes for document management is proposed. The VSM is fine-tuned for classifying documents generated during R&D processes. The document correlation values are computed with the VSM for efficient retrieval. Only documents with high correlation values are presented to meet the specific retrieval purpose, which results in efficient and effective document retrieval. We further design a document classification and retrieval prototype system. The prototype is implemented to facilitate R&D document management in semiconductor industries.

1. Introduction

A complete document management system is required for enterprises to deal with documents in a controlled way. However, traditional document management and classification methods cannot meet the demanding requirements pertaining to the growing number of electronic documents; therefore, automated technology is a necessity for better document classification and retrieval.

In regards to traditional document classification, it is firstly required to discover the specific classification characteristics, set up models for classification, and then identify their validity and suitability. For primitive data without classification characteristics, full-text retrieval technology is used to ensure individual comparison of all documents with inquiry keywords, thereby discovering intersections or matches. If document classification concepts are incorporated into the retrieval, and the special characteristics of each document are defined, only highly correlated documents are retrieved to ensure both a higher retrieval speed and improved accuracy.

R&D activities have a great impact upon the entire manufacturing process of the semiconductor industry, since the manufacturing process depends greatly upon the product design (Garraffo, 2000). To accumulate knowledge and experience from the documents generated by R&D processes in the semiconductor industry, literal or graphic documents are normally stored in the form of electronic documents. Since keywords are the only criterion used for searching the documents, the amount of retrieved documents is generally large, which increases difficulty in locating the specific documents. Thus, a suitable document classification structure is required to assist new product developers in terms of document retrieval. Moreover, a prototype system should be developed to provide assistance for document classification and reduce the influence of human factors, thereby improving the efficiency of document classification and retrieval. Specifically, the objectives of this research study are as follows:

1. Construct a document classification structure, and define R&D knowledge dimensions.
2. Investigate and establish proper models to improve document classification and retrieval efficiency.
3. Develop a document classification and retrieval prototype system for semiconductor industry R&D activities.

2. Literature review

2.1. Document classification and management

Document classification indicates that unclassified documents are assorted to predefined categories according to a specified principle. According to most automatic document classification methods, an unclassified document is often sorted into one specific category, which is called single document classification (SDC). Que (2000) stated that a document may be of concern to different topics or predefined categories are not fully independent from one other. Therefore, SDC is often not helpful. In certain cases, it is necessary to sort an unclassified document into several categories, which is referred to as multiple document classification (MDC). Fan (2002) employed MDC and developed knowledge management models for the R&D process of TFT–LCD products. However, Fan's approach did not take the keywords into consideration.

* Tel.: +886 4 23924505x7774; fax: +886 4 23929584.
E-mail addresses: sslin@ncut.edu.tw, yes.ebada@gmail.com

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.024

According to Que (2000), the keywords of intact document classification are collected from databases with classification characteristics. Therefore, a great amount of training and test data is often required to build correct models. Chou (2002) pointed out that the benefits of expert-defined keywords lie in the sense that classification structures conform to the ideas of organizational members. Wei, Yang, Hsiao, and Cheng (2006) focused on document clustering by exploiting users' document grouping preferences, and proposed a document clustering technique that combines preference and content-based approaches.

Additionally, it is necessary to avoid the influence of syntactical features and tenses of the English language when identifying the keywords. Therefore, stemming is required prior to document classification. Paice (1990) proposed a series of stemming regulations. Zhuang (1999) also stated that etymons should be located among the nouns and verbs in a document before keywords are found and document classified.

There are a variety of document classification methods, such as k-Nearest-neighbor (kNN), decision tree, Naïve Bayes, neural network, fuzzy correlation, genetic algorithms and support vector machines (SVM), and the vector space model (VSM). The major document classification methods are listed in Table 1.

kNN is used to store a certain amount of classification data, and to test the degree of similarity between documents and k training data, thereby determining the category of test documents (Tam, Santoso, & Setiono, 2002). Bang, Yang, and Yang (2006) proposed a new algorithm incorporating the relationship of concept-based thesauri into document categorization using a kNN classifier. Decision tree is also widely applied to document classification. The benefit is that it enables conversion into interpretable IF–THEN, and features rapid classification. Based on Bayes principle, Naïve Bayes is used to calculate the characteristics of a new document using keywords and joint probability of document categories (Que, 2000).

In recent years, neural network has been applied in document classification systems to improve efficiency. Deng and Wu (2001) investigated a new document classification method, growing cell structures (GCS), using neural network and took semantic micro-characteristics as the classification rule for newsletter classification. Lin (2003) established an electronic document management system using neural network-based automatic document classification technology. Trappey, Hsu, Trappey, and Lin (2006) developed a document classification and search methodology based on neural network technology that helps companies manage patent documents more effectively. This classification process begins by extracting key phrases from the document set by means of automatic text processing, and then determining the significance of key phrases according to their frequency in the text.

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification (Que, 2000). Genetic algorithms aim to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. Zhuang (1999) applied genetic algorithms into a classification of master theses and doctoral dissertations. However, Zhuang's results obtained using genetic algorithms easily located local optimum, rather than global optimal solutions. In addition, genetic algorithms require extensive calculation, leading to poorer efficiency during document classification.

Horng and Yeh (2000) proposed a novel approach to automatically retrieve keywords and then use genetic algorithms to adapt keyword weights. The proposed approach can retrieve any type of keyword, such as technical keywords or peoples' names. Effectiveness of the proposed approach is demonstrated by comparing how effective the keywords found by both this approach and the PAT-tree-based approach are. This comparison reveals that the authors' keyword retrieval approach is as accurate as the PAT-tree-based approach, yet their approach is faster and uses less memory. The study then applies genetic algorithms to tune the weight of retrieved keywords. Moreover, several documents obtained from web sites are tested, and the experimental results are compared with those of other approaches, indicating that the proposed approach is highly promising for applications.

Chen and Hsieh (2006) proposed a web page classification based on a support vector machine using a weighted vote schema for various features in order to effectively classify web pages, thus solving the synonymous keyword problem. The system uses both latent semantic analysis and the web page feature selection training and recognition through the support vector machine model. Latent semantic analysis is used to find the semantic relationships between keywords and between documents. The latent semantic analysis method projects terms and a document into a vector space to find latent information in the document. At the same time, text features are also extracted from the web page content. Through these text features, web pages are classified into suitable categories. These two features are sent to the SVM for training and testing, respectively. Based on the output of the support vector machine, a voting schema is used to determine the category of the web pages. Wang and Chiang (2007) presented a text categorization system to solve multi-class categorization problems. They specifically proposed a classifier to implement within a multi-class classification system. The performance of the classifier is evaluated by the macro-average performance index. The results of that empirical study show that the proposed method outperforms other multi-class text categorization schemes. Hao, Chiang, and Tu (2007) proposed a novel hierarchical classification method that generalizes support vector machine learning based on the results of the support vector clustering method with class hierarchy. Compared to previous non-hierarchical SVM classifier and famous document categorization systems, the proposed hierarchical SVM classification shows improvement in classification accuracy. However, the fact is that SVM ignore inter-class relationships. Also, it has been observed that obtaining a classifier that can discriminate between the two groups of classes is much easier than distinguishing simultaneously among all classes present.

Song, Lau, Bruza, Wong, and Chen (2007) developed effective information agents to autonomously classify and filter incoming

Table 1
Major classification methods

Document classification methods	Characteristics
k-Nearest-neighbor (kNN)	Good classification accuracy in the presence of little training data; otherwise, the calculation speed slowed down dramatically
Decision tree	Classification rule is interpretable, but classification accuracy is lowered in the presence of excessive categories
Na Bayes	Easy calculation, rapid classification, independent properties, continuous property values difficult to deal with
Neural network	Numerous training and test data available for correct modeling
Fuzzy correlation	Capable of dealing with fuzzy information or incomplete data; property value can be converted into fuzzy set without requiring complex calculation
Genetic algorithms	Good accuracy of classified results; calculated results are possibly local optimizations, rather than global optimizations
Support vector machines (SVM)	Better performance in document classification; ignore inter-class relationships in a document.
Vector space model (VSM)	Easy calculation and rapid classification

electronic data on behalf of human users. The information agents are innovative because they can quickly classify electronic documents based solely on the short titles of these documents. Moreover, supervised learning is not required to train the classification models of these agents. Document classification is based on information inferences conducted over a highly dimensional semantic information space. A belief revision mechanism continuously maintains a set of user preferred information categories and filters documents with respect to these categories. The authors claimed that their document classification and filtering mechanism outperforms the SVM model.

2.2. Vector space model

Salton and McGill (1983) established a document classification mechanism using the VSM, and explored the issue of multiple document classification. The model should convert every predefined category and unclassified document into space vector, and then determine the category of every unclassified document in a comparative way. The VSM is divided into the global vector space model (GVSM) (Benkhalifa, Bensaid, & Mouradi, 1999) and the local vector space model (LVSM) (Michael & Browne, 1999). The GVSM should define a unified vector space with a reference of individual keywords. The disadvantage is that the space dimension must be big enough to fully represent the characteristics of individual categories. In the absence of a unified vector space, the LVSM provides keywords for every category and defines a vector space individually, so the vector space of every category can fully reflect the characteristics of individual categories. Li (2003) stated that the VSM can enable a document to be represented by a multi-dimensional vector, of which every dimension represents different characteristics of a document, and as for word strings or documents for search, classification could be represented by a multi-dimensional vector.

According to VSM, a cosine coefficient is used to calculate the degree of similarity between the two documents. $X = (x_1, x_2, \dots, x_t)$ represents the vector of a document, $Y = (y_1, y_2, \dots, y_t)$ denotes the vector of the classification rule, where t is vector dimension of X and Y. The cosine coefficient is expressed by

$$\cos(X, Y) = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} x_i^2} * \sqrt{\sum_{i=1}^{t} y_i^2}} \qquad (1)$$

If the included angle of the two vectors is 0, the two vectors are in parallel, while the cosine coefficient is 1, showing a high degree of similarity. Otherwise, the two vectors are not highly similar if the cosine coefficient is lower.

Taghva, Borsack, and Condit (1995) investigated the performance of VSM in the presence of optical character recognition (OCR) errors and observed that cosine normalization plays a considerable role in the disparity seen between the documents. Zhang and Rasmussen (2001) compared two distinct similarity measures in a document vector space – the distance-based and angle-based similarity measures – and a newly developed similarity measure based upon both the distance and angle strengths of two compared objects was presented. The concept of the iso-extent contour, which facilitates understanding of the nature of the newly developed similarity measure, was also introduced. Tai, Ren, and Kita (2002) proposed a method to improve the retrieval performance of the VSM in part by utilizing user-supplied information regarding the documents relevant to the query in question. In addition to relevant user feedback information, original document similarities are incorporated into the retrieval model, built by using a sequence of linear transformations. Highly dimensional and sparse vectors are then reduced by singular value decomposition and transformed into a low-dimensional vector space, namely the space representing the latent semantic meanings of words. This method provided an approach that makes it possible to preserve user-supplied long term relevance information in the system in order to make use of it later. Kalczynski and Chou (2005) extended the VSM to the temporal document retrieval model, and introduced a novel approach to presenting temporal expressions. A user study was conducted to measure the degree of uncertainty for selected temporal expressions, and a method for presenting uncertainty based on fuzzy numbers was proposed. However, their approach did not adequately deal with R&D documents, where temporal manner is not significant.

2.3. Process-oriented structure

Jablonski, Horn, and Schlundt (2001) developed a system focusing on both the information structure and storage mechanism, whereby knowledge is stored into the system in the form of files, databases or other media. Since knowledge requires description (e.g. keywords) without a division into a basic knowledge element, it can be represented and stored in the form of documents or media. Jablonski et al. (2001) termed all knowledge media as knowledge carriers. Thereby, knowledge, referred to as particles, is collected and stored in a knowledge base, which is represented by a three-tier tree structure.

Bae and Kim (2002) proposed an XML-based approach to workflow management systems (WFMS) for form document-processing on the web, and implemented a prototype system to demonstrate the usefulness of the proposed model, called a document-process association model. However, the document needs to be partitioned into several different units that associate with some particular work processes, which is not possible with the R&D documents that our research targeted.

As an extension to the knowledge structure of Jablonski et al. (2001), we proposed a document classification structure integrating vector space and process attributes, and defined the important knowledge dimension for R&D in the semiconductor industry. It provides support to users performing document classification and retrieval by establishing the correlation between knowledge carriers and dimensions.

3. Knowledge management structure

3.1. Problems of R&D document management

We performed a field investigation of two leading semiconductor enterprises (hereinafter referred to as company T and U) in Taiwan. The business scales of these two companies have ranked as number one and two in Taiwan for the past decade (Chian, 2004).

The present status and shortcomings of R&D document management of both the companies T and U are listed in Table 2, where we find that the major difficulty for document management is the lack of proper criteria and experts to precisely assign document attributes for efficient retrieval. The major pitfall for document retrieval is that too many documents are retrieved solely by keyword, which results in overwhelming amounts of document screening.

To overcome the current pitfalls of the document classification and retrieval operation in the semiconductor industry, we thoroughly defined document classification categories and keyword sets for the semiconductor industry, and proposed a hybrid classification structure and methods to facilitate R&D document classification and retrieval tasks. We analyzed and designed a prototype system to validate our methodology and to assist the document management process in the semiconductor industry. The rest of this paper addresses the hybrid methodology for document

Table 2
Shortcomings of R&D document management in semiconductor industry

Co.	Present status of document classification and retrieval	Shortcomings
T	1. Documents stored in PC categorized by file folder 2. Documents evaluated by professional technology committee (PTC) 3. Document retrieval by keywords and sequencing by the rating from PTC	1. A huge amount of matched data displayed; screening of document names only for graphic files, resulting in overwhelming amounts of browsing 2. Documents evaluated by a technology committee, which may affect document retrieval precision
U	1. Documents stored in PC categorized by arbitrarily defined file folders 2. Text document retrieval by keywords 3. Graphic document retrieval by predefined keywords	1. Retrieval of document names by keywords only, with the results not reflecting the practical demands 2. No evaluative criteria for document retrieval, leading to inconvenient document inquiry

Table 3
Classification of project process and document

Classification of project process	Relevant activities and related documents/reports*
Programming	1. Quality development (Evaluation report) 2. Specification preparation (Manufacturing specification document, development plan) 3. Design principle preparation (Design manual) 4. Test principle preparation (Test specifications)
Engineering tests	1. Engineering change (Engineering change notice) 2. New module setting (Setting manual, feasibility acknowledgment) 3. Experimental design (Experimental design process, laboratory report) 4. Failure model analysis (Failure model analysis report) 5. Electrical simulation test (Electrical simulation test report)
Manufacturing certification	1. Specification review (Specification review report) 2. Experimental results analysis (Analysis report of experimental results)

* Shown in parenthesis.

Table 4
Classification of technology and related terminology

Manufacturing process	Technology
Lamination	Oxidation, CVD, PVD
Formation	Resist, exposure, develop, etch
Mixing	Diffusion, clean, ion implantation
Heating	Thermal, radiation

classification and retrieval, and the issues regarding analyzing and designing an information system.

3.2. Process-oriented document classification

The knowledge dimensions of the semiconductor industry are defined as the process knowledge dimension, the technology knowledge dimension, and the equipment and material knowledge dimension. The implemented three-tier knowledge structure is illustrated in Fig. 1.

3.2.1. Process knowledge dimension

According to our market analysis and feasibility study, a project team for new product development is set up to analyze and evaluate existing know-how and resources, and convert customer requirements into design principles and specifications, while holding regular project meetings for the purposes of tracking and control. The project process is divided into programming, engineering tests and obtaining manufacturing certification. Table 3 lists the project processes and documents generated from each process.

3.2.2. Technology knowledge dimension

The technology knowledge dimension focuses on the knowledge classification of manufacturing technology. The four major manufacturing processes include lamination, formation, mixing and heating. The R&D activities are intended for innovation and improvements in terms of new technologies or materials. The four manufacturing processes are divided into 12 technologies as shown in Table 4.

3.2.3. Equipment and material knowledge dimension

Semiconductor R&D technology is often subject to possession of certain equipment or materials. Under such considerations, the equipment and material dimension is defined as a knowledge dimension for document classification, which includes the 16 items as listed in Table 5.

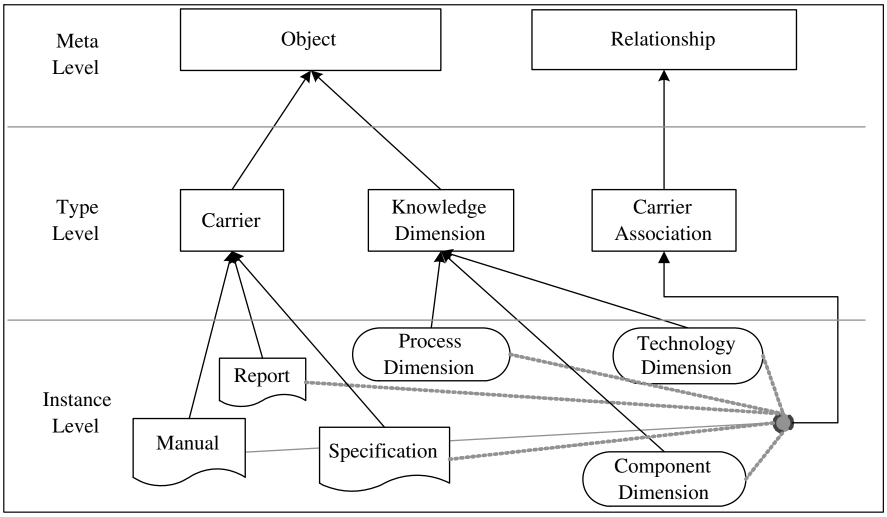


Fig. 1. Three-tier knowledge structure.

Table 5
Classification of equipment and material and related terminology

Manufacturing process	Classification of equipment and material
Oxidation	Oxidant, tube furnace
Formation	Resist, soft bake, exposure, develop, etch, strip
Mixing	Deposition source, cleaner, ion implantation
Lamination	Deposition system, deposition membrane, IMD, vacuum steam platter, sputter

3.3. Knowledge-related keywords analysis and stemming

The keywords for the classification and retrieval of electronic documents were then collected and listed. There were a total of 28 knowledge categories and 292 keywords defined such as dry oxidation, overshoot, turbulence, laminar flow, pattern shift, plug fill, virtual leak, thermal flow, diffraction, puddle development, incomplete etch, hydrophilic, acoustic stream, wafer charging, annealing, among others. The stemming process was introduced and performed. A wide range of words such as computer, computers, computing, compute, computational, computationally, computable, computes, computed, and computation that are etymologically related were treated with the stemming process proposed by Paice (1990).

3.4. Knowledge documents management methodology

3.4.1. Document classification

The proposed document classification process is defined in Fig. 2.

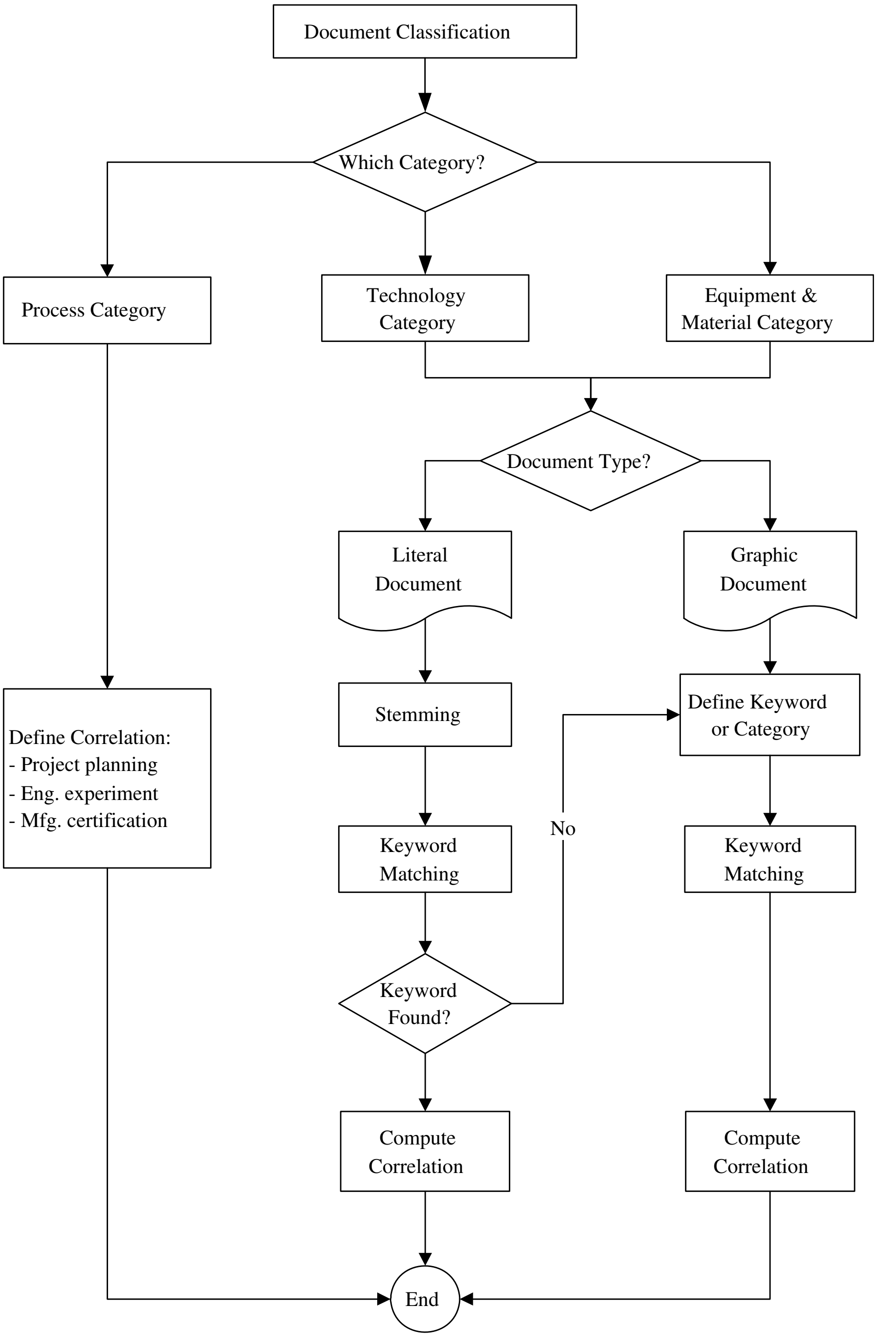


Fig. 2. Document classification process.

The R&D documents are classified into three knowledge dimensions, namely: the process dimension, the technology dimension and the equipment and material dimension. The process dimension includes the knowledge related to project processes. The project process dimension includes the processes of project planning, engineering experiments and manufacturing certifications. Each document in this dimension has one of four types of relationship, i.e. input, output, reference and none, to a specific process.

For technology dimension and equipment and material dimension documents, each electronic document is classified into literal and graphic documents. The literal documents are compared with keyword sets of various categories after stemming. Then the keyword matching process is carried out to define keywords. If no keywords can be constructed automatically, a document expert is required to set up keywords manually. As for the graphic documents, individual keywords are defined by the domain expert manually. The degree of correlation with various categories is then calculated to classify each document.

The detailed flow of document classification is as follows.

Let Y represent the technology dimension or equipment and material dimension keyword set, where $Y = (y_1, y_2, \ldots, y_t)$. If keyword weight is not considered and keyword i is present in specific documents, then $y_i = 1$; in the absence of the keyword i, $y_i = 0$.

Let X represent a keyword set $(x_1, x_2, \ldots, x_t)$ of a certain document; if $x_i = y_i$, then $x_i = 1$; if $x_i \neq y_i$, then $x_i = 0$, $i = 1, 2, \ldots, t$, where t is the total category number of the certain document.

- Define knowledge dimension and keyword set.
- Upload an electronic document to be classified, and select the knowledge dimension category.
- If process classification is selected, directly define three processes and four relationship items. Then the classification process is finished. If technology classification or equipment and material classification is selected, proceed with Step 4.
- Electronic documents are distinguished as literal or graphic documents. In the case of graphic documents, the categories and keywords are defined by the domain experts; in the case of literal documents, stemming is performed by the system, and the document contents are compared to see if there are keywords in the predefined keyword set. In the presence of category keywords, record the corresponding keyword set X_m, where m is the category of technology classification or equipment and material classification; otherwise, the categories and keywords are defined by domain experts.
- Calculate cosine values of every keyword set of the electronic documents using the VSM. If cos > 0, the documents are classified in the category; if cos = 0, the documents are not classified in the category.
- Continue with computations until all documents have been categorized.

To illustrate the computation process of the cosine value for classification, an example is given as follows.

There are 19 keywords in oxidation keyword set Y in the technology dimension, denoted by

$$Y = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1) \tag{2}$$

If a certain test design process includes the third and fourth keyword of this set, i.e. X represents the keyword set of a certain document, then

$$X = (0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \tag{3}$$

The cosine coefficient of these two vectors is as follows:

$$\cos(X,Y) = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} x_i^2}\sqrt{\sum_{i=1}^{t} y_i^2}} = \frac{2}{\sqrt{2}\sqrt{19}} = 0.3244 \tag{4}$$

The value of 0.3244 is defined as the degree of correlation between the document and the oxidation category in the technology dimension.

3.4.2. Document retrieval

Document retrieval allows for inquiry of relevant documents according to the category of knowledge dimension. The document retrieval process is defined in Fig. 3. It is depicted in two parts, of which the first part enables document retrieval via a project process. In such a case, the users pick directly from three process categories and four relationship items. The second part handles document retrieval involving the technology or equipment and material dimensions. The users can look up the documents according to the classifications of technology or equipment and materials, or retrieve the documents by selecting the category keywords. If an inquiry is made according to the classifications of technology or equipment and material, the system will calculate and sort the degree of correlation via VSM. If the document retrieval is made according to the keywords, the system will select and compare

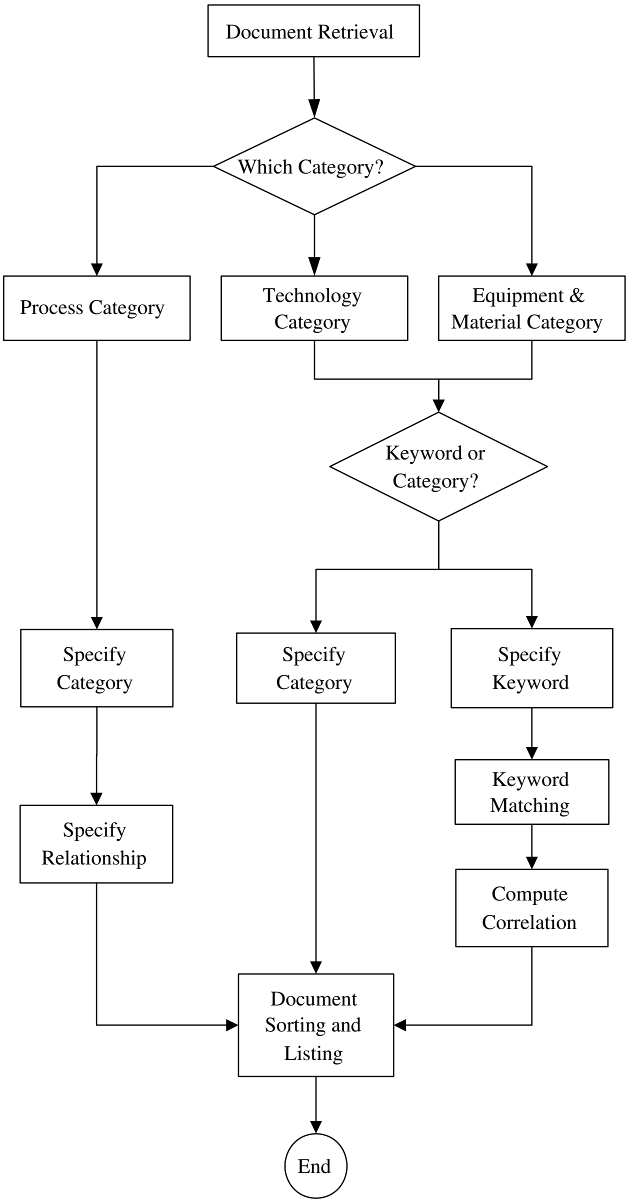


Fig. 3. Document retrieval process.

the keyword sets of the same category accordingly, and then calculate the degree of correlation.

The detailed flow of document retrieval is as follows:

- Select the knowledge dimension for document retrieval.
- If process classification is selected, define the category and relationship directly from categories of the three processes and four relationship items. If documents from the technology classification or equipment and material classification are wanted, proceed with Step 3.
- Select retrieval of technology classification or equipment and material classification. If selecting category inquiry, directly list the documents in this category and sort them according to the degree of correlation of categories. If selecting keywords in a category keyword set, proceed with Step 4.
- Locate keywords in a category keyword set, calculate the degree of correlation of all documents, and then list the documents in order of the higher degree of correlation.

4. System analysis and design

4.1. System analysis

The user demands regarding the R&D document classification and retrieval system were first analyzed, and system information requirement and data flow were then defined. Finally, the system structure, functional operations and system features were established and presented using IDEF0, a structured system analysis tool. The document classification and retrieval system were comprised of three sub-systems, namely the document classification system, the document retrieval system, and the maintenance and management system. Fig. 4 depicts the system structure constructed by IDEF0.

A document classification system (A1) enables the users to upload the documents for archiving. Then, the documents are categorized into a knowledge dimension for subsequent document retrieval. As a result of document classification, a document retrieval system (A2) allows users to locate the documents according to a knowledge dimension, category, or keywords; the maintenance and management system (A3) is designed to configure the system, such as updating keyword sets.

The document classification system consists of four modules, i.e. process classification (A11), document judgment (A12), stemming (A13) and technology, equipment and material classification (A14). The process classification module enables domain experts to define document categories according to the document relationships. The document judgment module distinguishes between literal and graphic documents. In the case of literal documents, stemming is triggered and performed; in the case of graphic documents, keywords set for the document must be defined manually. The keywords are then compared via the VSM. Then, the degree of correlation between the document and knowledge dimension is calculated, and consequently the document is categorized.

The document retrieval system consists of three modules, i.e. process dimension retrieval (A21), technology dimension retrieval (A22) and equipment and material dimension retrieval (A23). A document retrieval system primarily aims to retrieve documents according to the desired classification results of process, technology and also equipment and material. In the case of retrieval by process dimension, the system is required to list the property of process category; in the case of retrieval by technology or equipment and material dimension, only the degree of correlation between documents and categories need be listed.

The maintenance and management system is primarily used to manage users' access authorization (A31) and perform system configuration (A32). This sub-system empowers the system administrator to define different member groups, of which common member groups have only the right to update their own documents, while domain expert groups have the right to change the document property. The keyword set of technology and equipment and material dimensions can be updated and supplemented by domain experts (A33).

4.2. System design

The document classification and retrieval system is a three-tier information processing structure. The users' interface and system features are integrated by Javascript programs (JSP). The accessing

A0 R&D document classification & retrieval system
- A1 Document classification system
  - A11 Process classification module
  - A12 Document judgment module
  - A13 Stemming module
  - A14 Technology, equipment and material classification module
- A2 Document retrieval system
  - A21 Process dimension retrieval module
  - A22 Technology dimension retrieval module
  - A23 Equipment and material dimension retrieval module
- A3 Maintenance and management system
  - A31 User management module
  - A32 System configuration module
  - A33 Attribute administration module

Fig. 4. IDEF0 architecture of document classification and retrieval system.

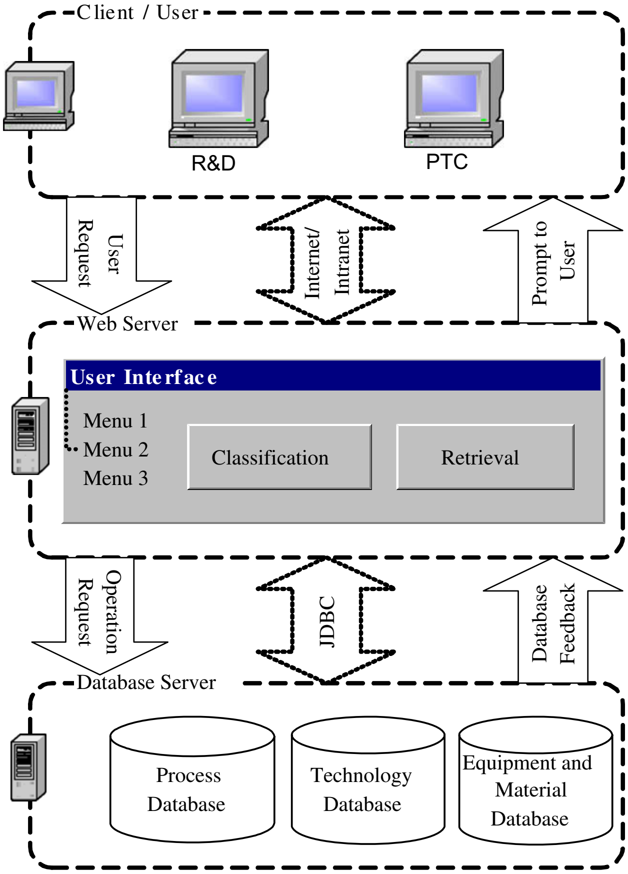


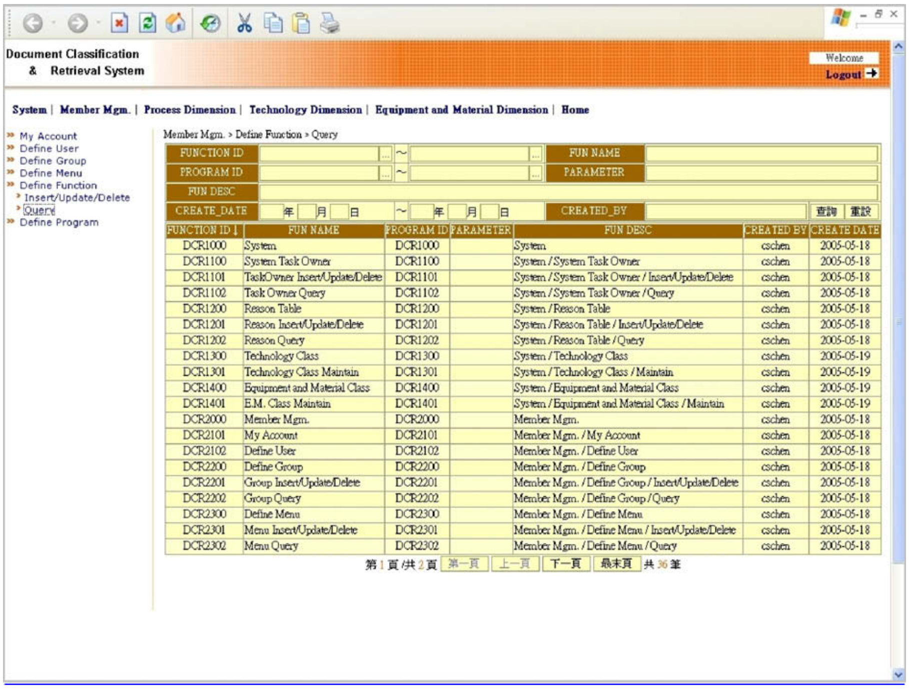
Fig. 5. System architecture.

to database is interpreted and performed by Tomcat and JDBC. The prototype system is configured and designed to be installed onto the servers at the R&D site, which is accessible via the Intranet or Internet. Fig. 5 illustrates the system architecture. The components are depicted as follows:

1 *Client*: The R&D researchers are the major system users. The Client is linked to the system via internet browsers and networked computers.
2 *Web server*: Responsible for integrating the clients with the database, playing the roles of data collector, manipulator and transferor.
3 *Database server*: Responsible for document storage and database operation.

4.3. Database design

The design of a database for the prototype system is mainly based on a relational database. The setting of primary keys and database normalization are inevitable to successfully implement a database application system. The primary key is a field that uniquely describes each record. In the document classification system, for instance, the document ID number is set to be the



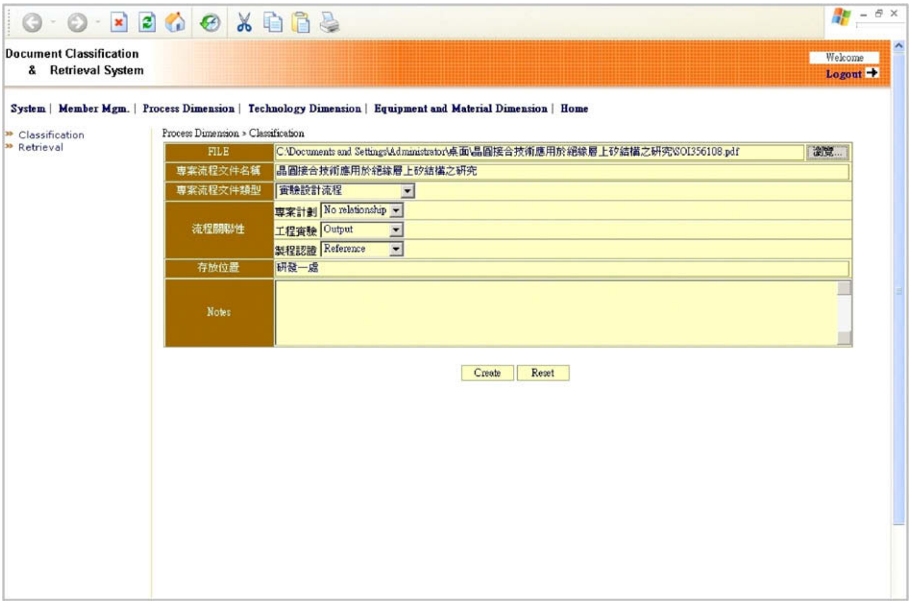
Member Mgm. > Define Function > Query

FUNCTION ID	~		FUN NAME	
PROGRAM ID	~		PARAMETER	
FUN DESC				
CREATE_DATE	年 月 日 ~ 年 月 日		CREATED_BY	查詢 重設

FUNCTION ID ↓	FUN NAME	PROGRAM ID	PARAMETER	FUN DESC	CREATED BY	CREATE DATE
DCR1000	System	DCR1000		System	cschen	2005-05-18
DCR1100	System Task Owner	DCR1100		System / System Task Owner	cschen	2005-05-18
DCR1101	TaskOwner Insert/Update/Delete	DCR1101		System / System Task Owner / Insert/Update/Delete	cschen	2005-05-18
DCR1102	Task Owner Query	DCR1102		System / System Task Owner / Query	cschen	2005-05-18
DCR1200	Reason Table	DCR1200		System / Reason Table	cschen	2005-05-18
DCR1201	Reason Insert/Update/Delete	DCR1201		System / Reason Table / Insert/Update/Delete	cschen	2005-05-18
DCR1202	Reason Query	DCR1202		System / Reason Table / Query	cschen	2005-05-18
DCR1300	Technology Class	DCR1300		System / Technology Class	cschen	2005-05-19
DCR1301	Technology Class Maintain	DCR1301		System / Technology Class / Maintain	cschen	2005-05-19
DCR1400	Equipment and Material Class	DCR1400		System / Equipment and Material Class	cschen	2005-05-19
DCR1401	E.M. Class Maintain	DCR1401		System / Equipment and Material Class / Maintain	cschen	2005-05-19
DCR2000	Member Mgm.	DCR2000		Member Mgm.	cschen	2005-05-18
DCR2101	My Account	DCR2101		Member Mgm. / My Account	cschen	2005-05-18
DCR2102	Define User	DCR2102		Member Mgm. / Define User	cschen	2005-05-18
DCR2200	Define Group	DCR2200		Member Mgm. / Define Group	cschen	2005-05-18
DCR2201	Group Insert/Update/Delete	DCR2201		Member Mgm. / Define Group / Insert/Update/Delete	cschen	2005-05-18
DCR2202	Group Query	DCR2202		Member Mgm. / Define Group / Query	cschen	2005-05-18
DCR2300	Define Menu	DCR2300		Member Mgm. / Define Menu	cschen	2005-05-18
DCR2301	Menu Insert/Update/Delete	DCR2301		Member Mgm. / Define Menu / Insert/Update/Delete	cschen	2005-05-18
DCR2302	Menu Query	DCR2302		Member Mgm. / Define Menu / Query	cschen	2005-05-18

第 1 頁 /共 2 頁 第一頁 上一頁 下一頁 最末頁 共 36 筆

Fig. 6. Interface design with member management function activated.


Document Classification
& Retrieval System
Welcome
Logout
System | Member Mgm. | Process Dimension | Technology Dimension | Equipment and Material Dimension | Home
Classification
Retrieval
Create
Reset

Process Dimension > Classification

FILE	C:\Documents and Settings\Administrator\桌面\晶圓接合技術應用於絕緣層上矽結構之研究\SOI356108.pdf 瀏覽...
專案流程文件名稱	晶圓接合技術應用於絕緣層上矽結構之研究
專案流程文件類型	實驗設計流程
流程關聯性	專案計劃 No relationship
	工程實驗 Output
	製程認證 Reference
存放位置	研發一處
Notes	

Fig. 7. Display of process document classification.

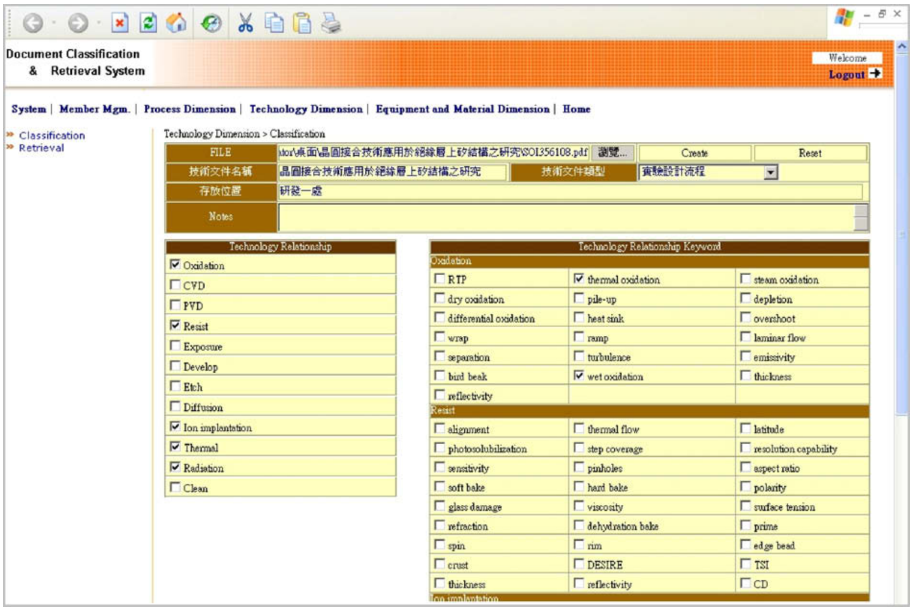
primary key in the data table storing document information. This field is unique and generated by the system automatically. The system has ten defined data tables such as equipment and material correlation, equipment and material keywords, equipment and material documents, technology correlation, technology keywords, technology documents, project process documents, document types, users and category vector. They are all defined in the SQL database server.

4.4. System process and interface design

The interface was constructed with a consistent layout. The function menu was located in the top portion of the display. When a main function was activated, its sub-menu items were deployed at the left side of the display. Fig. 6 illustrates the layout of the interface design with member management functions activated.

4.4.1. Classification process and interface design

One of the knowledge dimensions is the project process dimension. The treatment of process dimension documents is divided into classification and retrieval operations. The interface design of classification operations for project processes is shown in Fig. 7. If the documents are classified according to project processes, the documents are named automatically. After document names are generated and document categories are defined, the project processes and correlation categories are selected. The classification process terminates after all data fields are defined and documents uploaded without any warning or error messages.


Document Classification
& Retrieval System
Welcome
Logout
System | Member Mgm. | Process Dimension | Technology Dimension | Equipment and Material Dimension | Home
» Classification
» Retrieval

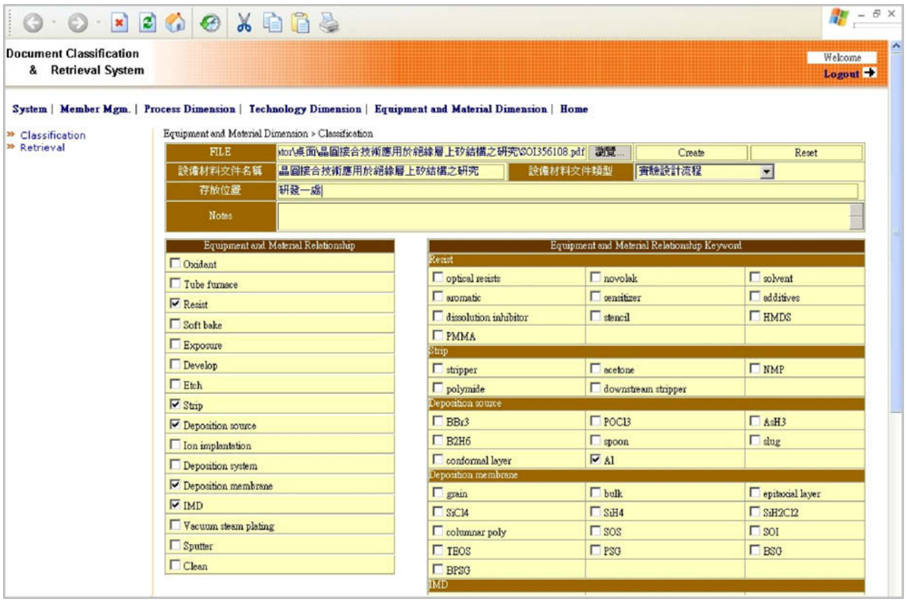
Technology Dimension > Classification

FILE	ator\桌面\晶圓接合技術應用於絕緣層上矽結構之研究\SOI356108.pdf 瀏覽...	Create	Reset
技術文件名稱	晶圓接合技術應用於絕緣層上矽結構之研究	技術文件類型	實驗設計流程
存放位置	研發一處		
Notes			

Technology Relationship
☑ Oxidation
□ CVD
□ PVD
☑ Resist
□ Exposure
□ Develop
□ Etch
□ Diffusion
☑ Ion implantation
☑ Thermal
☑ Radiation
□ Clean

Technology Relationship Keyword		
Oxidation		
□ RTP	☑ thermal oxidation	□ steam oxidation
□ dry oxidation	□ pile-up	□ depletion
□ differential oxidation	□ heat sink	□ overshoot
□ wrap	□ ramp	□ laminar flow
□ separation	□ turbulence	□ emissivity
□ bird beak	☑ wet oxidation	□ thickness
□ reflectivity		
Resist		
□ alignment	□ thermal flow	□ latitude
□ photosolubilization	□ step coverage	□ resolution capability
□ sensitivity	□ pinholes	□ aspect ratio
□ soft bake	□ hard bake	□ polarity
□ glass damage	□ viscosity	□ surface tension
□ refraction	□ dehydration bake	□ prime
□ spin	□ rim	□ edge bead
□ crust	□ DESIRE	□ TSI
□ thickness	□ reflectivity	□ CD
Ion implantation		

Fig. 8. Display of technology document classification.


Document Classification
& Retrieval System
Welcome
Logout
System | Member Mgm. | Process Dimension | Technology Dimension | Equipment and Material Dimension | Home
» Classification
» Retrieval

Equipment and Material Dimension > Classification

FILE	ator\桌面\晶圓接合技術應用於絕緣層上矽結構之研究\SOI356108.pdf 瀏覽...	Create	Reset
設備材料文件名稱	晶圓接合技術應用於絕緣層上矽結構之研究	設備材料文件類型	實驗設計流程
存放位置	研發一處		
Notes			

Equipment and Material Relationship
□ Oxidant
□ Tube furnace
☑ Resist
□ Soft bake
□ Exposure
□ Develop
□ Etch
☑ Strip
☑ Deposition source
□ Ion implantation
□ Deposition system
☑ Deposition membrane
☑ IMD
□ Vacuum steam plating
□ Sputter
□ Clean

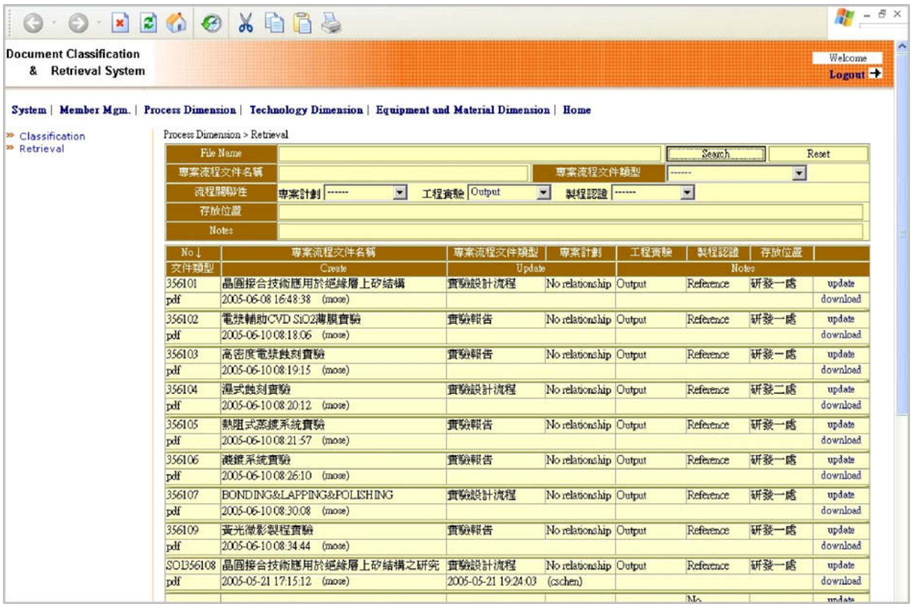
Equipment and Material Relationship Keyword		
Resist		
□ optical resists	□ novolak	□ solvent
□ aromatic	□ sensitizer	□ additives
□ dissolution inhibitor	□ stencil	□ HMDS
□ PMMA		
Strip		
□ stripper	□ acetone	□ NMP
□ polymide	□ downstream stripper	
Deposition source		
□ BBr3	□ POCl3	□ AsH3
□ B2H6	□ spoon	□ slug
□ conformal layer	☑ Al	
Deposition membrane		
□ grain	□ bulk	□ epitaxial layer
□ SiCl4	□ SiH4	□ SiH2Cl2
□ columnar poly	□ SOS	□ SOI
□ TEOS	□ PSG	□ BSG
□ BPSG		
IMD		

Fig. 9. Display of equipment and material document classification.

The other knowledge dimensions are the technology dimension and the equipment and material dimension, both of which have similar operational processes. The interface design of classification operations for the technology dimension is shown in Fig. 8, while Fig. 9 shows the display for equipment and material document classification. When documents are classified according to the technology dimension or the equipment and material dimension, the classification is required to distinguish if literal documents are included. In the case of document types such as doc or txt, the documents are judged by file extension names and the suitable process is triggered accordingly. After document names and types are traced, literal documents are directly uploaded and then compared with category keywords after stemming; graphic documents can be uploaded only after categories and keywords have been defined. The category correlation of uploaded documents is computed via the VSM, and thus the document is categorized.

4.4.2. Retrieval process and interface design

The interface design of the retrieval operation for project processes is presented in Fig. 10. Once the needed process and relationship are specified, the related documents are retrieved. The available processes include project planning, engineering experiments, and manufacturing certifications. The options for each process include input, output, reference, and none. The interface design of the retrieval operation for the technology dimension and the equipment and material dimension are shown in Figs. 11 and 12, respectively. The retrieval operation is similar for these two knowledge dimensions. Users can retrieve documents by spec-

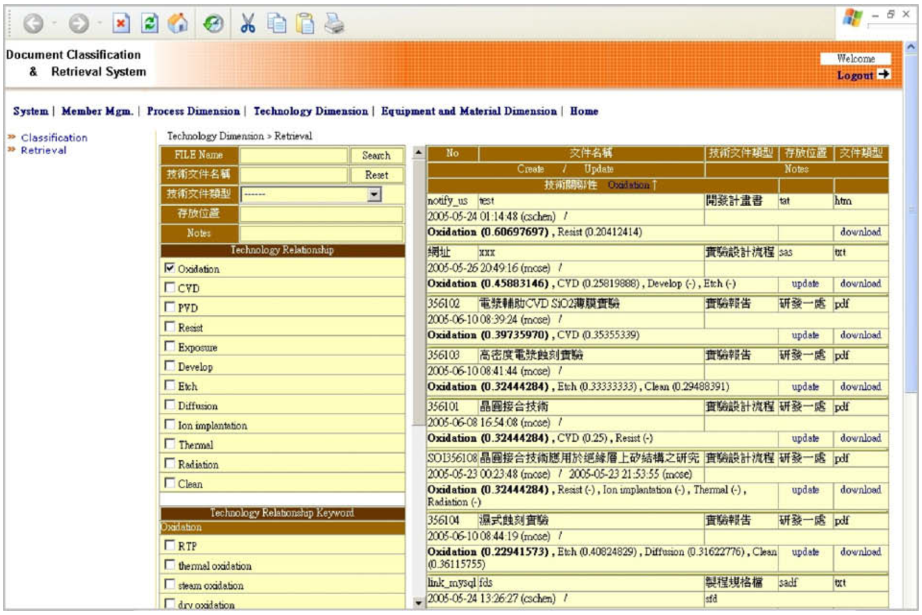

Document Classification
& Retrieval System
Welcome
Logout
System | Member Mgm. | Process Dimension | Technology Dimension | Equipment and Material Dimension | Home
Classification
Retrieval
Search
Reset

Process Dimension > Retrieval

File Name:
專案流程文件名稱: 專案流程文件類型: -------
流程關聯性: 專案計劃 ------- 工程實驗 Output 製程認證 -------
存放位置:
Notes:

No ↓	專案流程文件名稱	專案流程文件類型	專案計劃	工程實驗	製程認證	存放位置	
文件類型	Create	Update		Notes			
356101	晶圓接合技術應用於絕緣層上矽結構	實驗設計流程	No relationship	Output	Reference	研發一處	update
pdf	2005-06-08 16:48:38 (mose)						download
356102	電漿輔助CVD SiO2薄膜實驗	實驗報告	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:18:06 (mose)						download
356103	高密度電漿蝕刻實驗	實驗報告	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:19:15 (mose)						download
356104	濕式蝕刻實驗	實驗設計流程	No relationship	Output	Reference	研發二處	update
pdf	2005-06-10 08:20:12 (mose)						download
356105	熱阻式蒸鍍系統實驗	實驗報告	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:21:57 (mose)						download
356106	濺鍍系統實驗	實驗報告	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:26:10 (mose)						download
356107	BONDING&LAPPING&POLISHING	實驗設計流程	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:30:08 (mose)						download
356109	黃光微影製程實驗	實驗報告	No relationship	Output	Reference	研發一處	update
pdf	2005-06-10 08:34:44 (mose)						download
SOI356108	晶圓接合技術應用於絕緣層上矽結構之研究	實驗設計流程	No relationship	Output	Reference	研發一處	update
pdf	2005-05-21 17:15:12 (mose)	2005-05-21 19:24:03 (cschen)					download
					No		update

Fig. 10. Display of process document retrieval.


Document Classification
& Retrieval System
Welcome
Logout
System | Member Mgm. | Process Dimension | Technology Dimension | Equipment and Material Dimension | Home
Classification
Retrieval
Search
Reset

Technology Dimension > Retrieval

FILE Name:
技術文件名稱:
技術文件類型: -------
存放位置:
Notes:

Technology Relationship

☑ Oxidation
□ CVD
□ PVD
□ Resist
□ Exposure
□ Develop
□ Etch
□ Diffusion
□ Ion implantation
□ Thermal
□ Radiation
□ Clean

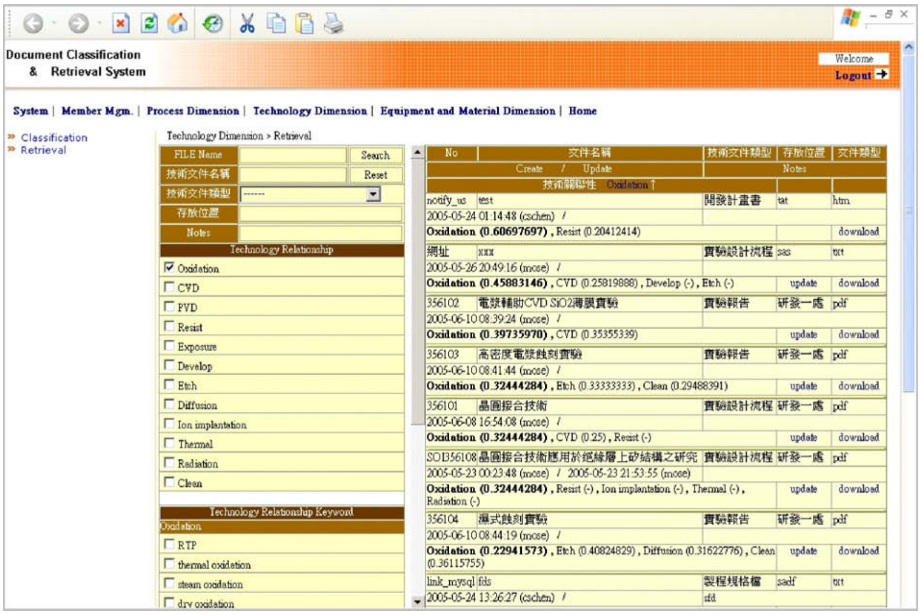
Technology Relationship Keyword

Oxidation

□ RTP
□ thermal oxidation
□ steam oxidation
□ dry oxidation

No	文件名稱	技術文件類型	存放位置	文件類型
Create / Update		Notes		
技術關聯性 Oxidation ↑				
notify_us	test	開發計畫書	tat	htm
2005-05-24 01:14:48 (cschen) /				
Oxidation (0.60697697) , Resist (0.20412414)				download
網址	xxx	實驗設計流程	sas	txt
2005-05-26 20:49:16 (mose) /				
Oxidation (0.45883146) , CVD (0.25819888) , Develop (-) , Etch (-)			update	download
356102	電漿輔助CVD SiO2薄膜實驗	實驗報告	研發一處	pdf
2005-06-10 08:39:24 (mose) /				
Oxidation (0.39735970) , CVD (0.35355339)			update	download
356103	高密度電漿蝕刻實驗	實驗報告	研發一處	pdf
2005-06-10 08:41:44 (mose) /				
Oxidation (0.32444284) , Etch (0.33333333) , Clean (0.29488391)			update	download
356101	晶圓接合技術	實驗設計流程	研發一處	pdf
2005-06-08 16:54:08 (mose) /				
Oxidation (0.32444284) , CVD (0.25) , Resist (-)			update	download
SOI356108	晶圓接合技術應用於絕緣層上矽結構之研究	實驗設計流程	研發一處	pdf
2005-05-23 00:23:48 (mose) / 2005-05-23 21:53:55 (mose)				
Oxidation (0.32444284) , Resist (-) , Ion implantation (-) , Thermal (-) , Radiation (-)			update	download
356104	濕式蝕刻實驗	實驗報告	研發一處	pdf
2005-06-10 08:44:19 (mose) /				
Oxidation (0.22941573) , Etch (0.40824829) , Diffusion (0.31622776) , Clean (0.36115755)			update	download
link_mysql	fds	製程規格檔	sadf	txt
2005-05-24 13:26:27 (cschen) /		sfd		

Fig. 11. Display of technology document retrieval.



No	文件名稱	技術文件類型	存放位置	文件類型
Create / Update		Notes		
技術關聯性 Oxidation ↑				
notify_us	test	開發計畫書	tat	htm
2005-05-24 01:14:48 (cschen) /				
Oxidation (0.60697697) , Resist (0.20412414)				download
網址	xxx	實驗設計流程	sas	txt
2005-05-26 20:49:16 (mose) /				
Oxidation (0.45883146) , CVD (0.25819888) , Develop (-) , Etch (-)			update	download
356102	電漿輔助CVD SiO2薄膜實驗	實驗報告	研發一處	pdf
2005-06-10 08:39:24 (mose) /				
Oxidation (0.39735970) , CVD (0.35355339)			update	download
356103	高密度電漿蝕刻實驗	實驗報告	研發一處	pdf
2005-06-10 08:41:44 (mose) /				
Oxidation (0.32444284) , Etch (0.33333333) , Clean (0.29488391)			update	download
356101	晶圓接合技術	實驗設計流程	研發一處	pdf
2005-06-08 16:54:08 (mose) /				
Oxidation (0.32444284) , CVD (0.25) , Resist (-)			update	download
SOI356108	晶圓接合技術應用於絕緣層上矽結構之研究	實驗設計流程	研發一處	pdf
2005-05-23 00:23:48 (mose) / 2005-05-23 21:53:55 (mose)				
Oxidation (0.32444284) , Resist (-) , Ion implantation (-) , Thermal (-) , Radiation (-)			update	download
356104	濕式蝕刻實驗	實驗報告	研發一處	pdf
2005-06-10 08:44:19 (mose) /				
Oxidation (0.22941573) , Etch (0.40824829) , Diffusion (0.31622776) , Clean (0.36115755)			update	download
link_mysql	fds	製程規格檔	sadf	txt
2005-05-24 13:26:27 (cschen) /		sfd		

Fig. 12. Display of equipment and material document retrieval.

ifying one or more inquiry fields, such as document names, document categories, technology correlations, and document locations. The retrieval results are listed on the right, while the correlation of all documents and categories is represented by the correlation value calculated from the VSM. The system sorts and lists the documents from the highest to the lowest according to their correlation values, subject to the maximum list number or total number of documents found. The desired document can be browsed online or downloaded.

5. Conclusion

Traditional document classification is heavily dependent upon manual efforts, which cannot deal extensively with the management of the electronic documents in tune with the growth of enterprises or information technology. This research study expands on the knowledge structure concept of the space vector model, and proposes a hybrid VSM-based document classification and retrieval structure for R&D activities pertaining to the semiconductor industry in Taiwan. It assists with document classification, and thus improves classification efficiency. It also provides precise retrieval of multi-dimension documents.

We define the knowledge dimension for document classification into the project process dimension, the technology dimension and the equipment and material dimension, allowing multiple document classifications. Only highly related documents are retrieved, thus improving upon the efficiency of traditional full-text retrieval via keywords. We also analyze and design a document classification and retrieval prototype system using the hybrid mechanism, such that the inquiry results are collated according to the degree of correlation. With the objective document evaluation mechanism, document classification and retrieval efficiency can be dramatically enhanced. The system was installed by a local semiconductor manufacturer, and the results were found to be promising. According to a series of pilot runs, the document management efficiency increased dramatically. For some specific knowledge dimensions and documents, the retrieval time was reduced from several minutes to seconds.

As for the continuation of the exploitable and extendable part of this study, we are seeking ways to improve some features of the document classification and retrieval system, such as (1) expansion and adjustment of the knowledge dimension categories, (2) automatic keyword-based screening technology and (3) connections to other business processes.

References

Bae, H., & Kim, Y. (2002). A document-process association model for workflow management. *Computers in Industry, 47*, 139–154.

Bang, S. L., Yang, J. D., & Yang, H. J. (2006). Hierarchical document categorization with k-NN and concept-based thesauri. *Information Processing and Management, 42*, 397–406.

Benkhalifa, M., Bensaid, A., & Mouradi, A. (1999). Text categorization using the semi-supervised fuzzy c-means algorithm. *International Fuzzy Information Processing Society*, 561–565.

Chen, R.-C., & Hsieh, C.-H. (2006). Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications, 31*, 427–435.

Chian, C.-S. (2004). *Semiconductor Industry Yearbook*. Hsinchu, Taiwan: Industrial Economy Information Center, Institute of Industrial Technology.

Chou, J. -Y. (2002). *Document classification based in-house knowledge management system*. Unpublished master thesis, Institute of Information Management, Yuan-Ze University, Taoyuan, Taiwan.

Deng, W., Wu, W. (2001). Document categorization and retrieval using semantic microfeatures and growing cell structures. In *12th international workshop on database and expert systems applications* (pp. 270–274).

Fan, H. -M. (2002). *Modeling and applications of process-oriented knowledge management system – Based on R&D process of TFT-LCD*. Unpublished master thesis, Industrial Engineering and Engineering Management Department, National Chiao-Tung University, Hsinchu, Taiwan.

Garraffo, F. M. (2000). Research and development investments and performance in high technology industries: Some evidence from semiconductor firms. In *2000 IEEE international conference* (pp. 234–239).

Hao, P.-Y., Chiang, J.-H., & Tu, Y.-K. (2007). Hierarchically SVM classification based on support vector clustering method and its application to document categorization. *Expert Systems with Applications, 33*, 627–635.

Horng, J.-T., & Yeh, C.-C. (2000). Applying genetic algorithms to query optimization in document retrieval. *Information Process and Management, 36*, 737–799.

Jablonski, S., Horn, S., & Schlundt, M. (2001). Process oriented knowledge management. In *11th international workshop on research issues in data engineering on document management for data intensive business and scientific applications* (pp. 77–84).

Kalczynski, P. J., & Chou, A. (2005). Temporal document retrieval model for business news archives. *Information Processing and Management, 41*, 635–650.

Li, J. -H. (2003). *Establishment of a XML-based content management system using web services technology and UNSPSC criteria*. Unpublished master thesis, Industrial Engineering and Engineering Management Department, National Tsing-Hua University, Hsinchu, Taiwan.

Lin, J. -Y. (2003). *Constructing electronic document management system using neural network based automatic document classification technology*. Unpublished master thesis, Industrial Engineering and Engineering Management Department, National Tsing-Hua University, Hsinchu, Taiwan.

Michael, W. B., & Browne, M. (1999). *Understanding search engines: Mathematical modeling and text retrieval*, NY.

Paice, C. D. (1990). Another stemmer. *SIGIR Forum, 24*(3), 56–61.

Que, H. -E. (2000). *Applications of fuzzy correlation on multiple document classification*. Unpublished master thesis, Information Engineering Department, Tamkang University, Taipei, Taiwan.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. NJ: McGraw-Hill.

Song, D., Lau, R. Y. K., Bruza, P. D., Wong, K. F., & Chen, D.-Y. (2007). An intelligent information agent for document title classification and filtering in document-intensive domains. *Decision Support Systems, 44*, 251–265.

Taghva, K., Borsack, J., & Condit, A. (1995). Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing and Management, 32*(3), 317–327.

Tai, X., Ren, F., & Kita, K. (2002). An information retrieval model based on vector space method by supervised learning. *Information Processing and Management, 38*, 749–764.

Tam, V., Santoso, A., & Setiono, R. (2002). A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization. *Proceedings of the 16th International Conference on Pattern Recognition, 4*, 235–238.

Trappey, A. J. C., Hsu, F.-C., Trappey, C. V., & Lin, C.-I. (2006). Development of a patent document classification and search platform using a back-propagation network. *Expert Systems with Applications, 31*, 755–765.

Wang, T. Y., & Chiang, H.-M. (2007). Fuzzy support vector machine for multi-class text categorization. *Information Process and Management, 43*, 914–929.

Wei, C.-P., Yang, C.-S., Hsiao, H.-W., & Cheng, T.-H. (2006). Combing preference- and content-based approaches for improving document clustering effectiveness. *Information Processing and Management, 42*, 350–372.

Zhang, J., & Rasmussen, E. M. (2001). Developing a new similarity measure from two different perspectives. *Information Process and Management, 37*, 279–294.

Zhuang, H. -M. (1999). *A study on document classification via intelligent computation*. Unpublished master thesis, Information Management Department, National Pingtung University of Science and Technology, Pingtung, Taiwan.