Electronic Document Secure Management System
Based on Content Analysis for Enterprise

Wen XIONG* and Zi-hui DING
China Great Wall Computer Shenzhen Company Limited, Beijing, China
*Corresponding author

Keywords: Electronic doc, Secure management, Content analysis, Partial submission.

Abstract. Electronic document management system is a critical component of enterprise. To enhance
the security of the management system was not only by Public Key Infrastructure (PKI), such as
public key, private key, and symmetric key based encryption and decryption, digital signature, and the
key and certificate automatic management, but also by intelligent text analysis, which extracted word
weight info automatically to generate security level of the doc using centroid-based classification and
integration with experts’ experience, and combined file’s attributes to classify the electronic docs into
the managed directory and database, and to locate the electronic docs, while checking the security
level of the doc to control the access. Meantime, we used the plugin for the doc transparent encryption
and decryption, and the digest signature with system-side private key for all docs. Furthermore, we
utilized the user interface (UI) for users’ convenience, which employed a partial submission strategy,
improving the efficiency of the doc management.

Introduction

Electronic official procedure often creates many electronic documents, from which specific
organization creates and distributes a large number of them, and individual user exchanges a good
number of them by using various of the instant messaging tools, such as ICQ, QQ, MSN, enterprise
Email-box, mobile Email-box, free Internet Email-box, micro message, blog, Skydive, cloud disk,
etc. Those documents always decentralize at different directories on hard disk or handheld devices,
which are difficult to manage manually.

Some methods of doc management based on search engine employ the similarity of the docs to
search the target docs and to sort them as the final results, which do not consider the security and
sensibility of contents in the procedure of creation, distribution, and storage.

As for the multi-user system, such as MS Window, UNIX, and Linux, a computer can be used by
different users with different privileges at different or the same period of time using telnet or remote
desktop to access the electronic docs stored on the remote host. Obviously, it is time-consuming and
high-demanding through access control lists to maintain the doc security. Meantime, it cannot face to
multifarious communications of the instant messaging tools. Similarly, as the cloud storage
environment for the enterprise, multi-users will not expect the leakage of the stored electronic docs
each other.

Many enterprises already have enterprise management systems, which also have work-flow
management sub-systems. Those work-flow sub-systems have customizable flows, which can
coordinate users of different levels to cooperate, to serial-sign, to countersign, to participate, to
interact, and to control those procedures, etc.

However, they are ordinary used for businesslike managements, such as financial managements,
human resources, information center, office automation, and business operation, etc., which involve
the information flows and interaction created in the procedure. The information has little special and
less important value due to it is created from the daily operations. Although it reflects the status of a
period of time of the organization, the knowledge and value cannot be presented by itself, which is
needed to mine and discover with extra tools.
However, more time, working members will communicate by other ways each other. Therefore, the diversification of the communication styles effectively makes-up the flaw and inflexible of the workflow, which improves the efficiency of the interflow and working result. Meantime, the docs created in these procedures will be more sensitive, which reflect the contents more concerned by the members in the organization.

However, many of the documents cannot be managed by the workflow management system due to the uncertain exchanging procedures. On the other hand, these docs will be more useful to the next working step for the members in the organization. Therefore, how effectively to manage and to utilize those electronic docs created by the exchange is the major problem cared by the system.

Many technologies whether decentralization management (DM) or centralization management (CM), only involves retrieve using simple file’s attributes, such as file name, create time, modification time, file size, etc., using keywords in pure texts, and using indexing from meta-data fields and invert index structure of the whole texts. The first two methods need little storage on the hard-disk, and can be easily implemented, but they need exact information from users, and utilize the contents of the documents in the shallow level. The latter usually needs biggish storage on the hard-disk, which applies an engine for full-text retrieval as its implement, and ignores the security of the documents.

However, the security of the documents is important. The CM for the electronic docs will be higher active and more secure than the DM for them, which avoids the non-controllability of the DM. Furthermore, CM based on content analysis (CMCA) will have more superiorities than pure texts retrieve, indexing specific fields, or the whole texts. Based on CMCA, a novel electronic document secure management system (EDSMS) is presented, which employs content analysis, security level generation of doc, and secure management methods, and will be suitable for the organization concerned with security more than other methods. The EDSMS can be implemented as equipment or software program that manages the classification, location, storage and secure access control based on content of electronic documents.

The rest of the paper is organized as follows: first, in Section 2, the novel EDSMS based on docs content analysis is presented; then, in Section 3, the related work for the electronic document management is introduced in brief; finally, in Section 4, conclusions are reached from the above discussion.

The EDSMS Based on Doc Content Analysis

We observed and studied some existing methods for electronic document management, and discovered that these methods have some common points that can be improved or enhanced to some extent.

Some Features for the Improvement of the System

1) High cost of hard-disk space for full-texts retrieves

In order to retrieve keywords in the full-texts, the retrieve engine must handle the original documents by computing word weights, and store them as well as extra information, such as word position, document id, etc. Therefore, the retrieve engine will have a high cost of hard-disk for the high speed response. To lower the cost, we employed a method to analyze, to extract, and to abstract critical info based on the Chinese contents of the electronic documents, which stored in the database first, and then, matched by using content similarity. Moreover, we integrated the attributes of doc, such as the file name, create time, modification time, size of the docs into the expression of a request to realize the automatic location, classification, security level management based on Chinese content analysis.

2) Long cost of retrieve time

Traditional retrieve employs a mode of from the end to end to search docs. If in this procedure, there are many noisy directories and files that are not belonged to the targets; the cost of retrieve time
will increase. Meanwhile, if the procedure is terminated by users to modify the conditional expression of the original search, and restarted for a new search, the contents of previous search will be abandoned, which resulting in the waste of the previous search, and a long cost of the retrieve time.

To avoid this flaw, we classify the relevant docs under the managed directory, which is effective to lower the cost of retrieve time due to that, there are not noisy directories and files under the managed directory, especially to some mime types, such as the compressive format RAR, GZIP, ZIP, TAR, etc.

3) The arbitrary copying and leakage of the electronic documents in the multi-user environment and client/server architecture

Normally, users can use privileges and roles of the special login account managed by the administrator of the operating system to avoid the arbitrary copying and leakage at the multi-user environment of the local host. However, there are so many operating system holes can be used for promotion authority, and users have feeble security consciousness for protecting the security of their electronic docs that the above leakage is obviously abused. In addition, complex operating actions confuse nonprofessional users.

To implement and enhance the security management of the electronic docs effectively, we employed automatic generation of the security level, grade management based on Chinese's natural language processing, and centralization management on the local host.

4) Security-level generation of the electronic docs

To handle the need of secret-involved electronic docs, the normal style is to add keywords, to form knowledge bases by the confidential employee manually. However, faced with large secret electronic docs of the organization, the style will be low efficiency, and large manual labor for maintaining the secrecy, and non-flexible change in pace with the new inputs of the secret electronic docs.

To avoid the above flaw, we employed the algorithms of machine learning (ML) and data mining (DM) for creating the grade knowledge base by training the set of electronic docs automatically, which utilized the centroid-based classification algorithm (e.g. nearest neighbors classifier, due to new attributes can be introduced easily).

5) Integration both experts’ experience and results from the ML and the DM

To enhance the efficiency of the ML and the DM, we permit that the system can be added to a few keywords, which are represented as experts’ experience.

The Steps and Mechanism of the EDSMS Based on Doc Content Analysis

The EDSMS includes the steps as follows:

**Step 1)** The location step based on content analysis adopts partial submission strategy, and provides users to input their words in docs’ name, keywords, and abstracts using user interface (UI) to improve the efficiency of the location, and to be convenience for the decision of the users. And it includes the sub-steps as follows:

**Sub-step 1.1)** Extracting effective keywords from the UI, such as user-define keywords, and sentences to form the foundation of the location, and learning experience from the interaction of users and equipment system to generate the keyword weights for the keywords, which are used for the comparison with directory info table, and user-define keyword table utilizing similarity algorithms to find a directory taken as the location directory by the maximum similarity value excessed a threshold $P_l$ (e.g. 0.98). The structures of those tables are shown as follows:

<table>
<thead>
<tr>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir-Id</td>
<td>Identify</td>
<td>Keywords</td>
<td>Set of top $N$ of the keywords</td>
<td>Create-time</td>
<td>Create time of the docs</td>
</tr>
<tr>
<td>Keyword</td>
<td>Keyword with the maximum weight</td>
<td>Dir</td>
<td>Full directory name</td>
<td>Modify-time</td>
<td>Modification time of the docs</td>
</tr>
</tbody>
</table>

where: the table records most important keyword and top N important keywords for the fast retrieve from the UI or doc to find a possible directory.
Table 2. User-define keyword.

<table>
<thead>
<tr>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-defined-keyword-Id</td>
<td>Identify Keywords</td>
<td>Set of top $N$ of the</td>
<td>Create-time</td>
<td>Create time of the docs</td>
<td></td>
</tr>
<tr>
<td>Keyword</td>
<td>Keyword with the</td>
<td>Keywords</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>maximum weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User-id</td>
<td>Users’ id</td>
<td>Dir-id</td>
<td>Directory id in the</td>
<td>Modify-time</td>
<td>Modification time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>table directory-info</td>
<td></td>
<td>of the docs</td>
</tr>
</tbody>
</table>

where: the table records most important user-define keyword and top $N$ important user-define keywords for the fast retrieve according to the user-define info from the UI to find a possible directory.

The similarity algorithm is based on the vector space model, and can be formulated as follows:

$$S = \sum_{i,j=1}^{n} w_{ij} \cdot w_{ij} / (\sqrt{|v(x)| \cdot |v(y)|})$$

(1)

where: $v(x_i) = v(y_j)$ indicates the $i$-th word surface of the $x$ vector is the same as that of the $y$ vector, then the algorithm will accumulate the inner multiply as $w_{ij} \cdot w_{ij}$, where the $|v(x)|$ represents the size of the $x$ vector, and $w_{ij}$ is the weight of the $i$-th word surface of the $x$ vector, which employs the popular term-frequency and inverse-document-frequency (TFIDF), and can be formulated as follows:

$$w_{ij} = \log \left( \frac{tf_i + 1}{1 + df_i} \right)$$

(2)

where: $tf_i$ is the term frequency of the word $i$ in the doc, and $df_i$ is the paragraph frequency in the doc, which is calculated by taking each paragraph as a doc to avoid the sparse data and dynamically joining in of the docs handled by the system. For the Chinese docs, a forward maximum segmentation algorithm was applied, and a feature selection algorithm was adopted in the directory info table and the user-define keyword table, which select top $N$ (e.g. 20) keywords with sorting descending according to their word weights.

Sub-step 1.2) Directly retrieving the managed directory to find docs according to the attributes of the files if the users input one or more attributes of files as the foundation of location.

Sub-step 1.3) Adding attributes of files into the structured query language (SQL) sentence in the Sub-step 1.1) if the users combine Sub-step 1.1) and 1.2) as the foundation of the location to narrow the range of the location.

Sub-step 1.4) Retrieving keyword indexing table according to time order descending if the above Sub-steps cannot find a location directory, which iterates the each docs in this table, where includes the attributes of docs, and calculates similarity according to the word weights and special similarity algorithm, and terminates the current location when the similarity value is greater than the threshold $P1$, and submits the docs’ full path name to the UI. The structure of the table is show as follows:

Table 3. Keyword indexing.

<table>
<thead>
<tr>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
<th>Columns</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Identify</td>
<td>Size</td>
<td>Size of the docs</td>
<td>User-id</td>
<td>Users’ id in the table user-login</td>
</tr>
<tr>
<td>Keywords</td>
<td>Set of keywords</td>
<td>Modify-time</td>
<td>Modification time of the docs</td>
<td>Dir-id</td>
<td>Directory id in the table directory-info</td>
</tr>
<tr>
<td>Full-path</td>
<td>Full path name of the docs</td>
<td>User-defined-keyword</td>
<td>Users’ definition keywords</td>
<td>User-defined-KeyWord-id</td>
<td>Id in the table user-defined-keyword</td>
</tr>
<tr>
<td>Create-time</td>
<td>Create time of the docs</td>
<td>Security-level</td>
<td>Security level of the docs</td>
<td>Insert-time</td>
<td>Insertion time to the table</td>
</tr>
</tbody>
</table>
where: the filed of keywords is the set of keywords of the doc, and the number of the keywords in the field is greater than that in the field of the directory info table (top $N$). Since the cost of similarity calculation is in proportion to the number of the keywords, the calculation of the directory info table first is better than that of the keyword indexing table first.

**Sub-step 1.5** Locating docs according to the similarity algorithm and new inputs from users if the returned info from sub-step 1.4) is not satisfied to the users till the user terminates the procedure of location, or modifies the condition of location and restarts from sub-step 1.1), or the user obtains the satisfied doc’ name and the abstracts. Then, the equipment will copy the doc into the temporary folder, and open the temporary folder for the user’s further usages. The similarity values of those located docs will less than $P1$ in this sub-step.

**Sub-step 1.6** Iterating the keyword indexing table using the returned directory from sub-step 1.3), and calculating the similarity value between keywords of the doc in this table and the inputs from the UI. Submit it to UI if the value is greater than $P1$.

**Sub-step 1.7** Checking the abstract of the doc from the UI by clicking the name of the doc to decide whether adopting this doc, locating next doc, modifying the condition of location and restarting, or terminating the current location.

**Sub-step 1.8** In the above sub-steps, the background process adopts the partial submission strategy to return location docs to UI instead of to return all location docs from the end to end, which reduces the cost of retrieve time and restarting time with a progressively interactive procedure of location, and takes the previous end point as the next start point, which avoid retrieve from the end to end each time, long waiting time, waste of the previous time, and improving the efficiency of location.

**Step 2** The classifying step of the electronic docs based on content analysis classifies the electronic docs or their copies under the managed directory according to the inputs from the UI, which will require the user to add possible needed info, such as user-define info, and content fragment when the info extracted from the classified doc is inefficiency or insufficient. The similarity algorithm is the same as that of the location sub-step 1.1), and the method will create new sub-directories under the managed directory when the new doc cannot be classified into the existing sub-directories each time of the classification.

**Step 3** The automatic assignment step of the electronic docs based on content analysis adopts ML method to generate the security-level for the docs, which integrates the experts’ experience by adding few weighted keywords manually to improve the accuracy of the security-level generation.

**Step 4** Strengthening the management security of electronic docs by developing the extending of the operating system, such as the file-filter drivers, kernel modules, and application program interface (API) hooks to control the access of the managed directory and embedded database.

**Step 5** Logging the operation info created by the users’ operations, and introducing encryption and decryption for the docs used in the steps from 1) to 3), which use the plugin developed for transparent encryption and decryption, and apply the digest signature to those docs under the managed directory using system-side private key.

**Related Work**

Over years, along with numerous documents accumulated at the hard disk, the users cannot explore easily and fast the complete information space inside the file system due to the large noisy directories and files, which occupies large disk space resulting in the retrieve time cost is very large when using the style from the end to the end. Therefore, the paper presented a prototypical inquiry system, which named DynaQ [1], and used a searching paradigm named Orienteering typically characterized by relatively small steps following one after another, having several advantages, such as reduction of cognitive load, sensitivity of the environment, and a better understanding of the results. This dynamic query enables the user to launch some small serial search queries in a context dependent way to get immediate middle results to help user explorer the docs’ sets satisfied the needs of the users.
The main target of the project of the electronic doc management system [2] was to create platform independent Web application to manage and distribute electronic docs just using Web browser, which provided an extra tool, that can speed up doc flow in any organization. Each user of the system would have a role, depended on that role user would have access to some options and electronic docs and others not. The project used Web services as authority methods by Web services interface Single Sign-On (SSO), and employed popular programming and secure domain technologies, such as Java 2 platform, enterprise edition (J2EE), Lightweight directory access protocol (LDAP), and Public key infrastructure (PKI) [3]. The introduction of the PKI has greatly enhanced the protection of the document content, such as the public key based encryption, digital signature, and the key and certificate automatic management to guarantee confidentiality, authenticity, integrity and non-repudiation. And a classic successful example of the PKI application based on the certificate to control the access is Security socket layer protocol (SSL) [4].

The document management plays an important role in the platform of cloud computing. The users exchange rich and varied information using electronic docs. The traditional methods based on the access control list are no longer satisfied the secure requirement of cloud computing. Therefore, the research [5] presented a novel user-based document secure management mechanism, which introduced the re-encryption. The re-encrypted method would combine the traditional access control lists to generate the re-encrypted key and to encrypt the document creation. They introduced the mature cryptographic technologies into their mechanism, and made those technologies automation and operating convenience for users, such as encryption for the confidentiality protection; message digests and digital signature for the data integrity protection, and random number introduction for the anti-replay attack defending. However, only combining with the general access control list is coarse grain size, and in our study, a content-based security level access control is presented, which can be used for improving the control grain size.

To prevent document leakage, the research [6] proposed a secure content protection and a secure management scheme about electronic docs, in which the client adopted the transparent encryption technology based on filter driver to encrypt files automatically under on-line and off-line modes. In addition, the scheme adopted different strategies in view of different receivers, in which the client encapsulated the permission and the cipher text or plaintext for the receiver internal or external while the internal receiver applied for decryption from the server. The security intensity of the system was depended on that of the encryption algorithms and cipher codes. However, in our system, enhanced sampling of hardware information of the client and the recognition of the client’s identity will strengthen the secure degree of the whole system. In addition, a Universal Serial Bus (USB) key stored user certificate with private key, which distributed by system administrator will improve the secure degree of the system instead of traditional simple username and password.

To enhance the security aspects during electronic docs' creation and distribution, the research [7] presented the design and implementation of a new PKI-based electronic document protection system, which concentrated on operating convenience for users when compared with the existing general electronic document management system.

Summary

To enhance the security of the electronic doc management system, not only the PKI was used in this system, but also the security level generation method based on content analysis was used, and the transparent encryption and decryption using plugin and USB key based method was adopted. The ML based, and DM based algorithm (e.g. centroid-based classification) was employed to generate the security level for the doc automatically, where the expert's experience was integrated. Meantime, for the convenience of the users, the UI guided the user to input the necessary word to narrow the range of the location, where the attributes of the file of file system also were utilized and combined into the SQL sentence, and a partial submission strategy was adopted to reduce the response time. In addition, the digest signature with system-side private key was applied for all docs under the managed
directory. In our internal testing and applying, the feedback of the users about the EDSMS has been quite good.

For future work, we plan to extend the back-end to access and handle more document types, e.g. open office and the Open document format (ODF), and will develop a new version for the cloud computing environment using client-server (C/S) and browser-server (B/S) architectures, where an SSL/Transport layer security protocol (TLS) based communication protocol will be used in the network transport security (e.g. Hypertext transfer protocol over secure socket layer (HTTPS)). As for the C/S architecture, Extensible markup language remote procedure call (XML-RPC) over SSL/TLS will be developed and adopted for the network security.

References


