Abstract

An office is termed as an information processing centre of an organization. The objectives of an office are to store and provide timely, accurate and relevant information for effective decision making. In a conventional office, information is captured in paper documents. But paper-based offices are failing to realize these objectives. With the advent of information technology, a paradigm shift has occurred in office automation. As a result, a transformation from a paper-based office to a paper-less office or an e-office is being observed. In an e-office, digital documents capture organizational information. Document production, storage and retrieval is a common work in almost all offices. The scope of the thesis is limited to this common work within an office. Inter-office document flow is beyond the scope of this work.

Document production in an office is based on a request-reaction-response paradigm. When a document containing a request is received in an office, the office reacts to the request. The reactions are recorded in the form of comments on the document and finally a response document is dispatched. We can term the process as Document Production Work-flow(DPW). The resultant document of a DPW is termed as a Multi-Part Multi-Signature Document(MPMSD). Therefore, a MPMSD is a case of the DPW. The first part of a MPMSD is the request document and the last part is the response document and the other parts in between are the comments of other review-
ers, that means, the reactions. Each part of a MPMSD is signed by the corresponding reviewer. The first reviewer is also termed as the originator. MPMSD is a generic framework. Parts belonging to different cases may also form a MPMSD. Moreover, multiple versions of a document may also be defined as a MPMSD. A reviewer creates a part of a MPMSD in the context of a set of existing documents constituted of rules, precedents and other support documents, which in turn may be MPMSDs. Rules, precedents and support documents constitute a reference space of an office.

A reviewer navigates through a subspace of the reference space before producing a new document and draws citations wherever necessary to substantiate the rules position, the precedent position etc. of the new document. This subspace is called the context of the DPW. The process of navigation through the context is called the case examination. The context of a DPW changes with time. As soon a new document, relevant to a DPW, is created or a new part is added to a document of the context, it is to be incorporated automatically. During case examination, when a part of a precedent is perused, the context of creation of the part, is to be recreated. Security, production and storage are the three aspects of the DPW problem, which are studied. Different issues of DPW are identified and solutions for addressing those issues are proposed.

The main objective of storage of information in an office is to keep track of the history of who did what, when, why and how. Thus storage in an office serves as the organizational memory, where the documents are the neurons. Therefore, the central issue is to store the documents in such a way that they can be identified, located and retrieved in an efficient way. Moreover, from a document thus retrieved, all the related documents, including the context, should be reachable in a simple and straightforward way. An office docu-
ment has three aspects: profile, content and a presentation. The profile is the bio-data of the office document. It contains meta-data of the document which provides a detailed description of the document. The profile comprises of a set of keywords and three types of records: production record, storage record and flow record. Most existing text retrieval techniques rely on indexing keywords or indexing terms. There are standard models for keyword based retrieval, like vector space model. We excluded keywords from the discussion of our model but it can be easily incorporated. Unfortunately, keywords alone cannot adequately capture the contents of office documents. We need other attributes, like record attributes to complement the keyword description of an office document. The production record consists of production related attributes, like class, type, topic, date of production etc., of an office document. The storage record consists of storage related attributes, like address, size, authorization, state etc. of a document. The flow record pertains to the flow of a document from one point to the other. The content of a document may be multimedia information. But, for the present work, we assume that it contains only text. The content may be unstructured, structured or semi-structured. The content of a document is presented for display or for printing in a layout framework. The layout framework provides the get up of a document. For storage and retrieval of office documents, the focus of our discussion is on the profiles of the documents. Therefore, content encoding and presentation aspects of a document is excluded from the rest of our discussion.

A model for storage and retrieval of documents in an e-office during document production workflow with the context as the main binding element is proposed. The office documents are considered here as pages. The model is termed as Page Cube (PC). A PC is a collection of registered pages of an
office. Here pages are the main entities. A page has a profile, which describes the page and is defined by a set of attribute-value pairs. Registration of a page means adding and recording a new page to the page cube and assigning a unique pageId to the new page. PC has two components: page space and page graphs. The page space is an n-dimensional space defined by n orthogonal dimensions. Each dimension represents a theme and is defined by an attribute. An attribute may have attributes and further attributes. Thus, attributes of a dimension form a dimensional hierarchy. Therefore, we can say that the page space is defined by n orthogonal hierarchies. A page is represented in this space as a point, whose coordinate is an n-tuple. The dimensions, for example are, type, topic, time, class, category, state etc. The pages of a PC are linked to a given page either implicitly or explicitly. Implicitly linked pages are those pages which satisfy a predicate defined over the dimensional values of the pages. Explicitly linked pages are those pages which are linked by explicit hyper links. Thus, the pages, which are explicitly linked, form a directed graph, where the pages are the vertices and the hyper links are the edges. In addition to the implicit links provided by the attributes of the dimensional hierarchy, the pages belonging to a dimension may be explicitly linked forming a dimensional graph(DG) of the concerned dimension. Thus the page graph component of PC is a set of DGs. The edges of DGs may be weighted. The construction of the DGs of dimension category is a mandatory one for a DPW and is discussed in detail. The algorithms for production of the DPW Context of a DPW and the Case Context of a precedent case are also proposed.

Next, the mechanism of querying a page cube is discussed. Two equivalent query languages are proposed. The first language is called Page Algebra (PA), an algebraic language which uses specialized operators on the sets of
pages to specify queries. The second one is called Page Structured Query Language (PSQL), a user-friendly pseudo-natural language with a simple means for expressing queries using a natural language form. The languages are similar to Relational Algebra and SQL respectively.

A framework for secure production and storage of digital documents is presented. DPW is a group work. Therefore, a distributed protocol without any central authority, like the SMTP of standard e-mail service, may prop up as a natural choice. In such a distributed approach, the digital document flows from mailbox to mailbox and the reviewers add their comments to the document. This mimics the flow of paper documents in manual offices. But this approach, fails to address the security issues pertaining to MPMSD, like part integrity, reuse of parts, non-repudiation of sending as well as receiving the document etc. To address these issues an in-line Trusted Third Party (TTP), called an arbiter is mandatory. The arbiter serves as the trusted intermediate agent in between the current reviewer and the next. To make a protocol as general as possible, researchers attempt to avoid the use of an arbiter. However, to provide non-repudiation with time information, an in-line TTP is necessary. To provide non-repudiation of a digital signature, time of the signature is essential as the key used in the signature may become public at a later time. In our environment, documents are persistent and so non-repudiation of a digital signature is essential. So an in-line TTP will be required. Further, to prevent the reuse of parts, an in-line TTP will evidently be required as immediate detection of such reuse will be necessary to prevent the taking of wrong decisions. Since an in-line TTP is mandatory for addressing issues like repudiation of sending and receiving documents, we can address other issues on MPMSD as well, with an arbiter as an in-line TTP. A protocol for production of MPMSD with an arbiter is proposed, where
a reviewer submits a comment (part) for a MPMSD to the arbiter and the arbiter adds the part to the MPMSD. This protocol addresses the security issues of MPMSD. A three-tier conceptual architecture with split logic is also developed.

In most of the office solutions, public-key based digital signatures are used. Commonly, the RSA algorithm is used. The computational complexity of the algorithm is high. Therefore, digital signatures based on RSA are slow. In an e-office a reviewer may have to sign many documents per day. This may lead to performance degradation of an e-office system. In our scheme, we have a novel idea of using signed session keys for the digital signatures of the reviewers on a MPMSD. Actually, this idea makes our protocols efficient. During a session, each reviewer has a session key, and a signed copy of the same, signed by the reviewer and certified and time-stamped by the arbiter, is available to only the reviewer and the arbiter. Now, since the session key is known only to the reviewer and the arbiter and the arbiter is trusted and will not cheat as per our assumption, any message encrypted with this session key during the session can be treated as the digital signature of the reviewer on the message.

Authorization is an important security aspect of DPW. All users of an office are not authorized to access all pages of the page cube. Authorization in DPW environment is dynamic in nature. Synchronization of authorization flow with the workflow is a fundamental security requirement in workflow environment. Other essential requirements include role based security policy, separation of duties and negative authorization etc. An event-based dynamic DPW Authorization Model (DPWAM) is proposed as an extension of the page cube model.

Finally, we tried to answer the question, where does our research work
stand in the space populated by the current academic as well as commercial solutions for office automation, like POLITeam project, Lotus Notes etc. followed by a discussion on implementation issues of the page cube in relational database as well as in XML document database models. Some conclusive remarks and future directions of work are also given.