IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7412

Chapter X

Using Weakly Structured Documents at the User-Interface Level to Fill in a Classical Database

Frederique Laforest and André Flory
National Institute of Applied Sciences, France

Electronic documents have become a universal way of communication due to Web expansion. But using structured information stored in databases is still essential for data coherence management, querying facilities, etc. We thus face a classical problem–known as "impedance mismatch" in the database world; two antagonist approaches have to collaborate. Using documents at the end-user interface level provides simplicity and flexibility. But it is possible to take documents as data sources only if helped by a human being; automatic document analysis systems have a significant error rate. Databases are an alternative as semantics and format of information are strict; queries via SQL provide 100% correct responses. The aim of this work is to provide a system that associates document capture freedom with database storage structure.

The system we propose does not intend to be universal. It can be used in specific cases where people usually work with technical documents dedicated to a particular domain. Our examples concern medicine and more explicitly medical records. Computerization has very often been rejected by physicians because it necessitates too much standardization and form-based user interfaces are not easily adapted to their daily practice. In this domain, we think that this study provides a viable alternative approach. This system offers freedom to doctors; they would fill in documents with the information they want to store, in a convenient order and in a freer way. We have developed a system that allows a database to fill in quasi-automatically from documents paragraphs.

The database used is an already existing database that can be queried in a classical way for statistical studies or epidemiological purposes. In this system, the document fund and the database containing extractions from documents coexist. Queries are sent to the database, answers include data from the database and references to source documents.

INTRODUCTION

Information capture is an important barrier for end-users software acceptance. In domains where the end-user is not compelled to use a computer or is demanding because of activity constraints, classical computerized systems have difficulty being accepted and widely used. Difficulties are more accurate when documents are the paradigm used to manipulate information. One can find many domains of this type: lawyers, doctors, etc., who use technical documents to store the information they need. These are domains where computerization is particularly little used. The example of the medical domain is obvious. Doctors are not satisfied by today's systems and prefer using paper-based medical records. Many trials have been and are still conducted in this field, but success has not completely come. The main barrier concerns information capture speed and facilities compared to computerized systems advantages. Capture forms have been chosen by computer scientists because they have the great advantage providing important querying capacities, as they are most often easily related to a database. Capture forms do not satisfy physicians as they cannot adapt to each case encountered. Forms impose data to be captured, the order in which to capture, and a strict format for each data.

With the current prevalence of the Web, and consequently, of electronic documents, the next idea that comes is to use electronic documents as a basis for the system. This idea has the first advantage of removing the mismatch between paper-based documents and capture forms. Information is captured in electronic documents and queries on documents can be made using a dedicated document querying language. To go forward in this idea, we have to note that one can find three different types of documents :

- Free-text documents only contain information and formatting instructions (these instructions are also known as physical structure). These documents are very easy to write, but very difficult to query. Free-text analysis is still a research domain; the results are not yet satisfying enough to be used in sensitive domains. Compared to paper-based documents, systems based on free-text documents still do not provide enough precision to be widely used.
- Strongly structured documents contain information and semantics guides (also known as logical structure). These guides are most of the time represented by tags circling information pieces. SGML (ISO, 1986) documents are good examples of such documents. These documents set a structure that the user has to follow. This structure is defined in a Document Type Definition (DTD) that provides tags, composition rules of tags, compulsory tags, attributes for tags, etc. This structure is not as rigorous as forms structure; no format is imposed for data. Moreover, only effectively captured information appears in each document. In forms, all fields are present even if not filled for the current case. Queries on strongly structured documents can be made using dedicated query languages like SgmlQL (LeMaitre, Murisasco, & Rolbert, 1998) or UnQL (Buneman, Davidson, & Hillebrand, 1996). These languages are currently under the form of prototypes and answers lack precision.

  In systems that link strongly structured documents to a database, each information to be stored in the database is tightly tagged so that there is a one-to-one relationship between data in the database and tagged information in documents. The database stores a representation of the tree structure of the document, without any treatment on information pieces; filling a database is thus rather easy but does not provide the same facilities as a real information system database filled through a form; queries are not as precise as queries on atomic data. This approach still does not satisfy end-users,