# Automatic Classification of Office Documents by Coupling Relational Data Bases and PROLOG Expert Systems

K. Woehl

Triumph-Adler AG, Research Department
Nuernberger Str. 159/161, D-8510 Fuerth, West Germany

Abstract This paper describes a component of an office automation system classifying office documents for later processing, filing and retrieval. This task is managed by a PROLOG knowledge-based system connected to a relational database management system. The main goal of this work is to get practical experience with the integration of expert system features in an existing software system.

## 1. Introduction

One of the main problems using artificial intelligence systems like expert systems or more general knowledge-based systems concerns the *integration with existing software*. Especially in the field of office automation we need a very close interaction with document editing and processing software, with traditional commercial software, and generally with database and information retrieval systems. There are only a few applications left being so self-contained that they can be realized by a stand-alone expert system.

A typical example is the automatic classification of office documents to prepare the later processing or filing. It is not enough to have an expert system analyse the documents only syntactically and find out, where the sender and the addressee can be mentioned in the letter, which words may be candidates for key words, etc. There also must be a substantial analysis: who is the sender, who is the adressee, which subject does the document concern, which data of the own company and of previous correspondence are referred to. While on the first aspect rather static rules about document syntax are used to deduce facts from the document, the complete analysis will substantially draw on data also being used by other software components. So the interface between data base management systems (DBMS) and expert systems (ES) will be of great importance in future office automation systems.

Vassiliou and others /1/,/2/,/3/ demonstrated ways to integrate expert systems and DBMS in their contribution to the last year's conference on Very Large Data Bases at Florence. Here we want to stress the practical relevance of such a combination of DBMS and artificial intelligence technologies by a special application. The presented classification system for office *documents is a component of the* experimental office automation system FOCUS of Triumph-Adler, which is mainly implemented in C and PASCAL on UNIX and the relational DBMS UNIFY. The PROLOG component for the classification is connected via a SQL interface to the main system.

## 2. Classification of Office Documents

The FOCUS-System (see picture 1) serves mainly for paperless document editing and processing. Paper-letters will be

**Proceedings of the Tenth International Conference on Very Large Data Bases.**

**Singapore, August, 1984**

read directly on receipt by an OCR-scanner, which recognizes the most popular kinds of typescripts. These letters further on can be processed by the computer as well as documents coming in by electronic mail, which the user creates with text software or selects from the filing system.

The next step classifies the documents according to

- document type
  (at this time German and English business letters, internal memos of Triumph-Adler)
- sender
- addressee
- date of the document
- reference to previous correspondance
- most important key words for later retrieval

The system files the documents according to the document type with its attributes /4/. The descriptors are maintained by a relational DBMS, whereas the documents are stored in the UNIX file system. The document filing and retrieval system maintains a strong relationship between them.

The classification system has access to the data of all customers and suppliers as well as to the corresponding data of the own company via a SQL interface. Moreover all descriptors and key words of the filed documents are available.

The classification system visualizes the actual state of the analysis to the user by marking the words being currently investigated on the screen (reverse video, frames, underlining etc.). The system also can ask the user questions, if the information in the letter and in the data base is not sufficient for the automatic classification process. The user can type in the information needed or point to it, if it is available in the letter. This happens especially, if the document came from a new business partner or if the letter contains important information in a logogram, which the OCR-scanner is unable to recognize so far.

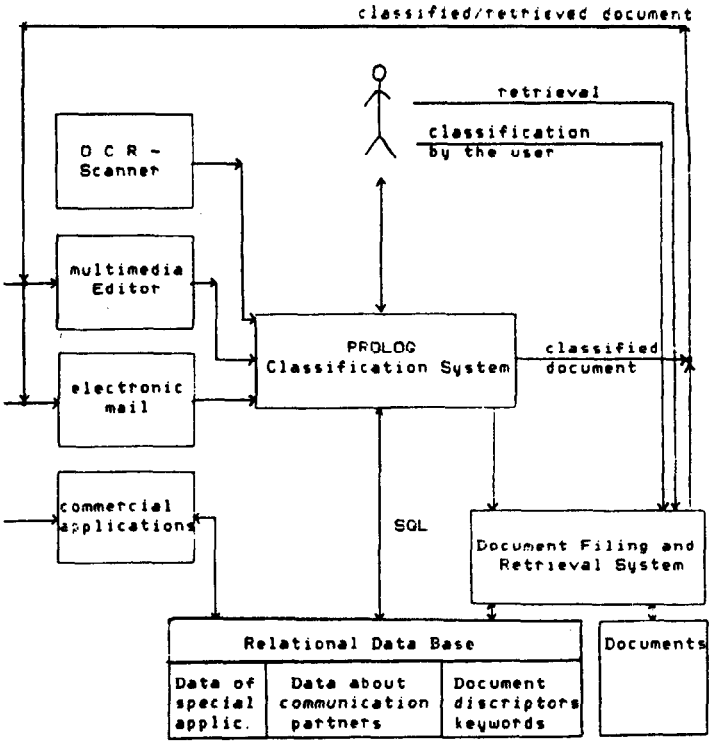


Figure 1

The user can check, add and modify the classification data. The classification process can be interrupted to enter data of new business partners with DBMS tools.

It is a very important feature that the expert system considers all data base entries, especially those, which came from other standard application software and previous classification processes. In this sense the system has a certain learning capability.

## 3. The expert system's task

The expert system contains facts and rules to

- identify the necessary special facts for the syntactic analysis
- recognize syntactic units, which may contain the desired descriptors
- create SQL-queries from the results of the syntactic analysis, to get additional semantic information about the document from the DBMS

- analyse the DBMS answers and to check the assumptions made formerly

- insert the created classification data into the DBMS via SQL insert and modify statements

- inform the user about the analysis state and ask him, if necessary, for additional information

Example: The letter-head consists of a logogram, which can not be interpreted by the OCR-scanner. While reading the document, the PROLOG system identifies all facts concerning syntactic units such as line, word, empty line, line with nonrecognized characters, half-filled line etc. Now the system can quickly match the discovered structures in the letter with a lot of predefined syntactic units as for instance:

```
<empty line>
Sincerely yours,
<empty lines or lines with
        nonrecognized characters>
.
.
.
(<word>) <empty>
```

The system assumes that <word> stands for the name of the sender, and puts a SQL-query about it to the DBMS. If there is really a business partner having this name known in the data base, the expert system will check the uniqueness of the name. Looking for further properties such as a reference to the previous correspondance, the system tries to verify the assumption.

Regarding the target data such as sender, addressee, key words etc. the expert system assumes that the descriptors of all these objects are already available in the data base. The system has to prove on the basis of the recognized letter information that there exists a unique relation to the stored data. If this cannot be verified because of missing data or contradictory statements, the system informs the user and tries to get the missing data from him.

Apart from the ability of the PROLOG inference machine to make deductions, to "reason", the advantage of PROLOG compared to other programming languages mainly consists here in the rather easy way to describe syntactic structures. This is not only needed for representing schemes of greeting phrases, address parts etc., but also for formulating SQL-queries. The advantages of PROLOG in programming, relational database management and knowledge representation are comprehensively discussed in /2/,/6/ and /7/.

## 4. Experiences, Further Development

Implementing our experimental office automation system FOCUS the tight coupling between expert system and DBMS was the most suitable integration way compared to the other possibilities, Vassiliou and others described in /1/,/2/,/3/. In contrast to their work our integration is more application dependant: A set of query fragments is prepared, which defines the class of possible queries needed by the application.

The main advantages are:

- The expert system remains quite small, because all substantial data are stored by the DBMS.

- Integration with every other DBMS application is possible.

- The system has a certain learning capability, because it adapts to the results of each previous classification as well as each other modification of the data base by the user or another application program.

Our further developement steps are

- Increasing and improving the rules for the classification of more different document types and for the identification of key words concerning separation or different postfixes

- Filing the documents according to predefined subjects

- Preparing the work of office workers by keeping all relevant documents ready

- Tuning the system with special regard to the CPU-time

- Extending the user interface

Acknowledgements: The author thanks the members of the FOCUS development team, his colleagues P. Frejek, K. Kreplin, R. Krueger, B. Mueller, N. Reithinger for many helpful discussions, contributions and comments to the realization and integration of the classification subsystem. I also thank the management of Triumph-Adler AG, especially Dr. Balzert (head of the research department) and Mr. Fauser for their support. Furthermore I am grateful to the referees for their constructive comments.

References.

/1/ Vassiliou, Y, Clifford, J., Jarke, M.,
"How does an Expert System get its Data?"
9th Conf. on VLDB 1983

/2/ Vassiliou, Y, Clifford J., Jarke M.,
"Access to Specific Declarative Knowledge by Expert Systems: The Impact of Logic Programming", in Decision Support Systems, Vol.1, No. 1, 1984

/3/ Jarke, M., Vassiliou, Y., "Databases and Expert Systems: Opportunities and Architectures for Integration" in New Applications for Databases, Academic Press 1984

/4/ Tsichritzis, D., "Form Management", Comm.ACM 25,7, July 1982

/5/ Brodie, M.L., Mylopoulos J., Schmidt J.W. (Ed.), "On Conceptual Modelling", Springer-Verlag 1984

/6/ Parsaye, K., "Database Management, Knowledge Base Management, and Expert System Development in PROLOG", Proc. Database Week, Business Applications, San Jose, Ca. 1983

/7/ Clark,K.L.,McCabe,F.G.,"PROLOG: a Language for Implementing Expert Systems", in Machine Intelligence 10, 1982



Appendix: Part of the Prolog-System used to analyse the sender of an letter

```
classifcation(LETTER) :-
   document_input(LETTER),
   ...
   sender_analysis(LETTER),
   ...

sender_analysis(LETTER) :-
   (sender_from_regards(LETTER,
                        NAME_OF_SENDER);
    sender_by_pointing(LETTER,
                        NAME_OF_SENDER);
    sender_by_input(LETTER,
                        NAME_OF_SENDER)),
   name_query(NAME_OF_SENDER),
   sender_query_answer(LETTER),
   show_sender_dates(LETTER).
sender_analysis(LETTER) :- ...

sender_from_regards(LETTER,NAME) :-
   regards(LETTER,LINE),
   name_in_regards(LETTER,LINE,NAME).
sender_from_regards(LETTER,NAME) :- ...

regards(LETTER,LINE) :-
  (word(LETTER,LINE,_,'Sincerely',1,_),
   ...                                  ),
   asserta(regards(LETTER,LINE)).
regards(LETTER,LINE) :- ...

name_in_regards(LETTER,REGARDS_LINE,
                              LINE,NAME) :-
   possible_line_with_name(LETTER,LINE),
   line(LETTER,LINE,_,NUMBER_OF_WORDS),
   word(LETTER,LINE,_,NAME,
                    NUMBER_OF_WORDS,_),
   asserta(name_in_regards(LETTER,
              REGARDS_LINE,LINE,NAME)).

name_query(unknown).
name_query(NAME) :-
   name_query_prefix(PREFIX),
   name_query_postfix(POSTFIX),
   tell(query),write(PREFIX),
   write(NAME),write(POSTFIX),told,
   system("SQL query > answer").

sender_query_answer(LETTER,unknown):-...
sender_query_answer(LETTER,NAME) :-
   see(answer),get0(C),
   read_name_query_answer(C,PERSON_ID,
        FIRM_ID,FIRM_NAME,ADDRESS,...),
   asserta(sender_person(LETTER,
                    PERSON_ID,NAME,...),
   asserta(sender_firm(LETTER,FIRM_ID,
                  FIRM_NAME,ADDRESS,...),
   ...

/* The relations line and word are
   asserted in document_input          */
```