Research and Development in Information Retrieval

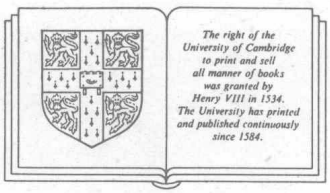
Proceedings of the third joint BCS & ACM symposium
King's College, Cambridge 2-6 July 1984

Edited by C. J. van RIJSBERGEN

Research and development in information retrieval

Proceedings of the third joint BCS and ACM symposium
King's College, Cambridge
2–6 July 1984

Edited by C. J. van Rijsbergen
Department of Computer Science, University College, Dublin



CAMBRIDGE UNIVERSITY PRESS
on behalf of the British Computer Society
Cambridge
London New York New Rochelle
Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
296 Beaconsfield Parade, Middle Park, Melbourne 3206, Australia

First published 1984

Printed in Great Britain at the University Press, Cambridge

Library of Congress catalogue card number: 84-45234

British Library cataloguing in publication data

Research and development in information retrieval.
(The British Computer Society workshop series)
1. Information storage and retrieval systems
I. van Rijsbergen, C. J.
II. British Computer Society III. Association
for Computing Machinery IV. Series
025'.04 Z699

ISBN 0 521 26865 6

CONFERENCE CHAIRMAN

A Steven Pollitt

Huddersfield Polytechnic

PROGRAMME CHAIRMAN

Keith van Rijsbergen

University College Dublin

LOCAL ARRANGEMENTS

J Ken M Moody

King's College Cambridge

TREASURER

Peter Willett

Sheffield University

PROGRAMME COMMITTEE

Tom Addis
Brunel University

Richard J Cichelli
American Newspaper
Publishers Association

W Bruce Croft
University of Massachusetts

Douglas R McGregor
University of Strathclyde

A Patricia Miller
Marathon Oil Company
Research Center

Najah Naffah
INRIA, Paris

Stephen E Robertson
City University

Hans J Schek
Technical University
Darmstadt

Roger Tagg
Independent Consultant

Clement T Yu
University of Illinois
at Chicago

PREFACE

The papers in this volume are those given at the symposium *Research and Development in Information Retrieval* at King's College, Cambridge in July 1984. This was the third joint BCS and ACM symposium in information storage and retrieval. The first joint conference was held in Cambridge in 1980 and the second one in Berlin in 1982. The 1984 conference was organised jointly by the specialist groups in information retrieval within The British Computer Society and The Association for Computing Machinery in co-operation with the Gesellschaft für Informatik e.V., the American Society for Information Science, and the Institute of Information Scientists.

The two themes of the conference were new research ideas in information storage and retrieval, and interesting new applications. The topics covered in the conference papers represent the following current areas of interest: office systems, integrated IR and DBMS systems, the user interface to IR, systems architecture, mathematical models, logical and physical clustering, evaluation, use of AI techniques in IR, and automatic indexing. This list highlights the way in which the problems of information storage and retrieval are central to many applications. Looking at the list in another way: it illustrates how techniques developed in other disciplines, for example AI, can be applied to problems in IR. Although IR is a discipline in its own right there can be no doubt that its development is inextricably linked to the developments in other disciplines. This makes research in IR especially exciting: a new approach to say DBMS can throw new light on some old problems in IR.

In selecting papers for this conference an attempt was made to represent the many and varied aspects of IR. Therefore, one will find papers arguing from an AI perspective, as well as papers approaching problems from a DBMS point of view. There are papers describing

applications in office information systems as well as in more traditional library systems. Some papers are devoted to hardware research aimed at solving some of the IR problems in hardware which have hitherto been solved in software. The overall architecture of IR systems is another theme covered. All in all the papers span a variety of important topics. This illustrates the vigour of the subject, in its ability to influence the research of other disciplines whilst at the same time taking on board new ideas from these same disciplines.

The conference owes much to the generous financial support from the BCS and from its sponsors: Acorn Computer Group plc, British Library Research & Development Department, IBM United Kingdom Ltd, International Computers Ltd, Plessey Telecommunications and Office Systems Ltd, and The General Electric Company plc. We on this side of the Atlantic would like to express our thanks to those on the other side who helped in the organisation. Finally, our thanks are extended to Kate Norman and Susannah Dean for their help.

C.J.v.R.

CONTENTS

Framework for the Development of an Experimental Mixed-Mode Message System

S. Christodoulakis *

Computer Systems Research Group
University of Toronto
10 King's College Road,
Toronto, M5S 1A4

* This work was done on behalf of the European Community as part of the ESPRIT pilot project on office filing.

Abstract

We describe a framework for the development of a mixed-mode message system for an office environment. Messages may be composed of attributes, text, images, and voice. Message retrieval is based on content. We discuss several issues related to the development of such systems. Text retrieval techniques are important for content retrieval in this environment.

1. Introduction

There is a growing interest among computer science researchers on office information systems that handle complex data such as text, attributes, graphics, images, and voice ([VLDB 83a], [VLDB 83b]). We will call the unit of multimedia information a *multimedia message*. Multimedia messages are composed of attribute text image and voice information. Some of the functions that these systems may provide are filing of multimedia information, content addressability of multimedia messages, and multimedia message transmition and reconstruction in a different site.

In a possible scenario an office worker uses the multimedia filing capability to find some information relevant to the interests of his company. He uses the extraction unit to extract some of this information, and the comparative interface unit to compare some images. When he selects an image with some statistical information he may want to alter its presentation form and/or further edit it. He uses the information extraction unit for it. Some information may be in a paper form. He uses

an image digitizer and an extraction capability to extract the relevant information. Finally he may want to use all the information that he has selected so far to compose a report. The report may be transmitted through communication lines to another station.

There are several important problems associated with the development of such systems. Some of these problems are identified next:

1. Query environment

Queries in this environment may be different than queries in traditional Data Base Management System Environments. Users may only have a vague idea of what they are looking for. Their understanding of what they want and how to specify it may increase as they look at other messages. Their query may prove to be inadequate. In that case they may want to reformulate it. Some other users may want to enhance their retrieval capability by specifying some characteristics of the message that have to do with the presentation form of the messages rather than the content. Finally queries on the image and text part of messages are not often adequately handled by traditionals DBMS's.

2. Content addressability in various data types

Content addressability in these diverse data types presents serious problems.

Content addressability in messages containing attribute value and text information can be achieved by allowing the user to specify expressions involving the attribute values of the message as well as regular expressions of words appearing within the text message ([Aho et al. 78], [Tsichritzis and Christodoulakis 83]). Structures for efficient retrieval of formatted data from single and multi-file environments have been studied extensively for various retrieval request types and frequencies ([Teorey and Fry 82], [Wiederhold 83], [Ullman 83], [Christodoulakis 84], [Christodoulakis 83a], [Christodoulakis 83b]). Content retrieval from text files has also been studied for various environments and efficient methods have been described ([Salton and McGill 83], [Rijsbergen 79], [Floyd and Ullman 80], [Haskin 81], [Tsichritzis and Christodoulakis 83], [Christodoulakis and Faloutsos 84]).

Content addressability of the image message and voice message part is much harder. One reason is that the current technology on image and voice recognition has serious limitations.

Picture recognition involves very expensive pattern recognition routines ([Tou and Gonzalez 73], [Duta and Hart 72]). In addition picture

recognition of general type pictures is still remote ([Ballard and Brown 82], [Rosenberg 76], [Fu 83], [Pavlidis 77]). Existing experimental and commercial systems based on high power machines (array processors) can be successful only when much knowledge about the scene is presented in a picture is available. A second, equally important problem related to pictures is that it is very difficult for the user to specify precisely the picture content of the pictures he wants.

Speech recognition presents similar problems ([Fu 83], [Reddy 75], [Reddy 76], [Erman et al 80], [Electronics 83]). Currently only speaker dependent, discrete speech, voice recognition devices with a limited vocabulary of words exists in the market. The speaker has to train the voice recognition system to recognize the limited vocabulary. Typically this involves repeating several times each word to be stored in the vocabulary. Discrete speech (words are separated by a pause) is divided into words and each word is compared with the words in the vocabulary. If it matches closely one of the words in the vocabulary the word is "recognized". Some systems allow more than one vocabularies to be stored but the size of each vocabulary is further reduced. The storage requirements of a vocabulary for speech recognition are very large and the algorithms for finding approximate matches expensive.

3. Information Organization and Access

In an office information environment files are seldom static. The information is usually diverse, the world changes fast, people do not like to spend time for organization and reorganization of information. In addition the good naming, structuring, consistency and quality of information which is assumed in data base environments is easy to be maintained in an office information system. Information may be inserted using error prone methods and the quality of information and the consistency of naming may be very difficult to be maintained. The query capability and the access methods used should be able to cope with these problems. Of course performance is of very high importance in this environment: By specializing the application environment we expect to provide better services (e.g. content addressability) and usable systems. Performance may become easily unbearable due to large volume of data and to the flexibility of the query environment.

4. Query Interfaces

Users of office information systems may have very diverse

backgrounds varying from the sophisticated type to the naive type. The precise syntax required by many DBMS may not be appropriate for this environment. The high quality screens, voice input output and other sophisticated devices that are now available have the potential for providing very effective interfaces. In this environment it would also be desirable that the query interface facilitates the user to express his queries in a better way.

5. Information extraction and internal representation

In a multimedia message environment several possible ways of message creation exist. Multimedia messages may be interactively generated in a given station and sent to another station via communication lines. In the receiving station additional editing of the message may take place. Alternatively messages or parts of messages (pictures) may be in a paper form in which case a powerful image segmentation and OCR capability may be used for extracting the information from the messages. Additional editing may also take place. In a different scenario an office worker synthesizes new messages from old ones. He uses the query capability of the system to identify some messages which contain the relevant information and he extracts this information to synthesize a new message. During this process he may change the presentation form of the information and/or further edit it. Finally, messages may be transmitted from a telephone and possibly edited later on.

Automatic extraction of information from documents is a difficult task. An important step in the automatic extraction of information from documents is image segmentation. The objective of the image segmentation process is to divide documents into segments containing text and segments containing various types of images as well as to identify objects within images.

Various automatic segmentation techniques have been proposed, but the problem in its full generality is very difficult [Ballard and Brown 82]. Segmentation into text regions and image regions is an easier problem and the performance of existing techniques is adequate [Wong et al. 82]. In addition optical character recognition techniques perform well for a variety of fonts. Thus existing techniques can be used for the automatic recognition of the text part of a generated from documents.

However, in many cases this may not be adequate. The information contained in various images may contain much redundancy [Pratt et al.

80]. For example if an image of a document contains a simple graph, this graph may be encoded in an internal representation form with much reduced storage requirements. Thus an internal representation may be used to reduce storage requirements as well as telecommunication costs. However, the internal representation may need be different from system to system depending on the availability of devices of various types, as well as on the workload characteristics of the system (e.g. CPU bound versus IO bound system, capacity of the communication medium).

6. Data presentation

Data base management systems have traditionally emphasized data organization and manipulation. However, presentation of data in devices with diverse capabilities has traditionally be deemphasized [Gray 83]. This problem is particularly important in the presence of bit map display capabilities with different number of gray levels, colors, display sizes, as well as facsimile input/output devices [Horak 83]. In addition the large storage requirements of image information and the large cost of transmitting this information through communication lines may impose an internal representation of some image types which is different than the presentation form of these images. Finally, voice input and output devices also present similar problems.

Information Retrieval Research has emphasized retrieval based on "relevance" rather than retrieval based on "occurrence" ([Rijsbergen 79], [Salton and McGill 83]). From this point of view techniques developed in this area could be very useful for this environment. A potential problem is that the office information environment is a dynamic one.

In this report we describe an approach for the development of an office information system which handles multimedia messages. Multimedia messages consist of attribute information, text information, voice information, picture information, graphic information or any mix of the above types. We assume an office information system environment which is equiped with existing high performance hardware and software such as: A set of personal computers with raster and vector display capabilities and a pointing device (mouse), magnetic disks, optical disks, directly addressable microfilm, directly addressable voice storage devices, a voice input and a voice output capability, a laser printer, an optical reader capability (OCR), and a local network.

We first present a conceptual framework for describing the logical structure of multimedia messages. Content addressability in the

attribute and text part of multimedia messages is achieved by allowing the user to specify combinations of attribute values and words appearing in the text part of the message [Tsichritzis and Christodoulakis 83]. This framework is an extension of this work in that it provides more content addressability in the image part of multimedia messages. In particular image text information, statistical information and spatial information is used to enhance content addressability. Since much of the information in offices may be coming from very diverse sources (letters, ads, publications, government statistics) and it possibly has very diverse formats [McLeod 83], it is important that a powerful query capability is used for content addressability. The logical structure of messages shows what are the components of the multimedia messages and how they are interrelated. This information is useful for the user's understanding of the system and its capabilities.

Multimedia messages are stored within the system using an internal representation. A mapping from the internal representation to the presentation form is maintained. Parameters kept with the mapping allow a close reproduction of multimedia messages in sites that may have screens with different display capabilities.

The user can retrieve multimedia messages by specifying contents of the logical message. To allow a powerful retrieval capability as well as reduced storage requirements important information is extracted and stored in the message's internal form.

Multimedia messages are stored in general files so that supervision and frequent reorganization due to changes in the environment are avoided. The retrieval of multimedia messages is achieved by using a powerful access mechanism based on message abstractions and a browsing query interface which is helped by miniatures and voice abstractions. To facilitate the users in formulating their queries the system provides a dynamic query reformulation capability.

2. A framework for multimedia messages

In this section we present a conceptual framework for multimedia messages. The conceptual framework describes the logical components of multimedia messages and their interrelationships. The framework is useful for describing the capabilities of the system to users.

The logical components of a multimedia message are shown in figures 1a and 1b. Multimedia messages have a *type* associated with them

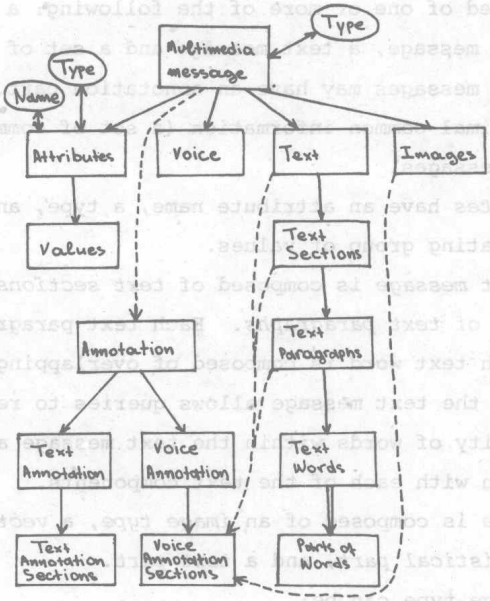


Figure 1a: Multimedia message structure

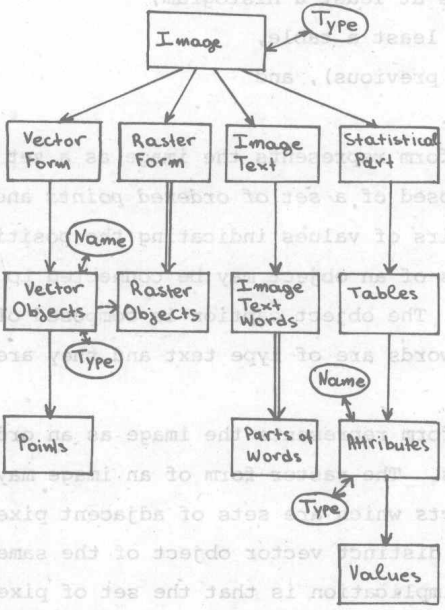


Figure 1b: Multimedia message structure (image part)

and they are composed of one or more of the following: a set of attributes, a voice message, a text message and a set of images. In addition multimedia messages may have an annotation part. The message type contains a minimal common information (a set of common attributes) in a large number of messages.

Attributes have an attribute name, a type, and a value. The value may be a repeating group of values.

The *text message* is composed of *text sections*. Each text section is composed of *text paragraphs*. Each text paragraph is composed of *text words*. Each text word is composed of overlapping *parts of words*. This structuring of the text message allows queries to restrict retrieval based on the proximity of words within the text message as well as to associate annotation with each of the text components.

An *image* is composed of an *image type*, a *vector form*, a *raster form*, a *statistical part*, and a *text part*.

The image type can be:
graph if it contains at least one graph,
pie chart if it contains at least a pie chart,
histogram if it contains at least a histogram,
table if it contains at least a table,
statistical (any of the previous), and
picture (anything else).

The vector form represents the image as a set of *image objects*. An image object is composed of a *set of ordered points* and an *object caption*. Points are pairs of values indicating the position of a point within an image. Points of an object may be connected to form lines, polygons, polylines, ... The object caption is composed of object *caption words*. Object caption words are of type text and they are composed of parts of words.

The raster form represents the image as an ordered set of pixels in two dimensions. The raster form of an image may contain possibly overlapping *raster objects* which are sets of adjacent pixels. Each raster object corresponds to a distinct vector object of the same picture which is a closed polygon. The implication is that the set of pixels composing the raster object is defined by the boundaries of the vector object when it is superimposed on the raster form of the image.

The statistical part of the image is composed of a set of *tables*. Each table has a set of *attributes*. Attributes have a *name*, a

type, and a set of *values*. Tables within an image are independent on each other. They are used to store the statistical information contained in a graph, pie chart, histogram or table of an image.

The *image text part* is composed of *image text words*. Image text words are composed of *parts of words*. The image text part is text related to a given image. The text part is formed by the following:
The image caption of a given image,
Text paragraphs related to the image,
Object caption words of objects within the image,
Attribute names of attributes in the statistical part of the image,
Attribute values of attributes of the type text in the statistical part of the image.

The voice message is composed of *voice words*.

Annotation is composed of *text annotation* and *voice annotation*. Text annotation is composed of *text annotation sections* and voice annotation is composed of *voice annotation sections*.

Annotation may be associated with a text message, text sections, text paragraph, text words, and images. (The lines from voice annotation are not shown in the figure.) Annotation is a further informal explanation about the contents of a message, paragraph, word, image or image object.

3. Internal Representation and Presentation Form of Multimedia Messages

The presentation form of the constituents of a message may be different from the internal representation of the message. For example the text message may be internally represented in a binary form but presented in a raster form.

The internal representation of an image does not have to have both an object form and a raster form. It may only have one of the two. An example of an image where both forms exist in the internal representation is a photograph where objects have been identified and stored in the object form for enhancing content retrieval. The actual photograph may be stored in high capacity devices like videodisks or directly addressable microfilm while the extracted information may reside in a disk and used for enhancing content addressability. An example of an image having only a raster internal representation is an uninterpreted photograph. An image having only an object form as internal representation can be an engineering design. (At the presentation level however, the object form may be used to display the design in a raster display.)