Digital Libraries Research at GMD-IPSI: Accessing Multimedia Documents by Knowledge Discovery Methods and Intelligent Retrieval

Reginald Ferber

GMD-IPSI (Integrated Publications and Information Systems Institute)
e-mail: ferber@darmstadt.gmd.de
http://www-cui.darmstadt.gmd.de/mind/

PART 1: Information Retrieval

A typical session with a conventional bibliographic database consists of a series of searches each based on the results of previous attempts. During this interaction the user elaborates her query and probably also her information need. While in this process the user is involved in many different cognitive actions, the only tasks of the IR system are to search for documents, to show documents and to store previous queries. For many inexperienced users the cognitive load for managing the search and scanning the documents found is rather high. Consequently they are not able to use all possibilities the system offers and they are often not satisfied with the results of their search.

1.1: Our View on IR

In our view Information Retrieval is a process that involves negotiation of the users information need and should be seen as a dialog between the user and the IR-system. Within this approach the goal is to move as much as possible of the cognitive load from the user to the IR-system.

This requires the analysis and modelling of the search process. For this purpose we employ a model of the dialog that is based on a speech act analysis and decomposition (Sitter & Stein 1992, Stein 1995), a rule based domain model to integrate external knowledge, and an abductive inference mechanism.

This view of the search process is inspired by van Rijsbergen's paradigm:

> *"It is my claim that, to design the next generation of IR systems, we will need to have a formal semantics for documents and queries. This semantic representation will interact with other types of knowledge in a controlled way, and this way is inference!"*
>
> van Rijsbergen, 1989 p. 81

In what follows we will discuss some of these points in more detail.

PART 2: Automated Indexing

Indexing is a specific kind of content extraction from documents. The natural way to do this with digital documents would be automated indexing. Since automated "understanding" of texts is far from being realistic, one has to look for other methods to enhance automated indexing methods. One way is to use structural information of multimedia documents.

2.1: Using Structural Information of Multimedia Documents

Several kinds of structural information of textual and non-textual documents can be distinguished. Textual documents can be coded in markups with a logical structure like SGML, latex, or (at least partially) HTML and more layout oriented markups like PostScript or other printer languages. SGML can be complemented by domain specific DTD's. Other textual information is bibliographic records or intellectually assigned key-terms. Nontextual information can be images or graphical elements in HTML documents like buttons, markers, links, ...

2.2: Inference Network

To extract information from a document we use an inference network. The basic assumption underlying this approach is that a document is relevant to a query, if the query can be deduced or inferred from the document. The advantages of inference networks are that they use probabilistic estimates and multiple paths of evidence. In this way partial evidence from various sources can be combined to a document's overall estimation of relevance for a specific query. The documents of a collection can be ranked according to these estimated relevance values. Sources of information can be specific parts of a document, bibliographic information, facts extracted by specific formats or DTD's (like names and date of birth in a collection of biographies) and partial information extracted from images and graphics.

The basic architecture of the inference network is a directed graph with nodes for the documents on top and nodes for indexing terms and features on the bottom (see Figure 1). In between there are nodes that represent different parts or views of the documents or specific information extracted from structural information. Each node has a value that represents it's amount of evidence. This "belief value" can be updated as a function of the values of it's parent nodes (i. e. those nodes that are connected to it) and an external belief value. Paths in the graph start from the documents and lead through the nodes for parts, views and features to the indexing nodes. If a document is presented to the network the belief value of it's node is set to 1 and the evidence is spread along the paths through the network to the indexing nodes.

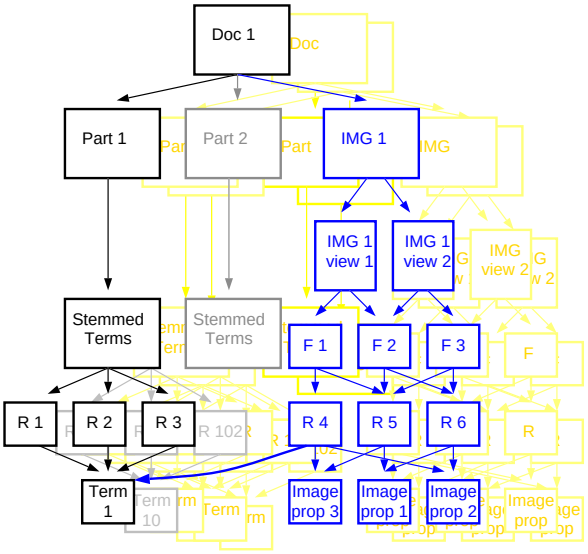
A special feature of our inference network MAGIC ("*Multimedia-based Automatic-Generation of Indexes and Clusters*") is a layer of rule nodes that is placed above the layer of indexing nodes. The rules associated to these nodes specify the roles of the terms in the documents. They can be used in the formulation of queries to specify the information need more precisely.

2.3: Query Processing

A query consists of a list of index terms, rules, and belief values. For each query the inference network constructs a list of documents ordered by estimated relevance. The estimated relevance values are calculated in the following way:

- the estimated relevance of a document is the sum of belief values of the nodes of the index terms in the list
- the belief values of an index term is the sum of the belief values of it's parent rule nodes
- the belief value of a rule node is zero if the rule is not in the list supplied in the query or if the rule is not satisfied. Otherwise it is a function of the rule nodes parent nodes and the belief value supplied in the query
- the belief values of a intermediate node is a function of the values of the parent nodes
- the belief value of the node of the document under consideration is 1, the values of all other document nodes is 0.

Figure 1: The architecture of MAGIC



(MAGIC = Multimedia-based Automatic-Generation of Indexes and Clusters)

On the left a path through the network for textual indexing is shown. The node "Part 1" may represent the abstract of an article, or biographic data of an artist. On the right a (planned) subnet for indexing of images is depicted. "View 1" and "View 2" can be exemplars of the image using different resolutions or a black and white vs. a color representation of a picture. "F 1" to "F 3" represent basic features of the picture like color distribution, texture values, or fractal dimension. "R 1" to "R 6" represent the rule nodes. At the bottom are the indexing nodes.

If no rules are given a set of default rules are applied.

2.4: Image Processing

Most IR methods for texts use words as atomic units of content. One of the main problems of image retrieval is that there are no such atomic units for the content of pictures. There are some features that can be determined for pixel images like texture, color histograms percentage of pixels in contours, and fractal dimension, but it is up to now not clear how they can be used for content based indexing.

On the other hand there are some basic "content features" that can be assigned to images by people like light (natural vs. artificial), percentage of shadow in a picture, source of picture (photograph, painting, ...), objects on the picture (natural, artificial, ...), or dimensionality of the objects. Our approach is to use a sample of manually classified pictures to predict the basic content features based on the basic pixel features. This can be done using statistical methods like discriminant analysis or methods from Machine Learning and Knowledge Discovery. The rules obtained by these methods can be integrated into the inference network, as shown in the right part of Figure 1.

PART 3: Advanced Retrieval Techniques

3.1: Abductive Retrieval

In the section on indexing we stated that a document is relevant to a query, if the query can be inferred from the document. This can be called the *document view* of a query. On the other hand a query can also be seen in the *information-need view*: a user deduces a query from her or his information need. In both cases the query is the consequence of a deduction process. What is searched for are the reasons or premises for this deduction process.

Abduction is an inference process that starts with a consequence and produces possible reasons for that consequence, based on the rules available to the system. Abduction can be used for both views of a query to generate hypotheses about it's background: In the information-need view abduction offers hypotheses on the users information need based on the knowledge of the system, in the document view abduction identifies documents that imply the query. In this way abductive reasoning allows us to infer possible interpretations of a query and to negotiate these interpretations with the user. The rule base is a way to integrate domain knowledge into the retrieval system.

The abductive approach is implemented in the MIRACLE system (see also Müller & Thiel 1994). This prototype operates on a sample of biographies from a dictionary of art. When a user enters a query the system generates the possible interpretations of that query based on a set of domain rules. The user can select a preferred interpretation or browse the documents found by these interpretations. MIRACLE uses MAGIC to index documents.

3.2: Knowledge Discovery

Apart from the selection of an appropriate interpretation of a query, it is often necessary to expand a query by relevant terms or to replace a term by a more specific one to enhance the search results. This can be done using thesauri. Such thesauri can be constructed automatically using methods of Knowledge Discovery: i. e. by extraction of understandable regularities from databases or corpora. Such methods have the advantage that they can be applied to documents of a specific domain. The regularities found are then specific for that domain. Such methods can also be applied to generate probabilistic rules for an inference network and to establish relations between institutions and topics, for example between repositories and the domains of interest they cover.

A simple method for finding regularities in textual documents is the use of co-occurrences. Associations between terms can be estimated by the deviation of their rate of co-occurrence from the rate expected in the case of statistical independence (see also Ferber, Wettler & Rapp 1995). A system that extracts such associative relations between terms is implemented in the prototype IMAGINE (*Interaction Merger for Associations Gained by Inspection of Numerous Exemplars*). It was applied to names occurring more than 30 times in a sample of some 14 000 documents of a dictionary of art. Figure 2 shows the terms associated to the term "Delos". It is worth noting that the relations are specific for the domain of an art dictionary (and not for the domain of computer science).

Figure 2: Names associated to "Delos" based on a Dictionary of Art with the IMAGINE System

Delos: 278.00	Cyclades: 48.65
Eretria: 139.00	Ganymede: 47.66
Kore: 130.82	Acropolis: 45.49
Pergamon: 127.73	Athena: 41.02
Pella: 121.62	Giambono: 39.71
Delphi: 116.58	Hellenistic: 38.99
Olynthos: 115.83	Antioch: 38.34
Samos: 113.26	Pio-Clementino: 37.07
Artemis: 99.92	Chios: 36.26
Hera: 92.67	Faun: 34.75
Hermes: 80.71	Herakleion: 33.10
Olympia: 68.43	Archaeol: 32.51
Ephesos: 66.72	Stag: 32.08
Lysippos: 65.41	Athenian: 31.95
Aphrodite: 64.87	Pliny: 30.47
Agora: 64.15	Macedonia: 29.82
Eros: 61.78	Melos: 29.79
Dionysos: 60.43	Phoenician: 29.29
Praxiteles: 57.52	Stourhead: 29.26
Pausanias: 55.60	Alexandria: 29.12
Orientalizing: 55.60	Pompeii: 28.76
Attica: 55.60	Attic: 28.76
Naxos: 53.81	Aigina: 27.80
Macedonian: 53.53	Beirut: 27.80
Dionysiac: 53.46	Herakles: 27.80
Zeus: 50.55	Rhodes: 27.12

In this example names were defined as terms that appear twice as often with a capital letter at the beginning than with a lowercase letter and are not found in a list of first names. The numbers given after the names are the strengths of the associations to "Delos".

PART 4: Digital Libraries and IR

Besides the general advantages of direct access and immediate delivery, Digital Libraries can offer additional benefits for effective IR techniques:

- machine readable documents can provide much more structural information (like logical markups) than paper documents, and this information is much more accessible for automated processing
- the bibliographic information is available on-line and will probably not be copyrighted. Thus it can be used for Knowledge Discovery processes
- the access to documents can be monitored and can be used as feedback information to optimize indexing and search processes. For example in the Dienst software access is performed in three steps: first only the titles are shown, then the user can ask for the bibliographic information including an abstract and finally she can view and download the complete document. It is likely that from this screening process valuable information for specific queries can be drawn

Of course there are also severe problems to be envisaged for a distributed Digital Library system. For example:

- due to the heterogeneity of such a system there will be no unified publishing policy across servers. Whereas some institutions will keep high standards of quality for publications other will strive for fast and exhaustive availability of information
- it will be hard to come up with unified structures (and the unified use of these structures) for documents and bibliographic records. This is especially the case if indexing and bibliographic processing are done by the authors and not by library professionals
- other problems will be the control of versions and updates of documents

Some of these problems can be attacked by the development of services that automatically monitor the collection and it's use and extract information to support users in their search. Such systems can be individualized for single users. For example we are developing a system that is able to infer potentially interesting servers from previous queries and their results. Further developments may use descriptions of servers (like the "Conspectus" concept under development at the University of Michigan) and information automatically extracted from sample documents of individual servers.

References

Ferber, R., Wettler, M., & Rapp, R. (1995). An associative model of word selection in the generation of search queries. *Journal of the American Society for Information Science (JASIS) 46*(9), 685-699.

Müller, A., & Thiel, U. (1994). Query expansion in an abductive information retrieval system. In *Proceedings of RIAO'94* (1994), 461-480.

Sitter, S., & Stein, A. (1992). Modeling the illucotionary aspects of information-seeking dialogues. *Information Processing and Management 28*(2), 165–180.

Stein, A. (1995). Dialogstrategien für kooperative Informationssysteme: Ein komplexes Modell multimodaler Interaktion. *Sprache und Datenverarbeitung 19*(1), 19–31.

Van Rijsbergen, C. J. (1989). Towards an information logic. In *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval* (1989), 77-86.