Expanding a Hypertext Information Retrieval System to Incorporate Multimedia Information

Rui Neto Marinheiro[1], Wendy Hall
The Multimedia Research Group
Department of Electronics and Computer Science
The University of Southampton, England, SO17 1BJ
E-mail rm95r@ecs.soton.ac.uk

Abstract

The integration of hypermedia with information retrieval is usually used to overcome user disorientation problems. However previous approaches have generally considered only text retrieval [7] *and if real content based multimedia information retrieval capabilities are used* [12], *effectiveness and efficiency are very low, when compared with text retrieval.*

This paper describes the development of a hypermedia information retrieval system that uses automatically created text descriptors to obtain multimedia information retrieval capability, using text retrieval techniques. There are two main aspects under consideration:

- Automatic creation of text descriptors for multimedia documents, using information taken from the hypermedia network, at the document and abstract classification level.

- Integration of open hypermedia systems, in this case the Microcosm system [4], *with text information retrieval tools to allow multimedia retrieval using text descriptors.*

At the end of this paper we present some conclusions about the effectiveness of such a system and improvements are suggested for further research.

1. Introduction

Disorientation in hypermedia systems is a problem that must be avoided and many previous systems have attempted to find a solution by integrating hypermedia systems and information retrieval systems. These systems are related in terms of the similarity of information they contain but they have a different way of navigating through it. Integrating these two types of systems is possible, in order to compensate the disadvantages of one system with the advantages of the other. A hypermedia system may allow the user an independent non-sequential reading of the information, that is, the freedom to browse through the information. But this freedom may not be so beneficial if the user is not able to find the relevant information that he/she seeks. So, if a usable hypermedia system is required it is necessary, in addition, to have a good way of providing clues to relevant information, and this usually leads us into the area of information retrieval. Traditionally, in these systems, a way of "linking" to the information is offered after the user makes an analytical query that he/she specifies in accordance with his/her search strategy. By this mean, it is possible to satisfy the information needs of the user by giving him/her more goal oriented tools. The most interesting aspect is that this integration not only diminishes user disorientation problems, but provides new ways of viewing and changing applications and improving retrieval recall by considering all the information associated in the hypermedia network of the application.

The integration between these two systems depends on how the researcher views the retrieval of information and the aspects of navigation, and at which level that integration must be implemented. In some information retrieval systems a two layer model is adopted, separating the document level from the index/abstract level. However, on integration, there are usually different ways of viewing the hypermedia network in relation to the information retrieval tool: at the document level, at the index level or at both. If the hypermedia links are seen as only relations between nodes at the document level, then hypermedia information is used to improve the definition of the index level and the search quality of the retrieval mechanism. If hypermedia is seen as a means of interconnecting indexes, that is, as a way of navigation at the index level to reach documents, then hypermedia is considered more as an interface tool and as a way of semantically relating indexes for the information retrieval tool.

In some works [1, 2, 3], hypermedia functionality was used to retrieve documents through navigation using the concept of index classification space. The cognitive

[1] This research was supported by PRAXIS XXI - PORTUGAL under grant BD/5399/95

overhead required by the user to operate the information retrieval mechanism is diminished by this kind of integration, since it is not necessary to implement a detailed retrieval strategy. In these approaches, it is the hypermedia functionality that is used for the benefit of information retrieval systems.

On the other hand, in different works, hypermedia information at the document level is used as a way of improving text retrieval [8, 7], and or as a suggestion for the retrieval of multimedia documents [7]. The information retrieval mechanism is then used for example to give a starting node for navigation at the document level, to suggest a next node of navigation, or to construct a guided tour [9]. Thus, the information retrieval uses hypermedia information but it is its functionality that is applied to improve browsing.

Many of these systems are restricted to the text world, and thus cannot really be called hypermedia systems, or if they are hypermedia systems with content based multimedia information retrieval tools, these tools are usually time consuming and consequently not efficient to use in PCs [12], and sometimes with a disappointing lack of effective retrieval. For these reasons it is possible to see that there is a need for efficient and effective information retrieval techniques to use with hypermedia systems.

One of the most important contributions was made by Dunlop et al. [5] and further developed by Harmandas et al. [11] who introduced the automatic creation of non-text media descriptors to allow efficient and effective multimedia retrieval. Of course, text descriptions for non-text media could be built in directly by human indexers, but this is a time consuming job and is also a technique that is very much based on the indexer's personal understanding of the media.

Our approach attempts to retrieve multimedia files through the use of text descriptors but our approach to building the description is however different from the one used in Dunlop's and Harmandas' work. At the document level, information such as link anchor selection, link type and link description of the document is used. To avoid the problem of the lack of links in some hypermedia applications, logical type classification, brief authored text description and document keywords at the index/abstract level are used as well.

In the next section we discuss a number of methods for integrating information retrieval techniques with hypertext/hypermedia systems. The third section describes the open hypermedia system used in our approach and in section four, the techniques for the text description of multimedia files in this open model are presented. In the fifth section, integration with an information retrieval tool is presented. In the last section some results and conclusions are presented as regards problems and achievements, with some suggestions for implementation for the future.

2. Reference to previous efforts of information retrieval integration with hypermedia systems

Different approaches have already been implemented by many research groups. Generally they have tried to use information in nodes or in typed or semantic links to improve information retrieval and navigation through the hypermedia/text network. One of the earliest to apply this concept was Frisse [8] with his hypertext medical handbook. In this work, the information contained in the handbook was divided into individual small fixed size cards using a hierarchical structure. In order to preserve the top down hierarchy structure of the handbook, links were made between each card and its parent. Labels were also associated with every card to make navigation through the hypertext easier. With a hierarchical structure, it can be difficult to move across the hierarchy and to find the relevant information, because at times the structure is not apparent to the user. Therefore an information retrieval process was designed in which a starting card, the starting point for browsing, was created following a user query. After the creation of this starting card, the user was free to browse through the hypertext.

In Frisse [8], the information retrieval was based on a statistical approach, where the results were dependent on the hypertext structure. This dependency was achieved by basing the score (importance of relevance) of each card on a statistical comparison between the query and the card itself and on the scoring of the immediately following cards.

The hierarchical inter-dependence of this system is a facet to be considered because it connects the information in a tight way. However, the retrieval of interdependent nodes was pioneering work. As the author states, more work is to be done on a network structure.

Li [13, 14] also developed a method for integrating information retrieval and open hypertext systems.

In his approach, Li used classical methods of statistical retrieval [16] by pre-indexing the content of all text nodes in order to achieve a quicker response time for retrieval using well-known algorithms. But it must be pointed out that he improved retrieval by introducing concepts such as break words and developing further the use of phrase weighting. Unfortunately his approach has the major disadvantage of being static, i.e. it is only efficient in hypertext systems if there are no changes in the nodes, as often occurs in other information retrieval systems. If a new file is inserted, removed or changed, into the

hypermedia collection of documents, all the pre-indexed information must be recalculated for all nodes.

One good advantage of this proposed approach however is that the information retrieval is considered as a different kind of link - a dynamic computed link - making the retrieval more transparent to the user. In this way the retrieval is no longer considered an independent tool.

Previous approaches only considered text information retrieval in hypertext/hypermedia systems. Although the work carried out by Frei et al. [7] was also based on a hypertext system, they used the semantic content of hypertext links to retrieve information, in the hope that this philosophy could be applied in future hypermedia systems.

They proposed a system which improves information retrieval by considering for retrieval not only the content of the nodes but also the information stored in the semantics of links. The content of each semantic link was associated with a link description as well as a source and destination anchor and a set of structured link attributes such as creation time and author name. The description takes into account the neighbouring nodes of the link, but if the author (or the user during browsing) thinks that it is important, the description may be extended or completely changed.

To retrieve information, two different methods of searching are considered: the exhaustive search, which helps a user who does not know much about the hypertext collection (this help is given by means of indicating some interesting starting points), and secondly the navigational search, which helps the user navigate through the hypertext collection until the next interesting node is reached.

The suggested system incorporates good techniques such as the automatic construction of descriptors that can improve the quality of retrieval, helps the user on navigation, and, very importantly, allows the retrieval of non-text media. Because of the physical amount of the information kept in the semantic links, problems in storage may arise in a large hypertext system, and these problems will increase if the system has a considerable number of links. On the other hand, if the system is sparsely connected, it is necessary to use greater maximum analysis distances, increasing the time for retrieval to an unacceptable level. However, this is one of the systems that takes the hypertext network more into account without too much extra effort on the part of the author.

Currently there is no efficient system that can retrieve non-text information as well as text information. Several efficient (and some of them effective) retrieval techniques have already been developed for text information that use natural language sentences as a query to retrieve the relevant text documents. However, as far as other media are concerned, it is easy to conclude that there is yet a long way to go to find a feasible method. This is because text information is already 'encoded' [15], i.e., the information is already segmented into a discrete set of symbols such as words. This characteristic facilitates the indexing and classification of information and therefore facilitates the efficiency of retrieval. Since with non-text files, such as pictures, the information is unencoded, the apparently simple techniques of retrieval from text information cannot be applied.

Lewis et al. [12] proposed an approach - MAVIS - that allows content based navigation and content based retrieval using a hypermedia link service for documents of any media type. With the introduction of the idea of signatures, rather than anchors as in hypermedia systems, it is possible to expand the model of navigation and retrieval. To navigate and retrieve information, signatures were compared, for example: text, colour, shape, etc, rather than the content directly. This pre-processed information improves the efficiency of the system. The structure of the system is generic enough to allow the expansion of models and permit improvements, although direct analysis of non-text media still has a long way to go to reach the efficiency of the text medium. The number of models is still limited and it is difficult to represent non-text media properly. Lewis et al. [12] pointed out that prior knowledge of the media may be important to improve the efficiency of navigation and retrieval, but such an implementation will be application dependent

Dunlop et al. [5] presented an important step forward in the use of text information retrieval techniques to retrieve multimedia information in a hypermedia system, using text retrieval techniques. The early work of Frisse [8], in which the retrieval was dependent on the structure, and previous work that used text descriptors for multimedia retrieval presented the case that multimedia retrieval could be possible by analysing the linking network. Since a non-text node can have several links to and from it, that start at text nodes and finish at text nodes, then it is possible to build a text descriptor node, based on the content of the neighbouring nodes. This will allow the indexing of text information, and therefore the implementation of relatively non-complex computational retrieval mechanisms. This is based on the assumption that the neighbouring nodes would be related to the described node with a common information topic, since links usually connect related information.

To test this theory, the retrieval efficiency of a hypertext application was compared using normal indexing techniques based on the content of all nodes with the same hypertext application where a few text nodes were replaced by descriptor text files. In this way they had a means for comparison of retrieval using text descriptors with retrieval using the content of the nodes. The results

show that the retrieval efficiency using text descriptor files is not that much less than the efficiency with the original nodes. However that could not happen if the system were not rich in links. The quality would degrade when only one or two links existed from, or to, the described node.

3. The Open Hypermedia System Approach

Microcosm is an open hypermedia system developed at the University of Southampton. In the Microcosm model [6, 4, 10] there are two ways of navigating through the information: navigation on the abstract level and hypermedia navigation on the document level, as can be seen in Fig 1. At the index/abstract level, the navigation is made through the classification of the documents until the relevant document is reached. The information about the classification of these nodes is kept in a separate database that is managed by the Document Management System: information like logical index, document description, keywords, author name, etc. It is therefore the use of all this information that allows abstract navigation in a file manager style where directories are logical indexes and where the files are identified by a text description.

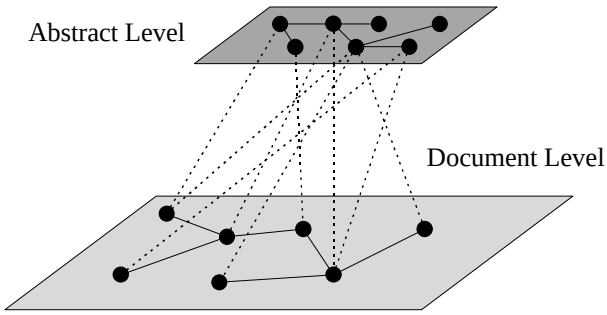


Fig 1 - The two layer model for navigation on Microcosm

On the other hand, the basic Microcosm model presents different kinds of links at the document level that allow the author, or the user, to relate information:

- *Specific links*, a tight connection between two fixed anchors. This link may be operated by a button highlighted by the system to indicate its presence. This type of link is important to create the back-bone link structure of the application at the document level that will allow the user to navigate between different items of information that may be obviously related or not
- *Local links*, a link with a fixed destination anchor, but with a source anchor, in a particular document, dependent on rules such as pattern matching, i.e. any occurrence of a particular object such as a text string. This kind of link is important for the definition of words or names, to give examples, or to show obviously related information, valid inside the node.
- *Generic links*, a link with a fixed destination anchor, but with the source anchor, in any document, dependent on rules such as pattern matching , i.e. any occurrence of a particular object such as a text string. This kind of link is important for the definition of words or names, to give examples, or to show obviously related information, valid for all the application.
- *Dynamic computed links*, a link with a computed destination anchor, and a selected source. This last kind of link was described by Li [13, 14] in order to integrate the Microcosm hypermedia system with some information retrieval techniques and suggests starting points for navigation.

With the first three types of links in addition to the link functionality type, other information is kept, such as text link description showing the motivation for the link, the selection anchor of the source and destination of the link, etc. All this relevant information is used for the maintenance of the document links and to help the user in document level navigation.

Microcosm is an open hypermedia system since it does not use any mark-up link information in the document data files and because it can use data that comes from a variety of third-party applications running under the operating system, among other facilities. The lack of mark-up within the data, is achieved by keeping all link information in link databases. So it is possible to have different sets of links applied to the same information, allowing the co-existence of an author's link database (linkbase) and several user linkbases that each user can add to, independently. The integration with third-party applications, is achieved by using a model in which a number of autonomous processes communicate with each other by messages. The message passing system also allows the development of new tools that may easily be integrated into the system. Microcosm also separates the front end of the open hypermedia system from the back end. The front end consists of viewers and is responsible for user interaction. The back end consists of a chain of filters (processes) and is responsible for many of the operations requested by the user. Connecting all this there is a "dipole" Document Control System - Filter Management System that is responsible for managing and integrating the whole system, as shown in Fig 2. Even with third-party applications without message passing capabilities it is possible to integrate the hypermedia link service provided by Microcosm, using the clipboard.

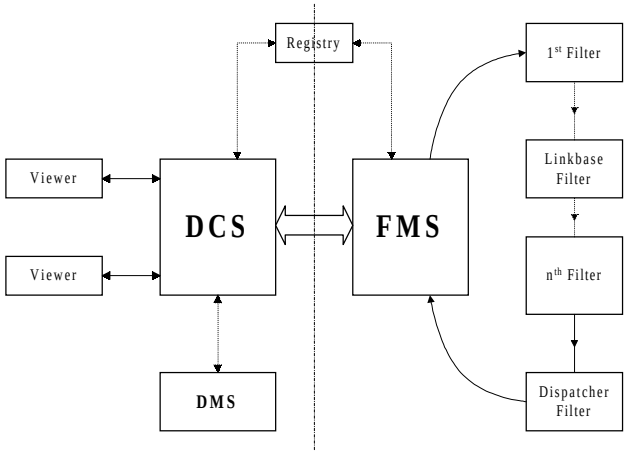


Fig 2 - The communicating processes that build up Microcosm system

4. Text description of multimedia files using context information from the hypermedia network

In many hypermedia systems that use information retrieval tools, such as Microcosm [4, 14], to reach non-text information through retrieval one has to hope that when text nodes are retrieved, it is possible to get to non-text nodes (related to the text nodes) through navigation from the retrieved text nodes. This assumption is based on the fact that usually hypermedia links connect related information.

As we have seen, Frisse [8] in his hypertext medical handbook developed the concept of retrieval to aid navigation, using the dependency of the retrieval on the hypertext structure to improve the retrieval. In further work, Frei et al. [7] used not only text files but also semantic links, and, as has been seen before, the semantics links were built with keywords retrieved from the source node and destination node of the link. This system considers not only the nodes but also considers the link network between the nodes. These previous approaches support the conclusion that the hypertext network at the document level is something to be considered for supporting information retrieval.

As was proposed by Dunlop et al. [5] the existence of links could also be used to produce a computed text descriptor of the non-text node that could permit it to be retrieved directly from a text query. In their approach, they applied the cluster (set) centroid algorithm to calculate a descriptor that would be the average meaning of all text nodes represented in the set, and as they claimed to prove, that descriptor represents the non-text node in a reasonable way (see Fig 3).

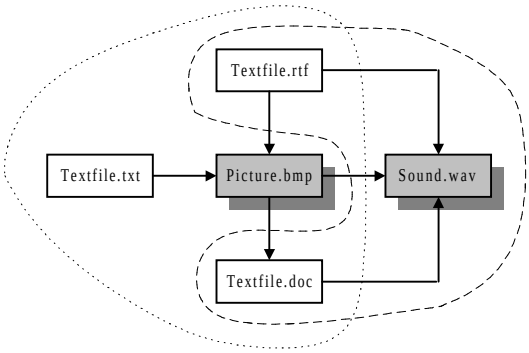


Fig 3 - Links between nodes aggregate them on a set of related information

We propose extending this model of text description to represent multimedia information by using more than simply content information from neighbouring nodes. Using the open hypermedia system Microcosm [4], where there is much more information available about the hypermedia network, a more precise text description of multimedia information is possible. An improved description can be achieved by extracting the relevant information from the hypermedia application at the document level but extending the Dunlop et al. [5] approach and extracting the relevant information about the documents that is maintained at the abstract classification level for user browsing.

4.1. Extending the model for the text description of multimedia information using information from the document level

As a first step we can produce some improvements by refining the way the content information is extracted from the neighbouring nodes of the multimedia file being described. In our approach the neighbouring nodes of a multimedia node are the ones connected with at least one direct link to that multimedia node.

In their extraction operation Dunlop et al. [5] took into consideration all the information kept in the neighbouring nodes and not simply the contextual information in the node that could be related to the file being described. However, Salton et al. [17, 18] suggested a different approach to retrieve information from text nodes through decomposition of the text in 'text segments' and 'text themes'. It seems reasonable that a text segment, or theme, could be much more relevant to the construction of the non-text node descriptor than the whole text node, since it is possible to have in the same document (node) a lot of information that may not be related to a particular destination node of a link that starts from a marginal sub-topic of the original node.

With this previous idea we can consider a different integration in the Microcosm open hypermedia system [4] when extracting content information. So, instead of considering all the information kept in the nodes, as Dunlop et al. [5] did, it could be better to consider only the segment(s), or the theme(s), parts of the nodes that have links to the non-text node to be described. In this way the quality of the descriptor would increase and as a result the information retrieval efficiency would increase as well. The Salton approach is however time consuming and in large applications this might be a significant disadvantage. We will then consider the information (text) kept only in the link anchor selection to build a more reasonable descriptor.

Further development may also be achieved by considering the phrase or the paragraph where the anchor is located. This is justified by the work developed in Harmandas et al. [11] on the World Wide Web. There, they exploit the link nature of the Web to build the text description. This is achieved by using the information from the image caption around each image, the image caption of neighbouring images, the full text where the image is inserted and the full text of other links to that page. They show that considering the full text where the image is inserted is the option that gets the worst recall/precision on the retrieval, and that the information that most improves recall/precision is the image text caption. Nevertheless, for now, it is possible to suggest that the anchor information will be more closely related to the described node than all the text information stored in the text file.

Another way of extending the model of text description is possible by considering other valid information that is available in an open hypermedia system such as Microcosm. In Microcosm there is a small text description about each link, and this kind of link description may also help in the building of the file descriptor for each non-text file, in the same way as the link anchor will. Frei et al. [7] have already used link description to improve recall/precision with good results mainly with author description. However in their work the use of the link description for information retrieval required a time consuming computational model. But since we will only use the link description for the construction of the text descriptors then this problem will not emerge.

Finally another way of extending the model of text description can be achieved by distinguishing between different kinds of links. An advantage of Microcosm is that there are different kinds of links for different functionality. Usually, the generic links and the local links are used more for definitions, examples, and accessory related information, and specific links are preferentially used for the back-bone link structure of the application at the document level that will allow the user to navigate between different items of information that may be obviously related or not. It is then clear that if the links relate information in a different way, it should be possible to differentiate the use of the information associated with each link in the description file.

Generic links usually relate information that is semantically closer to the content and are not so much dependent on the browsing information, whereas specific link may give more information about the hypermedia network functionality of an application, and not so much about the content. We then have the possibility of choosing a weighting, dependent on the type of each link, which will allow us to attribute more relevance to generic links. This weight will be then the number of repeated times the information referent to a link will be present at the descriptor file.

4.2. Extending the model for the text description of multimedia information using information from the abstract level

Until now we have only considered the construction of the text descriptor from the information on links that the system has at the document level. A lack of links at this level may degrade the retrieval efficiency of non-text media, as was shown by Dunlop et al. [5], and sometimes the links to images are not relevant or relate information in a non descriptive manner, as discussed by Harmandas et al. [11].

There has been a significant amount of research on “hypermedia like navigation” through a semantically connected set of concepts at the abstract level [1, 2, 3]. In their work, multimedia information retrieval was possible with the same efficiency irrespective of whether the media to be retrieved was text or non-text. The classification of documents against a semantically connected thesaurus could allow retrieval through the concepts of classification. In works like Arents et al. [2] the creation of trails through the information using semantic coupled hyperindexes has also been suggested.

In Microcosm further valid information is kept at the abstract level and this suggests its use for the building of the text descriptor. To implement navigation in Microcosm at the abstract level the author is usually required to create a small text description of each file, some keywords, and its classification under an abstract concept - a logical index. Thus, this important information can be used in the same way as the information taken from the links to build the text descriptor. The new system will then always have a small text description that allows the retrieval of non-text media, independent of whether there are any relevant links.

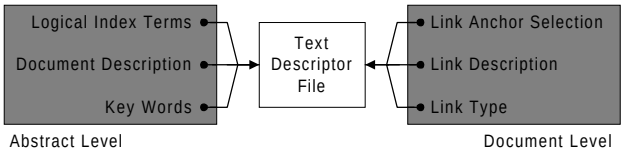


Fig 4 - Information that is used to build the text descriptor file for a non-text node

In summary, for the construction of the text description files, our approach uses not only the information kept at the document level, as the majority of research has done until now, but also information maintained at the abstract level, improving the possibilities of better retrieval. Another novelty is the consideration of different link types that allow the distinction between different kinds of information taken from the document level and from the abstract level, as shown in Fig 4. In the formula bellow used to build the text descriptor files, it can be seen that it is possible to change several parameters to improve the retrieval quality of the information retrieval tool to integrate with the Microcosm system:

$$l_g \sum_{L \subset Ge} \left(a \cdot L_{iS} + b \cdot L_{iD} \right) +$$

$$\sum_{n \subset N} \left(\begin{array}{l} l_l \sum_{L \subset Lc} \left(a \cdot L_{iS} + b \cdot L_{iD} + c \cdot n_{jD} + d \cdot n_{jA} \right) + \\ l_s \sum_{L \subset Sp} \left(a \cdot L_{iS} + b \cdot L_{iD} + e \cdot n_{jD} + f \cdot n_{jA} \right) \end{array} \right)$$

where:

L___ link

Ge _ set of generic links connecting to the non-text node

Lc__ set of local links connecting to the non-text node

Sp__ set of specific links connecting to the non-text node

L_{iS} _ link anchor selection on link i

L_{iD} _ link anchor description on link i

l_s___ integer weight given to the specific link

l_l___ integer weight given to the local link

l_g___ integer weight given to the generic link

n___ node

N __ set of text neighbouring nodes around on the non-text node to be described

n_{jD} _ description of node j

n_{jA} _ logical index of node j

a…f different user /author integer weights

Besides the different weighting depending on different link types, that has already been justified, there is also the possibility of changing the relative weighting between information coming from the node description, the node logical index, the link description and the link anchors. This allows the system to have different text description file construction strategies for different applications. Some hypermedia information in Microcosm is author/user dependent and its importance on the construction of the descriptor must reflect the author/user behaviour.

5. Integration of an open hypermedia system with text information retrieval to facilitate multimedia information retrieval.

In our work there were two main goals too be met when we considered the extension of the information retrieval tool to incorporate the retrieval of multimedia information within the open hypermedia system Microcosm:

-The first goal was to design an integration model that allows us to reuse an existing text information retrieval tool behaving like a “black box”. This allows for the future reuse of different text information retrieval tools that can easily be inserted into the modular architecture of the Microcosm system;

-The second goal was to integrate the information retrieval tools in a transparent way for the user/author, i.e. to do it in such a way that the user/author would be hardly aware of its existence, apart from times when it is necessary to change parameters. To implement this it is necessary that the system tracks down all the information about the hypermedia network that may change in the application and update the affected text descriptor automatically.

5.1. The reuse of an existing text information retrieval tool

In Microcosm, each file that exists in a specific application must be registered under the DMS (Document Management System). The registration information is kept in a database record that holds all the information necessary to control the file. The system permits the creation of a new field in that record and this allowed us to create one extra filed for each text descriptor file record, in a way that saves the physical reference of the file that is being described. The reverse situation was implemented as well by creating a reference field in the described file record that points to its descriptor. Consequently, there is already a physical reference association between the descriptor and the described file, and it is therefore simple to get one if we have the other (see Fig 5). This registration information allows the

retrieval of a non-text file when its text descriptor is retrieved by a text retrieval tool, as will now be shown.

In the third section we saw that in the Microcosm system the management of all information relevant to the link services is at the back end and is processed independently from the front end, where the viewers interact with the user. These back end link services are managed by a sequence of filter modules that communicate with each other by messages. Each filter process the messages relevant to the information that it is responsible for, and then pass on the processed message to the next filter, either in a new message or not.

One such filter in the filter chain responsible for the text information retrieval capabilities inside Microcosm is the Computed Links filter developed by Li et al. [13, 14] as described in section two. The output of this filter is a set of ranked text files from the set of all text files in the application. This ranked set of text files contains suggestions for a starting point for navigating at the document level of the hypermedia application, similar to that obtained through a text selection that the user has specified. For each suggested file a message with its reference is transmitted, that will be received by the Dispatcher Filter. This last filter will show the brief abstract level description of the suggested files, in the same order as the messages were received. Then, the user just has to select the desired file(s) to be shown.

In order to satisfy our first objective of integration, a new filter - the Descriptor filter - was developed and inserted into the Microcosm filter chain after the Computed links filter and just before the Dispatcher filter.

Since Microcosm messages are passed from each filter to another in a sequential order, the Descriptor filter, inserted in the filter chain after the Computed links filter, only has to keep track of the file registration information part of the messages that goes through it. In this way it is sure to catch all non-text files represented by descriptor text files. After the Computed link filter returns the text descriptor files or normal text files the Descriptor filter only has to check in the message for references to the described files. In the messages where this information is found, it is only necessary to replace them with a different one where the respective described file is returned (see Fig 5).

In this approach, the Computed Links filter does not have to be aware that it is retrieving a descriptor file on the one hand, nor does the Descriptor filter have to know how to retrieve information in response to a query. So, the independence of the two processes is assured. Finally, the Dispatcher Filter shows the user all the retrieved links, and between them links to every kind of media, for example: text, images, sound, videos, etc.

So, as it can be seen, it is possible to reuse an existing text retrieval tool, extending it to incorporate multimedia retrieval. In principle, the Computed links Filter could be replaced by any text retrieval tool that is able to receive and output messages. This approach benefits from the use of the file registration, the flexibility of the filter chain structure, and the existence of one filter that already retrieves text files.

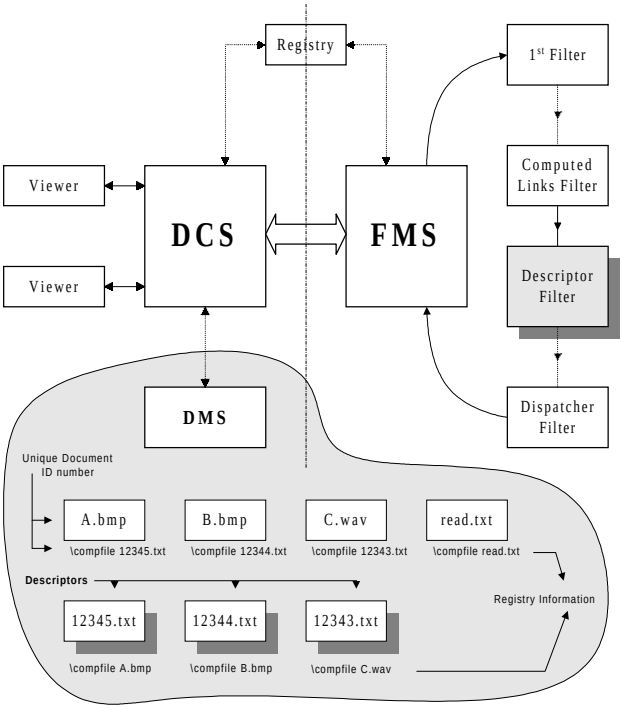
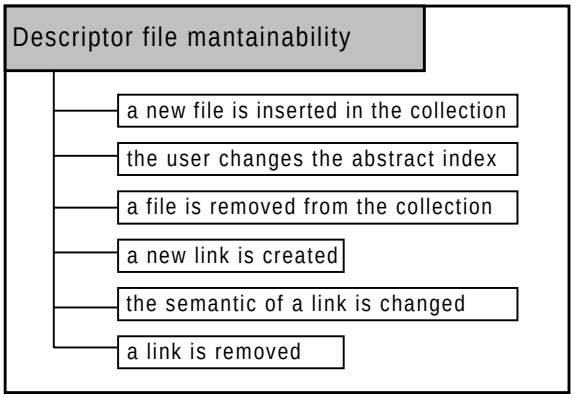


Fig 5 - Integration of the Descriptor Filter with the Microcosm System for the retrieval of multimedia information

5.2. The transparent integration of an existing text information retrieval tool

Now it is necessary to define how the Descriptor filter should behave in order to satisfy the second objective of transparency. One of the major problems with the Computed Link filter developed by Li et al. [13] was its static behaviour, i.e., actualisation of the content or structure of each application was not considered by the filter, unless the author specified it. In more detail, the Computed Links filter does not change the inverted file information necessary for retrieval when files are inserted or removed from the collection. So every time the user inserts a new file or removes an old one, he/she has to recreate the inverted file. To avoid a similar problem in this new filter, it is necessary to carry out some other procedures to allow a transparent integration, for the user/author, with the Microcosm system, see Table 1.

Table 1 - Situation to which the Descriptor filter reacts in order to maintain consistent descriptors



When a new file is created, a notification message is put onto the filter chain. So, after the Descriptor filter acknowledges that the new file has been inserted, it selects a proper name and creates the descriptor text file with the information provided by the document registration attributes. If this filter is inserted for the first time in the system, on the creation of the text description file, it also asks the linkbases about all the links that are pointing to the described file, and gathers all the text in the source anchors and descriptions and includes it in the text description file, and the formula described previously is applied. However, special care is taken in the storage of this information since all of it must be recognised by the Computed Links filter. Since only text information should be saved, separation between different parts of the text descriptor is achieved through the use of a different number of linefeeds. This is important to allow the distinction between the different links in order to permit their maintainability when they change. When a file is removed from the collection it is only necessary to remove the descriptor text file.

When a new link is created, the Descriptor filter has only to catch the message that created the link in order to withdraw the text selection and link description from the message and store it properly using the formula in the text description file, together with information taken from the registry concerning the linked files. If a link is removed, the text concerning the source anchor stored in the text description file is taken out and when the document registration attributes or the user description is changed, the text description file is also updated.

6. Results and conclusions

With this approach, some ad-hoc qualitative experiments were undertaken, since with the text information retrieval tool used (the Computed Links Filter) it was not possible to access the analytical information about the ranking of the retrieved nodes.

We decided to choose two different applications showing different strategies in the building of applications. In the first experiment we tried to study the situation where the author usually does not plan the construction of the hypermedia application, by considering the extra functionalities that the Microcosm system provides. In this application the logical type index and text description at the abstract level were not valid or related to the content of the nodes. At the document level the only valid information considered was the link anchor selection. In the second experiment, we tried to test the potential of the Microcosm system for the automatic construction of text descriptors by using a better planned hypermedia application with valid information both at the document and at the abstract level.

In the first experiment, the recall of multimedia files was not very high in subjects related to the one addressed in the non-text node to be retrieved. In the second one however, since we had considered information from the abstract level and used the link description, the recall was improved with the possibility of retrieving multimedia files even when non-text files had a small number of links connecting them.

With these first results, it is already possible to conclude that in applications developed for Microcosm the integration works well, but some improvements must be made in the future. In the ad-hoc experiments, we noticed that the text description is mainly dependent on two factors. The first concerns the quality of information used at the abstract level. This information improves the retrieval of multimedia information through the use of text descriptors if the author has built the applications in a planned way. If care is not taken in the insertion of information at the abstract level, that information can not be considered for retrieval. However, for a planned hypermedia application, the good results obtained justify a better study of the abstract level in the future.

The second factor is the number of links connecting non-text nodes with text nodes. In a poorly connected hypermedia application the efficiency of retrieval diminished. The consideration of segments (themes) as well as link anchors is therefore something that should be considered in the future, as discussed in section 4.1. But the use of information taken from the abstract level allows the retrieval of multimedia files even in a situation where the hypermedia application is poorly connected.

Currently, the integration of a more powerful information retrieval tool is under research. With this new tool it will be possible to get more analytical results, and then obtain improved results in the importance of the hypermedia network information such as link type and

description and, abstract level logical index and description, on the construction of descriptors.

7. References

[1] Maristella Agosti and Pier g. Marchetti, “User Navigation in the IRS Conceptual Structure through a Semantic Association Function”, *The Computer Journal*, Vol. 35, No. 3, 1992, pp 194-199

[2] Hans C. Arents and Walter F.L. Bogaerts, “Concept-Based Retrieval of Hypermedia Information: from Term Indexing to Semantic Hyperindexing.” *Information Processing and Management*, Vol. 29, No. 3, 1993, pp 373-386

[3] Daniel Cunliffe, Carl Taylor and Douglas Tudhope, “Query-based Navigation in Semantically Indexed Hypermedia”, *The Eighth ACM Conference on Hypertext*, ACM Press, April 1997, pp 87-95,

[4] H. Davis, W. Hall, I. Heath, G. Hill and R. Wilkins, “Towards an Integrated Information Environment with Open Hypermedia Systems.”, *Proceedings of the ACM Conference on Hypertext*, ACM press, 1992, pp 181-190

[5] M.D. Dunlop and C.J. van Rijsbergen, “Hypermedia and Free Text Retrieval”, *Information Processing and Management*, Vol. 29, No 3, May 1993, pp 287-298

[6] Andrew M. Fountain, Wendy Hall, Ian Heath and Hugh C. Davis, “MICROCOSM: An Open Model for Hypermedia With Dynamic Linking”, *Hypertext: Concepts, Systems and Applications (Proceedings of ECHT’90)*, Cambridge University Press, 1990, pp 298-311

[7] H.P. Frei and D. Stieger, “Making use of hypertext links when retrieving information.” *Proceedings of the ACM Conference on Hypertext*, ACM press, 1992, pp 102-111

[8] Mark E. Frisse, “Searching for information in a hypertext medical handbook.” *Communications of the ACM*, Vol. 31, No. 7, July 1988, pp 880-886

[9] Catherine Guinan and Alan F. Smeaton, "Information Retrieval from Hypertext using Dynamically Planned Guided Tours.", *Proceeding of the ACM Conference on Hypertext*, ACM Press, 1992, pp 122-130

[10] Wendy Hall, “Ending the Tyranny of the Button”, *Multimedia*, IEEE, Spring 1994, pp 60-68

[11] V. Harmandas, M. Sanderson and Mark D. Dunlop, “Information Retrieval by Hypertext Links.”, *http://www.dcs.gla.ac.uk/~sanderso/papers/my_papers/SIGIR97.pdf, To be published on the Proceedings of SIGIR 97*, 27-31 July 1997

[12] Paul Lewis, Hugh Davis, Steve Griffiths, Wendy Hall and Rob Wilkins, “Media-based Navigation with Generic Links”, *The Seventh ACM Conference on Hypertext (Hypertext’96)*, ACM Press, March 1996, pp 215-223

[13] Z. Li, H. Davis and W. Hall, “Hypermedia Links and Information Retrieval.”, *British Computer Society 14th Information Retrieval Colloquium*, Lancaster University, 13-14 April 1992

[14] Zhuoxun Li , “Information Retrieval for automatic link creation in hypertext systems”, *PhD thesis*, University of Southampton, Department of Electronics and Computer Science, 1993

[15] Desai Narasimhalu and Mun-Kew Leong, "Experiences with Content Based Retrieval of Multinedia Information", *Proceedings of the Final Workshop on Multimedia Information Retrieval (MIRO Glasgow ‘95)*, Ian Ruthven (Ed) , September 1995, pp 6

[16] C.J. van Rijsbergen, *Information Retrieval*, Butterworths, London 1979

[17] Gerard Salton, James Allan, Chris Buckley and Amit Singhal, “Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts” *Science*, Vol. 264, 3 November 1994, pp 1421-1426

[18] Gerard Salton, Amit Singhal, Chris Buckley and Mandar Mitra, “Automatic Text Decomposition Using Text Segments and Text Themes”, *The Seventh ACM Conference on Hypertext (Hypertext’96)*, ACM Press, March 1996, pp 53-65