Dynamic Background Libraries as an improved Way for Web-Based Learning using HIKS (Hierarchical Interactive Knowledge System)

A Working Prototype of a Knowledge Based Information Retrieval System used as an Dynamic Background Library for Hyperwave's WBT

Christian Gütl, Thomas Dietinger, Dietmar Neussl, Bernhard Knögler, Klaus Schmaranz
Institute for Information Processing and Computer Supported New Media (IICM)
Graz University of Technology
Schißstattgasse 4a, A-8020 Graz, Austria
E-mail: {cguetl, tdieting, dneussl, bknoeg, kschmar} @iicm.edu

Dynamic Background Libraries as an improved Way for Web-Based Learning using HIKS (Hierarchical Interactive Knowledge System)

A Working Prototype of a Knowledge Based Information Retrieval System used as an Dynamic Background Library for Hyperwave's WBT

Abstract

This paper describes in brief HIKS, a working prototype of a Web-based interactive knowledge system which could be perfectly used as a dynamic background library in a web based training environment. Experience in the field of web based learning at the IICM had shown that courseware and static background library only does not satisfy the courseware and the learners needs. A additional dynamic background library will provide additional knowledge which will always guarantee up-to-date background knowledge. Relevant knowledge spaces from the biggest knowledge store, the Internet, will be extracted by HIKS. The core of this system is a sophisticated information gatherer and knowledge area broker system which will be combined with a Hyperwave-based web based training system. This paper describes the technique of the gatherer and broker and its interaction to the learning process as well as to the possibility to produce new courseware. Then experiences from working with the system will be discussed. Furthermore a way to build up knowledge hierarchy to specific topics and co-operations between organisations is shown.

Keyword:

Web based training, dynamic background library, intelligent knowledge broker, information structuring, information relevance assessment, Hyperwave

1. Introduction

In the last two decades at the Institute for Information Processing and Computer supported new Media a lot of research has been done in the field of information and document management, computer based training, digital libraries and electronic publishing [Guetl98] [Maurer96b] [Marchionini95] [Maurer94]. The research results and intense co-operation with the industry even led to commercial products like e.g. Hyperwave [Maurer96a] [Kappe94]. Based on these experiences a group is currently developing GENTLE [Dietinger98] [Maurer97], an electronic lecturing system combined with a digital lecture library for teachers and students. Gentle has already been tested by hundreds of students at the Graz University of Technology. The feedback gained during the lectures is leading to continuous improvements of the system invoking as well as new research concepts.

One of the experiences we made was that only the introduction of new technologies like multimedia or hyperlinked objects doesn't necessarily lead to more effective and more efficient learning. Technical environment could support learners needs for interpersonal communication, the opportunity to ask questions and discuss problems with the tutors and co-learners. On the other hand technologies are also very useful and necessary for finding proper information and creating courseware.

This paper will discuss aspects of getting the proper information to improve the learning process as well as possibilities allowing the user to solve complex knowledge depending problems. A pure courseware alone does not sufficient the named requirements. Therefore a proper background library has to be provided. A static background library (a collection of electronic books and journals) alone will not satisfy the users' needs because it cannot deal with the growing stream of information. Therefore a further component - a dynamic one - has to be

provided. The obvious thing to do is to integrate a world-wide information repository, the information stored in the Internet. About 20,000,000 computers and more than 150,000,000 documents are publicly available at the time of writing. It represents the biggest information source mankind ever had but unfortunately it is also the most unstructured one.

One approach taken to handle chaos is the use of huge index servers like AltaVista or Yahoo just to mention a few. Although today's index servers provide at least a way of searching in the information jungle they do not solve the problems' complexity: First it can happen that queries are not narrow enough and results of 50,000 hits and more are quite usual. For this reason some sort of automatic ranking is performed by the index servers, mostly just by counting the number of hits in a document. Second it can happen that a query is narrow enough but still returns undesired results because keyword context information is missing.

Let us consider a simple example of somebody looking for information on computer viruses. The obvious query is to look for all documents containing both the word computer and the word virus. This query will also find medical documents on viruses if the author happens to use the word computer in it in a different context in the same document. On the other hand it will not find documents where a Computer scientist just uses the word virus implicitly assuming that a computer virus is meant. By using context information and an intelligent knowledge broker this situation can be solved quite elegantly as we will describe in the following sections of this paper.

To add a philosophical aspect - [Busch 97] brings the problem to the point- *"To obtain or process information people in an information society must have an important competence: they have to be able to structure information, to assess it and to make it available ... "*. Also [Rieder 97] addresses this deficit by saying *"Not only is the gathering of information demanded; this information must also have meaning ... "*.

These specified problems affect the learning process and the search of relevant information within a web based training environment. Learners as well as course authors will be frustrated in case of getting almost endless lists of ridiculous web sites by using common search engines. The basic idea is to combine the dynamic of brand-new information on the web and important additional information, the quality aspects. This concept will allow always up-to-date information in background library with the restriction that only high-quality and reliable documents will be provided.

The approach in HIKS is influenced by the philosophy as stated in [Guetl98]: *"Information harvesting and reliable search will reach increasing importance in our society, but we must change from the information society to the knowledge society"*. To make the step from an information base towards a knowledge base we are using Hyperwave [Maurer96] as the core information system for a web-based learning system. For the dynamic background system a combination of hierarchical gathers together with intelligent knowledge broker are built. In the next chapter the user requirements will be discussed in brief.

2. System Requirements from the Users' Perspective

Generally parsing and indexing documents is a 100 percent objective process but on the other hand present search attempts *"lead to almost endless lists of ridiculous web sites, which contain the searched words purely by chance but have nothing in common with the desired topics - hits are a pure matter of luck"* [Fichtner 97]. The question which arises is whether 100 percent objectiveness is really useful. From our point of view it is much more efficient to let subjective human expert knowledge find its way into the system too. Subjective in this context means classification of documents by topics, but also by quality, up-to-data'ness, relevance for concerned knowledge areas, etc. Subjectivity goes even further: users (courseware authors and learner) can define their own profiles including preferred language, areas of interest and skills, etc.

The introduced system HIKS combines document contents, extracted information and server or site information together with additional human expert knowledge and user profiles. This additional information is used by queries to narrow down results as well as it is presented to the user together with the query hits to give a better understanding how relevant a hit really is. Decisions on relevance based on objective and subjective criteria are made by the intelligent knowledge area brokers.

User profiles cannot only be relevant for queries and understanding of query results, there is another important point where profiles come into play - the use of background libraries itself. The use of background libraries (static and dynamic one) is also a result of user feedback from previous projects. The following describes the problem: query hits in highly specific knowledge areas very often contain words which the users simply do not know partially because they are in foreign language, partially because the users are laymen in the appropriate

knowledge area. Our first approach to solve this problem was the integration of searchable digital encyclopaedia's and dictionaries into the system. Utilising the power of user profiles there is a simple and efficient solution: Take it as an example: a French speaking experienced chemist looking up an English chemical term in the background library. There is no use to provide translations to other languages than French, so if not explicitly desired by the user only the French translation is returned. Furthermore there is no use to show a very basic explanation from an encyclopaedia for beginners in chemistry, so only high level explanations are given.

The current implementation of user profile management stores information given by the users themselves. Some future research will be done to implement a dynamic, self-learning user profile management system which adapts to the users' behaviour over time.

Besides profile management we already discussed expert knowledge integration above in this section. From the experts' point of view it has to be as easy as possible to insert categorisation and quality assessments into the system, otherwise this is useless in huge systems. For this reason the user interface of HIKS has to be adaptive to present different interface objects to authorised experts than to normal readers. Whenever authorised experts of a knowledge area download a document in their area they automatically also get a set of additional buttons for their quality assessments including style of writing, relevance for the research area, up-to-data'ness, etc. They also get the possibility to add textual annotations to the document or a part of it alerting readers of mistakes or new results in that research area. Any information gained from experts' opinions then become part of the document classification yielding a self expanding dynamic knowledge base over time supporting courseware authors as well as learners.

As a conclusion HIKS as a working prototype for a knowledge based information retrieval system has to gather all information it can get from different information sources as there are the document data itself, the documents' meta data (e.g. title, author, creation time, language, knowledge area, etc.) and the system's data (e.g. time of indexing, original document location, etc.). Further it has to integrate human expert knowledge and make it part of the meta data. Query results and user interfaces have to be adapted dynamically according to the users' preferences as well as relevance rating of search hits has to be provided. The following chapters describe the specialised gatherer hierarchy and the intelligent broker system needed to achieve this.

3. Extended Information Gatherer

To provide a solid base for the intelligent knowledge area brokers we first have to collect as much information as possible about indexed documents. The backend for achieving this is the extended information gatherer which is indexing documents from arbitrary many sites which are given by authorised users of the learning system. Besides simply indexing documents the gatherer also extracts descriptions, context information and other relevant information (quality, language, etc.) if available, for later use with the intelligent knowledge broker. As will be discussed later the gatherer is also the collecting point for expert knowledge input. The Gatherer itself is divided into several modules: content analyser, keyword extractor, description extractor and expert knowledge integrator.

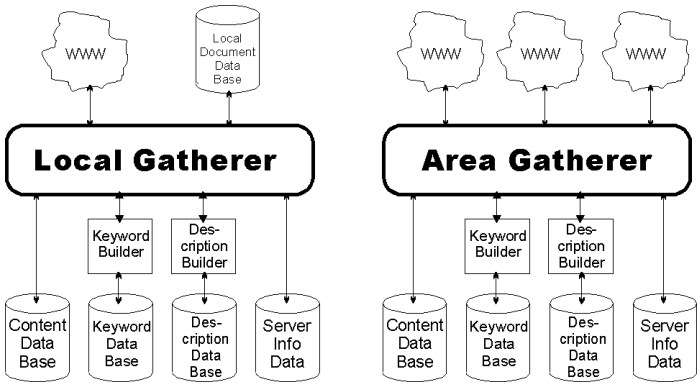

Figure 1: Architecture of the Information Gatherer

First of all gathered information is filtered by the content analyser which extracts meta information and recognises whether Java Applets, ActiveX objects, JavaScript, multimedia objects, etc. are embedded in a document. The content analyser also collects relevant standard data included in documents (title, headings, etc.). Second step in information gathering is the keyword extractor. It builds a list of keywords describing a

document. As we have seen in the computer virus example above this is certainly not enough. For this reason the keyword extractor also stores the context together with the keywords, in our case at the moment the paragraph where the word is located. This information can then be evaluated later in the intelligent knowledge broker to decide the relevance of documents. Some future work will be done to improve the context extraction and evaluation, but up to now this simple approach works very well.

Keywords and their context are not always enough to describe a document. Therefore the gatherer also consists of a third module which gets even more relevant information out of a document: the description builder. It tries to find authors and abstract in the paper if available. This approach is based on the fact that most scientific papers contain this information either in a section called "Abstract" or have it in the meta-information. If nothing like that can be found the headings of the document are extracted as they provide more information than the usual first few lines of the document. Only if no headings can be found the system fall back on extracting an excerpt of several lines at the top. Further work will be done in future integrating AI modules to find proper descriptions.

Another interesting point to consider about site locations comes from document update rates: Over time the gatherer modules can find out about update rates and the description builder can also add this attribute to a document to provide the user with the knowledge whether a document could already be outdated and a more recent version could be found on the original site. That is why the gatherers' behaviour is adapted to the update rates: servers with a long mean document life time are visited more seldom than others changing their contents often.

After running through all the steps of the gathering process the whole bunch of information is ready to be handled by our intelligent knowledge broker described in the following section. Some future work on information gathering will be done in the area of indexing not only Web sites but several different kinds of databases, e.g. Oracle or Informix, which are widely used to store information.

4. Intelligent Knowledge Broker

The intelligent knowledge broker in HIKS is both, the front end for presenting knowledge to the courseware authors and learners obtained periodically by its gatherers and the channel for bringing expert knowledge back into the system. An adaptive interface provides facilities for creating user profiles, formulating search queries and adding supplementary expert knowledge. Further there is not only one single broker in the system, but a hierarchical collection of several knowledge area brokers integrated transparently to the user. This concept allows to build up proper knowledge spaces for particular courses.

As shown above ordinary search services lack the ability to handle more user information than just the search string or Boolean combinations of strings. This often leads to unwanted, frustrating large, partly irrelevant results. By giving the system a user description (user profile), a query result can be narrowed down more precisely. For this reason the intelligent knowledge broker presents a form to the users where they can fill out their preferences. These include preferred language, foreign language knowledge, fields of interest, profession, qualifications, education, quality of expected document content, etc. In order not to be forced to fill out the form every time the user profiles are stored together with chosen nicknames for future use.

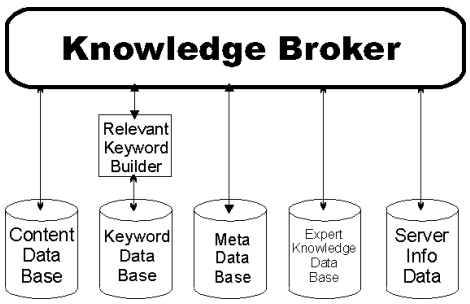

Figure **2: Architecture of the Information Broker System**

Present robots do not give information about a Web server, its owner, its purpose, location, etc. This often leads to questionable results. Our intelligent knowledge broker has access to additional information from the extended gatherers. Thus it allows a better categorisation of the documents by quality and relevance on one hand and semantic context on the other. So web servers from universities, public libraries and governmental departments

are categorised differently from private web sites. This concept allows much more exact search results by using this system at the learning process.

A single knowledge area broker analyses keywords and contexts extracted by the Gatherer to build relevance-based keyword lists. This can be done in quite a simple way as shown in the following example: let us assume a server dedicated to the research area of physics. Most of the documents will have the keyword "physics' somewhere in their body, therefore it cannot be considered a relevant description for the contents of a document. For this reason the classification of search hits will not contain the keyword "physics' when found on this server. On the other hand the keyword "medicine" will not be found in many documents, therefore it will be presented to the user to show that this document has some medical context. The threshold for considering keywords irrelevant is defined by the server administrator. This means that each knowledge broker will provide relevant keywords which depends on the whole set of the document base on its system.

Back again to our virus example it is obviously not enough to provide only keywords to describe documents, even if the keywords are considered relevant. The intelligent knowledge broker takes this into account by evaluating context and user profiles and providing additional features. A medicine student looking for "virus' is rather interested in articles dealing with virology than in computer science. We are not trying to implement some a priori behaviour of the broker, because this would be a highly complicated artificial intelligence problem. Instead HIKS implements a user dialog to make it easy to narrow down the result. In case of the "virus" example the medicine student is able to tell the system that having "virus" in the same context with "computer" is undesirable and the broker will narrow down the search. In order to prevent to be forced to enter the same restrictions every time, such behaviour can also be added to the user profile and is then assumed valid for this user from then on.

Besides keywords and keyword contexts the broker also handles levels of knowledge (beginners, advanced, experts) and desired quality of papers very much in the same way as described above. Users can simply tell the broker that no basic papers will be provided in specific knowledge area, but on the other hand not present very high-level papers in another area, etc.

There are several approaches to automatically categorise documents but until now these systems do not yield satisfying results. To supply a high quality human expert knowledge is necessary to classify documents. For this reason the broker provides an interface allowing a distinct group of authorised experts in specific areas of knowledge to link quality assessments and annotations. Since it is mostly not possible to simply say that a document is good or bad a more detailed and configurable quality assessment takes place. Authorised experts get "good", "medium' and "bad" buttons for different aspects like style of writing, depth of knowledge, seriousness etc. as well as the ability to add textual annotations.

Instead of a simple search result list, a more logically ordered representation of information objects is often desirable. The context categorisation of the knowledge system can describe the relations between relevant documents in a hierarchically structured way. So information retrieval can also be a step by step walk through knowledge hierarchies. Furthermore additional information like web site description, keywords as well as quality information, etc. will be provided by the system.

5. HIKS as a Cascaded Distributed Network System

Besides the aspects considered so far in this paper there is another very important requirement: it must be possible to build up a distributed network database instead of building one huge database on a single server. It is also desirable to import information from existing information gatherers and index servers or even treat them as subsystems. For this reason knowledge brokers can be cascaded and HIKS is conceived for building up sub systems and putting them together to huge knowledge bases. One major advantage is that specialised topics and geographical areas can be built. For example a campus-wide knowledge system could be installed which consists of several subsystems and itself represents a subsystem of a province-wide knowledge base.

This means, that a knowledge space related to a specific course could be provided to other organisation units, training organisations or institutes. The advantages of distributed architecture take effect. That are, for example, reduced gathering and indexing work which may lead to higher update rate of the indexes' databases and therefore to a higher grade of consistency between information supply and the brokers' indexes. Further work should be include a pricing system for the interchange of such knowledge spaces. This concept will allow for example the co-operation between institutes and universities to provide particular knowledge and consume other knowledge spaces.

Another possibility would be to combine all departments of computer science from one country or even countries to create a big specific knowledge topic. Distributed information systems cause less maintenance for each sub system and specific expert knowledge can be managed easier by local experts. This concept also leads to further advantages. On the one hand each sub system gathers only a small amount of the whole huge information space and is therefore able to provide always up-to-date data. On the other hand the introduced system reduces server and network load. The cascaded brokers than deliver all information transparently to the user. Such knowledge topics will also provide a perfect background library for courseware authors as well as learners.

Basically HIKS is also able to include sets of information from other gatherers as well as it can present gathered information in common formats like SOIF, XML or MCF to already existing services. Therefore already existing web based training systems could use HIKS by integrating the knowledge base in their own system.

6. HIKS and Web Based Training System Integration

As already mentioned HIKS could be used as a dynamic background library in a learning environment. At present we are going to make some field experience by using HIKS combined to GENTLE, a Hyperwave-based web based training system. GENTLE not only provides a set of lessens for learners, but also synchronous as well as asynchronous discussions between course trainers and learners. GENTLE is also a question and answer system and a annotation system concerned to specific section at lessens. HIKS as a dynamic background library allows the GENTLE system (the annotation system, the question and answer system, the lessons itself) to reference to the background knowledge. HIKS will help courseware authors to produce new lessons or update such. Furthermore HIKS will make knowledge available which will help learners to solve problems, to enlarge the content of the lessons and to produce presentations.

7. Conclusions and Future Work

By adding dynamic background libraries we do not only present a profound additional knowledge base for learners to widen their horizons but also a huge data collection for courseware authors to create new lessens. Apart from that the contents of these libraries are automatically organised and updated. Some future research will be done to implement a dynamic, self-learning user profile management system which adapts to the users' behaviour over time and to improve the context extraction and evaluation. Further work will be done in the future integrating AI modules to find proper descriptions. At present we are going to work in the area of indexing not only Web sites but several different kinds of databases, e.g. Oracle or Informix, that are widely used to store information.

8. Literature

[Busch97] Busch R.: "Lernen aus dem Netz - Multimedia und Internet, Anlass für eine Bildungsreform", c´t Computermagazin6/97 (1997).

[Guetl98] Gütl, C.; Andrews, K.; Maurer,H. "Future Information Harvesting and Processing on the Web", European Telematics: advancing the information society, Barcelona and http://www.iicm.edu/-cguetl/papers/fihap (1998)

[Dietinger98] Dietinger T., Maurer H.: "GENTLE - General Network Training an Learning Environment", Proc. of EDMEDIA (1 998), Freiburg, Germany, -to be published-

[Kappe94] Kappe, F., Andrews, K., Faschingbauer, J., Gaisbauer, M., Maurer, H., Pichler, M., Schipflinger, J. (1994). "Hyper-G: A New Tool for Distributed Multimedia", Proc. of the IASTED/ISMM Conf. on Distributed Multimedia Systems and Applications, Honolulu, USA,209-214.

[Marchionini95] Marchionini, G., Maurer, H.(1995). "The Role of Digital Libraries in Teaching and Learning", Communications of the ACM 38,4, 67-75.

[Maurer97] Maurer H., Dietinger T.: How Modern WWW Systems Support Teaching and Learning. Proceedings of International Conference on Computers in Education 1997 (Ed. Z. Halim, T. Ottinann, Z. Razak), Kuching, Sarawak Malaysia, December 2-6, 1997, p. 37 - 5 1.

[Maurer96a] Maurer, H. (1996). "HyperWave: The Next Generation Web Solution", Addison Wesley Pub.Co., UK.

[Maurer96b] Maurer, H., Scherbakov, N.(1996). "Multimedia Authoring for Presentation and Education: The Official Guide to HM-Card", Addison Wesley Pub.Co., Germany.

[Maurer94] Maurer, H., Schmaranz, K.(1994). "J.UCS -- The Next Generation in Electronic Journal Publishing". Proc. of the Electronic Publ. Conference (1 994), London; also in: Computer Networks for Research in Europe 26, Supplement 2-3, 63-69.

[Rieder97] Rieder J.: "Found highways, lost memory", Internet Professional 10/97 (1997).