Targeted Information Retrieval

Thomas Dietinger[*,], Christian Gütl[*], Hermann Maurer[*,**], Maja Pivec[*,***],**

**Institute for Information processing and Computer supported new Media (IICM), Graz University of Technology, Austria, (tdieting@iicm.edu), (cguetl@iicm.edu) (hmaurer@iicm.edu), (mpivec@iicm.edu)*

***Austrian Web Application Center (AWAC), Austrian Research Centers Seibersdorf (tdieting@awac.at)*

****Faculty of Mechanical Engineering, University of Maribor, Slovenia*

The goal of targeted information retrieval is to get more relevant information, which is task and user specific and may be used for adaptive problem solving. Possible problems and their solutions are discussed with emphasis on collaborative knowledge transfer environments. Distributed knowledge gathering and knowledge clustering as a quality improvement of information gained is described in the main part of the paper. The impact of relevant information obtained for the improvement of the learning process is presented.

Keywords: **Data Mining Technology in Education, information retrieval, information structuring, documents clustering, Web-Based ITS/CAI, information relevance improvement**

1 Introduction

"We are drowning for information but starving for knowledge." (- John Naisbett) Our present-day society could be described as the so-called information society. It is characterized by a very huge unstructured knowledge repository as well as rapid increase of information. In the last few years the Internet has become a very interesting area for publishing and gathering information and must be included in a future-oriented learning environment. In the early 1998 the current number of Web pages was estimated to exceed 150 million [4], in the summer of 1999 the number of indexable Web pages are estimated about 800 million [6]. But only gathering this information does not satisfy the users' needs!
The main problem is to get the right information with proper quality, reliability and timeliness and to get only information that has been requested: we will call this 'knowledge'. Rieder [7] addresses this situation by saying *"Not only is the gathering of information demanded; this information must also have meaning [...]"*. From the users' point of view, obtaining the right information, which is needed to solve a problem or accomplish a task, increases the value of the Web decisively. This means that not only is the access to information important, but also the relevance of quality itself matters.

2 Document Clustering System

By definition [3], clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe data. The categories may be mutually

exclusive and exhaustive, or consist of a richer representation such as hierarchical or overlapping categories. Examples of clustering applications are the discovery of homogeneous sub-populations (e.g. for consumers in marketing data bases) or identification of sub-categories (e.g. of spectra of infra red sky measurements). In real life clusters may overlap allowing that data, documents or part of documents (e.g. sections) belong to more than one cluster. To get better results we can use conceptual clustering. The difference between conventional and conceptual clustering is that in conceptual clustering the entities are grouped based on a conceptual cohesiveness (e.g. set of neighboring examples). Unlike statistical clustering methods, these algorithms rely on a search for objects within same or similar concepts.

As an example of conceptual clustering we can use the term *virus*. If we search for *virus* we can get as a result different documents which describe the topic in computer science or documents which deal with viruses in medicine. Using the conceptual clustering, those documents are put in different clusters because they belong to different concepts. What we need to be able to accomplish this is a network of meta data, where each meta data object correlates on one hand with the information it describes and on the other hand with other instances of meta data. We suggest to use pure meta data (without a document content), so called base terms which are related to other base terms to describe main concepts and thus work as a seed for new clusters. To improve effectiveness and versatility relations themselves should also contain some meta data: A *type*, that specifies what kind of relation exists between two nodes (e.g. sub- and super-concept, cause or result, opposite or synonym, prerequisite, introductory, etc.), a *weight value*, specifying to what degree this type applies to that relation (e.g. fuzzy values like perfect, good, average, bad etc. expressed by a certain percentage) and a *quality value* that specifies the reliability of the connection (also percentage). With that users can give feedback about the correctness of the relation during browsing and searching of the cluster and thus influence weight and reliability of the relation! This has the advantage that a new relation need not be completely correct right from the beginning, but can converge to a commonly accepted status by collaborative voting. Thus it is not important any more whether the creator of the relation is completely trustworthy or not, it just influences the starting reliability value. I.e., not all connections have to be created by domain experts but can also be created by other (e.g. novice) users or algorithms!

3 A possible Approach: Knowledge gathering process

As already shown, mankind today has to handle highly dynamic knowledge structures. Consequently such highly dynamic knowledge structures have to be taken into account for getting up-to-data and relevant information [1]. E.g. in the field of Web based training or distance learning environments particular domain knowledge can be built by the course material as well as by a static and dynamic background library [1]. The distinction made between the dynamic and static part is based on the premise that we have the control and influence over the static part whereas our influence is minimal in the dynamic component of the background library. The dynamic part may consists of relevant Web sites or Web areas, news forum, annotation systems etc. The static library will include electronic books, electronic journals, question-answer dialogs, exercises, student papers, etc. Static as well as the dynamic knowledge sources are gathered, processed and stored by a knowledge broker point. Such a broker point can process content and meta data (e.g. keywords, reader level, quality ratings). However, former experience [5] [8] has shown that especially meta data are very important for the categorization and structuring of knowledge as well as for

finding relevant information by users. Possible ways to enrich content information will be discussed as follows. Authors and publishers cannot guarantee sufficient and correct generation of meta data because the creation of meta data will mostly not be consequent and objective. Nearly exponential increase of information will make it impossible for human domain experts to categorize all information. A possible way for useful and sufficient meta data can be achieved by a combination of human knowledge and computer-added techniques (see also Fig. 1). It should also be noted that there are existing systems like PHOAKS, Referral Web, GroupLens, Siteseer and Alexa, which can be used to get additional useful information about documents [9].

Human domain experts and users are able to define a set of terms - the base terms and their relations (similarity, hierarchy, etc.) concerning their specific domain subject. Furthermore, experts are able to categorize and rate a small subset of available documents. Rating and categorization of knowledge sources (e.g. a Web site) or their subset (e.g. a Web area) can also be of great value. This may allow a rough overview and can be a source for meta data. However, a human domain expert will get a higher weight than a normal user. The weight should be able to dynamically adapt the learning system. Automated processes and usefull AI techniques have to be taken into account as well. The automatic process component e.g. may use meta data from domain knowledge experts and users to categorize knowledge clusters. However, base term relations, categorization and ratings can be the basis for automatic processes. Quite similar to the human learning process automated system components need information from human experts or users to improve their internal knowledge. A solution for such feedback process may be an output rating by users. Automatic process modules can also help to categorize new information, e.g. to position it as new pieces of knowledge in the system. Automatic processing modules may also be used for detection of new base terms and possible relations to other terms as well as new knowledge sources and connections between them.

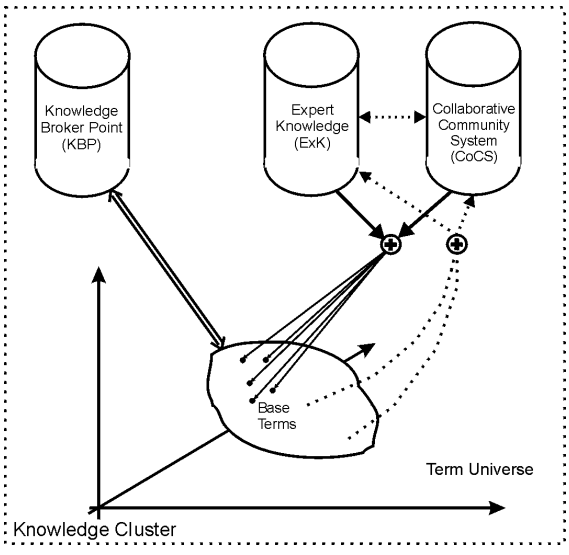

Figure 1: The Knowledge Cluster

One of the automatic processing modules which has been already implemented in a first prototype covers terms for describing and clustering text documents. Former experience [1] [2] has shown that only “relevant” terms depending on the proper domain subject or environment have to be taken into account. In the gathering process keywords from meta data as well as automatically extracted terms build basic terms. The dynamic discriminator filter module holds back terms with high frequency that are based on a whole set of

documents. The remaining terms represent the named discriminators which will be used for the clustering process. The main advantage is a dramatic reduction of process time building clusters. Furthermore, a stop word list for additional static filtering is implemented for improved results. Similar discriminators (relevant terms) may infer similar content and meaning of the documents and will build up knowledge clusters.

4 Conclusions and Future Work

The increased availability of information requires a proper structure to interlinks relevant information and to give these relations a meaning by adding reliability values and descriptions. Human efforts by experts and users as well as automatic processing are needed for this process. Base terms and document clustering may a possible way for an improved information gathering process. First prototype implementation will be tested in the Web based training environment GENTLE and its influence of the quality of information retrieval for students will be evaluated.

References

[1] Dietinger T., Gütl C., Maurer H., Pivec M., Schmaranz K.: Intelligent knowledge gathering and management as new ways of an improved learning process . Proc. WebNet 98 (1998), Orlando, Florida, 244-249.

[2] Dietinger T., Gütl C., Knögler B., Neussl D., Schmaranz K.: Dynamic Background Libraries - New Developments In Distance Education Using HIKS (Hierarchical Interactive Knowledge System), J.UCS (1999) Vol.5, No.1, 2- 10.

[3] Advances in Knowledge Discovery and Data Mining (Ed.:Fayyad U.M., Piatetsky – Shapiro G., Smyth P., Uthrusamy R.), AAAI Press (1996), ISBN 0-262-56097-6.

[4] Gütl, C.; Andrews, K.; Maurer, H. "Future Information Harvesting and Processing on the Web"; Presented at European Telematics: advancing the information society, Barcelona (1998) and http://www2.iicm.edu/~cguetl/papers/fihap.

[5] Hodgins W., Wason T., Duval E.: Learning Object Metadata (LOM), Draft Document v2.1, 25 June1998, IEEE Learning Technology Standards Committee (LTSC), http://www.manta.ieee.org/p1484/ltscdocs/wgc/LOMdoc2_1.html

[6] Rötzer, F.: “Suchmaschinen sind einseitig”, Verlag Hans Heise Online (1999), http://www.heise.de/tp/deutsch/inhalt/te/5059/1.html

[7] Rieder, J.: "Found Highway, lost memory", Internet Professional (1997), No. 11, 111.

[8] Weibel S., Kunze J., Lagoze C., Wolf M.: Dublin Core Metadata for Resource Discovery, The Internet Society, (September 1998), ftp://ftp.isi.edu/in-notes/rfc2413.txt.

[9] Horwath, J.: Personalised Recommender System, Theses, TU-Graz (1999), 30 – 46.