CONF-9605124--1
ANL/DIS/CP--89339

MODELING THE EFFICIENT ACCESS OF FULL-TEXT INFORMATION

Michelle Bernard, Robert Kero, Peter Korp, Michele McCusker-Whiting, Shalom Tsur
Argonne National Laboratory

Kelly Dunlap, Lorrie Johnson
DOE - Office of Scientific and Technical Information

RECEIVED
MAR 27 1996
OSTI

Keywords: Visualization, Full-Text Searching, Contextualization, World Wide Web, Deductive Database

Abstract: The title of this paper describes a research goal set by many offices within the U.S. Department of Energy. The paper will reviews efficient full-text searching techniques being development to better understand and meet this goal. Classical computer human interaction (CHI) approaches provided by commercial information retrieval (IR) engines fail to contextualize information in ways that facilitate timely decision making. The uses of advanced CHI techniques (e.g., visualization) in combination with deductive database technology augment the weaknesses found in the presentation capabilities of IR engines and therefore are discussed. Various techniques employed in a Web-based prototype system currently under development are presented.

1. INTRODUCTION

The rate of full-text information intake throughout the U.S. Department of Energy (DOE) complex, and throughout industry in general, is constantly increasing. In DOE's case, most of this information is received in the form of scientific and technical documents on paper or in electronic format. In its present form, this ever-increasing flow poses a "glut-of-information" problem to the decision maker, who is forced to sift through it to select relevant pieces of information. The fact that most decisions need to be taken in some limited time compounds the problem. Typically, it becomes impossible to review all of the potentially relevant documentation within the time frame allotted to decide. The present form of full-text searching support raises thus the serious concern that the quality of decision making may be impaired because decisions may be based on incomplete or random knowledge. This problem will be exacerbated as budget constraints force DOE to do more with less.

Besides the information-glut problem, another factor that also contributes to the present unsatisfactory state of affairs is the *decontextualization* of the stored electronic full-text documents. This term deserves an explanation. Briefly, documents are created for a specific reason and within a specific context. The reason is often to meet a legislative or policy requirement, while the context is based on the organization(s) involved, people who created the document, time and place of authorship, and relationships to other documents and issues. For example, a DOE "Finding of No Significant Impact" document possesses at least one closely coupled specific relationship (i.e., lineage: child to parent) to its respective environmental impact statement. This type of additional knowledge, often vital to the decision-making process, is lost or made difficult to discern when each document's textual content is all that is preserved. Presently, the burden of remembering (or the cost of dynamically determining) contextual information is the sole responsibility of each user. This

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

3. ADVANCING THE STATE OF THE ART: ARCHITECTURE OF THE INTELLIGENT QUERYING (IQ) SYSTEM

This section elaborates on a system architecture designed to remedy the problems described in section 2 and on the additional knowledge required to provide more comprehensive responses to user queries. Figure 1 depicts the architecture of the proposed IQsystem. It is composed of three major components:

1. document repository. This component is a traditional public domain or commercial IR system of the type described in section 2.

2. context base. This component contains all of the additional knowledge about the corpus of documents kept in the document repository.

3. user interface. This component allows users to access the information via the context base. They can observe the information in various abstract graphical representations that will give them a "bird's eye" view. The interface outlines and abstracts the information so users can reason with the contents of the documentation without having to scan or read it at its actual text-level.

We envision that when the IQ system is used, the process of obtaining information from the repository will change from the traditional cycle: *query→response (hitlist)→analysis (read contents of each hit) → new-query* cycle. It will become a modified cycle of interaction with the context bases' meta-information. Only during the last phase will the actual retrieval and reading of the underlying relevant documentation occur. In some cases, this last step may be entirely omitted. We also expect that the interaction with meta-information rather than with actual text will be vastly more efficient and contribute significantly to timely, high-quality decision making.

The knowledge kept in the context base is of two types: interdocument and intradocument knowledge. Intradocument knowledge represents the logical structure of the document: the relationships among its major components (paragraphs, sections, subsections, tables, figures, etc.) and the relationships between a component and its physical layout on the page. This knowledge will enable the visualization of those components of a document relevant to a certain query. It will enable a user to visualize the immediate context of a document component (e.g., the section containing a relevant paragraph component) and, in general, to "zoom in" or "pan out" of a document at will.

Interdocument knowledge represents relationships among different documents and document classes, such as relationships indicating that documents were created at the same site, by the same authors, or in the same organizations, or that a class C document was derived from a class A and B document. The advanced version of the IQ system will allow the definition of relationships that are customized to the needs of individual users as well as the use of those relationships for querying and visualization purposes. Use of this knowledge will enable users to visualize documents relevant to a query from a particular perspective, such as a geo-spatial perspective, temporal perspective (e.g., "all relevant documents published between 1Q88 and 4Q90," laid out on a time line) or other user-specific perspectives. Advanced versions will enable the documents to be linked by issue, where the issue is user-defined. In addition, dynamically generated views or other arbitrary grouping constructs will be supported to allow for the grouping (strong ··· weak) of documents and document classes for

arrangement can quickly result in an uneconomical cost to the organization. Domain experts leverage their individual *a-priori* knowledge of the corpus and its contents to better control their information searching and gathering. Novice users drown in what falsely looks like a predominately unrelated sea of documents. In either case, there is no capitalization upon an organizationally defined contextual understanding of the corpus, which would (1) increase searching efficiency, (2) increase searching effectiveness, and (3) reduce uncertainty in the decision-making process for all classes of users. Instead, present technology normally allows the recall of the documentation if certain keywords that appear in their text are used. Often the context is not an explicit part of the documentation; hence, it cannot be used in the search for relevant documentation. Consequently, the decision maker who must make decisions solely on the basis of document contents, without understanding their context, can be at a huge disadvantage.

The reengineering process proposed here is to restore the lost context as well as to store the actual documentation by making the context available through use of graphical interfaces accessible via the World-Wide Web.

2. STATE OF THE ART: FULL-TEXT SEARCHING

This text briefly describes the cost-benefit considerations of the reengineering process, main components of a software architecture required for the task, and main steps of the reengineering process itself.

Commercial systems for document storage and retrieval store the full text of the document and group sets of documents in a corpus. These systems provide support for extensive indexing or pattern recognition methods that allow for the retrieval of documents relevant to a user-posed query. Queries are usually formulated as conjunctions, disjunctions, and other Boolean expressions using *keywords* of interest. The result of a query is in the form of a *hit-list*: a list, usually ranked by some measure of relevancy, of summarized information (e.g., document titles) from the documents containing one or more of the keywords that satisfy the query. The measures of success typically used with this technology include *precision* and *recall,* which respectively register the extent to which those documents appearing in the hit list should have been included and the extent to which *all* of the relevant documents have been hit. Over the years, the statistical methods used with regard to resolving user queries have been refined to achieve higher degrees of these measures.

The basic limitation of this technology is that in order for a document to be retrieved, its text must contain the exact keywords used in the query. Thus, if a query uses the term "automobile," but the document contains the term "car," it will not be included in the hit list. This situation is also true for misspelled words (British vs. American, slang, etc.) Another limitation, in its most general case, is that all that is returned by a hit list is the fact that a certain document contains a query keyword. For example, a hit list does not reveal (1) specifics about the distribution of the keyword throughout the document, (2) if the word was used in various logical pieces of the document, or (3) if the word was used in specific classes of documents. This knowledge is often very important to a user in assessing the relevance of a document in specific situations. The statistical measures mentioned earlier usually contribute very little in this respect, since they normally are based on frequency of occurrence weighted in a particular way.

specific and potentially transient types of work.

A third form of knowledge included is linguistic knowledge, in the form of a thesaurus. The use of the thesaurus will enable the "narrowing" or "broadening" of query terms. Thus the terms "cardiologist," "endocrinologist," and "urologist," could be broadened to "medical doctor" and substituted in the query, manually or automatically. Alternatively, a general term could be replaced with more specific instances. This process will be driven by the structure and content of the thesaurus used. The combination of document knowledge and linguistic knowledge is referred to as the "document ontology" or, "ontology" for short.

Figure 1 shows the query processing method: a query, constructed after a cycle of broadening or

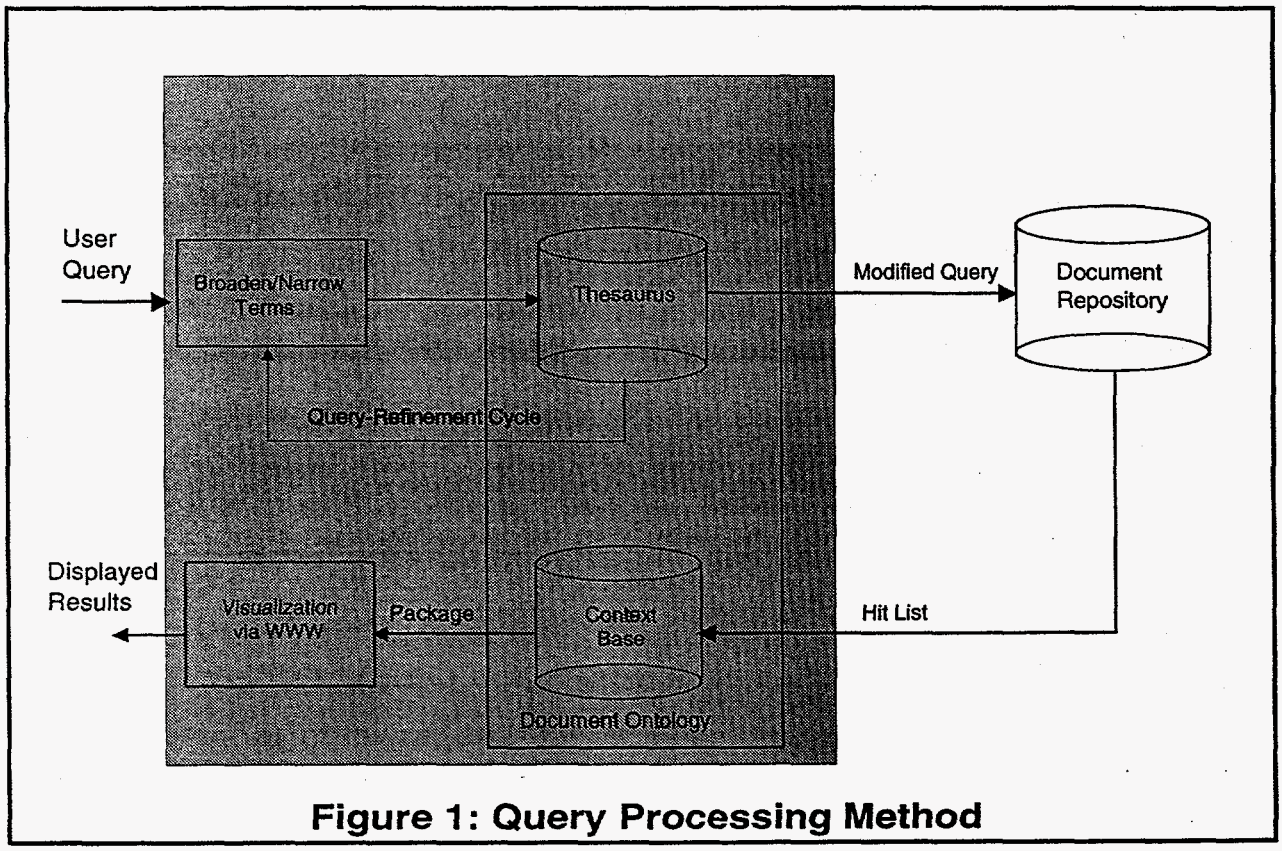

Figure 1: Query Processing Method

narrowing, is sent to the document repository. The result contains a hit-list of *individual document components*, each individually identified. This list is sent to the context base, where it is translated to physical page information, and the relationships to other components and documents are determined. The translated "package" is sent to the visualization system for display.

4. REENGINEERING PROCESS

This section discusses what must be performed to transform an existing corpus of documentation, as well as future documentation, into a format that lends itself to the treatment described in the previous section.

First the documentation must be marked up (e.g., a document in ASCII format is marked up according to the SGML standard). Tags are assigned to each of the documents' components. This step, which clearly is the most labor-intensive step in the process, can be performed manually (when a person uses a tool such as a general-purpose editor to scan the document, identify the components, and insert the appropriate tags) or automatically (when a custom markup or commercial tool is used to identify the components and attach the tags to them).

Documents produced by means of a popular word processing system such as Microsoft Word or other markup languages such as TeX or LaTeX already contain an internal markup, and a variety of filters are available to transform this markup into an equivalent SGML set of tags.

The tagged documents must conform to a DTD (document type definition). An SGML parser can therefore be used to infer the structure of the document and check if it is compatible with the declared DTD. The result of this operation is a parse tree identifying document components corresponding to each of its nodes. Incompatible documents are rejected.

At this point, the document can be decomposed into its component parts and saved by means of an IR engine. The parse tree constructed can provide the initial information necessary to populate the intradocument knowledge part of a context base. The context base being implemented in the prototype will be based on deductive database technologies.

5. COST-BENEFIT DISCUSSION

The simplest and the most common method of information storage and retrieval requires only the initial entry of the whole document into an IR system; no further processing is required. By comparison, the proposed method requires a higher investment in initial document processing (i.e., the performance of the reengineering process for each new document). Although the proposed method constitutes an overhead cost, this cost (which we expect to decrease over time with the advent of more automated and better markup tools) would be clearly offset by the additional long-term benefit that would accrue from this enhanced form of information storage. Specifically, the following benefits can be achieved:

1. The user has control over the level of information displayed. Thus, this system implements the idea of an "information lens" that can be used to adjust the level and volume of information to meet the user's needs.

2. The maintenance of a separate context base allows information to be abstracted and preserved at different levels of detail. This feature makes the long-term archiving of information more flexible and cost-effective.

3. The context base can be augmented with additional information on its usage. In particular, it can record changing trends in the use of the information and can be used in decisions that need to be taken in the *management of change* as the information content changes over time. Ultimately, we believe that this approach is an effective answer to the management of the "information-glut" problem. This feature is a direct consequence of the separation of the text and the declarative meta-data in the context base. Over time, the only dynamically changing entity is the context base, which

can be updated and, as such, can be used to provide a dynamically changing view of the underlying corpus of documentation.

6.0 ACKNOWLEDGMENTS

Work supported by the U.S. Department of Energy, Office of Information Management and Office of Scientific and Technical Information, under contract W-31-109-Eng-38.

7.0 REFERENCES

1 Beavers, E., et al., "Efficient Full-Text Searching Based on Advanced Visualization and Cooperative Query Construction," *12th Office Information Technology Conference*, Augusta, Georgia, July 12-14, 1995

2 CACM, 38, 4, *Digital Libraries*, April 1995.

3 Croft, W. Bruce, et al., "Providing Government Information on the Internet: Experiences with THOMAS," *Digital Libraries '95*, June 1995.

4 Government Accounting Office, GAO/OCG-93-5TR, *Information Management and Technology Issues, December 1992.*

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.