Knowledge-based Information Retrieval

Bruce Porter and **Art Souther**
Department of Computer Science
University of Texas at Austin
porter@cs.utexas.edu

Abstract

Unlike the world wide web or general libraries, digital libraries typically serve a specialized community of experts sharing a relatively narrow focus, such as some aspect of law, science, technology, or business. Moreover, these experts are not "casual users"; they have stringent information requirements. For these reasons, digital libraries increasingly invest in sophisticated methods for indexing and retrieving their information assets.

The goal of this project is to develop and test a method of *knowledge-based information retrieval*, in which a request for information is posed as a question, and information sources are identified that pertain to steps in the logical process of answering the question. We aim to develop this technology by leveraging our results from fifteen years of research on building knowledge bases and developing automated methods for using them to answer questions. While our previous research required extensive knowledge bases that are costly to build and maintain, our current research will significantly reduce this requirement with a novel combination of symbolic reasoning and more conventional information retrieval.

To evaluate our results, we plan to build an information retrieval system for the wide variety of users needing information on the effects of global climate change, and to measure its success compared with human experts and conventional systems.

This paper introduces knowledge-based information retrieval. While our work on an integrated system is preliminary, the components of our solution (an extensive knowledge base and methods of using it to answer questions) are well developed.

Knowledge-based Information Retrieval

We aim to develop automated methods that perform an important new information-retrieval task:

- *Given* a user's question requesting information,
- *Find* relevant information sources, and *explain* their relevance to the user's request.[1]

As an example of this task, consider a question that might be asked by a user concerned with the effects of climate on one of the seven major food crops (wheat, rice, potato, cassave, sorghum, groundnut, and maize) (Tinker & Ingram 1996): "How will the net primary productivity of rice in Madagascar change if the concentration of CO_2 in the atmosphere doubles?" Hypothetically, a computer system (like the one we propose to build) might find a variety of information sources that bear on this question, and explain why each is relevant, as shown in Figure 1.

As this example illustrates, with our approach "relevance" is a sophisticated notion. An information source is deemed relevant to a question if it pertains to a step in *the logical process of answering the question.* For example, the journal paper entitled "The Effects of Temperature on the Metabolic Rate in various strains of Rice" is deemed relevant because net primary productivity (the dependent variable of the query) is influenced by metabolic rate via growth rate, and atmospheric CO_2 is known to influence temperature.

In contrast, contemporary methods of information retrieval have a weak metric for measuring relevance (for overviews, see (Kowalski 1997; Frants, Shapiro, & Voiskunskii 1997; Goldman 1992; Salton & McGill 1983)). For most common search engines, a request for information is expressed with a boolean expression of keywords (not a question), and relevance is measured by matching words, either exactly or statistically (Brenner 1996).[2] Efforts to improve search engines – e.g. by adding, to the list of keywords, their synonyms (Ruge 1992) and stems (Frakes & Baeza-Yate 1992) – show the benefits of using semantic knowledge, even within the inherent confines of keyword systems. We expect to

[1] Information sources are digital documents, databases, images, multimedia objects, and software, as well as on-line references to non-digital information.

[2] For some specialized search engines, such as WESTLAW, queries may be expressed in English (Turtle 1994; 1990)), and they are automatically converted to boolean expressions.

user: How will the net primary productivity of rice in Madagascar change if the concentration of CO_2 in the atmosphere doubles over the next 60 years?
system: The information sources below are probably relevant to your question. For more information on the contents, structure, or accessibility of any information source, simply click on it.

1. *Projected atmospheric CO_2 concentrations over the next 100 years* a database, is relevant because it provides predicted magnitudes and rates of change of CO_2.
2. *Crop distribution in Madagascar*, a digital map, is relevant because it provides the location of rice growing areas.
3. *Carbon dioxide absorption rate as a function of temperature in rice* a database, is relevant because it provides data on the amount of CO_2 available to rice plants.
4. *The mean annual temperature by latitude* and *The mean monthly temperature by latitude*, both databases, are relevant because they provide temperature data needed by (3).
5. *The Effects of Temperature on the Metabolic Rate in Various Strains of Rice* a journal paper, is relevant because growth rate is influenced by metabolic rate, and atmospheric CO_2 influences temperature.
6. *The effect of atmospheric CO_2 on atmospheric temperature*, a global climate model, is relevant because it estimates the strength of the influence needed by (5).

Figure 1: **A hypothetical scenario illustrating the functionality of our proposed system for knowledge-based information retrieval.** The user poses a request for information as a question. The system uses a knowledge base to step through the logical process of answering the question. Typically, only a small portion of the knowledge base is relevant to any particular question. The system collects references to information sources, such as databases and digital maps, that are associated with that portion of the knowledge base, and explains their relevance in terms of steps in the question-answering process.

improve substantially on these results by utilizing more knowledge in the information retrieval process.

Our approach, called *knowledge-based information retrieval*, will use four component technologies. First, it will include a *knowledge base* of basic information important in a subject area. For example, in the domain of global warming, the knowledge base might include the biological classification of rice as the genus Oryza of the family Poaceae (grasses), which is in turn a family of flowering plants, and that the rate of diffusion of CO_2 into plant leaves is a function of temperature. Second, the knowledge base will include summaries of *information sources*. For example, a database on the net primary productivity of rice over the last 30 years might be referenced, and linked to the `rate` attribute of the `growth` process of the `rice` genus of the `grass` family. Third, our approach will include an *answer-template generator*, a program that identifies the portion of the knowledge base relevant to answering the question, including the pathways of influence between independent variables (such as atmospheric CO_2) and dependent variables (such as net primary productivity). Those information sources linked to the template are potentially relevant to the query. Fourth, we will develop an *explanation generator*, a program that explains the relevance of each information source to the user's question.

Knowledge Bases

Background We have built a series of large knowledge bases in a variety of domains, including botany, law, and distributed computing. Each encodes a substantial body of information in a form that enables a computer to go beyond merely reciting "canned answers" to anticipated questions by performing inferences to derive information that is not explicitly encoded.

In contrast to the knowledge bases for most expert systems, which include only the information needed to perform a single, narrow task, our knowledge bases cover broad domains without "task boundaries." For example, our Botany Knowledge Base contains over 200,000 explicit facts, with several million more facts readily derivable by automatic inferencing techniques (Sojoodi-Haghighi 1992). These facts are organized into descriptions of 47 spaces (e.g. roots, stems, leaves), 172 substances in those spaces (e.g. oxygen in the leaves), and 313 processes covering physiology and development (Rickel 1995, p. 133).

The information in our knowledge bases is encoded as *KM-terms*[3], a variant of conceptual graphs (Sowa 1984). As with conceptual graphs, each term is a graph of concepts and relations (called sorts and features in psi-terms (Ait-Kaci, Podelski, & Goldstein 1993)), organized in a taxonomic hierarchy. However, as compared with conceptual graphs, the language of KM-terms significantly improves the efficiency of automated reasoning (Clark & Porter 1997). We add the construct of *access paths*, previously used in the Algernon language (Crawford & Kuipers 1991). Each access path expresses co-reference of sub-terms in a KM-term, and is used to encode domain-specific information that guides inference by specifying preferred chains of subgoals for each inference goal (and hence, other chains will not be tried).

Extending the Knowledge Base We aim to extend the Botany Knowledge Base to support knowledge-based information retrieval for an important, and very active, community concerned with the effects of climate change on botanical systems. Working with experts in this area, we will add to the knowledge base functional information about the effects on plant distribution of

[3]for 'Knowledge Manager', the name of the software managing our KB.

temperature, rainfall, insolation, and soil nutrient concentrations.

Past projects in areas of medicine (Findler 1991) and law (Griffith 1993; Hafner 1981) demonstrate that knowledge-based approaches to information retrieval are both feasible and successful, at least for specialized archives. Because these archives offer the best potential for high quality indexing and retrieval of information, we predict they will be increasingly common. Future archives might focus on particular domains (e.g. molecular biology) or particular resources (e.g. a database of satellite imagery). Other researchers share the vision that information will be widely distributed into specialized archives, and some are developing information-retrieval methods in which the first step is determining which archives to search (Chakravarthy & Haase 1995; Callan, Lu, & Croft 1995).

Although building knowledge bases is expensive, this cost should not deter the development of knowledge-based information retrieval. Our experience building the Botany Knowledge Base offers a point of comparison. While that project required about 25 man-years of effort, our proposed project will be simpler in two important ways. First, much of the work – such as developing the knowledge representation formalism – was a one-time effort yielding results that we can reuse.[4] Second, other aspects of our past work – such as building fully automated question answering methods (Lester & Porter 1997; Rickel & Porter 1997; Mallory 1998) – are not necessary for the current task. Moreover, the expense of building a knowledge base can be amortized among a variety of applications, such as: helping users navigate digital libraries (Adam *et al.* 1996, pp. 121–8); expanding users' queries (Voorhees 1994); interpreting natural-language queries (Turtle 1994); and summarizing texts (Paice 1990).

Indexing Information Sources

As an "overlay" to the knowledge base, we will encode descriptions of information sources in the same representational formalism used in the Botany Knowledge Base. Figure 2 is an example of this encoding. The source is a journal paper describing a model that relates climate variables, such as temperature, to plant growth variables, such as net-primary productivity (Woodward, Smith, & Emanuel 1995). The abstract of the paper is given on the top half of the figure. The bottom half of the figure shows the encoding of the abstract and its integration into the Botany Knowledge Base, which provides essential background information to "tie together" the terms used in the abstract. For example, the dependent variables in the model — `Leaf Area Index` (LAI) and `Net Primary Productivity` (NPP) — are not connected to the independent variables (e.g. `temperature`) using only the terms in the abstract. The Knowledge Base supplies the relevant connections.

As this figure illustrates, the encoding of the summary of an information source is a graph of terms and relations, embedded in the knowledge base. We will exploit this fact extensively in our methods for finding information sources relevant to a user's question, as described in the next section. However, before we turn to the issue of *using* such indices, we first discuss the cost of creating them.

Encoding summaries of *every* information source in a digital library – or those added during the continuous growth of the library – would be prohibitively expensive. Fortunately, this is unnecessary, we believe. As part of this research, we will investigate a novel organization of digital libraries that promises to significantly reduce the cost of using *any* sophisticated indexing method, such as ours.

Our proposed organization is motivated by two observations. First, methods of indexing vary in quality and cost,[5] and these dimensions are (usually) directly related. For example, consider two indexing methods, one that extracts and uses key phrases from titles and abstracts, and another that extracts this information from entire documents. The former is likely to provide lower quality results, but at less expense, than the latter.

Second, information retrieval is best thought of as a dialogue between a user and an (automated) assistant, not a "one-shot request." The dialogue begins when the user requests information, then alternates between the assistant suggesting candidate information sources and the user selecting ones that are most (or least) interesting and requesting "more (or none) like these". By providing this *relevance feedback* (Salton & Buckley 1990), the user hillclimbs to increasingly relevant information sources.

In light of these observations, we propose that accurate, yet costly, indexing methods be applied to only a small fraction of the information sources in a digital library. When a user begins a dialogue requesting information, these methods will select the information sources that serve as the initial states for the hillclimbing search. After the user ranks these sources, indices are generated from the best ones in order to search a larger fraction of the digital library, a subcollection that has been indexed by a method that is less costly, but of lower quality, than the first. With each iteration of the dialogue, as the user hillclimbs to the information sources of greatest interest, the assistant searches a larger fraction of the library using weaker indexing methods.

As a first step, we plan to test this search process with only two levels of indexing, and to apply it to the *Virtual Library on Earth Sciences and the*

[4] For example, we will reuse our knowledge representation formalism; see `http://www.cs.utexas.edu/users/mfkb/km.html`

[5] Quality is measured, for example, by recall and precision rates; cost is measured by the time and expense of indexing each information source.

Abstract: A global primary productivity and phytogeography model is described. The model represents the biochemical processes of photosynthesis and the dependence of gas exchange on stomatal conductance, which in turn depends on temperature and soil moisture. Canopy conductance controls soil water loss by evapotranspiration. The assignment of nitrogen uptake to leaf layers is proportional to irradiance, and respiration and maximum assimilation rates depend on nitrogen uptake and temperature. Total nitrogen uptake is derived from soil carbon and nitrogen and depends on temperature. The long-term average annual carbon and hydrological budgets dictate canopy leaf area. The model is used to simulate the global distributions of leaf area index and annual net primary productivity.

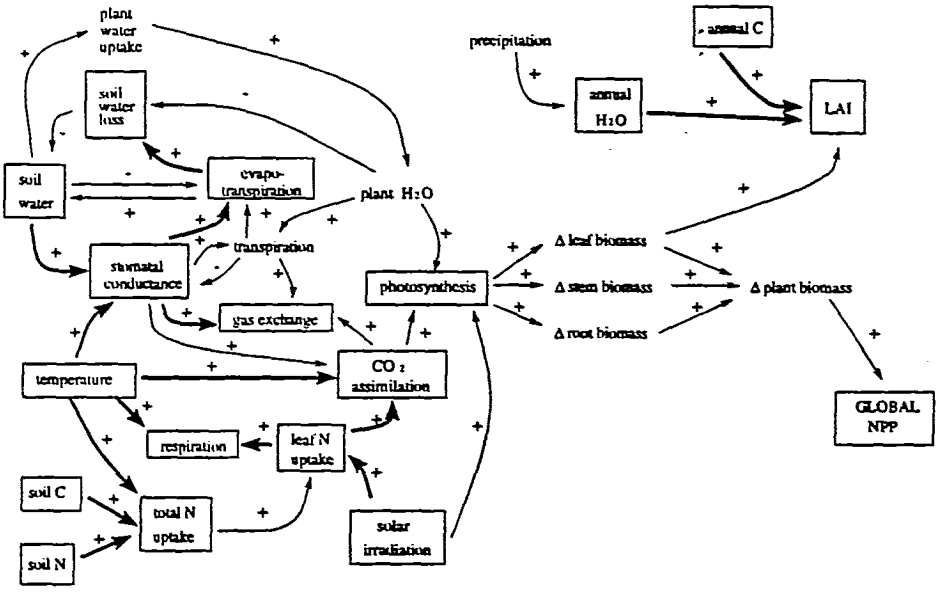

Figure 2: **An example of the encoding of an information source in the knowledge base.** Top: The abstract for a journal paper describing a model that relates changes in climate to plant growth. Bottom: The encoding of the abstract and its integration into part of the Botany Knowledge Base. The boxed terms and the bold arrows come from the abstract itself. The rest of the graph – the unboxed terms, the non-bold arrows, and the nature of the qualitative influence (positive or negative) – is supplied by the Knowledge Base.

Environment.[6] This library contains about 1200 information sources indexed by conventional word-based methods. To this base, we plan to index about 100 sources more extensively, by summarizing each with an overlay of the Botany Knowledge Base, as described above. Knowledge-based information retrieval, applied to this two-level library, will work as follows. After the user poses a question requesting information, our methods (which are described in the next section) will select some of the 100 well-indexed information sources. After the user ranks them, indices are generated from the best ones in order to search the second (and final) level of the library. Since the second level is indexed by conventional word-based methods, the documents selected from the first level (or selected portions of them) may be used as examples to provide relevant keywords and phrases (this commonly used method is called "query by example" (Frants, Shapiro, & Voiskunskii 1997)).

In later research, we plan to test other types of indexing that might be intermediate – in terms of quality and cost – between the two described above. Promising types include: (semi-automatic) parsing of selected portions of documents, such as abstracts and introductions; automatic summarization of documents; and automatic extraction of key phrases.

Finding Information Sources Relevant to Users' Questions

With our approach to information retrieval, a request is posed as a question. This has two major advantages over conventional approaches, in which a request is posed as a boolean expression of keywords. First, because it retains useful structure, a question can be more precise. For example, a question may include a preface (such as, "in the context of increased levels of precipitation ...") or ask for a causal influence (such as, "what is the effect of elevated atmospheric CO_2 on plant growth rate?"). Ignoring this structure is a major factor contributing to the ambiguity of conventional queries.

Second, the structure of a question suggests the steps required for answering it. For example, consider a question about the causal influence between an independent variable and a dependent variable. Answering the question requires finding the pathway(s) connecting the variables, then determining when they are active. Questions of a different form require different steps. By identifying these steps, a system for knowledge-based information retrieval can expand users' questions to search for information pertinent to the subgoals in answering them.

For questions of many types, our goal is to develop domain-independent methods to identify information sources relevant to answering each one. In this phase of our research, we plan to focus on two broad types of questions:

- questions that ask for descriptions
- questions that ask for predictions

In general, *answering* questions of these forms, automatically or with an expert's assistance, is very difficult, requiring considerable domain knowledge and two steps of automated reasoning:

1. identify the portion of the knowledge base that bears on that question,
2. apply techniques of automated reasoning and natural-language generation to just that portion of the knowledge base to compute an answer to the question

In contrast, our proposed research focuses on the more tractable task of helping users answer their own questions, by *identifying* relevant information sources and then *explaining* their relevance. Therefore, it emphasizes the first step, and not the second.

Description Questions

A description question has numerous surface forms, but they all ask for a description of an object or process, or some aspect of it. A description of an object typically includes its function, structure, location, and history, while a description of a process typically includes its purpose, steps, and time and place of occurrence. In addition to this information, a comprehensive description often contains embedded descriptions, generated by recursive invocations of the description generator. For example, the description of an object might include the description of the process in which the object participates, and *vice-versa*.

In previous work, which we plan to build on here, we studied the content and organization of descriptions in order to construct a computational theory of how descriptions are generated, and we implemented it in a computer program called KNIGHT (Lester & Porter 1997; 1996). We found that many descriptions follow a similar pattern, which we generalized and cast in a data structure called an *explanation design plan*, or EDP. As an example, Figure 3 shows an EDP for describing processes.

An EDP is only a template for a description. When KNIGHT applies an EDP to a particular concept, the template is instantiated with information specific to that concept, thereby generating a description. The description's content is specified by the leaf nodes of the EDP, and its organization is specified by the internal nodes.

The leaf nodes of an EDP specify content by describing the subgraphs of the knowledge base, termed *viewpoints*, that should be included (Acker & Porter 1994; Acker 1992). For example, in the EDP for describing processes (Figure 3), the left-most leaf node contains the specification for the viewpoint "As-a-kind-of Process Description". This viewpoint contains all, but only, the subgraph of the knowledge base describing why one process is a taxonomic child of another.

[6]see: `http://earthsystems.org/Environment.shtml`

When KNIGHT applies the EDP to the process *photosynthesis* (a specialization of *production*), it extracts from the knowledge base the viewpoint "photosynthesis as-a-kind-of production", as shown in Figure 4.

An instantiated EDP selects a subgraph of the knowledge base, i.e. the graph union of the viewpoints it specifies. This subgraph may intersect with the descriptions of information sources that have been encoded in the knowledge base, as described in Section . In this way, an instantiated EDP identifies the information sources relevant to a user's description question.

After our methods instantiate an EDP, and thereby identify relevant information sources, we will enable the user to alter this subgraph using the following operations:

1. extending the graph by asking for descriptions of terms it contains, and incorporating into the graph their instantiated EDP's.
2. constricting the graph by selectively omitting EDP nodes on irrelevant topics

Each alteration of the graph can change the information sources that the graph identifies.

Prediction Questions

Although they have many surface forms, in the simplest case, prediction questions ask "How will a change in variable X affect variable Y?" For example, Figure 1 gives a prediction question which asks for the effect of increased atmospheric CO_2 on net primary productivity. We have developed a modeling program, called TRIPEL (Rickel & Porter 1997; Rickel 1995), that answers prediction questions by first constructing qualitative models, then simulating their behavior to yield predictions. As with KNIGHT, it is the first of these capabilities, and not the second, that we plan to reuse for knowledge-based information retrieval.

Constructing a qualitative model requires finding the relevant portion of the knowledge base: i.e. the *influence diagram* that includes all, and only, the important connections between the independent and dependent variables of the question. In TRIPEL's case, the variables are continuous and real-valued quantities and the connections are qualitative influences, each specifying a positive or negative correlation between a pair of variables. As an example, Figure 2 shows an influence diagram connecting climate variables (e.g. `temperature`) to plant growth variables (e.g. `net-primary productivity`).

Building an influence diagram can be difficult because questions are typically incomplete. For example, the prediction question from Figure 2 mentions none of the "structural conditions" (such as the fact that the plant leaves are in the atmosphere) or "initial conditions" (such as the fact that the plants are mature and healthy). The automated modeler must infer these structural and initial conditions by the process of *scenario elaboration*.

Potentially, this process could produce a very large influence diagram, especially in domains in which subsystems are highly integrated, such as biology. TRIPEL employs various techniques to limit its size (Rickel & Porter 1994). For example, TRIPEL is able to determine when variables are exogenous to the system being modeled. An exogenous variable is one which influences the dependent variable, or an intervening variable, but is not itself influenced by any variable within the system, at least within the time-frame of interest. Continuing with our example, the concentration of atmospheric CO_2 influences the rate of photosynthesis of plants on a time scale of days, but is not itself influenced by the photosynthetic removal of atmospheric CO_2 except over a time scale that is too long to be relevant to the question.

We will not attempt to describe *how* TRIPEL constructs an influence diagram — this detail is in our *Artificial Intelligence* journal article. The important point is this: an influence diagram references an information source when the encoding of the source (a subgraph of the knowledge base) intersects the influence diagram (also a subgraph of the knowledge base). Again, Figure 2 illustrates this graph intersection.

Building on this work, we propose to develop domain-independent ways of handling a wider variety of prediction questions, which will require better methods of controlling scenario elaboration. TRIPEL handled questions in which both the dependent and independent variables were stated. Many prediction questions, however, are more open-ended, such as "what happens if < *event* > occurs?" and "what causes (caused) < *event* >?" In addition to using time scale information, we will explore new ways of bounding the model. The most promising method involves representing, in the knowledge base, the relative strength of influences and the conditions under which an influence is dominant. We will extend TRIPEL to use this knowledge to build a series of models, beginning with the one comprising only the dominant influences. We will enable the user to steer this process, interjecting assumptions and preferences as the model is constructed.

Expanding a Query

The information-retrieval methods described above may discover too few information sources. In this event, the user's question can be expanded using the same knowledge base that the information-retrieval methods use. For example, consider a question for which the photosynthetic rate of a particular plant species is relevant. (Or, more simply, consider this to be a query posed directly by a user.) If this query fails to identify enough relevant information from the archive, it can be expanded with a variety of heuristic strategies, such as:

- inheritance: look at more general classes of plants (e.g. the genus or tribe), or more general metabolic processes
- transductive inheritance: look at "siblings" in those

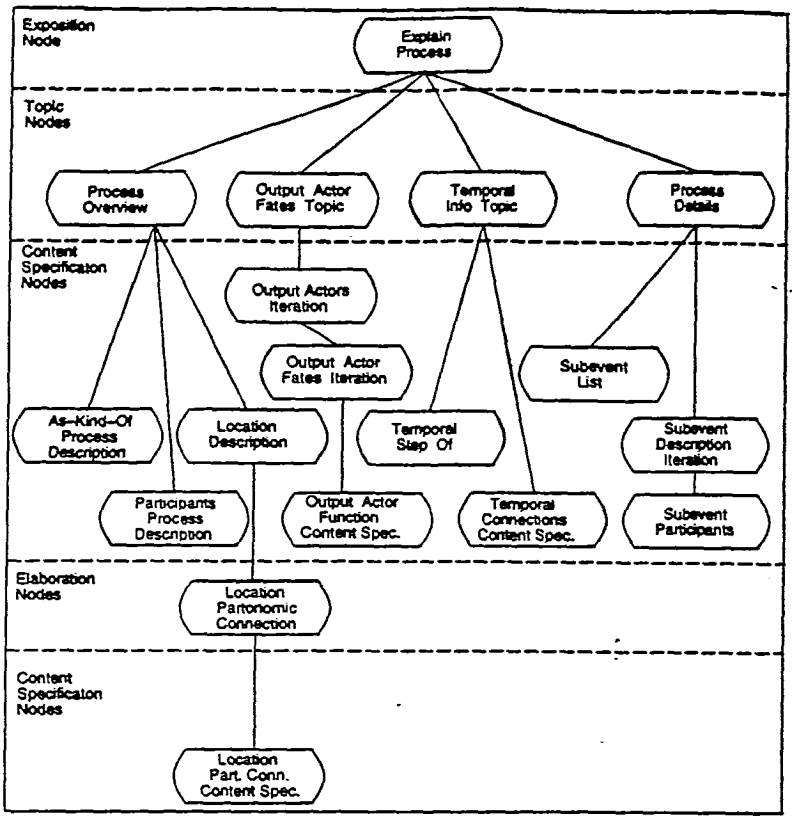

Figure 3: **The Explanation Design Plan for describing processes.** Each leaf node specifies a *viewpoint*, i.e. a subgraph of the knowledge base containing information about some aspect of the process. The internal nodes determine the organization of the viewpoints into a comprehensive description.

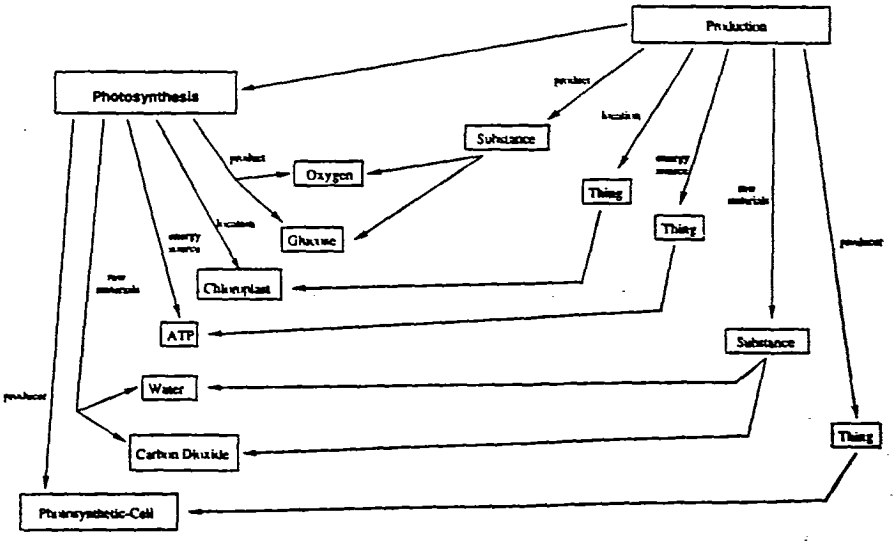

Figure 4: **The subgraph of the Botany Knowledge Base comprising the viewpoint of photosynthesis as a production process.** The viewpoint was extracted by the View Retriever program, a version of which was used in KNIGHT. The input to the View Retriever is a specification of the desired viewpoint. In this case, the viewpoint includes all, and only, the information about `photosynthesis` for the roles (or relations) unique to `production`. The unlabeled arcs are `specialization` relations.

same classification hierarchies

- induction: look at specific members of the plant species (e.g. the varieties of the species)

These methods of heuristic reasoning are domain-independent: they work in many domains, but not without errors. However, these errors are not too serious. Because they are errors of commission (not omission), in the worst case they recommend information sources which are not relevant, thereby lowering the system's precision. We expect that users will more easily detect these errors with our system than with conventional systems because our system will explain the purported relevance of each information source. The explanations will include the role of heuristic strategies in expanding the query.

Explanation Generation

After our system identifies information sources that are relevant to a user's question, it will be able to explain why each is relevant. The explanation will describe the relationships between information sources and steps in the logical process of answering the question. Because a summary of each information source is encoded in the knowledge base, an explanation need only state why information from a summary was selected by the answer-template generator.

We will develop domain-independent ways of presenting explanations to users both graphically and textually. The graphical presentation will show the overall structure of the explanation. In response to description questions, the structure will be the instantiated EDP; In response to prediction questions, the structure will be the generated model. We will build a browser to enable the user to navigate through these structures, querying our system for more detailed explanations of the relevance of each information source.

Detailed explanations will be presented textually. Each will include a summary of the information source and an explanation of how it relates to one or more steps in answering the question. These explanations should be generated, not simply recited from "canned text", because the relevant aspects of an information source depend on the user's question and should be related to the overall structure of the explanation.

For description questions, we have considerable experience with automatically generating text from information encoded in our knowledge-base formalism, and we plan to leverage these results (Lester & Porter 1997; 1991). However, for prediction questions, our past work is less applicable and new methods are required. We have developed ways of describing the behavior of a computational model, as determined by a simulator (Mallory, Porter, & Kuipers 1996; Mallory 1998), but we need a way of describing the model itself. A facility that describes qualitative models in English would benefit everyone developing and using qualitative reasoning.

Empirical Evaluation

We will empirically test two conjectures: (1) For users' questions requesting information, knowledge-based information retrieval selects appropriate information sources and explains their relevance correctly; (2) The precision and recall of our two-level search (as described in Section) is superior to conventional one-level search.

To test the first conjecture, since there are no comparable computer systems, we will compare our system's performance with human experts, with these steps:

1. assemble two panels of researchers who are interested in the effects of climate change on botanical systems, drawn from the large and diverse population of graduate students and post-doctoral researchers at UT-Austin. Call the panels the *examiners* and the *examinees*.
2. describe to the *examiners* the scope of the extended Botany Knowledge Base. Instruct the panel to generate about fifty information-retrieval requests (in the form of questions, of the types described in Section).
3. give the *examinees* these questions plus the descriptions of the 100 information sources (i.e., the same information that human indexers typically receive). Instruct the panelists, working individually, to identify information sources pertinent to each question, and to explain the relevance of each one.
4. present the same questions to our computer system.
5. instruct the *examiners* to evaluate the output from the *examinees* and our computer system, without knowing the source of the output (or even that a computer system is involved). Evaluations will be based on errors of omission, errors of commission, and the quality of the explanations. To control for factors that are irrelevant to this research, we will instruct the *examiners* to judge errors with respect to only the 100 information sources in the knowledge base, and we will normalize the writing quality of the explanations produced by the *examinees* and our computer system.[7]

To test the second conjecture, we will compare the two-level search — i.e. knowledge-based information retrieval applied to a small portion of the library, followed by "query by example" applied to the complete library — with conventional keyword-based, one-level search, with these steps:

1. Randomly select about twenty of the fifty questions generated in step 2 above, along with the information sources and explanations produced by our system.
2. Search the complete library for relevant information sources, in two ways:

[7] We have successfully used this basic experimental design in four previous projects (Porter, Bareiss, & Holte 1990; Branting & Porter 1991; Lester & Porter 1997; Rickel & Porter 1997).

- *Produce the "2-level results:"* Ask a panel of experts to consult the explanations and rank the information sources for relevance to each question. Then, use the most relevant sources as examples to query the complete library.
- *Produce the "1-level results:"* To normalize inputs, ask professional librarians to cast each question as a boolean query, then pose each query to a conventional search engine.

3. ask the other panel of experts to judge the recall and precision for the 2-level results and the 1-level results.

While these experiments will yield rankings, our further analysis will attempt to explain *why* and *when* one approach performs better than another. One experimental technique we plan to use is ablation studies, which test the relative contribution of specific features of our approach.

Summary

Because digital libraries typically serve a specialized community of experts who have stringent information requirements, sophisticated methods for indexing and retrieving information are both necessary and feasible. We propose to address a challenging issue at the core of information retrieval: selecting, from all the information in an archive, that portion which is relevant to a user's request. While many researchers have worked on this issue, our approach is completely novel: a request for information is posed as a question, and information is deemed relevant if it pertains to a step in the logical process of answering the question.

Our long-term goal is to develop automated methods for handling a variety of types of questions. However, our first step focuses on two types – description questions and prediction questions. We will develop automated methods that identify the information sources relevant to each question, then explain their relevance to the user's question. We aim to develop this technology by leveraging our results from fifteen years of research on building knowledge bases and developing automated methods for using them to answer questions. While our previous research required extensive knowledge bases that are costly to build and maintain, our proposed research will significantly reduce this requirement with a novel combination of symbolic reasoning and more conventional information retrieval.

As a testbed for our work, we plan to use an archive of information about the effects of climate change on botanical systems – a burgeoning concern of scientists, economists, agronomists, and many others – and to thoroughly evaluate the performance of our information-retrieval system as compared with human experts and conventional methods.

References

Acker, L., and Porter, B. 1994. Extracting viewpoints from knowledge bases. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 547–552. Menlo Park, CA: AAAI Press.

Acker, L. E. 1992. *Access Methods for Large, Multifunctional Knowledge Bases.* Ph.D. Dissertation, Department of Computer Sciences, University of Texas at Austin.

Adam, N.; Bhargava, B.; Halem, M.; and Yesha, Y., eds. 1996. *Digital Libraries: Research and Technology Advances.* Springer.

Ait-Kaci, H.; Podelski, A.; and Goldstein, S. C. 1993. Order-sorted feature theory unification. Tech Report PRL-RR-32, Digital Paris Research Labs. (http://www.isg.sfu.ca/ftp/pub/hak/prl/PRL-RR-32.ps.Z).

Branting, K., and Porter, B. 1991. Rules and precedents as complementary warrants. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 3–9.

Brenner, E. 1996. *Beyond Boolean — New Approaches to Information Retrieval.* The National Federation of Abstracting and Information Services.

Callan, J.; Lu, Z.; and Croft, B. 1995. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21–28.

Chakravarthy, A., and Haase, K. 1995. Netserf: Using semantic knowledge to find internet information archives. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4–11.

Clark, P., and Porter, B. 1997. Using access paths to guide inference with conceptual graphs. In Iukose, D.; Delugach, H.; Keeler, M.; Searle, L.; and Sowa, J., eds., *Proceedings of the 5th International Conference on Conceptual Structures – ICCS'97 (Lecture Notes in AI vol 1257)*, 521–535. Berlin: Springer. (http://www.cs.utexas.edu/users/pclark/papers/cg97.ps).

Crawford, J. M., and Kuipers, B. J. 1991. Algernon – a tractable system for knowledge-representation. *SIGART Bulletin* 2(3):35–44.

Findler, N. 1991. *An Artificial Intelligence Technique for Information and Fact Retrieval: An Application in Medical Knowledge Processing.* MIT Press.

Frakes, W., and Baeza-Yate, R. 1992. *Information Retrieval Data Structures and Algorithms.* Prentice Hall Publishers.

Frants, V.; Shapiro, J.; and Voiskunskii, V. 1997. *Automated Information Retrieval.* Academic Press Publishers.

Goldman, N. 1992. *Online Information Hunting.* McGraw Hill Publishers.

Griffith, C. 1993. WESTLAW's winning ways. *Law Office Computing* 31–38.

Hafner, C. 1981. *An Information Retrieval System Based on a Computer Model of Legal Knowledge.* UMI Research Press.

Kowalski, G. 1997. *Information Retrieval Systems: Theory and Implementation.* Kluwer Academic Publishers.

Lester, J., and Porter, B. 1991. A revision-based model of instructional multi-paragraph discourse production. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society.*

Lester, J., and Porter, B. 1996. Scaling up explanation generation: Large-scale knowledge bases and empirical studies. In *Proceedings of the National Conference on Artificial Intelligence.*

Lester, J., and Porter, B. 1997. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics* 23(1):65–101.

Mallory, R.; Porter, B.; and Kuipers, B. 1996. Comprehending complex behavior graphs through abstraction. In *Proceedings of the International Workshop on Qualitative Reasoning about Physical Systems.*

Mallory, R. 1998. *Generating Structured, Causal Explanations of Qualitative Simulations.* Ph.D. Dissertation, Department of Computer Sciences, University of Texas at Austin.

Paice, C. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management* 26:171–186.

Porter, B.; Bareiss, R.; and Holte, R. 1990. Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence Journal* 45(2):229–263.

Rickel, J., and Porter, B. 1994. Automated modeling for answering prediction questions: Selecting the time scale and system boundary. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 1191–1198. Menlo Park, CA: AAAI Press.

Rickel, J., and Porter, B. 1997. Automated modeling of complex systems to answer prediction questions. *Artificial Intelligence* 93(1–2):201–260.

Rickel, J. 1995. *Automated Modeling of Complex Systems to Answer Prediction Questions.* Ph.D. Dissertation, Department of Computer Science, University of Texas at Austin. Technical Report AI95-234.

Ruge, G. 1992. Experiments on linguistically based term associations. *Information Processing and Management* 28(3):317–332.

Salton, G., and Buckley, C. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* 41:288–297.

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

Sojoodi-Haghighi, O. 1992. An implementation of access-limited logic for the KM system. Master's thesis, University of Texas at Austin.

Sowa, J. F. 1984. *Conceptual structures: Information processing in mind and machine.* Addison Wesley.

Tinker, P., and Ingram, J. 1996. The work of focus 3. In Walker, B., and Steffen, W., eds., *Global change and terrestrial ecosystems.* Cambridge: Cambridge University Press.

Turtle, H. 1990. *Inference Networks for Document Retrieval.* Ph.D. Dissertation, Computer Science Department, University of Massachusetts, Amherst.

Turtle, H. 1994. Natural language vs. boolean query evaluation: A comparison of retrieval performance. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 212–220.

Voorhees, E. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 61–69.

Woodward, F.; Smith, T.; and Emanuel, W. 1995. A global land primary productivity and phytogeography model. *Global Biogeochemical Cycles* 9(4):471–490.