Classification of Information Sources using Visual Graphic Structures

Kenneth Baclawski, Waltham, MA

Abstract

A computer system including a processor, display device and main memory, which may include local disks, or may, alternatively or additionally, be connected to a network. The computer system is used to classify information sources according to their relevance to a query. The classifications are labeled with visual graphic structures. The query may be specified using a variety of languages including natural language and visual graphic structures. The classification of an information source is determined by the best graphic structure that is contained within or inferable from the information source and that is contained within or inferable from the query. The information source classes are labeled with a visual representation of the graphic structure defining the class. The classes are arranged in the order of relevance that is determined by how close the graphic structure of the class is to the graphic structure of the query.

References

[1] G. Birkhoff. *Lattice Theory, Third Edition*, volume 25. American Mathematical Society, 1973.

[2] C. Cleverdon and E. Keen. Factors determining the performance of indexing systems. Vol. 1: Design, Vol. 2: Results. Technical report, Aslib Cranfield Research Project, Cranfield, UK, 1966.

[3] C. Cuadra and R. Katter. Experimental studies of relevance judgments: Final report. I: Project summary. Technical Report NSF Report No. TM-3520/001/00, System Development Corporation, Santa Monica, CA, 1967.

[4] C. Cuadra and R. Katter. Opening the black box of "relevance". *Info. Proc. and Management*, 21(6):489–499, 1967.

[5] B. Davey and H. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, UK, 1990.

[6] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer Verlag, Heidelberg, Germany, 1999.

[7] A. Rees and D. Schultz. A field experimental approach to the study of relevance assessments in relation to documents searching. I: Final report. Technical Report NSF Contract no. C-423, Case Western Reserve University, Cleveland, OH, 1967.

[8] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *J. Amer. Soc. Info. Sci.*, 26:321–343, 1975.

[9] J. Sowa, editor. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* PWS Publishing, 2000.

[10] B. Vickery. The structure of information retrieval systems. In *Proc. Intern. Conf. Sci. Info.*, volume 2, pages 1275–1289, 1959.

[11] R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In J. Rival, editor, *Ordered Sets*, pages 445–470. D. Riedel Publishing Company, Dordrecht, Holland, 1982.

1 Field of the Invention

The invention relates to the classification of information sources by a computer system.

2 Background of the Invention

The increasing popularity of high speed computer networking has made large amounts of data available to individuals. Methods used in the past for dealing with information were adequate when information was scarce, but they do not scale up to handle the enormous amount of information that is now easily accessible.

Research is a fundamental activity of knowledge workers, whether they are scientists, engineers or business executives. While each discipline may have its own interpretation of research, the primary meaning of the word is "a careful and thorough search." In most cases, the thing one is searching for is information. In other words, one of the most important activities of modern educated individuals is searching for information. Whole industries have arisen to meet the need for thorough searching. These include libraries, newspapers, magazines, abstracting services and online search services.

Not surprisingly, the search process itself has been studied at least since the 1930s [8], and a standard model was developed by the mid 1960s [2]. In this model, the

searcher has an "information need" which the searcher tries to satisfy using a large collection or *corpus* of information sources. The information sources that satisfy the searcher's needs are the "relevant" information sources. The searcher expresses an information need using a formal statement called a *query*. Queries may be expressed using topics, categories and/or words. The query is then given to a search intermediary. In the past the intermediary was a person who specialized in searching. It is more common today for the intermediary to be a computer system. Such systems are called *information retrieval systems* or *online search engines*. The search intermediary tries to match the topics, categories and/or words from the query with information sources in the corpus. The intermediary responds with a set of information sources that, so it is hoped, satisfy the searcher's needs.

Queries are certainly not the only way to find information in a corpus. Another very commonly used technique is to follow citations or *references* within the documents in the corpus. This technique is called *browsing*. Online browsing tools are now becoming very popular. Such a tool allows a searcher to follow references contained in information sources, often by simply "clicking" on a word or picture within the information source. In the standard model for information retrieval, a sharp distinction is made between searching using queries and searching using references.

In the standard model, the quality of a search is measured using two numbers. The first number represents how thorough the search was. It is the fraction of the total number of relevant information sources that are presented to the searcher. This number is called the *recall*. If the recall is less than 100%, then some relevant information sources have been missed. The second number represents how careful the search was. It is the fraction of the information sources presented to the searcher that are judged to be relevant. This number is called the *precision*. If the precision is less than 100%, then some irrelevant information sources were presented to the searcher.

One can always increase the recall by adding many more information sources to those already presented, thereby ruining the precision. One would like to balance the recall and precision so as to achieve a search that is as careful and thorough as possible. Typical online search engines can achieve only about 60% recall and 40% precision. Surprisingly, these performance rates have not changed significantly in the last 20 years.

Relevance is the central concept in human (as opposed to computer) communication. This was recognized already in the 1940s when information science was first being formed as a discipline. The first formal in-depth discussion of relevance occurred in 1959 [10], and the topic was discussed intensively during the 1960s and early 1970s. As a result of such discussions, researchers began to study relevance from a human perspective. The two best known studies were by Cuadra and Katter [3, 4] and by Rees and Schultz [7] both of which appeared in 1967. The main conclusions of these studies are that the standard model for information retrieval is not compatible with

how people perceive relevance.

The present invention eliminates this defect of current search tools.

3 Summary of the Invention

The invention relates to a computer system which includes a processor, display device, main memory and one or more secondary storage devices, and which operates as a system for classifying information sources.

In the following description, numerous specific details are set forth describing specific representations of data such as graphical displays and hierarchical displays, in order to provide a thorough understanding of the present invention. However, it will be apparent to one of ordinary skill in the art to which the present invention pertains, that the present invention may be practiced without the specific details disclosed herein. In other instances, well known systems or processes have not been shown in detail in order not to obscure the present invention unnecessarily.

3.1 Relating information sources with visual graphic structures

When the knowledge expressed by an information source is represented as a computer data structure, the data structure is called a *knowledge representation* [9]. A knowledge representation can be visualized using a graphic structure. The graphic structure consists of vertices joined by directed edges. A vertex represents a concept. The concept can be represented using a word, phrase and/or icon. A vertex may also contain a category. The category can be represented either textually or by a distinct shape, color and/or icon. An edge may be labeled by an edge type. Different types of edges can be distinguished by using a textual label or by using a distinct shape, color and/or icon. Two vertices that are joined by an edge are called *adjacent vertices.*

The categories, concepts and edge types are specified by the *ontology.* An ontology models knowledge within a particular domain. In addition to the categories, concepts and edge types, an ontology can include specialized vocabulary, syntactic forms and inference rules. In particular, an ontology specifies the features that information sources can possess as well as how to extract features from information sources. When the extracted features are represented as a computer data structure, they form the knowledge representation of the information source.

1. The visual graphic structure is an alternative view of the knowledge representation of the information source. Information sources can be processed to extract the corresponding knowledge representation. The extracted knowledge representation can be visualized using a visual graphic structure.

2. The relationship between the vertices of the visual graphic structure and vertices (such as words, phrases and visual features) of the information source is visually represented by using the same feature (such as the same color or the same location on the screen) for corresponding parts of the two views. The vertices of the two views that are participating in the correspondence are said to be *highlighted*.

3. Selecting a vertex of either view causes the selected vertex and vertices adjacent to the selected vertex to be highlighted in both views. By selecting a succession of vertices in the visual graphic structure, one can perform *relevance navigation* of the information source. By selecting vertices of the information source, one can perform *relevance exploration* of the information source.

4. A *query* is an information requirement. Like any other information source, a query can be processed to extract the corresponding knowledge representation. The extracted knowledge representation can be visualized using a visual graphic structure. The knowledge representation of the query is compared with the knowledge representations of the available information sources. Those information sources that have a substructure which matches the query in full or in part, are classified by the largest matching substructure of the query. Substructures are also called *subgraphs*. The matching of substructures in the query with substructures in the information sources is called *subgraph isomorphism*.

5. If a query is highly specific, then subgraph isomorphism suffices as a technique for classification of relevant information sources. When the query is unspecific a different strategy is employed, because an unspecific query matches far too many information sources for a user to process. This happens, for example, when the query consists of a single, commonly occurring word. The strategy for unspecific queries is to classify information sources using structures (called *supergraphs*) that contain more features than the original query. Supergraphs are constructed by a process of adding new vertices, so that each vertex being added is adjacent to another vertex in the supergraph. In addition, each supergraph must occur in at least one information source as part or all of its knowledge representation. In this way the large set of relevant information sources is subclassified into smaller sets of information sources. The user is presented a list of relevant supergraphs rather than a set of information sources. The subclassifications may also be subclassified in the same manner. The classifications and subclassifications form a hierarchical structure, called a *taxonomy* or *classification hierarchy*.

6. If a query is neither highly specific nor very unspecific, then both subgraphs and supergraphs may be employed to classify the information sources relevant to the query.

7. When one requests the *Next Occurrence* of a visual graphic structure, the system searches the current information source knowledge representation for another substructure equivalent to the visual graphic structure occurring later in the information source. If such an equivalent substructure is found, then the corresponding vertices of the information source are highlighted. Similarly requesting the *Previous Occurrence* searches for an equivalent substructure occurring earlier in the information source.

3.2 Classification

Information sources are classified using a lattice of structures. A *lattice* [1, 5] is an ordered set for which each pair of elements has a *least upper bound* and a *greatest lower bound*. *Formal concept analysis* is a branch of mathematics that studies hierarchies of concepts based on attributes [11, 6]. The hierarchies that occur in formal concept analysis are lattices of concepts ordered by generality, i.e., a concept A is less than a concept B if A is less general (more specific) than B. Formal concept analysis is an established field with applications in knowledge processing. The concepts and lattices that occur in the present invention differ from those that occur in formal concept analysis, since the concepts of the present invention are based on knowledge representations, not on attributes.

In the case of information source classifications, the lattice of structures may be constructed in several ways. The simplest construction is obtained by using the knowledge representation of the query as the top of the lattice. The structures in the lattice are substructures of the query. Such structures are called *subgraphs* of the query. The subgraphs of the query are arranged by containment of one subgraph in another. This construction method is best suited for highly specific queries.

The lattice of structures may also be constructed by using the knowledge representation of the query as the bottom of the lattice. Structures in the lattice, in this case, are structures that contain the query. Such structures are called *supergraphs*. The supergraphs of the query are arranged by containment of one supergraph in another. This construction method is best suited for very simple, unspecific queries.

In general, the lattice of structures is constructed by using both subgraphs and supergraphs of the query. Each information source is classified by the largest structures in the lattice that are contained in the knowledge representation of the information source. A single information source can belong to more than one classification.

4 Description of the Drawings

This invention is pointed out with particularity in the appended claims. The above and further advantages of the invention may be better understood by referring to the

following description taken in conjunction with the accompanying drawing, in which:

FIG. 1 is a block diagram of an overview of an embodiment of the system of the invention;

FIG. 2 is an overview of the steps used by the embodiment of the system to respond to a query.

5 Detailed Description of the Invention

Referring to FIG. **1**, in broad overview, one embodiment of a system of the invention includes a user computer **101** which is in communication with a classification engine **102,103,104** through a network **103**. The individual computer nodes may include local disks, or may, alternatively or additionally, obtain data from a network disk server.

The computer nodes of the classification engine may be of several types, including home nodes **102** and index nodes **104**. The nodes of the classification engine need not represent distinct computers. In one embodiment, the classification engine consists of a single computer which takes on the roles of all home nodes and index nodes. In another embodiment, the classification engine consists of separate computers for each home node and index node. Those skilled in the art will realize many variations are possible which will still be within the scope and spirit of the present invention.

Considering the processing of a query, and referring also to FIG. **2**, in one embodiment when a user transmits a query **201** to the classification engine, a home node **102** receives the query **201**. The home node **102** is responsible for establishing the connection with the user computer **101** to enable the user to transmit a query and to receive a response in an appropriate format. The home node **102** is also responsible for any authentication and administrative functionality. In one embodiment, the home node **102** is a World Wide Web server communicating with the user computer **101** using the HTTP protocol.

After verifying that the query **201** is acceptable, the home node **102** performs any reformatting necessary to make the query **201** compatible with the requirements of the search engine. The home node **102** then transmits **202** the query to a knowledge extractor **203** which utilizes the ontology **215** to extract a knowledge representation from the query **201** The knowledge representation **205** may be transmitted back to the user computer **101** where the knowledge representation **205** may be edited by the user. Alternatively, the user may transmit a knowledge representation **205** directly to the classification engine **102** without the step of knowledge extraction.

Upon receiving confirmation from the user, if necessary, the home node **102** transmits **206** the query knowledge representation **205** to a high recall retrieval engine **207** which produces **208** a collection of information source knowledge representations **209**.

This collection **209** is transmitted **210** to the graph matching processor **211** along with the query knowledge representation **205,214**. The graph matching processor **211** makes use of the ontology **215** to perform any appropriate inference rules during isomorphism testing. The graph matching processor **211** produces **212** an ordered list of categories **213** each containing a subgraph or supergraph of the query knowledge representation **205** and a set of relevant information sources and/or knowledge representations. This aforementioned ordered list **213** is transmitted to the user computer **101** and displayed appropriately.

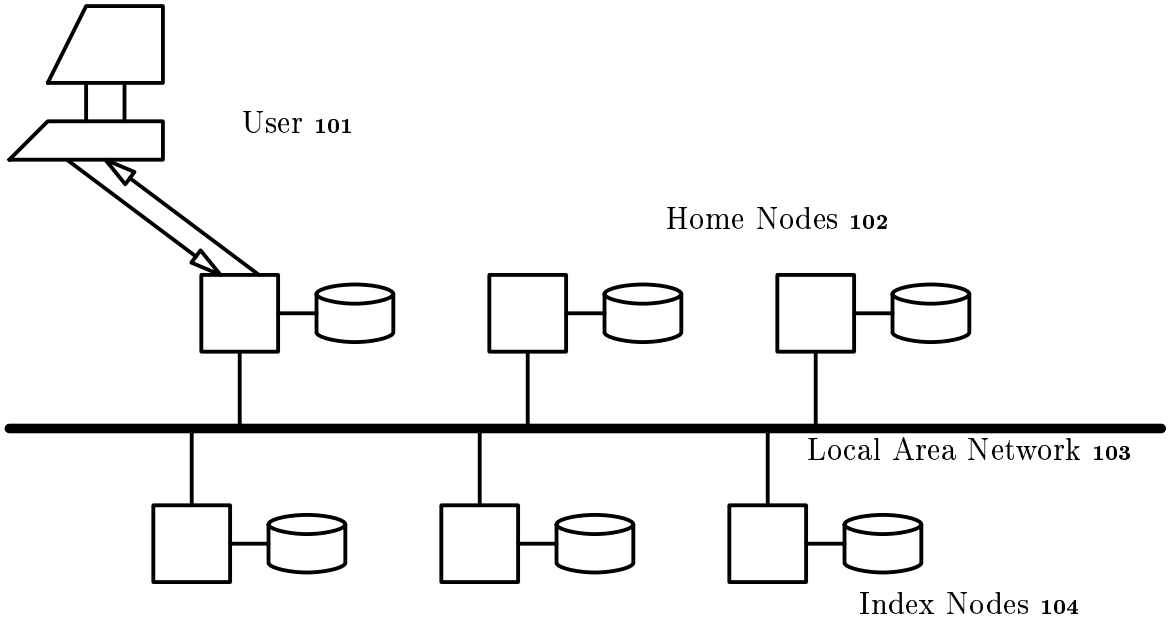
User 101
Home Nodes 102
Local Area Network 103
Index Nodes 104

FIG. **1**

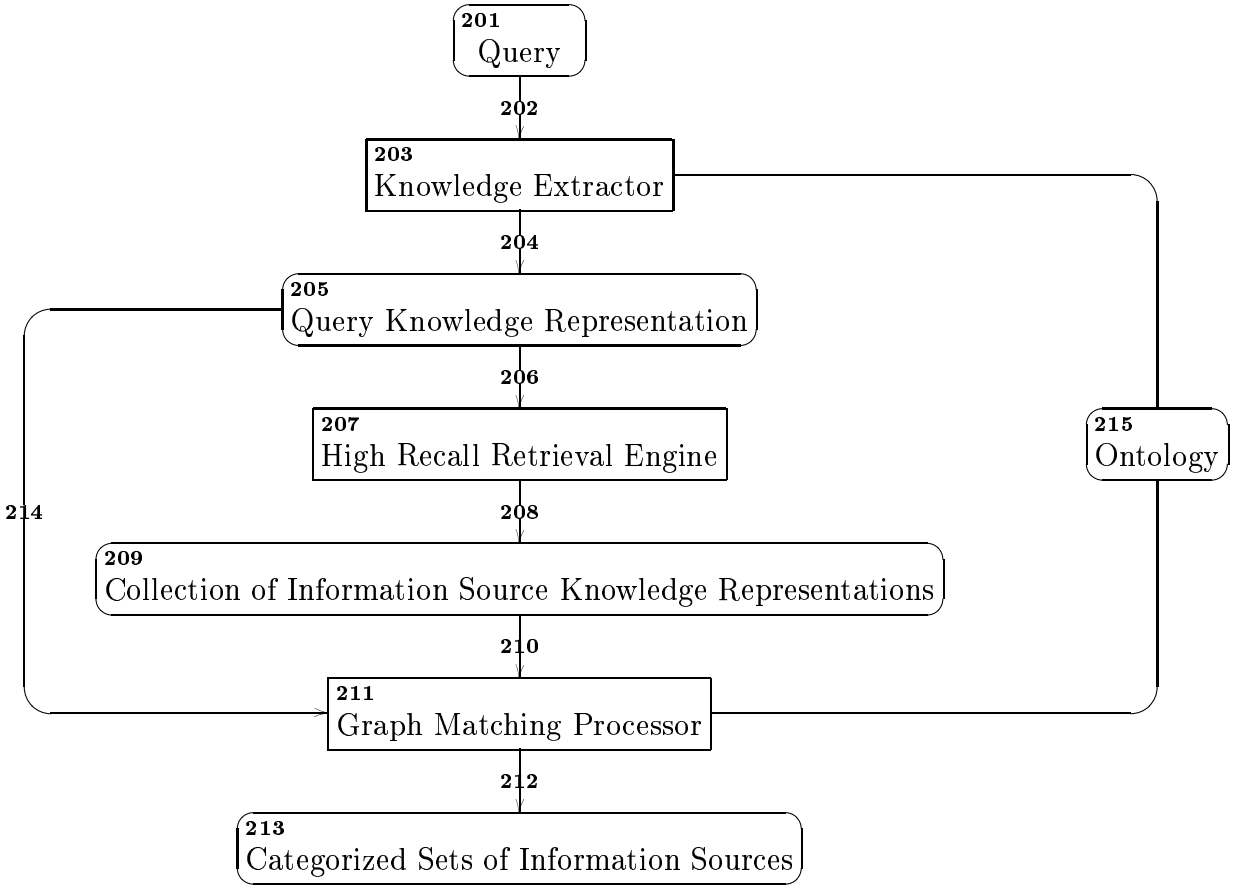
201
Query
202
203
Knowledge Extractor
204
205
Query Knowledge Representation
206
207
High Recall Retrieval Engine
208
209
Collection of Information Source Knowledge Representations
210
211
Graph Matching Processor
212
213
Categorized Sets of Information Sources
214
215
Ontology

FIG. **2**