ISBN: 978-93-90847-20-4

2. Information Storage and Retrieval System: An Evaluation

Govind Kumar Gautam

Librarian,
JNV, Baanswara, Raj.

Maya Gautam

Librarian,
JNV, Baanswara, Raj.

Abstract:

For a wide variety of sensor network environments, location information is unavailable or expensive to obtain. We propose a location-free, lightweight, distributed, and data-centric storage/retrieval scheme for information producers and information consumers in sensor networks. Our scheme is built upon the Gradient Landmark-Based Distributed Routing protocol, a two-level routing scheme where sensor nodes are partitioned into tiles by their graph distances to a small set of local landmarks so that localized and efficient routing can be achieved inside and across tiles. Our information storage and retrieval scheme uses two ideas on top of the GLIDER hierarchy — a distributed hash table on the combinatorial tile adjacency graph and a double-ruling scheme within each tile. Queries follow a path that will probably reach the data replicated by the producer. We show that this scheme compares favorably with previously proposed schemes, such as Geographic Hash Tables, providing comparable data storage performance and better locality-aware data retrieval performance.

More importantly, this scheme uses no geographic information, makes few assumptions on the network model, and achieves better load balancing and structured data processing and aggregation even for sensor fields with complex geometric shapes and non-trivial topology.

2.1 Introduction:

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications. Making effective use of the vast amount of data gathered by large-scale sensor networks requires scalable and energy-efficient data storage and data dissemination algorithms. Queries on sensor networks may be content-based, in that users are primarily interested in data satisfying certain attributes, not in the details of which node currently contains the data.

An information producer generates data that may be of interest to multiple information consumers located in other parts of the network at a much later time. An information

brokerage scheme is a mechanism that carries out data publication, data replication for the information producers and data retrieval for the information consumers in the sensor network setting, we formulate it as a mechanism to enable a network of nodes to self-organize and store the sensed data, cooperate to route and answer the query. The utility of a sensor network derives primarily from the data it gathers. Previous work has addressed data-centric routing and data-centric storage as efficient data management schemes for sensor networks. In data-centric routing, low-level communications are based on names that are external to the network topology and relevant to the applications. A typical data-centric routing protocol, directed diffusion, uses a flooding algorithm to distribute interests to match with data obtained at source nodes.

Matched data are delivered back to the sink (consumer) on reversed paths, the best of which will be reinforced for continuing future use. Little collaborative preprocessing is performed on the data gathered by the sensors in such schemes. Thus the discovery of the desired information has to rely on flooding the network. Information retrieval is the method of searching information in documents, documents themselves or metadata that describes these documents.

This definition is not dependent on method of document storage, or their type which determines the content of information being searched. This can be a search in the local database or in the Internet for text, images, sound, or data. Information retrieval is a loosely-defined term and the problem of information retrieval can be investigated under different aspects. This paper deals with the automatic information retrieval tasks of the information represented as text.

2.1.1 Information as a Resource:

Is considered an economic resource, somewhat on par with other resources such as labor, material, and capital. This view stems from evidence that the possession, manipulation, and use of information can increase the cost-effectiveness of many physical and cognitive processes. The rise in information-processing activities in banking industry as well as in human problem solving problem has been remarkable.

2.1.2 Information as a Commodity:

Complementary to definitions of information as a commodity is the concept of an information production chain through which information gains in economic value. The notion of information as a commodity incorporates “the exchange of information among people and related activities as well as its use” [8] implies buyer, sellers and a market. In contrast to the absence of power of information as a resource, information as a commodity has economic power.

2.1.3 Information as Perception of Pattern:

Here the concept of information is broadened by the addition of context. Information “has a past and a future, is affected by motive and other environmental and casual factors, and itself has effect [8].

The concept of information and its processes is broadened so much so that information in this sense can be applied to a highly articulated social structure. Information has a power of its own although its effects are isolated. The example given is of information reducing uncertainty but only in regard to a single question.

2.1.4 Information as process:

That is, when someone is informed, or what he or she knows is changed. Information in this sense refers to the act of informing or communicating knowledge or "news" of some fact.

2.1.5 Information as Knowledge:

That is, information, being new to a recipient, serves to reduce uncertainty and improves existing knowledge. Information in this sense refers to the knowledge communicated concerning some particular facts, subject or event, which, when assimilated, changes the recipients existing knowledge.

2.1.6 Information as Thing:

Used attributively for objects, such as data in documents, because they are regarded as being informative, or having the quality of communicating information or impacting knowledge. Usage of the concept of "information" also appears to have changed over time along revolutions in computer technology. In the 1950s and 1960s, it meant the amount of reduction on uncertainty, particularly in the context of communication signals and symbols. In the 1980s it meant decision-relevant data. Hence, focused on the effective use of information by humans to solve social problems. Later, and probably in conformity with the use of „data process in to mean everything processed by computer, the word „information came to be widely used to denote „processed dat. Today, with the computer being more and more widely used to support decisions through database systems, spreadsheets and graphics, information has, once again, come to mean information that can help its users to make better decisions.

2.1.7 Information Retrieval:

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. Information retrieval systems are everywhere: Web search engines, library catalogs, store catalogs, cookbook indexes, and so on. Information retrieval (IR), also called information storage and retrieval (ISR or ISAR) or information organization and retrieval, is the art and science of retrieving from a collection of items a subset that serves the user's purpose; for example:

- Web pages useful in preparing for a trip to Europe;
- Magazine articles for an assignment or good reading for that trip to Europe;
- Educational materials for a learning objective;
- Digital cameras for taking family photos;

- Recipes that use ingredients on hand;
- Facts needed for deciding on a company merger.

The main trick is to retrieve what is useful while leaving behind what is not. Need for information retrieval tools. Information retrieval system consists of three sub system, i.e. information sub-system, user's subsystem and retrieval sub-system. Information system contains a bundle of information. User subsystem is user interface through which users submit their query and retrieval sub-system acts as a mediator that takes the search query and matches it up with existing database and provides relevant information to the users. Fourth Law of Library Science states that "Save the time of the readers." This is possible only if information form databases and other sources are clearly analyses, classified, and organized in a proper manner. In an IR system, users submit their query in form of keywords or phrases and retrieval system matches it with indexed objects and serves such matched objects to the users.

2.2 Objectives of Information Storage and Retrieval System:

Traditionally, IR has concentrated on finding whole documents consisting of written text; much IR research focuses more specifically on text retrieval – the computerized retrieval of machine-readable text without human indexing. But there are many other interesting areas:

- Speech retrieval, which deals with speech, often transcribed manually or (with errors) by automated speech recognition (ASR).
- Cross-language retrieval, which uses a query in one language (say English) and finds documents in other languages (say Chinese and Russian).
- Question-answering IR systems, which retrieve answers from a body of text. For example, the question who won the 1997 World Series? Finds a 1997 headline World Series: Marlins are champions.
- Image retrieval, which finds images on a theme or images that contain a given shape or color.
- Music retrieval, which finds a piece when the user hums a melody or enters the notes of a musical theme.
- IR dealing with any kind of other entity or object: works of art, software, courses offered at a university, people (as experts, to hire, for a date), and products of any kind. Text, speech, and images, printed or digital, carry information, hence information retrieval. Not so for other kinds of objects, such as hardware items in a store. Yet IR methods apply to retrieving books or people or hardware items, and this article deals with IR broadly, using "document" as stand-in for any type of object. Note the difference between retrieving information about objects (as in a Web store catalog) and retrieving the actual objects from the warehouse.

2.2.1 Utility of Information Storage and Retrieval System:

Utility and relevance underlie all IR operations. A document's utility depends on three things, topical relevance, pertinence, and novelty. A document is topically relevant for a topic, question, or task if it contains information that either directly answers the question or can be used, possibly in combination with other information, to derive an answer or perform

the task. It is pertinent with respect to a user with a given purpose if, in addition, it gives just the information needed; is compatible with the user's background and cognitive style so he can apply the information gained; and is authoritative. It is novel if it adds to the user's knowledge. Analogously, a soccer player is topically relevant for a team if her abilities and playing style fit the team strategy, pertinent if she is compatible with the coach and novel if the team is missing a player in her position. Utility might be measured in monetary terms: "How much is it worth to the user to have found this document?" "How much is this player worth to us?" "How much did we save by finding this software?" In the literature, the term "relevance" is used imprecisely; it can mean utility or topical relevance or pertinence. Many IR systems focus on finding topically relevant documents, leaving further selection to the user. Relevance is a matter of degree; some documents are highly relevant and indispensable for the user's tasks; others contribute just a little bit and could be missed without much harm (see ranked retrieval in the section on Matching). Evaluation studies commonly use recall and precision or a combination; whether these are the best measures is debatable. With low precision, the user must look at several irrelevant documents for every relevant document found. More sophisticated measures consider the gain from a relevant document and the expense incurred by having to examine an irrelevant document. For ranked retrieval, performance measures are more complex. All of these measures are based on assessing each document on its own, rather than considering the usefulness of the retrieved set as a whole; for example, many relevant documents that merely duplicate the same information just waste the user's time, so retrieving fewer relevant documents would be better.

2.3 Database:

The term or expression of database originated within the computer industry. A possible definition is that a database is a structured collection of records or data which is stored in a computer so that a program can consult it to answer queries. The records retrieved in answer to queries become information that can be used to make decisions. The computer program used to manage and query a database is known as a Database Management System (DBMS). The properties and designs of database systems are included in the study of information science. The central concept of a database is that of a collection of records, or pieces of knowledge. Topically, for a given database, there is a structural description of the type of facts held in that database: this description is known as a schema. The schema describes the objects that are represented in the database, and the relationships among them. There are a number of different ways of organizing a schema, that is, of modeling the database structure: these are known as database models (or data models). The model in most common use today is the relational model, which in layman's terms represents all information in the form of multiple related tables each consisting of rows and columns (the true definition uses mathematical terminology). This model represents relationships by the use of values common to more than one table. Other models such as the hierarchical model and the network model use a more explicit representation of relationships. The term database refers to the collection of related records, and the software should be referred to as the database management system or DBMS. When the context is unambiguous, however, many database administrators and programmers use the term database to cover both meanings.

Many professionals would consider a collection of data to constitute a database only if it has certain properties: for example, if the data is managed to ensure its integrity, if it allows shared access by a community of users, if it has a schema, or if it supports a query language.

However, there is no agreed definition of these properties. Database management systems are usually categorized according to the data model that they support: relational, object-relational, network, and so on. The data model will tend to determine the query languages that are available to access the database. A great deal of the internal engineering of a DBMS, however, is independent of the data model, and is concerned with managing factors such as performance, concurrency, integrity, and recovery from hardware failures. In these areas there are large differences between problems.

2.3.1 Types of Databases:

Databases are usually categorized according to their various models. Various techniques are used to model data structure. Most database systems are built around one particular data model, although it is increasingly common for products to offer support for more than one model. For any one logical model various physical implementations may possible, and most products will offer the user some level of control in tuning the physical implementation, since the choices that are made have a significant effect on performance.

An example of this is the relational model: all serious implementations of the relational model allow the creation of indexes which provide fast access to rows in a table if the values of certain columns are known. A data model is not just a way of structuring data: it also defines a set of operations that can be performed on the data. The relational model, for example defines operations such as select; project; and join. Be explicit in a particular query language, they provide the foundation on which a query language is built.

a. Flat Model:

This may not strictly qualify as a data model, as defined above. The flat (or table) model consists of a single, two-dimensional array of data elements, where all members of a given column are assumed to be similar values, and all members of a row are assumed to be related to one another. For instance, columns for name and password that might be used as a part of a system security database; each row would have the specific password associated with an individual user. Columns of the table often have a type associated with them, defining them as character data, date or time information, integers, or floating point numbers. The model is, incidentally, a basis of the spreadsheet.

b. Hierarchical Model:

In a hierarchical model, data is organized into a tree-like structure, implying a single upward link in each record to describe the nesting, and a sort field to keep the records in a particular order in each same-level list. Hierarchical structures were widely used in the early mainframe database management systems, such as the Information Management System (IMS) by IBM, and now describe the structure of XML documents. This structure allows one 1: N relationship between two types of data. This structure is very efficient to describe many relationships in the real world; recipes, table of contents, ordering of paragraphs/verses, any nested and sorted information. However, the hierarchical structure is inefficient for certain database operations when a full path (as opposed to upward link and sort field) is not also included for each record.

c. Network Model:

The network model (defines by the CODASYL specification) organizes data using two fundamental constructs, called records and sets. Records contain fields which may be organized hierarchically, as in the programming language COBOL). Sets (not to be confused with mathematical sets) define one-to-many relationships between records: one owner, many members. A record may be an owner in any number of sets, and a member in any number of sets. The operations of the network model are navigational in style: a program maintains a current position, and navigates from one record to another by following the relationships in which the record participates. Records can also be located by supplying key values. Although it is not an essential feature of the model, network databases generally implement the set relationships by means of pointers that directly address the location of a record on disk. This gives excellent retrieval performance, at the expense of operations such as database loading and reorganization.

2.4 Challenges of effective Information and Storage and retrieval System:

The intensive penetration of computers and other information and telecommunication technology in the former socialist countries- countries in transition has contributed to the automation of many information activities in organizations and enterprises and has changed considerably their ability to access and use information from distant information resources. This, however, has not changed the information behavior in these countries; it has not automatically triggered higher information awareness or changed the attitudes towards information and communication activities. In spite of the technological possibilities and an increased number of computer engineers and information systems professionals, most organizations continue to face serious information problems due to the lack of interdisciplinary knowledge required for an integrated approach to the complex information activities involved in every aspect of work and doing business. Neither librarians nor other information professionals have the interdisciplinary knowledge needed for organizing and managing information activities in a broader context.

Thus they cannot fully understand the information phenomenon and the implications of the global information societies and information highway trends. Under these circumstances an abundance of money is spent on expensive technology and gadgetry that is not exploited to the greater benefit of the organization. All organizations and enterprises, regardless of the socioeconomic and political systems in which they operate, need enormous amounts of information. This is particularly true for those in transitional economies. Many organizations and enterprises in countries in transition suffer from inefficient and ineffective administration and exploitation of their information resources. They have no organized special libraries, information centers or services of any kind and also suffer from a lack of suitably trained professional library information manpower. Relevant information, whether internally generated or externally available, is still not tapped. Management and operational functions both the macro and micro levels (in government and in private organizations) are performed without the benefit of timely, relevant and reliable information. In many organizations we find a great number of different information resources managed in a diffuse way. There are no vertical or horizontal connections and the resources are not applied in a synergistic way toward the fulfillment of strategic objectives [30].

As far as special libraries and information services are concerned, managers do not, as yet, recognize that locating; accessing, retrieving, processing and disseminating information are activities of great importance for the successful functioning of their organizations. Managers still tend to see the library/information Centre as a cost, rather than as a strategic, resource. The lack of appreciation of the role, functions and services of special libraries and information centers has led to a situation in which organizations have no proper instruments to make them aware of the wealth of domestic and foreign information sources that technically are now available in the countries in transition. Such a situation is very symptomatic when it is widely recognized that organizations need specialized help in dealing with information, as noted in a recent statement by Peter Ducker:

To think through what the business needs requires somebody who knows and understands the highly specialized information field. There is far too much information for any but specialists to find their way around. The sources are totally diverse. Companies can generate some of the information about themselves, such as information about customers and non-customers or about technology in one's own field. But most of what enterprises need to know about the environment is obtainable only from outside sources-from all kinds of data banks and data services, from journals in many languages, from trade associations, from government publications, from Would Bank reports, etc.

2.4.1 How Information Retrieval Systems Work:

IR is a component of an information system. An information system must make sure that everybody it is meant to serve has the information needed to accomplish tasks, solve problems, and make decisions, no matter where that information is available. To this end, an information system must actively find out what users need, acquire documents (or computer programs, or products, or data items, and so on), resulting in a collection, and match documents with needs. Determining user needs involves studying user needs in general as a basis for designing responsive systems (such as determining what information students typically need for assignments), and actively soliciting the needs of specific users, expressed as query descriptions, so that the system can provide the information Figuring out what information the user really needs to solve a problem is essential for successful retrieval. Matching involves taking a query description and finding relevant documents in the collection; this is the task of the IR system. Relevant items correctly retrieved all relevant items in the collection relevant items retrieved all items retrieved irrelevant items correctly rejected all irrelevant items in the collection.

2.4.2 Indexing: Creating Document Representations:

Indexing (also called cataloging, metadata assignment, or metadata extraction) is the manual or automated process of making statements about a document, lesson, person, and so on, in accordance with the conceptual scheme. We focus here on subject indexing – making statements about a document's subjects. Indexing can be document-oriented – the indexer captures what the document is about, or request-oriented – the indexer assesses the document's relevance to subjects and other features of interest to users; for example, indexing the testimonies in with Jewish-Gentile relations, marking a document as interesting for a course, or marking a photograph as publication quality.

Related to indexing is abstracting – creating a shorter text that describes what the full document is about (indicative abstract) or even includes important results (informative abstract, summary). Automatic summarization has attracted much research interest. Automatic indexing begins with raw feature extraction, such as extracting all the words from a text, followed by refinements, such as eliminating stop words (and, it, of), stemming (pipes Y pipe), counting (using only the most frequent words), and mapping to concepts using a thesaurus (tube and pipe map to the same concept). A program can analyze sentence structures to extract phrases, such as labor camp (a Nazi camp where Jews were forced to work, often for a company; phrases can carry much meaning). For images, extractable features include color distribution or shapes. For music, extractable features include frequency of occurrence of notes or chords, rhythm, and melodies; refinements include transposition to a different key. Raw or refined features can be used directly for retrieval. Alternatively, they can be processed further: The system can use a classifier that combines the evidence from raw or refined features to assign descriptors from a pre-established index language. To give an example from the classifier uses the words life and model as evidence to assign bioinformatics (a descriptor in Google's directory). A classifier can be built by hand by treating each descriptor as a query description and building a query formulation for it as described in the next section. Or a classifier can be built automatically by using a training set, such as the list of documents for bioinformatics in, for machine learning of what features predict what descriptors. Many different words and word combinations can predict the same descriptor, making it easier for users to find all documents on a topic Assigning documents to (mutually exclusive) classes of a classification is also known as text categorization. Absent a suitable classification, the system can produce one by clustering – grouping documents that are close to each other (that is, documents that share many features).

2.4.3 Query Formulation: Creating Query Representations:

Retrieval means using the available evidence to predict the degree to which a document is relevant or useful for a given user need as described in a free-form query description, also called topic description or query statement. The query description is transformed, manually or automatically, into a formal query representation (also called query formulation or query for short) that combines features that predict a document's usefulness. The query expresses the information need in terms of the system's conceptual schema, ready to be matched with document representations. A query can specify text words or phrases the system should look for (free-text search) or any other entity feature, such as descriptors assigned from a controlled vocabulary, an author's organization, or the title of the journal where a document was published. A query can simply give features in an unstructured list (for example, a "bag of words") or combine features using Boolean operators (structured query).

2.4.4 Matching the Query Representation with Entity Representations:

The match uses the features specified in the query to predict document relevance. In exact match the system finds the documents that fill all the conditions of a Boolean query (it predicts relevance as 1 or 0). To enhance recall, the system can use synonym expansion (if the query asks for pipe, it finds tubes as well) and hierarchic expansion or inclusive searching (it finds capillary as well).

Since relevance or usefulness is a matter of degree, many IR systems (including most Web search engines) rank the results by a score of expected relevance (ranked retrieval).

2.5 Relevance Feedback and Interactive Retrieval:

Once the user has assessed the relevance of a few items found, the query can be improved: The system can assist the user in improving the query by showing a list of features (assigned descriptors; text words and phrases, and so on) found in many relevant items and another list from irrelevant items. Or the system can improve the query automatically by learning which features separate relevant from irrelevant items and thus are good predictors of relevance. A simple version of automatic query adjustment is this: increase the weights of features from relevant items and decrease the weights of features from irrelevant items.

2.6 Experimental Design:

To support the workflow proposed in the beginning of the paper, which is shown in, experimental design should bind preparation phase with available workload, configured system and acquired measurements. Even having such a narrow scope, it has many aspects. And it is not possible to generalize on the experimental design for information retrieval systems evaluation in the wide sense. Most of such existing experiments use just one set of requests/queries to evaluate or compare a number of systems. It is called 'matched pairs' procedure when the efficiency of the systems is compared on the same request. Moreover, there is a clear statistical reason for such approach.

Any statistical significance testing will be much more efficient with this method. Again, this approach is oriented to decrease the influence of the bottle-neck of the whole evaluation process which is the amount of requests. With this approach it is possible to reuse the requests decreasing the need in the larger number of distinct requests. Therefore, experimental design can be quite simple. Each request/query is searched against every system or every system configuration. Since the searching part of the system is controlled by simple rules, there is no problem in relation to replicating searches or the order in which the systems are tried. The only matter of convenience in case of a single system evaluation is performing the evaluation for the whole request set for a single configuration, further reconfiguring the system and performing the complete course of evaluation for a new setup.

2.7 Measurements:

Now, we know how to perform experiment. In order to be able to answer the question we have posed at the beginning we need to perform statistical evaluation of the measurements taken in the course of experiment. There are number of ways to measure how well the retrieved information matches the intended one. We will use the standard recall, precision and measures.

2.8 Evaluation of Information Storage and Retrieval System:

IR systems are evaluated with a view to improvement (formative evaluation) or with view to selecting the best IR system for a given task (summative evaluation).

IR systems can be evaluated on system characteristics and on retrieval performance. System characteristics include.

The following:

- The quality of the conceptual schema (Does it include all information needed for search and selection?);
- The quality of the subject access vocabulary (index language and thesaurus) (Does it include the necessary concepts? Is it well structured? Does it include all the synonyms for each concept?);
- The quality of human or automated indexing (Does it cover all aspects for which an entity is relevant at a high level of specificity, while avoiding features that do not belong?);
- The nature of the search algorithm;
- The assistance the system provides for information needs clarification and query formulation; and
- The quality of the display (Does it support selection?).

Measures for retrieval performance (recall, discrimination, precision, novelty) were discussed in the section Relevance and IR system performance. Requirements for recall and precision vary from query to query, and retrieval performance varies widely from search to search, making meaningful evaluation difficult. Standard practice evaluates systems through a number of test searches, computing for each a single measure of goodness that combines recall and precision, and then averaging over all the queries. This does not address very important system ability: the ability to adapt to the specific recall and precision requirements of each individual query. The biggest problem in IR evaluation is to identify beforehand all relevant documents (the recall base); small test collections have been constructed for this purpose, but there is a question of how well the results apply to large-scale real-life collections. The most important evaluation efforts of this type today are TREC and TDT (see Further Reading).

2.9 Conclusions:

There is no such thing as a watertight method for evaluating an information retrieval system. Any existing approach to evaluating or comparing information retrieval systems will have to deal with heuristics to some extend only for the reason of this process been highly dependent on human factor. In this paper we discussed the approach to conducting evaluation of information retrieval system starting from preparation of workload, conducting experiment and finally statistical data analysis. This approach is suitable for comparison of two information retrieval models or evaluation of a single system under different configurations of the model used. The work reported in this paper can be treated as the set of instructions to take in order to perform quantitative evaluation of any information retrieval system. It is also possible to reuse some of the parts of the proposed approach or extend it to suit specific requirements. However, theoretical value of this work is in its completeness, thus future pure realization of the proposed approach is highly encouraged. Researchers in the information retrieval field have devoted a significant amount of time in developing good, standardized evaluation techniques.