Automated Generation of Metadata for Studying and Teaching about Africa from an Africancentric Perspective: Opportunities, Barriers and Signs of Hope

Abdul Karim Bangura

Abstract

As ironical as it may seem, most of what I know today about African history I learned in the West, and the opportunities availed to me to travel back and forth to conduct research in Africa were made possible by living in the United States. Yet, as I have demonstrated in a relatively recent work (Bangura, 2005), after almost three centuries of employing Western educational approaches, many African societies are still characterized by low Western literacy rates, civil conflicts and underdevelopment. It is obvious that these Western educational paradigms, which are not indigenous to Africans, have done relatively little good for Africans. Thus, I argued in that work that the salvation for Africans hinges upon employing indigenous African educational paradigms which can be subsumed under the rubric of *ubuntugogy*, which I defined as the art and science of teaching and learning undergirded by humanity towards others. Therefore, *ubuntugogy* transcends *pedagogy* (the art and science of teaching), *andragogy* (the art and science of helping adults learn), *ergonagy* (the art and science of helping people learn to work), and *heutagogy* (the study of self-determined learning). As I also noted, many great African minds, realizing the debilitating effects of the Western educational systems that have been forced upon Africans, have called for different approaches. One of the biggest challenges for studying and teaching about Africa in Africa at the higher education level, however, is the paucity of published material. Automated generation of metadata is one way of mining massive datasets to compensate for this shortcoming. Thus, this essay raises and addresses the following three major research questions: (1) What is automated generation of metadata and how can the technique be employed from an Africancentric perspective? (2) What are the barriers for employing this approach? (3) What signs are on the horizon that point to possibilities of overcoming these barriers? After addressing these questions, conclusions and recommendations are offered.
Keywords: Metadata, Data Mining, Information and Communication Technology (ICT), Africa, Africancentrism, *Ubuntugogy*

Introduction

While a great deal of attention has been paid to the "digital divide" within developed countries and between those countries and the developing ones, most Africans do not even have such luxury as access to books, periodicals, radio and television channels, which is precisely why information and communication technology (ICT) is so important to Africa. ICT has the potential to have a positive impact on Africa's development. So, how can Africans transform that potential into reality? And how can Africans access that technology? Without access, that technology cannot do much for Africans—thus, the essence of digital technology.

The digital often refers to the newest ICT, particularly the Internet. There are, of course, other more widely available forms of ICT, such as radio and telephones. But there are many problems concerning the generally abysmal state of networks of every kind on the continent that make it

difficult to fully utilize the development potential of even this technology. Africa's electrical grid is grossly inadequate, resulting in irregular or nonexistent electrical supplies. The biggest problem is that in many countries, significant power distribution networks are non-existent in rural areas.

Africa's phone systems are spotty and often rely on antiquated equipment, and progress is hamstrung by bureaucracy and, in most instances, state-owned monopolies. But African governments have the power to alter these circumstances and, gradually, some are doing so. The signs of progress are unbelievable. A few years ago, a couple of countries had Internet access. Today, all 54 countries and territories in Africa have permanent connections, and there is also rapidly growing public access provided by phone shops, schools, police stations, clinics, and hotels.

Although Africa is becoming increasingly connected, access to the Internet, however, is progressing at a limited pace. Of the 770 million people in Africa, only one in every 150, or approximately 5.5 million people in total, now uses the Internet. There is roughly one Internet user for every 200 people, compared to a world average of one user for every 15 people, and a North American and European average of about one in every two people. An Internet or E-mail connection in Africa usually supports a range of three to five users. The number of dial-up Internet subscribers now stands at over 1.3 million, up from around one million at the end of 2000. Of these, North Africa accounts for about 280,000 subscribers and South Africa accounts for 750,000 (Lusaka Information Dispatch, 2003). Kenya now has more than 100,000 subscribers and some 250 cyber cafes across the country (BBC, 2002). The widespread penetration of the Internet in Africa is still largely confined to the major cities, where only a minority of the total population lives. Most of the continent's capital cities now have more than one internet service provider (ISP); and in early 2001, there were about 575 public ISPs across the continent. Usage of the Internet in Africa is still considered a privilege for a few individuals and most people have never used it (Lusaka Information Dispatch, 2003).

In Zambia, for example, there are now about five ISPs, which include Zamnet, Microlink, Coppernet, Uunet, and Sambia Telecommunication Service (ZAMTEL), which is government owned. Most people in Lusaka go to Internet cafes to check for their E-mail unlike surfing the Internet to conduct research (Lusaka Information Dispatch, 2003).

Indeed, ICT can play a substantial role to improve access to all forms of education (formal schooling, adult literacy, and vocational educational training) and to strengthen the economic and democratic institutions in African countries. It can also help to address the major issue of this essay: i.e. one of the biggest challenges for studying and teaching about Africa in Africa at the higher education level being the paucity of published material. I suggest in this paper that automated generation of metadata is one way of mining massive datasets to compensate for this shortcoming. Thus, this essay raises and addresses the following three major research questions: (1) What is automated generation of metadata and how can the technique be employed from an Africancentric perspective? (2) What are the barriers for employing this approach? (3) What signs are on the horizon that point to possibilities of overcoming these barriers? After addressing these questions, conclusions and recommendations are offered.

Automated Generation of Metadata

The capabilities of generating and collecting data, observed Alshameri (2006), have been increasing rapidly. The computerization of many business and government transactions with the attendant advances in data collection tools, he added, has provided huge amounts of data. Millions of databases have been employed in business management, government administration, scientific and engineering management, and many other applications. This explosive growth in data and databases

has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge (Chen et al., 1996). This chapter explores the nature of data mining and how it can be used in doing research on African issues.

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing upon such areas as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

Data mining denotes a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis to engineering design and scientific exploration.

There are also many other concepts, appearing in some literature, carrying a similar or slightly different definitions, such as knowledge mining from databases, knowledge extraction, data archaeology, data dredging, data analysis, etc. By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, thereby serving as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity for major revenue generation. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications. Researchers in many different fields, including database systems, knowledge-based systems, artificial intelligence, machine learning, knowledge acquisition, statistics, spatial databases, and data visualization have shown great interest in data mining. Furthermore, several emerging applications in information providing services, such as on-line services and the World Wide Web, also call for various data mining techniques to better understand user behavior in order to ameliorate the service provided and to increase business opportunities.

Mining Massive Data Sets

Recent years have witnessed an explosion in the amount of digitally-stored data, the rate at which data is being generated, and the diversity of disciplines relying upon the availability of stored data. Massive data sets are increasingly important in a wide range of applications, including observational sciences, product marketing, and the monitoring and operations of large systems. Massive datasets are collected routinely in a variety of settings in astrophysics, particle physics, genetic sequencing, geographical information systems, weather prediction, medical applications, telecommunications, sensors, government databases, and credit card transactions. The nature of these data is not limited to a few esoteric fields, but, arguably, to the entire gamut of human intellectual pursuits, ranging from images on Web pages to exabytes ($\sim 10^{18}$ bytes) of astronomical data from sky surveys (Hambrusch et al., 2003).

There is a wide range of problems and application domains in science and engineering that can benefit from data mining. In several of these fields, techniques similar to data mining have been used

for many years, albeit under different names (Kamath, 2001). For example, in the area of remote sensing, rivers and boundaries of cities have been identified using image understanding methods. Much of the use of data mining techniques in the past has been for data obtained from observations of experiments, as one-dimensional signals or two-dimensional images. These techniques, however, are increasingly attracting the attention of scientists involved in simulating complex phenomena on massively parallel computers. They realize that, among other benefits, the semi-automated approach of data mining can complement visualization in the analysis of massive datasets produced by the simulations.

There are different areas which provide for the use of data mining. The following are some examples:

(a) *Astronomy and Astrophysics* have long used data mining techniques such as statistics that aid in the careful interpretation of observations that are an integral part of astronomy. The data being collected from astronomical surveys are now being measured in terabytes ($\sim 10^{12}$ bytes), because of the new technology of the telescopes and detectors. These datasets can be easily stored and analyzed by high performance computers. Astronomy data present several unique challenges. For example, there is frequently noise in the data due to the sensors used for collecting data: atmospheric disturbances, etc. The data may also be corrupted by missing values or invalid measurements. In the case of images, identifying an object within an image may be non-trivial, as natural objects are frequently complex and image processing techniques based on the identification of edges or lines are inapplicable. Furthermore, the raw data, which are in high-dimensional space, must be transformed into a lower-dimensional feature space, resulting in a high pre-processing cost. The volumes of data are also large, further exacerbating the problem. All these characteristics, in addition to the lack of ground truth, make astronomy a challenging field for the practice of data mining (Grossman et al., 2001).

(b) *Biology, Chemistry, and Medicine*—informatics, chemical informatics and medicine are all areas where data mining techniques have been used for a while and are increasingly gaining acceptance. In bioinformatics, which is a bridge between biology and information technology, the focus is on the computational analysis of gene sequences (Cannataro et al., 2004). Here, the data can be gene sequences, expressions, or protein information. Expressions mean information on how the different parts of a sequence are activated, whereas protein data represent the biochemical and biophysical structures of the molecules. Research in bioinformatics related to sequencing of the human genome evolved from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. In medicine, image mining is used on the analysis of images from mammograms, MRI scans, ultrasounds, DNA micro-arrays and X-rays for tasks such as identifying tumors, retrieving images with similar characteristics, detecting changes, and genomics. In addition to these tasks, data mining can be employed in the analysis of medical records.

In the chemical sciences, the information overload problem is becoming staggering as well, with the chemical abstract service adding about 700,000 new compounds to its database each year. Chemistry data are usually obtained either by experimentation or by computer simulation. The need for effective and efficient data analysis techniques is also being driven by the relatively new field of combinatorial chemistry, which essentially involves reactivating a set of starting chemicals in all possible combinations, thereby producing large datasets. Data mining is being used to analyze chemical datasets for molecular patterns and to identify systematic relationships between various chemical compounds.

(c) *Earth Sciences, Climate Modeling, and Remote Sensing* are replete with data mining opportunities. They cover a broad range of topics, including climate modeling and analysis, atmospheric sciences, geographical information systems, and remote sensing. As in the case of astronomy, this is another area in which the vast volumes of data have resulted in the use of semi-automated techniques for data analysis. Earth science data can be particularly challenging from a practical view point, and they come in many different formats, scales and resolutions. Extensive work is required to pre-process the data, including image processing, feature extraction, and feature selection. It suffices to say that the volumes of earth sciences data are typically enormous, with the NASA Earth Observing System expected to generate over 11,000 terabytes of data upon completion. Much of these data is stored in flat files, not databases.

(d) *Computer Vision and Robotics* are characterized by a substantial overlap. There are several ways in which the two fields can benefit each other. For example, computer vision applications can benefit from the accurate machine learning algorithms developed in data mining, while the extensive work done in image analysis and fuzzy logic for computer vision and robotics can be used in data mining as well, especially for applications involving images (Kamath, 2001). The applications of data mining methodologies in computer vision and robotics are quite diverse. They include automated inspection in industries for tasks such as detecting errors in semiconductor masks and identifying faulty widgets in assembly line productions; face recognition and tracking of eyes, gestures, and lip movements for problems such as lip-reading; automated television studios, video conferencing and surveillance, medical imaging during surgery as well as for diagnostic purposes, and vision for robot motion control. One of the key characteristics of the problems in computer vision and robotics is that they must be done in real time (Kamath 2001). In addition, the data collection and analysis can be tailored to the task being performed as the objects of interest are likely to be similar to one another.

(e) *Engineering*—with sensors and computers becoming ubiquitous and powerful, and engineering problems becoming more complex, there is a greater focus on gaining a better understanding of these problems through experiments and simulations. As a result, large amounts of data are being generated, providing an ideal opportunity for the use of data mining techniques in areas such as structural mechanics, computational fluid dynamics, material science, and the semi-conductor industry. Data from sensors are being used to address a variety of problems, including detection of land mines, identification of damage in aerodynamic systems (e.g., helicopters) or physical structures (e.g., bridges), and nondestructive evaluation in manufacturing quality control, to name just a few. In computer simulation, which is increasingly seen as the third mode of science, complementing theory and experiment, the techniques from data mining are yet to gain widespread acceptance (Marusic et al., 2001). Data mining techniques are also employed in studying the identification of coherent structures in turbulence.

(f) *Financial Data Analysis*—most banks and other financial institutions offer a wide variety of banking services such as checking, savings, and business and individual customer transactions. Added to that are credit services like business mortgages and investment services such as mutual funds. Some also offer insurance and stock investment services. Financial data collected in the banking and financial industries are often relatively complete, reliable and of high quality, which facilitate systematic data analysis and data mining. Classification and clustering methods can be used for customer group identification and targeted marketing. For example, customers with similar behaviors regarding banking and loan payments may be grouped together by multidimensional clustering techniques (Han et al., 2001). Effective clustering and collaborative

filtering methods such as decision trees and nearest neighbor classification can help in identifying customer groups, associate new customers with an appropriate customer group, and facilitate targeted marketing. Data mining can also be used to detect money laundering and other financial crimes by integrating information from multiple databases, as long as they are potentially related to the study.

(g) *Security and Surveillance* comprise another active area for data mining methodologies. They include applications such as fingerprint and retinal identification, human face recognition, and character recognition in order to identify people and their signatures for access, law enforcement or surveillance purposes. Data mining techniques can also be used in tasks such as automated target recognition.

The preceding areas have benefited from the scientific and engineering advances in data mining. Added to these are various technological areas that produce enormous amounts of data, such as high energy physics data from particle physics experiments that are likely to exceed a petabyte ($\sim 10^{15}$ bytes) per year and data from the instrumentation of computer programs run on massively parallel machines that are too voluminous to be analyzed manually. What is becoming clear, however, is that the data analysis problems in science and engineering are getting more complex and more pervasive, giving rise to a great opportunity for the application of data mining methodologies. Some of these opportunities are discussed in the following subsection.

Requirements and Challenges of Mining Massive Data Sets

In order to conduct effective data mining, one needs to first examine what kind of features an applied knowledge discovery system is expected to have and what kind of challenges one may face at the development of data mining techniques. The following are some of the challenges:

(a) *Handling of different types of high-dimensional data.* Since there are many kinds of data and databases used in different applications, one may expect that a knowledge discovery system should be able to perform effective data mining on different kinds of data. Massive databases contain complex data types, such as structured data and complex data objects, hypertext and multimedia data, spatial and temporal data, remote sequencing, transaction data, legacy data, etc. These data are typically high-dimensional, with attributes numbering from a few hundreds to the thousands. There is an urgent demand for new techniques for data retrieval and representation, new probabilistic and statistical models for high-dimensional indexing, and database querying methods. A powerful system should be able to perform effective data mining on such complex types of data as well.

(b) *Efficiency and scalability of data mining algorithms.* With the increasing size of data, there is a growing appreciation for algorithms that are scalable. To effectively extract information from a huge amount of data in databases, the knowledge discovery algorithms must be efficient and scalable to large databases. Scalability refers to the ability to use additional resources such as the central processing unit (CPU) and memory in an efficient manner to solve increasingly larger problems. It describes how the computational requirements of an algorithm grow with problem size.

(c) *Usefulness, certainty and expressiveness of data mining results*. Scientific data, especially data from observations and experiments, are noisy. Removing the noise from data, without affecting the signal, is a challenging problem in massive datasets. Noise, missing or invalid data, and exceptional data should be handled elegantly in data mining systems. The discovered knowledge should accurately portray the contents of the database and be useful for certain applications.

(d) *Building reliable and accurate models and expressing the results*. Different kinds of knowledge can be discovered from a large amount of data. These discovered kinds of knowledge can be examined from different views and presented in different forms. This requires the researcher to build a model that reflects the characteristics of the observed data and to express both the data mining requests and the discovered knowledge in high-level languages or graphical user interfaces, so that the discovered knowledge can be understandable and directly usable.

(e) *Mining distributed data*. The widely available local and wide-area computer networks, including the Internet, connect many sources of data and form huge distributed heterogeneous databases, such as the text data that are distributed across various Web servers or astronomy data that are distributed as part of a virtual observatory. On the one hand, mining knowledge from different sources of formatted or unformatted data with diverse data semantics poses new challenges to data mining. On the other hand, data mining may help disclose the high-level data regularities in heterogeneous databases which can hardly be discovered by simple query systems. Moreover, the huge size of the database, the wide distribution of data, and the computational complexity of some data mining methods motivate the development of parallel and distributed data mining algorithms.

(f) *Protection of privacy and data security*. When data can be viewed from many different angles and at different abstraction levels, it can threaten the goal of ensuring data security and guarding against the invasion of privacy (Chen et al., 1996). It is important to study when knowledge discovered may lead to an invasion of privacy and what security measures can be developed to prevent the disclosure of sensitive information.

(g) *Size and type of data*. Science datasets range from moderate to massive, with the largest being measured in terabytes. As more complex simulations are performed and observations over long periods at higher resolution are conducted, the data will grow to the petabyte range. Data mining infrastructure should support the rapidly increasing data volume and the variety of data formats that are used in the scientific domain.

(h) *Data visualization*. The complexity and noise of massive data affect data visualization. Scientific data are collected from variant sources by using different sensors. Data visualization is needed to use all available data to enhance an analysis. Unfortunately, a difficult problem may emerge when data are collected on different resolutions, using different wavelengths, under different conditions, with different sensors (Kamath, 2001). Collaborations between computer scientists and statisticians are yielding statistical concepts and modeling strategies to facilitate data exploration and visualization. For example, recent work in multivariate data analysis involves ranking multidimensional observations based on their relative importance for information extraction and modeling, thereby contributing to the visualization of high dimensional objects such as cell gene expression, profile and image.

Mining Spatial Databases

The study and development of data mining algorithms for spatial databases are motivated by the large amount of data collected through remote sensing, medical equipment, and other instruments. Managing and analyzing spatial data became an important issue due to the growth of the applications that deal with geo-reference data. A spatial database stores a large amount of space-related data, such as maps, pre-processed remote sensing or medical imaging data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques. Another difference is the query language that is employed to access spatial data. The complexity of the spatial data type is another important feature (Palacio et al., 2003).

The explosive growth in data and databases used in business management, government administration and scientific data analysis has created the need for tools that can automatically transform the processed data into useful information and knowledge. Spatial data mining is a subfield of data mining that deals with the extraction of implicit knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases (Koperski et al., 1995). Such mining demands an integration of data mining with spatial database technologies. It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and non-spatial data, constructing spatial knowledge databases, reorganizing spatial databases, and optimizing spatial queries. It is expected to have wide applications in geographic imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are employed (Han et al., 2001).

A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods. Challenges in spatial data mining arise from different issues. First classical data mining is designed to process numbers and categories, whereas spatial data are more complex and include extended objects such as points, lines, and polygons. Second, while classical data mining works with explicit inputs, spatial predicates and attributes are often implicit. Third, classical data mining treats each input independently of other inputs, while spatial patterns often exhibit continuity and high autocorrelation among nearby features (Shekhar et al., 2002).

Related Work

Statistical spatial data analysis has been a popular approach used to analyze spatial data. This approach handles numerical data well and usually suggests realistic models of spatial phenomena. Different methods for knowledge discovery, algorithms and applications for spatial data mining are created. Classification of spatial data has been analyzed by some researchers. A method for classification of spatial objects has been proposed by Ester et al. (1997). Their proposed algorithm is based on ID3 algorithm, and it uses the concept of neighborhood graphs. It considers not only non-spatial properties of the classified objects, but also non-spatial properties of neighboring objects: objects are treated as neighbors if they satisfy the neighborhood relations. Ester et al. (2000) also define topological relations as those which are invariant under topological transformations. If both objects are rotated, translated, or scaled simultaneously, the relations are preserved. These scholars present a definition of topological relations derived from the nine intersections model: i.e. the topological relations between two objects are (1) disjoint, (2) meet, (3) overlap, (4) equal, (5) cover,

(6) covered-by, (7) contain, and (8) inside; the second type of relations refers to (9) distance relations. These relations compare the distance between two objects with a given constant using arithmetic operators like <, >, and =. The distance between objects is defined as the minimum distance between them. The third type of relations they define are the direction relations. They define a direction relation *A R B* of two spatial objects using one representative point of the object A and all points of the destination object B. It is possible to define several possibilities of direction relations depending on the points that are considered in the source and the destination objects. The representative point of a source object may be the center of the object or a point on its boundary. The representative point is used as the origin of a virtual coordinate system, and its quadrants define the directions.

Fayyad et al. (1996) used decision tree methods to classify images of stellar objects to detect stars and galaxies. About three terabytes of sky images were analyzed. Similar to the mining association rules in transactional and relational databases, spatial association rules can be mined in spatial databases. Spatial association describes the spatial and non-spatial properties which are typical for the target objects but not for the whole database (Ester et al., 2000). Koperski et al. (1995) introduced spatial association rules that describe associations between objects based on spatial neighborhood relations. An example can be the following:

is_a(X,"African_countries")^receiving(X,"Western_aid")→highly(X,"corrupt")[0.5%,90%]

This rule states that 90% of African countries receiving Western aid are also highly corrupt, and 0.5% of the data belongs to such a case.

Spatial clustering identifies clusters or densely populated regions according to some distance measurement in a large, multidimensional dataset. There are different methods for spatial clustering such as the k-mediod clustering algorithms (Ng et al. 1994) and the Generalized Density Based Spatial Clustering of Applications with Noise (GDBSCAN) that relies on a destiny-based notion of clusters (Sander et al., 1998).

Visualizing large spatial datasets became an important issue due to the rapidly growing volume of spatial datasets, which makes it difficult for a human to browse such datasets. Shekhar et al. (2002) have constructed a Web-based visualization software package for observing the summarization of spatial patterns and temporal trends. The visualization software will help users gain insight and enhance their understanding of the large data.

Mining Text Databases

Text databases consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, E-mail messages, and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic forms, such as electronic publications, E-mail, CD-ROMs, and the World Wide Web (which also can be considered as a huge interconnected dynamic text and multimedia database).

Data stored in most text databases are semi-structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as a title, author's name(s), publication date, length, category, etc., and also contain some largely unstructured text components, such as an abstract and contents.

Traditional information retrieval techniques have become inadequate for the increasingly vast amounts of text data (Han et al., 2001). Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the

documents, it is difficult to formulate effective queries for extracting and analyzing useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining.

Text mining has also emerged as a new research area of text processing. It focuses on the discovery of new facts and knowledge from large collections of texts that do not explicitly contain the knowledge to be discovered (Gomez et al., 2001). The goals of text mining are similar to those of data mining, since it attempts to find clusters, uncover trends, discover associations, and detect deviations in a large set of texts. Text mining has also adopted techniques and methods of data mining, such as statistical techniques and machine learning approaches. Text mining helps one to dig out the hidden gold from textual information, and it leaps from old fashioned information retrieval to information and knowledge discovery (Dorre et al., 1999).

Basic Measures for Text Retrieval

Information retrieval is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, however, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text based documents. A typical information retrieval problem is to locate relevant documents based on user input, such as keywords or example documents. This type of information retrieval system includes online library catalog systems and online document management systems (Berry et al., 1999).

It is vital to know how accurate or correct a text retrieval system is in retrieving documents based on a query. The set of documents relevant to a query can be called "{Relevant}," whereas the set of documents retrieved is denoted as "{Retrieved}." The set of documents that are both relevant and retrieved is denoted as "{Relevant}∩{Retrieved}." There are two basic measures for assessing the quality of a retrieval system: (1) precision and (2) recall (Berry at al., 1999).

The precision of a system is the ratio of the number of relevant documents retrieved to the total number of documents retrieved. In other words, it is the percentage of retrieved documents that are in fact relevant to the query—i.e. the correct response. Precision can be represented as follows:

$$\text{Precision} = \frac{\left| \{R\textit{elevant}\} \cap \{\textit{Retrieval}\} \right|}{\left| \{R\textit{etrieved}\} \right|}$$

The recall of a system is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection. Stated differently, it is the percentage of documents that are relevant to the query and were retrieved. Recall can be represented the following way:

$$\text{Recall} = \frac{\left| \{R\textit{elevant}\} \cap \{R\textit{etrieval}\} \right|}{\left| \{R\textit{elevant}\} \right|}$$

Word Similarity

Information retrieval systems support keyword-based and/or similarity-based retrieval. In keyword-based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A good information retrieval system should consider synonyms when answering the

queries. For example, synonyms such as "automobile" and "vehicle" should be considered when searching the keyword "car." There are two major difficulties with a keyword-based system: (1) synonymy and (2) polysemy. In a synonymy problem, keywords such as "software product" may not appear anywhere in a document, even though the document is closely related to a software product. In a polysemy problem, a keyword such as "regression" may mean different things in different contexts.

The similarity-based retrieval system finds similar documents based on a set of common keywords. The output of such retrieval should be based on the degree of relevance, where relevance is measured in terms of the closeness and relative frequency of the keywords.

A text retrieval system often associates a stop list with a set of documents. A stop list is a set of words that are deemed "irrelevant." For instance, "a," "the," "of," "for," "with," etc. are stop words, even though they may appear frequently. The stop list depends on the document itself: for example, together, "artificial intelligence" could be an important keyword in a newspaper; it may, however, be considered a stop word on research papers presented at an artificial intelligence conference.

A group of different words may share the same word stem. A text retrieval system needs to identify groups of words in which the words in a group are small syntactic variants of one another, and collect only the common word stem per group. For example, the group of words "drug," "drugged," and "drugs" share a common word stem, "drug," and can be viewed as different occurrences of the same word.

Panel et al. (2002) computed the similarity among a set of documents or between two words, w_i and w_j, using the cosine coefficient of their mutual information vectors:

$$sim(w_i, w_j) = \frac{\sum_c {}^{mi}w_ic \times {}^{mi}w_jc}{\sqrt{\sum_c {}^{mi}w_ic^2 \times \sum_c {}^{mi}w_jc^2}}$$

where ${}^{mi}w_ic$ is the positive mutual information between context (c) and the word (w). $F_c(w)$ be the frequency count of the word, w, occurring in context c:

$$Mi_{wic} = \frac{\frac{F_c(w)}{N}}{\frac{\sum_i F_i(w)}{N} \times \frac{\sum_j F_c(J)}{N}}$$

where $N = \sum_i \sum_j F_i(j)$, is the total frequency counts of all words and their context.

Related Work

Text mining and applications of data mining to structured data derived from texts have been the subjects of many research projects in recent years. Most text mining has used natural language processing to extract key terms and phrases directly from documents.

Pantel at al. (2002) have proposed a clustering algorithm, Clustering By Committee (CBC), in which the centroid of a cluster is constructed by averaging the feature vectors of a subset of the

cluster members. The subset is viewed as a committee that determines which other elements belong to the cluster. Pantel and his partners divided the algorithm into three phases. In the first phase, they found top similar elements. To compute the top similar words of a word *w*, they sorted *w*'s features according to their mutual information with *w*. They computed only the pairwise similarities between *w* and the words that share high mutual information features with *w*. In the second phase, they found committees—a set of recursively tight clusters in the similarity space—and other elements that are not covered by any committee. An element is said to be covered by a committee if that element's similarity to the centroid of the committee exceeds some high similarity threshold. Assigning elements to clusters is the last phase on the CBC algorithm. In this phase, every element is assigned to the cluster containing the committee to which it is most similar.

Wong et al. (1999) designed a text association system based on ideas from information retrieval and syntactic analysis. In their system, the corpus of narrative text is fed into a text engine for topic extractions, and then the mining engine reads the topics from the text engine and generates topic association rules which are sent to the visualization system for further analysis. There are two text engines developed on this system. The first one is word-based and results in a list of content-bearing words for the corpus. The second one is concept-based and results in concepts based on the corpus.

Dhillon et al. (2001) designed a vector space model to obtain a highly efficient process for clustering very large collections exceeding 100,000 documents in a reasonable amount of time on a single processor. They used efficient and scalable data structures such as local and global hash tables. In addition, a highly efficient and effective *spherical k-means* algorithm was used, since both the document and concept vectors lie on the surface of a high-dimensional sphere. The basic idea of the vector space model is to represent each document as a vector of certain weighted word frequencies. Each vector component reflects the importance of a particular term in representing the semantics or meaning of that document. The vectors for all documents in a database are stored as the columns of a single matrix (Berry et al., 1999).

A database containing a total of *d* documents described by *t* terms is represented as a $t \times d$ *term-by-document matrix A*. The *d* vectors representing the *d* documents form the columns of the matrix. Thus, the matrix element a_{ij} is the weighted frequency at which the term *i* occurs in document *j*. In the vector space model, the columns of *A* are the *document vectors*, whereas the rows of *A* are the *term vectors*.

To create the vector space model, there are important parsing and extraction steps needed, such as all unique words from the entire set of documents. Eliminate all "stop words" such as "a," "and," "the," etc. For each document, count the number of occurrences of each word. Eliminate "high-frequency" and "low-frequency" non-content-bearing words by using the heuristic or information-theoretic criteria. Finally, for each word, *w*, assign a unique identifier between one to *w* to each remaining word and unique identifier between one to *d* to each document. The geometric relationship between document vectors and also the term vectors can help to identify the similarities and differences between the document's content and also in term usage.

In the vector space model, in order to find the relevant documents, the user queries the database by using the vector space representation of those documents. Query matching is finding the documents most similar to the query in use and weighting the terms. In the vector space model, the documents selected are those geometrically closest to the query according to some measure.

Mining Remote Sensing Data

The data volumes of remote sensing are rapidly growing. National Aeronautics and Space Administration's (NASA) Earth Observing System (EOS) program alone produces massive data

products with total rates more than 1.5 terabytes per day (King et al., 1999). Application and products of Earth observing and remote sensing technologies have been shown to be crucial to global social, economic and environmental well being (Yang et al., 2001).

In order to help scientists search massive remotely sensed databases and find data of interest to them, and then order the selected data sets or subsets, several information systems have been developed for data ordering purposes to face the challenges of the rapidly growing volumes of data, since the traditional method where a user downloads data and uses local tools to study the data residing on a local storage system is no longer helpful. To find interesting data, scientists need an effective and efficient way to search through the data. Metadata are provided in a database to support data searching by commonly used criteria such as spatial coverage, temporal coverage, spatial resolution, and temporal resolution. Since metadata search itself may still result in large amounts of data, some textual restrictions, such as keyword searches, could be employed for interdisciplinary researchers to select data of interest to them. The usual data selection procedure is to specify a spatial/temporal range and to see what datasets are available under those conditions.

Yang and his colleagues (1998) developed a distributed information system, the Seasonal International Earth Science Information Partner (SIESIP), which is a federated system that provides services for data searching, browsing, analyzing and ordering. The system will provide not only data, but also data visualization, analysis and user support capabilities.

The SIESIP system is a multi-tiered, client-server architecture, with three physical sites or nodes, distributing tasks in the areas of user service, access to data and information products, archiving as needed, ingest and interoperability options, and other aspects. This architecture can serve as a model for many distributed Earth system science data. There are three phases of user interaction with the data and information system; each phase can be followed by other phases or it can be conducted independently:

Phase 1, Metadata Access: using the metadata and browse images provided by the SIESIP system, the user browses the data holding. Metadata knowledge is incorporated into the system and a user can issue queries to explore this knowledge.

Phase 2, Data Discovery/Online Data Analysis: the user gets a quick estimate of the type and quality of data found in Phase 1. Analytical tools are then utilized as needed, such as statistical functions and visualization algorithms.

Phase 3, Data Order: after the user locates the dataset of interest, s/he is now ready to order datasets. If the data are available through SIESIP, the information system will handle the data order; otherwise, an order will be issued to the appropriate data provider such as Earth Science Distributed Active Archive Center (GES DAAC), on behalf of the user, or necessary information will be forwarded to the user for this task for further action.

The database management system is used for the system to handle catalogue metadata and statistical summary data. The database system supports two major kinds of queries. The first query is used to find the right data files for analysis and ordering based on catalogue metadata only. The second one queries data contents which are supported by the statistical summary data.

Data mining techniques help scientific data users not only in finding rules or relations among different data but also in finding the right datasets. With the rapid growth of massive data volumes, scientists need a fast way to search for the data in which they are interested. In this case, scientists need to search data based not only on metadata but also actual data values. The main goal of the data mining techniques in the SIESIP system is to find spatial regions and/or temporal ranges over

which parameter values fall into certain ranges. The main challenge is the speed and the accuracy, because they affect each other inversely.

Different data mining techniques can be applied on remote sensing data. Classification of remotely sensed data is used to assign corresponding levels with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from one another within the image. Also, several methods exist for remote sensing image classification. Such methods include both traditional statistical or supervised approaches and unsupervised approaches that usually employ artificial neural networks.

Mining Astronomical Data

Astronomy has become and immensely data-rich field, with numerous digital sky surveys across a range of wavelengths and many terabytes of pixels and billions of detected sources, often with tens of measured parameters for each object. The problem with the astronomical database is not only the very large size, but also the variable quality of the data and the nature of astronomical objects with their very wide dynamic range in apparent luminosity and size present additional challenges. The great changes in astronomical data enable scientists to map the universe systematically, and in a panchromatic manner. Scientists can study the galaxy and the large scale structure in the universe statistically and discover unusual or new types of astronomical objects and phenomena (Brunner et al., 2002).

Astronomical data and their attendant analyses can be classified into the following five domains:
(1) *Imaging data* are the fundamental constituents of astronomical observations, capturing a two-dimensional spatial picture of the universe within a narrow wavelength region at a particular epoch or instance of time.

(2) *Catalogs* are generated by processing the imaging data. Each detected source can have a large number of measured parameters, including coordinates, various flux quantities, morphological information, and a real extant.

(3) *Spectroscopy, polarization, and other follow-up measurements* provide detailed physical quantification of the target systems, including distance information, chemical composition, and measurements of the physical fields present at the source.

(4) *Studying the time domain* provides important insights into the nature of the universe by identifying moving objects, variable sources, or transient objects.

(5) *Numerical simulations* are theoretical tools which can be compared with observational data.

Handling and exploring these vast new data volumes, and actually making real scientific discoveries, pose considerable technical challenges. Many of the necessary techniques and software packages, including artificial intelligence techniques like neural networks and decision trees, have been already successfully applied to astronomical problems such as pattern recognition and object classification. Clustering and data association algorithms have also been developed.

As early as 1936, Edwin Hubble established a system to classify galaxies into three fundamental types. First, elliptical galaxies had an elliptical shape with no other discernible structure. Second, spiral galaxies had an elliptical nucleus surrounded by a flattened disk of stars and dust containing a

spiral pattern of brighter stars. And third, irregular galaxies have irregular shapes and did not fit into the other two categories.

Humphrey and his partners (2001) created an automated classification system for astronomical data. They visually classified 1,500 galaxy images obtained from the Automated Plate Scanner (APS) database in the region of the north galactic pole. Given the size and brightness of galaxy images taken into consideration, images that were difficult to classify were removed from this sample.

Grossman and company (2001) developed an application which simultaneously works with two geographically distributed astronomical source catalogs: (1) the Two Micron All Sky Survey (2MASS) and (2) the Digital Palomar Observatory Sky Survey (DPOSS). The 2MASS data are in the optical wavelengths, whereas the DPOCC data are in the infrared range. These scientists created a virtual observatory supporting the statistical analysis of many millions of stars and galaxies with data coming from both surveys. By using the data space transfer protocol (DSTP), a platform independent way to share data over a network, they built a query for finding all pairs from the DPOSS and 2MASS. The query visualized by coloring the data as red, if they appear in 2MASS; blue, if they appear in DPOSS; or magenta, if they appear in both surveys. The client DSTP application formulates the fuzzy join and sends the resulting stars and galaxies back to the client application. The DSTP protocol enables an application in one location to locate, access, and analyze data from several other locations.

Mining Bioinformatics Data

Bioinformatics is described by Cannataro and his colleagues (2004) as a bridge between the life sciences and computer science. It has also been described by Barker and his associates (2004) as a cross-disciplinary field in which biologists, computer scientists, chemists, and mathematicians work together, each bringing a unique point of view. The term *bioinformatics* has a range of interpretations, but the core activities of bioinformatics are widely acknowledged: storage, organization, retrieval and analysis of biological data obtained by experiments or by querying databases.

The increasing volume of biological data collected in recent years has prompted increasing demand for bioinformatics tools for genomic and proteomic (the set of proteins encoded by the genome to define models representing and analyzing the structure of the proteins contained in each cell) data analysis. Bioinformatics applications' design should represent the biological data and databases efficiently; contain services for data transformation and manipulation such as searching in protein databases, protein structure prediction, and biological data mining; describe the goals and requirements of applications and expected results; and support querying and computing optimization to deal with large datasets (Wong et al., 2001).

Bioinformatics applications are naturally distributed, due to the high number of datasets involved. They require higher computing power, due to the large size of datasets and the complexity of basic computations; they may access heterogeneous data; they require a secure software infrastructure because they could access private data owned by different organizations.

Cannataro et al. (2004) show technologies that can fulfill bioinformatics requirements. The following are some of these technologies:

(a) *Ontologies* to describe the semantics of data sources, software components and bioinformatics tasks. An *ontology* is a shared understanding of well defined domains of interest, which is realized as a set of classes or concepts, properties, functions and instances.

(b) *Workflow Management Systems* to specify in an abstract way complex (distributed) applications, integrating and composing individual simple services. A workflow is a partial or total automation of a business process in which a collection of activities must be executed by some entities (humans or machines) according to certain procedural rules.

(c) *Grid Infrastructure* to show its security, distribution, service orientation, and computational power.

(d) *Problem Solving Environment* to define and execute complex applications, hiding software development details.

These researchers developed two types of ontologies: (1) *OnBrowser* for browsing and querying ontologies and (2) DAMON for the data mining domain describing resources and processes of knowledge discovery in databases. The latter is used to describe data mining experimentations on bioinformatics data.

Cannataro et al. (2004) designed PROTEUS, a Grid-based Problem Solving Environment (GPSE), for composing and running bioinformatics applications on the Grid. They used ontologies for modeling bioinformatics processes and Grid resources and workflow techniques for designing and scheduling bioinformatics applications.

PROTEUS assists users in formulating bioinformatics solutions by choosing among different available bioinformatics applications or by composing a new one as collections on the Grid. It is used to present and analyze results and then compare them with past results to form the PROTEUS knowledge base. PROTEUS combines existing open source bioinformatics software and public-available biological databases by adding metadata to software, modeling applications through ontology and workflows, and offering pre-packaged Grid-aware bioinformatics applications.

Web-based bioinformatics application platforms are popular tools for biological data analysis within the bioscience community. Wong et al. (2001) developed a prototype based on integrating different biological databanks into a unified XML framework. The prototype simplifies the software development process of bioinformatics application platforms. The XML-based wrapper of the prototype demonstrated a way to convert data from different databanks into XML format and be stored in XML database management systems.

DNA data analysis is an important topic in biomedical research. Recent research in the area has led to the discovery of genetic causes for many diseases and disabilities, as well as the discovery of new medicines and approaches for disease diagnosis, prevention, and treatment. Data mining has become a powerful tool and contributes substantially to DNA analysis in the following ways, according to Han et al. (2001):

(a) *Semantic integration of heterogeneous, distributed genome database*: due to the highly distributed, uncontrolled generation and use of a wide variety of DNA data, the semantic integration of such heterogeneous and widely distributed genome databases becomes a pivotal task for systematic and coordinated analysis of DNA databases.

(b) *Similarity search and comparison among DNA sequences*: one of the most important search problems in genetic analysis is similarity search and comparison among DNA sequences. Gene sequences isolated from diseased and healthy species can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes and then finding and comparing the frequently occurring patterns of each class.

(c) *Association analysis and identification of co-occurring gene sequences*: association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships between them.

(d) *Visualization tools and genetic data analysis*: complex structures and sequencing patterns of genes are most effectively presented in graphs, trees, cuboids, and chains by various kinds of visualization tools. Visualization, thus, plays an important role in biomedical data mining.

Research Methodology

The following is a discussion of the proposed approach for mining massive datasets for studying Africa from an Africancentric perspective. It depends on the METANET concept: a heterogeneous collection of scientific databases envisioned as a national and international digital data library which would be available via the Internet. I consider a heterogeneous collection of massive databases such as remote sensing data and text data. The discussion is divided into two separate, but interrelated, sections: (1) the automated generation of metadata and (2) the query and search of the metadata.

Automated Generation of Metadata

Metadata simply means data about data. They can be characterized as any information required to makeg other data useful in information systems. Metadata are a general notion that captures all kinds of information necessary to support the management, query, consistent use and understanding of data. Metadata help users to discover, understand and evaluate data, and help data administrators to manage data and control their access and use. Metadata describe how, when and by whom a particular set of data was collected, and how the data are formatted. Metadata are essential for understanding information stored in data warehouses.

In general, there exist metadata to describe file and variable types and organization, but they have minimal scientific content data. In raw form, a dataset and its metadata have minimal usability. For example, not all the image datasets in the same file form that are produced by the satellite-based remote sensing platform are important to scientists. In fact, only the image datasets that contain certain patterns will be of interest to a scientist. Scientists need metadata about the image datasets' content to enable scientists to narrow their search time, taking into consideration the size of the datasets: i.e. terabyte datasets (Wegman, 1997).

Creating a digital object and linking it to the dataset will make the data usable, and at the same time the search operation for a particular structure in a dataset will be a simple indexing operation on the digital objects linked to the data. The objective of this process it to link digital objects with scientific meaning to the dataset at hand and make the digital objects part of the searchable metadata associated with the dataset. Digital objects will help scientists to narrow the scope of the datasets that the scientists must consider. In fact, digital objects reflect the scientific content of the data, but they do not replace the judgment of the scientists.

The digital objects will essentially be named for patterns to be found in the datasets. The goal is to have a background process, launched either by the database owner or, more likely, via an applet created by a virtual data center, examining databases available on the data-Web and searching within datasets for recognizable patterns. Once a pattern is found in a particular dataset, the digital object

corresponding to that pattern is made part of the metadata associated with that set. If the same pattern is contained in other distributed databases, pointers would be added to that metadata pointing to metadata associated with the distributed databases. The distributed databases will then be linked through the metadata in the virtual data center.

At least one of the following three different methods is to be used to generate the patterns for which a researcher can search. The first method is to delineate empirical or statistical patterns that have been observed over a long period of time and may be thought to have some underlying statistical structure. An example of an empirical or statistical pattern is a certain pattern in DNA sequencing. The second method is to generate the model-based patterns. This method is predictive if verified on real data. The third method is to tease out patterns found by clustering algorithms. With this method, patterns are delineated by purely automated techniques that may or may not have scientific significance (Wegman, 1997).

Query and Search

The notion of the automated creation of metadata is to develop metadata that reflect the scientific content of the datasets within the database rather than just data structure information. The locus of the metadata is the virtual data center where it is reproduced. The general *desideratum* for the scientist is to have a comparatively vague question that can be sharpened as s/he interacts with the system. Scientists can create a query, and this query may be sharpened to another vague one, but the data may be accessible from several distributed databases for the later query. The main issues of the retrieval process are the browser mechanism for requesting data when the user has a precise query and an expert system query capability that would help the scientist reformulate a vague question into a form that may be submitted more precisely.

Query and search would comprise four major elements: (1) a client browser, (2) an expert system for query refinement, (3) a search engine, and (4) a reporting mechanism. These four elements are described in the following subsections.

Client Browser

The client browser would be a piece of software running on a scientist's client machine, which is likely to be a personal computer (PC) or a workstation. The main idea is to have a graphical user interface (GUI) that would allow the user to interact with a more powerful server in the virtual data center. The client software is essentially analogous to the myriad of browsers available for the World Wide Web (WWW).

Expert Systems for Query Refinement

A scientist interacts with a server via two different scenarios. In the first scenario, the scientist knows precisely the location and type of data s/he desires. In the second scenario, the scientist knows generally the type of questions s/he would like to ask, but has little information about the nature of the databases with which s/he hopes to interact. The first scenario is relatively straightforward, but the expert system would sill be employed to keep a record of the nature of the query. The idea is to use the query as a tool in the refinement of the search process. The second scenario is more complex. The approach is to match a vague query formulated by the scientist to

one or more of the digital objects discovered in the automated generation of metadata phase. Disciplined experts give rules to the expert system to perform this match. The expert system would attempt to match the query to one or more digital objects. The scientist has the opportunity to confirm the match when s/he is satisfied with the proposed match or to refine the query. The expert system would then engage the search engine in order to synthesize the appropriate datasets. The expert system would also take advantage of the interaction to form a new rule for matching the original query to the digital objects developed in the refinement process. Thus, two aspects emerge: (1) the refinement of the precision of an individual search and (2) the refinement of the search process. Both aspects share tactical and strategic goals. The refinement would be greatly aided by the active involvement of the scientist. S/he would be informed about how his/her particular query was resolved, allowing him/her to reformulate the query efficiently. The log files of these iterative queries would be processed automatically to inspect the query trees and, possibly, improve their structure.

Also, two other considerations of interest emerge. First, other experts not necessarily associated with the data repository itself may have examined certain datasets and have commentary in either informal annotations or in the refereed scientific literature. These commentaries should form part of the metadata associated with the dataset. Part of the expert system should provide an annotation mechanism that would allow users to attach commentary or library references (particularly digital library references) as metadata. Obviously, such annotations may be self-serving and potentially unreliable. Nonetheless, the idea is to alert the scientist to information that may be useful. User derived metadata would be considered secondary metadata.

The other consideration is to provide a mechanism for indicating data reliability. This would be attached to a dataset as metadata, but it may in fact be derived from the original metadata. For example, a particular data collection instrument may be known to have a high variability. Thus, any set of data that is collected by this instrument, no matter where in the data it occurred, should have as part of the attached metadata an appropriate caveat. Hence, an automated metadata collection technique should be capable of not only examining the basic data for patterns, but also examining the metadata themselves; and, based on collateral information such as just mentioned, it should be able to generate additional metadata.

Search Engine

As noted earlier, large scale scientific information systems will likely be distributed in nature and contain not only the basic data, but also structured metadata: for example, sensor type, sensor number, measurement data, and unstructured metadata such as a text-based description of the data. These systems will typically have multiple main repository sites that together will house a major portion of the data as well as some smaller sites and virtual data centers containing the remainder of the data. Clearly, given the volume of the data, particularly within the main servers, high performance engines that integrate the processing of the structured and unstructured data are necessary to support desired response rates for user requests.

Both database management systems (DBMS) and information retrieval systems provide some functionality to maintain data. DBMS allow users to store unstructured data as binary large objects (BLOB), and information retrieval systems allow users to enter structured data in zoned fields. DBMS, however, offer only a limited query language for values that occur in BLOB attributes. Similarly, information retrieval systems lack robust functionality for zoned fields. Additionally, information retrieval systems traditionally lack efficient parallel algorithms. Using a relational database approach for information retrieval allows for parallel processing, since almost all

commercially available parallel engines support some relational DBMS. An inverted index may be modeled as a relation. This treats information retrieval as an application of a DBMS. Using this approach, it is possible to implement a variety of information retrieval functionality and achieve good run-time performance. Users can issue complex queries including both structured data and text.

The key hypothesis is that the use of a relational DBMS to model an inverted index will (a) permit users to query both structured data and text via standard Structured Query Language (SQL)—in this regard, users may use any relational DBMS that support standard SQL; (b) permit the implementation of traditional information retrieval functionality such as Boolean retrieval, proximity searches, and relevance ranking, as well as non-traditional approaches based on data fusion and machine learning techniques; and (c) take advantage of current parallel DBMS implementations, so that acceptable run-time performance can be obtained by increasing the number of processors applied to the problem.

Reporting Mechanism

The most important issue for a reporting mechanism is not only to retrieve datasets appropriate to the needs of the scientist, but scaling down the potentially large databases the scientist must consider. Put differently, the scientist would consider megabytes ($\sim 10^6$ bytes) instead of terabytes ($\sim 10^{12}$ bytes) of data. The search and retrieval process may still result in a massive amount of data. The reporting mechanism would, therefore, initially report the nature and magnitude of the datasets to be retrieved. If the scientist agrees that the scale is appropriate for his/her needs, then the data will be delivered by a file transfer protocol (FTP) or a similar mechanism to his/her local client machine or to another server where s/he wants the synthesized data to be stored.

Implementation

To help scientists search for massive databases and find data of interest to them, a good information system should be developed for data ordering purposes. The system should be performing well based on the descriptive information of the scientific datasets or metadata, such as the main purpose of the datasets, the spatial and temporal coverage, the production time, the quality of the datasets, and the main features of the datasets.

Scientists want to have an idea of what the data look like before ordering them, since metadata searching alone cannot meet all scientific queries. Thus, content-based searching or browsing and preliminary analysis of data based on their actual values will be inevitable in these application contexts. One of the most common content-based queries is to find large enough spatial regions over which the geophysical parameter values fall into certain intervals given a specific observation time. The query result could be used for ordering data as well as for defining features associated with scientific concepts.

For researchers of African topics to be able to maximize the utility of this content-based query technique, there must exist a Web-based prototype through which they can demonstrate the idea of interest. The prototype must deal with different types of massive databases, with special attention being given to the following and other aspects that are unique to Africa:

(a) African languages with words encompassing diacritical marks (dead and alive)

(b) Western colonial languages (dead and alive)

(c) Other languages such as Arabic, Russian, Hebrew, Chinese, etc.

(d) Use of desktop software such as Microsoft Word or Corel WordPerfect to type words with diacritical marks and then copy and paste them into Internet search lines

(e) Copying text in online translation sites and translating them into the target language

The underlying approach must be pluridiscipinary, which involves the use of open and resource-based techniques available in the actual situation. It has, therefore, to draw upon the indigenous knowledge materials available in the locality and make maximum use of them. Indigenous languages are, therefore, at the center of the effective use of this methodology.

What all this suggests is that the researcher must revisit the indigenous techniques that take into consideration the epistemological, cosmological and methodological challenges. Hence, the researcher must be culture-specific and knowledge-source-specific in his/her orientation. Thus, the process of redefining the boundaries between the different disciplines in our thought process is the same as that of reclaiming, reordering and, in some cases, reconnecting those ways of knowing, which were submerged, subverted, hidden or driven underground by colonialism and slavery. The research should, therefore, reflect the daily dealings of society and the challenges of the daily lives of the people. Towards this end, at least the following six questions should guide pluridisciplinary research:

(1) How can the research increase indigenous knowledge in the general body of global human development?

(2) How can the research create linkages between the sources of indigenous knowledge and the centers of learning on the continent and the Diaspora?

(3) How can centers of research in the communities ensure that these communities become "research societies"?

(4) How can the research be linked to the production needs of the communities?

(5) How can the research help to ensure that science and technology are generated in relevant ways to address problems of the rural communities where the majority of the people live and that this is done in indigenous languages?

(6) How can the research help to reduce the gap between the elite and the communities from which they come by ensuring that the research results are available to everyone and that such knowledge is drawn from the communities? (For more on this approach, see Bangura 2005.)

In the collection of remote sensing and text databases, one must implement a prototype system that contains at least a four-terabyte storage capability with high performance computing. Remote sensing data are available through NASA JPL, NASA Goddard, and NASA Langley Research Center.

The prototype system will allow scientists to make queries against disparate types of databases. For instance, queries on remote sensing data can focus on the features observed in images. Those

features may be environmental or artificial features which consist of points, lines, or areas. Recognizing features is the key to interpretation and information extraction. Images differ in their features, such as tone, shape, size, pattern, texture, shadow, association, etc.

Tone refers to the relative brightness or color objects in the image. It is the fundamental element for distinguishing between different targets or features. Shape refers to the general form, structure, or outline of an object. Shape can be a very distinctive clue for interpretation. Size of objects in an image is a function of scale. It is important to assess the size of a target relative to other objects in a scene, as well as the absolute size, to aid in the interpretation of that target. Pattern refers to the spatial arrangement of visibly discernible objects. Texture refers to the arrangement and frequency of tonal variation in a particular area of an image. Shadow will help in the interpretation by providing an idea of the profile and relative height of a target or targets which may make identification easier. Association takes into account the relationship among other recognizable objects or features in proximity to the target of interest.

Other features of the images that also should be taken into consideration include percentage of water, green land, cloud forms, snow, and so on. The prototype system will help scientists to retrieve images that contain different features; the system should be able to handle complex queries. This calls for some knowledge of African fractals, which I have defined as a self-similar pattern—i.e. a pattern that repeats itself on an ever diminishing scale (Bangura, 2000:7).

As Ron Eglash (1999) has demonstrated, first, traditional African settlements typically show repetition of similar patterns at ever-diminishing scales: circles of circles of circular dwellings, rectangular walls enclosing ever-smaller rectangles, and streets in which broad avenues branch down to tiny footpaths with striking geometric repetition. He easily identified the fractal structure when he compared aerial views of African villages and cities with corresponding fractal graphics simulations. To estimate the fractal dimension of a spatial pattern, Eglash used several different approaches. In the case of Mokoulek, for instance, which is a black-and-white architectural diagram, a two-dimensional version of the ruler size versus length plots were employed. For the aerial photo of Labbazanga, however, an image in shades of gray, a Fourier transform was used. Nonetheless, according to Eglash, we cannot just assume that African fractals show an understanding of fractal geometry, nor can we dismiss that possibility. Thus, he insisted that we listen to what the designers and users of these structures have to say about it. This is because what may appear to be an unconscious or accidental pattern might actually have an intentional mathematical component.

Second, as Eglash examined African designs and knowledge systems, five essential components (recursion, scaling, self-similarity, infinity, and fractional dimension) kept him on track of what does or does not match fractal geometry. Since scaling and self-similarity are descriptive characteristics, his first step was to look for the properties in African designs. Once he established that theme, he then asked whether or not these concepts had been intentionally applied, and started to look for the other three essential components. He found the clearest illustrations of indigenous self-similar designs in African architecture.

The examples of scaling designs Eglash provided vary greatly in purpose, pattern, and method. As he explained, while it is not difficult to invent explanations based on unconscious social forces—for example, the flexibility in conforming designs to material surfaces as expressions of social flexibility—he did not believe that any such explanation can account for its diversity. He found that from optimization engineering, to modeling organic life, to mapping between different spatial structures, African artisans had developed a wide range of tools, techniques, and design practices based on the conscious application of scaling geometry. Thus, for example, instead of using the Koch curve to generate the branching fractals used to model the lungs and acacia tree, Eglash used passive lines that are just carried through the iterations without change, in addition to active lines that create a growing tip by the usual recursive replacement.

For the text database, the prototype system must consider polysymy and synonymy problems in the queries. Polysymy means words having multiple meanings: e.g. "order," "loyalty," and "ally." Synonymy means multiple words having the same meaning: e.g., "jungle" and "forest," "tribe" and "ethnic-group," "language" and "dialect," "tradition" and "primitive," "corruption" and "lobbying."

The collected documents will be placed into categories depending on the documents' subjects. Scientists can search into those documents and retrieve only the ones related to queries of interest. Scientists can search via words or terms, and then retrieve documents on the same category or from different categories as long as they are related to the words or terms in which the scientists are interested.

Barriers

Many barriers to gain access to the Internet can hamper the automated generation of metadata for studying and teaching about Africa in the continent. These barriers include bandwith, copyright laws, costs, politics and bureaucracy, training and personnel, and unreliability or system glitches. These obstacles are discussed sequentially in the following subsections.

Bandwith

Bandwith is broadly defined as the rate of data transfer: i.e. the capacity of the Internet connection being used. The greater the bandwith, the greater the capacity for faster downloads from the Internet. It is among the biggest problems African universities face in accessing the Internet. The universities have been forced to buy bandwith from much more expensive satellite companies. What is worse is that the purchases have been made through middlemen, increasing the costs even more. Nonetheless, the evolving technology and steadily falling satellite prices could help to ameliorate this barrier. In fact, various organizations in Africa are getting together to buy bandwith as a "bridging" strategy to obtain Internet access via satellite until terrestrial fiber-optic cable becomes available. With new technology, along with the continued growth of satellite companies, this could be a leap-frog technology that will enable Africa to avoid laying much of the terrestrial cable that was essential in the developed world (Walker, 2005).

Universities involved in the Partnership for Higher Education in Africa (PHEA)—a collaborative effort involving Carnegie Corporation of New York, along with the MacArthur, Ford, and Rockefeller foundations, which have pledged $100 million over a five-year period to help strengthen African universities—tend to be among the leaders in adapting ICT. But even among them, there are considerable differences, as the pace of change can be dizzying. The University of Jos in Nigeria, for example, has blazed a trail in ICT among West African universities, generally regarded as the region that has the most problems. At Jos, one could get a broadband connection from several labs on campus and download large research files. About 12 years ago, a group of individuals at the university who were dedicated to making ICT flourish took on the challenge to make the institution gain a high level of connectivity. They did not have much money, but they had strong institutional support, going over the tenure of three vice chancellors. In 1979, Jos had a student and staff population of 15,000, but it had no ICT staff. There were less than 10 computers and fewer than 10 people who had any computer skills. But today, Jos has over 3,000 E-mail users, over 400 networked computers, all three of the campuses are linked by 15 local area networks (LANs) utilizing fiber optics and Cisco switches, and an established tradition of training (Walker, 2005).

At Makerere University in Uganda, there exists a notable program of ICT advances. Nonetheless, the university is still coping with a bandwith problem. The fundamental challenge is that satellite access is quite expensive. For every $500 an American institution pays for access per month, Makerere pays $28,000 (Walker, 2005).

Most African universities do not have enough bandwiths to access the millions of digitized images of the pages of the most prestigious journals. Organizations which offer such resources such as JSTOR have explored putting much of their material on local servers for African universities, but publishers object to the suggestion arguing that the kind of security controls and abuse monitoring they have on the Internet would not be maintained on local servers (Walker, 2005).

Copyright Laws

A strenuous challenge that has emerged in managing and accessing the Internet in Africa has to do with copyright laws. African scientists and scholars are being denied access to World Wide Web (WWW) resources because of increasingly contentious intellectual property rights debates. Among the groups involved in the debates is the PHEA, whose goal on this issue is to improve African universities' capacity to utilize technology. Currently, as I mentioned earlier, the PHEA is in the midst of an effort to help the universities gain control of the cost and training issues surrounding online access. Its initial focus has been to facilitate the formation of a coalition of African universities that will be better positioned to negotiate lower bandwith prices (Walker, 2005).

The development of limited "virtual libraries" highlights what has become the largest obstacle to African scholars' access to WWW resources. In the developed countries, there is a tradition of each student having his/her own textbooks, copies of journals, etc. The tradition just does not exist in Africa. The reality is that people in the continent copy books for 3,000 students who cannot afford to buy them. This is certainly what publishers do not want, since it violates their copyright to photocopy books and journals. The situation might be different if these publishers were to provide their products to African students, teachers and scholars at lower costs. Publishers must ask themselves just how much potential income they lose in poor African countries. Copyrights on Western business and computer science books, for example, are exorbitantly expensive for African institutions. The publishers are charging $800 for the books compared to $1,000 the universities charge for tuition. The idea of textbooks cost as much as tuition is staggering for many people. Part of the solution may hinge upon a collective bargaining process similar to the universities' bandwith consortium that would negotiate lower fees from publishers (Walker, 2005).

Publishers and database vendors are requiring African institutions to guarantee that they will know exactly who is accessing the information at all times at all the universities so that appropriate fees can be charged. African institutions are asking why vendors should care about how many students will have access to the information. They rhetorically ask how vendors can be sure that if they bought a hard copy book half the village will not just copy it. African institutions also perceive copyright-related problems as a true barrier to development. They believe that people in developing countries cannot afford to get caught up in the trap of the global intellectual property regime, which strongly favors the West, where most of the laws are made and enforced. They believe that for Africa, ICT has become a tool for survival. Anything that stands in the way is unacceptable, especially when African states are confronted with so many other challenges. They also believe that the worst thing that could happen to African nations is to become the total consumers of information, as the property rights issues are becoming barriers that prevent Africans from placing their knowledge in the global stream. Africans have unique cultures, languages, histories, environments, fauna, flora, archaeology and increasingly valuable information in the hard sciences,

and they must figure out a way to barter African intellectual property for access to others. Meanwhile, many African scholars are voicing growing concern about their inability to access Internet resources because of copyright issues. Professor I. S. Diso, vice chancellor of Nigeria's Kano University of Technology, is heading a French-funded investigation into the alleged harmful restrictions copyrights are placing on African scholars, scientists, and researchers (Walker, 2005).

In Africa, concern about copyright as a barrier is reflected in the growing realization of the role it can play in preserving unique African knowledge and protecting its ownership. Out of fear that the works of African scientists and scholars might be stolen in developed countries, university officials across the continent are reluctant to have those works placed on the WWW (Walker, 2005).

Costs

In Africa, the average total cost of using a local dial-up Internet account for 20 hours a month is about $68 per month. This includes usage fees and local telephone time, but not telephone line rental. ISP subscription rates vary between $10 and $100 per month. The prevailing high level of poverty in most of Africa is among the factors hindering people to use the Internet for sustainable development. Most Africans prefer to spend their hard earned money on buying food to fill their stomachs as opposed to surfing the Internet. The cost of computers is another factor that is still plaguing access to the Internet in Africa because governments have not reduced duty on computers (Lusaka Information Dispatch, 2003).

When they gained independence, many African countries embarked upon serious efforts to build telephone lines, but the attempts have been stymied by inefficient state-owned monopolies and high rates of theft of the copper wire used in telephone installations. Consequently, universities interested in gaining access to the Internet almost all had to do so by buying bandwith from satellite companies, often at more than 100 times the cost in developed countries. The bandwith price has started to come down, with the average African university paying $4 less per kilobit per second. And now that African universities have formed a "bandwith club," they are expecting the prices to come down even more. In the new tender they are floating, they are looking at a target price of no more than $2.50 per kilobit per second. This will help some universities to get up to six times more bandwith for what they are currently paying (Walker, 2005).

Access to the Internet in Africa is hampered by unfair costs. The continent is being ripped off to the tune of some $500 million a year for hooking up to the WWW. This extra cost is partly to blame for slowing the spread of the Internet in Africa and helping to sustain the "digital divide." The continent is being forced by Western companies to pay the full cost of connecting to worldwide networks, leading to the exploitation of the continent's young Internet industry. The problem is that International Telecommunications Union regulations, which should ensure the costs of telephone calls between Africa and the West are split 50:50, are not being enforced with regard to the Internet. British Telecom or American Online does not pay a single cent to send E-mail to Africa. The total cost of any E-mail sent to or received by an Internet user in Africa is paid entirely by African ISPs. Consequently, the current and latent demand for bandwith in Africa cost about $1 billion a year (BBC, 2002).

No matter how well it is managed, however, many critics argue that the amount of money universities are spending on bandwith is inappropriate. They say that universities are worshipping at the altar of the Internet when the money could be better spent to remedy grossly overcrowded classes, dilapidated infrastructure, and water and sewer systems that sometimes do not work. Supporters of spending on bandwith counter by stating that with proper bandwith, an instructor might be able to teach 10,000 more students through distance learning, as opposed to hiring 24

more professors. They add that bandwith allows, for example, all the courses at the Massachusetts Institute of Technology (MIT) and other Internet learning resources to be online and adapted to meet the needs of universities in Africa. It should be mentioned, however, that in addition to its monetary cost, Internet use has another major cost: i.e. someone in Europe or in the United States can download 1,000 abstracts in a brief period of time; in Africa, it will take the person two days. That slows down the process of research and discourages users from relying on the system (Walker, 2005).

Politics and Bureaucracy

There exist principled objections and bureaucratic pathologies towards the ICT focus within African educational and political institutions. Some professors just do not believe that distance learning is an effective educational strategy. Many people do not believe that those who push the ICT focus make a lot of sense, and others are scared that the technology might take away their jobs (Walker, 2005).

Another major reason why African universities have lagged so far behind in accessing the Internet has to do with the history of authoritarian and repressive governments on the continent. Many in the last generation of African leaders viewed mass communications as a security risk. Some current leaders hold this same view as well, believing that it could be dangerous to allow many people access to a communications tool like the WWW that is not easily monitored or controlled. They see them like private radio stations which opposition parties have used as tools to overthrow governments. As a consequence, many African governments have been very slow to provide the kind of regulatory and funding assistance taken for granted by universities in developed countries (Walker, 2005). Endless red tape, lack of clear policy, unreliable power supplies and monopoly by banks hold back the IT sector. Adding to this is inter-country rivalry for the best contender of an African Internet hub: it is argued that Nigeria has too much corruption, Senegal is francophone, and Ghana is not ready yet (Hale, 2003).

Also, wars and military incursions sabotage African universities, which are supposed to be the places of open inquiry, as they become completely muzzled. In Sierra Leone, for example, the educational institutions became targets for destruction during its 11-year civil war (Kargbo, 2002). In Nigeria, the universities came to have the same kind of ethnic and religious intolerance and corruption that the military bred into the larger society. The result has been people not trusting one another and being afraid to cooperate on projects or even ask questions. Even with the restoration of democracy in Nigeria, there has been no concerted and coordinated approach to provide universities with the necessary technological infrastructure needed for advancement. While in South Africa, for instance, there are special rates on all kinds of things for educational institutions, in Nigeria, universities are charged the full commercial rate for telephone and electric services, on which the government has monopolies. Nigeria has two national telecommunications carries owned by the government. While both of them have a fiber-optic backbone, none has connected any university or even asked (Walker, 2005).

Another example of how outdated and uncoordinated Nigerian government regulations continue to hamper ICT development has to do with bureaucratic pathologies. For example, when computers and other ICT equipment are sent by overseas donors, they will sit at customs for ten or more months while the tax on them accumulates with time. Once all that is added together, the donations tend to make no sense, for they become just as expensive as buying new ones. Even vendors of Voice Over Internet Protocol (VOIP) services who have been quite successful in gaining numerous customers are worried that eventually the Nigerian government will try to stifle VOIP because it means the end of the state-owned telecommunications service (Walker, 2005).

In Zambia, where the government has no policy on the use of ICT, the opinion leaders see no need to entice and educate people on the need to use the Internet and to make the full use of computers. Some people who have computers just keep them for prestige rather than enhancing them for the benefit of improving their economic aspect or business. Most politicians are concentrating on improving agriculture, fighting corruption, diversifying the economy but forgetting that ICT usage could help to double the effort of achieving economic development that most people require. A challenge Zambians are facing is that most politicians who speak on their behalf are ignorant about Internet usage; if they are aware of it, they just know it by virtue of being elected as members of parliament (Lusaka Information Dispatch, 2003).

Despite the political and bureaucratic malaise, progress is being made. For example, in Uganda, after much lobbying by universities and other members of the ICT community, the government has agreed, in principle, to lay fiber-optic cables whenever it builds new roads. Building a road costs $3 million per kilometer, and adding fiber optics would only be an additional $100,000 per kilometer (Walker, 2005).

Insufficient Training and Personnel

Most people in Africa say they have heard about the Internet, while others say they are ignorant about its usage (Lusaka Information Dispatch, 2003). Training a new generation in managing and accessing the Internet is a major issue in Africa. Managing bandwith involves training and cultural transformations at nearly all universities. Much of the bandwith universities are buying is being wasted, as it is difficult to prevent students from doing selfish things like downloading videos and music. Building firewalls and monitoring use on a per-student basis is also required. Universities must also establish local area networks (LAN) that can be used instead of the more expensive Internet connections for many activities. For instance, a great deal of the research and materials which individuals need can be placed on a LAN and shared through it. Better management of LANS can substantially reduce Internet costs (Walker, 2005).

Unreliability or System Glitches

Many of the ICT systems in African universities are unreliable, either because of power outages, a modem is broken, or someone did not pay the bill. Hardware systems are "pinged" at African universities all the time, and many of them run for only four hours a day. Researchers have been known to wait two days to download a large file and only find out later that the file was corrupted because online access was interrupted. When this happens several times, some researchers simply give up (Walker, 2005). Irregular or non-existent electricity supply, high intensity lighting strikes during the rainy season, and creaking infrastructure are other major barriers to Internet usage, especially outside the big cities and towns (Lusaka Information Dispatch, 2003; Kargbo, 2002; Hale, 2003).

Hopeful Sings on the Horizon

Despite all of the preceding barriers to gaining access to the Internet in Africa, there are hopeful signs on the horizon. To begin with, there is the "leap-frog" potential of evolving technology, as ICT is becoming more reliable and easier to use every day. Africa can be one case where the last can

become the first. The technology is changing so rapidly that having had it for a long time is no longer a significant advantage. One money-saving potential for African users is VOIP, which could greatly reduce many universities' telephone bills (Walker, 2005).

Universities are also exploring alternatives such as placing as much information as possible on virtual libraries. These media have opened university libraries to large populations, making libraries renewed places in the universities' lives (Kargbo, 2002; Walker, 2005). Virtual libraries offer one way of allowing the dissemination of knowledge while still maintaining some measure of control over its distribution. One virtual library already being used by some African universities is a project run out of the University of Iowa in the United States called eGranary, which is described as an "Internet substitute." Project participants use very large hard drives to store nearly two million documents that publishers and authors are willing to share. Once the hard drives are installed locally, users can access the material much faster than trying to use the Internet. These media have everything from a virtual hospital with thousands of pieces of patient literature to full textbooks. There are nearly 50 eGranaries installed in sub-Saharan Africa (Walker, 2005).

JSTOR, as I stated earlier, was originally developed and funded by the Andrew W. Mellon Foundation and now receives support from Carnegie Corporation, is another organization working to provide copyrighted scholarly material to African universities. JSTOR started as a way to electronically store back issues of many scholarly journals so that American university libraries could free up some shelf space. Eventually, it realized that it had a valuable resource for developing countries. It has digitized 17 million images of the pages of 400 of the most prestigious journals in 45 different disciplines. Scholars around the world use JSTOR for research. When it comes to African universities, most of them just do not have enough bandwith to effectively access JSTOR's resources (Walker, 2005).

The debate about how to make copyrighted Internet resources available to African scholars has become prominent in the international controversy over access to information. The Open Access Movement, a growing collection of intellectuals, academics, personnel at nongovernmental organizations (NGOs) and government officials who believe that knowledge should be free, is advancing Africa's case. The movement is calling for owners of copyrights to be more generous in making information available to African institutions. Universities in developed and developing countries are battling academic and scientific journal publishers for their regular, substantial subscription price increases, and their use of bundling, which forces libraries to subscribe to journals they do not want in order to get the ones they do. Committees in both the British and American legislatures have passed resolutions demanding that government-sponsored research be available for free. Responding to the growing pressure, Reed Elsevier, the world's largest publisher of scientific journals, for example, announced that authors publishing in its journals would be allowed to post articles in institutional repositories. Another development favoring the Open Access Movement is the announcement of Google, the world's most popular Internet search engine, that it has reached an agreement with some of America's leading university research libraries to begin converting their holdings into digital files that would be freely searchable over the Internet. In addition, Google's competitor, Yahoo, as well as others such as Amazon.com, is in a mad rush to get their shares of the $12 billion scholarly journal business. Google has developed a separate site called Google Scholar for academic researchers. These rapid developments may ultimately work to the advantage of developing nations like those in Africa (Walker, 2005).

An initiative called the Halfway Proposition urges fellow African countries to develop national exchanges and then interconnect regional ones, as has been done in other parts of the developing world. This would at least mean that revenues generated from intra-African E-mail will stay in the continent as opposed to going to Western nations. While no one really know just how much intra-

African traffic exists, it will certainly grow and become significant. And if even only five percent of the traffic is intra-regional, it would add up to a sizeable amount (BBC, 2002).

The most hopeful sign on the ICT horizon in Africa is the exuberant optimism. Every study or report that has been done on the continent shows that everyone who works on ICT issues is optimistic. African universities feel that they have been shut off for so long from the global knowledge community that they are now hungry and thirsty, and are just full speed ahead (Walker, 2005). Despite the many woes of ICT on the continent, there is the steadfast belief in the potential of Africa to become a Silicon Valleyesque hi-tech hub. It wants to take a slice of the outsourcing that has been won by India's Bangalore and win the foreign investors that have shunned Africa for so long. And there is certainty that technology must be the way to improved economic prosperity. When computer fairs are held in Africa, long queues form at the stands. Would-be students wait patiently to fill out registration forms for computer courses in anything from basic word processing to diplomas in programming. Some potential students admit afterwards that they have no idea how or whether they will be able to pay the fees to attend such course, but they, like Africa, are determined to find a way of using technology to enter the arena of the global economy (Hale, 2003).

Conclusions and Recommendations

Data mining techniques and visualization must play a pivotal role in retrieving substantive electronic data to study and teach about African phenomena in order to discover unexpected correlations and causal relationships, and understand structures and patterns in massive data. Data mining is a process for extracting implicit, nontrivial, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data in massive databases. The goals of data mining are (a) explanatory—to analyze some observed events, (b) confirmatory—to confirm a hypothesis, and (c) exploratory—to analyze data for new or unexpected relationships. Typical tasks for which data mining techniques are often used include clustering, classification, generalization, and prediction. The most popular methods include decision trees, value prediction, and association rules often used for classification. Artificial Neural Networks are particularly useful for exploratory analysis as non-linear clustering and classification techniques. The algorithms used in data mining are often integrated into Knowledge Discovery in Databases (KDD)—a larger framework that aims at finding new knowledge from large databases. While data mining deals with transforming data into information or facts, KDD is a higher-level process using information derived from a data mining process to turn it into knowledge or integrate it into prior knowledge. In general, KDD stands for discovering and visualizing the regularities, structures and rules from data, discovering useful knowledge from data, and for finding new knowledge.

Visualization is a key process in Visual Data Mining (VDM). Visualization techniques can provide a clearer and more detailed view on different aspects of the data as well as results of automated mining algorithms. The exploration of relationships between several information objects, which represent a selection of the information content, is an important task in VDM. Such relations can either be given explicitly, when being specified in the data, or they can be given implicitly, when the relationships are the result of an automated mining process: for example, when relationships are based on the similarity of information objects derived by hierarchical clustering.

Understanding and trust are two major aspects of data visualization. Understanding is undoubtedly the most fundamental motivation behind visualizing massive data. If scientists understand what has been discovered from data, then they can trust the data. To help scientists understand and trust the implicit data discovered and useful knowledge from massive datasets concerning Africa, it is imperative to present the data in various forms, such as boxplots, scatter

plots, 3-D cubes, data distribution charts, as well as decision trees, association rules, clusters, outliers, generalized rules, etc.

The software called Crystal Vision is a good tool for visualizing data. It is an easy to use, self-contained Windows application designed as a platform for multivariate data visualization and exploration. It is intended to be robust and intuitive. Its features include scatter plot matrix views, parallel coordinate views, rotating 3-D scatter plot views, density plots, multidimensional grand tours implemented in all views, stereoscopic capability, saturation brushing, and data editing tools. It has been used successfully with datasets as high as 20 dimensions and with as many as 500,000 observations (Wegman 2003). Crystal Vision is available at the following Internet site: <ftp://www.galaxy.gmu.edu/pub/software/CrystalVisionDemo.exe>

In light of all these possibilities, it is imperative that there be on-the-ground commitment on the part of implementers, as well as university and government authorities, in order to achieve sustainable ICT in Africa. Only through their participation will the Internet transform the classroom, change the nature of learning and teaching, and change information seeking, organizing and using behavior.

Finally, as I have suggested elsewhere (Bangura, 2005), the provision of education in Africa must employ *ubuntugogy* (which I define as the art and science of teaching and learning undergirded by humanity toward others) to serve as both a given and a task or *desideratum* for educating students. *Ubuntugogy* is undoubtedly part and parcel of the cultural heritage of Africans. Nonetheless, it clearly needs to be revitalized in the hearts and minds of some Africans. Although compassion, warmth, understanding, caring, sharing, humanness, etc. are underscored by all the major world orientations, *ubuntu* serves as a *distinctly African rationale* for these ways of relating to others. The concept of *ubuntu* gives a distinctly African meaning to, and a reason of motivation for, a positive attitude towards the other. In light of the calls for an African Renaissance, *ubuntugogy* urges Africans to be true to their promotion of peaceful relations and conflict resolution, educational and other developmental aspirations. We ought never to falsify the cultural reality (life, art, literature) which is the goal of the student's study. Thus, we would have to oppose all sorts of simplified or supposedly simplified approaches and stress instead the methods which will achieve the best possible access to real life, language and philosophy

References

Alshameri, F. J. 2006. Automated generation of metadata for mining image and text data. Doctoral dissertation, George Mason University, Fairfax, Virginia.

Bangura, A. K. 2005. Ubuntugogy: An African educational paradigm that transcends pedagogy, andragogy, ergonagy and heutagogy. *Journal of Third World Studies* xxii, 2:13-54.

Bangura, A. K. 2000. *Chaos Theory and African Fractals*. Washington, DC: The African Institution Publications.

Bangura, A. K. 2000. Book Review of Ron Eglash's *African Fractals: Modern Computing and Indigenous Design*. *Nexus Network Journal* 2, 4.

Barker, J. and J. Thornton. 2004. Software engineering challenges in bioinformatics. *Proceedings of the 26th International Conference on Software Engineering* (ICSE '04).

BBC. April 15, 2002. The great African internet robbery. *BBC News Online*. Retrieved on April 18, 2002 from http://news.bbc.co.uk/2/hi/africa/1931120.stm

Berry, M., Z. Drmac and E. Jessup. 1999. Matrices, vector spaces, and information retrieval. *Society for Industrial Applied Mathematics* (SIAM) 41, 2:335-362.

Brunner, R., S. Djorgovsky, T. Prince and A. Szalay. 2002. Massive datasets in astronomy. In J. Abello et al., eds. *Handbook of Massive Datasets*. Norwell, MA: Kluwer Academic Publishers.

Cannataro, M., C. Comito, A. Guzzo and P. Veltri. 2004. Integrating ontology and workflow in PROTEUA, a grid-based problem solving environment for bioinformatics. *Proceedings of the International Conference on Information Technology: Coding and Computing* (ITCC '04).

Chen, M., J. Han and P. Yu. 1996. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8, 6:866-883.

Dhillon, I., J. Han and Y. Guan. 2001. Efficient clustering of very large document collection. In R. Grossman et al., eds. *Data Mining for Scientific and Engineering Applications*. Norwell, MA: Kluwer Academic Publishers.

Dorre, J., P. Gerstl and R. Seiffert. 1999. Text mining: Finding nuggets in mountains of textual data. KDD-99:398-401. San Diego, CA.

Eglash, Ron. 1999. *African Fractals: Modern Computing and Indigenous Design*. New Brunswick, NJ: Rutgers University Press.

Ester, M., H. Kriegel and J. Sander. 2001. Algorithms and applications for spatial data mining. *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*. Taylor and Francis, 167-187.

Ester, M., H. Kriegel and J. Sander. 1997. Spatial data mining: A database approach. *Proceedings of the International Symposium on Large Spatial Databases*. SSD '97:47-66. Berlin, Germany.

Ester, M., A. Frommelt, H. Kriegel and J. Sander. 2000. Spatial data mining: Database primitives, algorithms and efficient DBMS support. *Data Mining and Knowledge Discovery* 4, 2/3:193-216.

Fayyad U. M., G. Piatetsky-Shapiro and P. Smyth. 1996. From data mining to knowledge discovery: An overview. U. M. Fayyad et al., eds. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press.

Gomez, M., A. Gelbuhk, A. Lopez and R. Yates. 2001. Text mining with conceptual graphs. *IEEE* 893-903.

Grossman, R., E. Creel, M. Mazzucco and R. Williams. 2001. A dataspace infrastructure for astronomical data. In R. Grossman et al., eds. *Data Mining for Scientific and Engineering Applications*. Norwell, MA: Kluwer Academic Publishers.

Hale, Briony. May 16, 2003. Africa's tech pioneers play catch up. *BBC News Online*. Retrieved on May 17, 2003 from http://news.bbc.co.uk/2/hi/business/3033185.stm

Hambrusch, S., C. Hoffman, M. Bock, S. King and D. Miller. 2003. Massive data: Management, analysis, visualization, and security. A School of Science Focus Area at Purdue University Report.

Han, J. and M. Kamber. 2001. *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufman Publishers.

Humphreys, R., J. Cabanela and J. Kriessler. 2001. Mining astronomical databases. In R. Grossman et al., eds. *Data Mining for Scientific and Engineering Applications*. Norwell, MA: Kluwer Academic Publishers.

Kafatos, M., R. Yang, X. Wang, Z. Li and D. Ziskin. 1998. Information technology implementation for a distributed data system serving earth scientists: Seasonal to International ESIP. *Proceedings of the 10th International Conference on Scientific and Statistical Database Management* 210-215.

Kamath, C. 2001. On mining scientific datasets. In R. Grossman et al., eds. *Data Mining for Scientific and Engineering Applications*. Norwell, MA: Kluwer Academic Publishers.

Kargbo, John Abdul. March 2002. The internet in schools and colleges in Sierra Leone. *First Monday* 7, 3. Retrieved on March 18, 2002 from http://firstmonday.org/issues/issue7_3/kargbo/index.html

King, M. and R, Greenstone. 1999. *EOS Reference Handbook*. Washington, DC: NASA Publications.

Koperski, K. and J. Han. 1995. Discovery of spatial association rules in geographic information databases. *Proceedings of the 4th International Symposium on Advances in Spatial Databases*. 47-66. Portland, ME.

Koperski, K. and J. Han and N. Stefanovic. 1998. An efficient two-step method for classification of spatial data. *Proceedings of the Symposium on Spatial Data Handling*. 45-54. Vancouver, Canada.

Lusaka, Information Dispatch. January 07, 2003. Internet access still a nightmare in Africa. Retrieved on February 13, 2003 from
http://www.dispatch.co.zm/modules.php?name=News&file=article&sid=180

Marusic, I., G. Candler, V. Interrante, P. Subbareddy and A. Moss. 2001. Real time feature extraction for the analysis of turbulent flows. In R. Grossman et al., eds. Data Mining for Scientific and Engineering Applications. Norwell, MA: Kluwer Academic Publishers.

Ng, R. and J. Han. 1994. Efficient and effective clustering methods for spatial data mining. *Proceedings of the 20th International Conference on Very Large Databases*. 144-155, Santiago, Chile.

Palacio, M., D. Sol and J. Gonzalez, 2003. Graph-based knowledge representation for GIS data. *Proceedings of the Fourth Mexican International Conference on Computer Science* (ENC '03).

Pantel, P. and D. Lin. 2002. Discovering word senses from text. In *Proceedings of SIGKDD-01*. San Frncisco, CA.

Sander, J. M. Ester and H. Kriegel. 1998. Density-based clustering in spatial databases: A new algorithm and its applications. *Data Mining and Knowledge Discovery* 2, 2:169-194.

Shekhar, S., C. Lu, P. Zhang, and R. Liu. 2002. Data mining and selective visualization of large spatial datasets. *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI '02).

Walker, Kenneth. Spring 2005. Bandwith and copyright: Barriers to knowledge in Africa? *Carnegie Reporter* 3, 2:1-5.

Wegman, E. 2003. Visual data mining. *Statistics in Medicine* 22:1383-1397 plus 10 color plates.

Wegman, E. 1997. A Guide to Statistical Software. Available at http://www.galaxy.gmu.edu/papers/astr1.html

Wong, P., P. Whitney and J. Thomas. 1999. Visualizing association rules for text mining. In G. Wills and D, Keim, eds. *Proceedings of IEEE Information Visualization '99*. Los Alamitos, CA: IEEE CS Press.

Wong, R. and W. Shui. 2001. Utilizing multiple bioinformatics information sources: An XML database approach. *Proceedings of the Second IEEE International Symposium on Bioinformatics and Bioengineering.*

Yang, R., X. Deng, M. Kafatos, C. Wang and X. Wang. 2001. An XML-based distributed metadata server (DIMES) supporting earth science metadata. *Proceedings of the 13th International Conference on Scientific and Statistical Database Management* 251-256.

Acknowledgment

This essay benefitted greatly from the insights of my colleague, Professor Faleh J. Alshameri, albeit all shortcomings herein lie with me.

About the Author

Abdul Karim Bangura is Professor of Research Methodology and Political Science at Howard University in Washington, DC, USA. He holds a PhD in Political Science, a PhD in Development Economics, a PhD in Linguistics, and a PhD in Computer Science. He is the author and editor/contributor of 57 books and more than 450 scholarly articles. He has served as President, United Nations Ambassador, and member of many scholarly organizations. He is the winner of numerous teaching and other scholarly and community awards. He also is fluent in about a dozen African and six European languages and studying to strengthen his proficiency in Arabic and Hebrew.