A Metadata Based Parametric Comparison of Information Retrieval Models for Satellite Images

Uzma Ashiq[1], Amjad Farooq[2] and Muhammad Abuzar Fahiem[1]

[1]Department of Computer Science, Lahore College for Women University, Lahore, Pakistan
[2]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

Abstract– **The domain of information retrieval is becoming more and more complicated with the ever increased volume of data. The domain gets even more complicated when spatial databases are to be processed. Spatial databases contain images which require heavy processing to retrieve desired information. A solution to reduce this heavy processing and to get lower computations is in tagging the images on the basis of metadata. In this paper, we have focused on satellite images and metadata tags associated with these images, for efficient retrieval of information. We have compared various information retrieval models suitable for satellite images. The models compared are; Linear Model, Finite State Model, Bayesian Network Based Model, and Rigorous Model. These models are compared on Region, Weather, Agriculture, Building, Road, Water and Land metadata tags.**

Keywords– **Information Retrieval, Satellite Images, Metadata, Indexing and Spatial Data**

I. INTRODUCTION

Metadata is data about data which refers to machine understandable information of structured data about data. Usually metadata is used to keep catalogued records in an indexed library environment which contains descriptive information that defines resources to identify and retrieve the relevant information about data from selected index [1]. Metadata based information retrieval (IR) models of images contain information about contents of images [2]. There are different types of metadata for images, such as technical metadata, content metadata and embedded metadata. Technical metadata describes the technical aspects of images such as its height and width in pixels, and the compression scheme used to store the image. Content metadata saves the contents of images such as name of the photographer, and the date and time when it was created. Embedded metadata provides the facility of encapsulating both type of data as one unit for indexed storage and retrieval in XML format [3]. Interoperability of metadata plays a vital role in exchanging data with minimum loss of content and functionality.

There are three types of metadata: descriptive metadata, structural metadata and administrative metadata. The purpose of descriptive metadata is to identify the resources such as title, abstract, element, author etc. Structural metadata combines the objects and keep them together e.g., chapters (all pages are put together in a chapter) and administrative metadata helps to manage resources and keep the record. It creates file type and other metadata related technical information. Metadata resources define the purpose of metadata of a particular object. In this paper, we have derived metadata tags of satellite images (SI). These metadata tags containing properties of SI help to search similar and dissimilar resources. The tags define location, storage and retrieval of information. Interoperability of metadata plays a vital role in exchanging data with minimum loss of content and functionality, which is collected using different software and hardware platforms, data strictures and interfaces. Preservation and archiving of data is the main problem. First problem is how to maintain the metadata resources which can access digital information, because it may be corrupted when format of storage technology with respect to software and hardware is changed. So, preservation needs metadata schema and set of elements to keep the flow. Metadata behaves differently in SI. SI are the photographs of earth, other planets and clouds etc. that are used in different fields such as agriculture, geology, forestry, monitoring atmospheric events, regional planning, education, intelligence and warfare, etc. SI have valuable data in the form of images that contains a lot of information about different aspects. Combination of different types of SI is stored in spatial database which requires metadata for retrieving the images according to the query of user. There are different types of SI used in operational and research meteorology such as Visible Imagery, InfraRed, Water Vapour, etc. These types work on different ranges of wavelength. The first generation visual information retrieval system is bases on metadata model and it searches images through metadata from linked database. The mathematical model of meaning is used to compute the specific meaning of key words for retrieving images unambiguously and dynamically. This retrieval is on the basis of identification of similarity between the metadata items of the images and keywords. The grid is an emerging computing model which provides dynamic access to large amount of computing and storage resources and defines design and validation of a metadata catalogue to access satellite image metadata on the Grid. Information retrieval systems manage the information on the internet like textual documents, images, videos, online services etc. Internet systems depend on information retrieval system by filtering the user requirements and showing the relevant results after matching indexed documents. Accurate information retrieval systems search for exact and most relevant documents that fulfill the user requirement. There are three types of processes of information 1) representation of content of documents, 2) representation of user information need and 3) comparison of two representation/ query. In index processing user does not interact with system. Index processing is based on algorithms which searches for required information from index and stores the documents in split form like title, abstracts and images. So metadata helps in retrieval of image's information as required by the user. User needs some information that refers to query database to search image match. The query with index documents results in ranked list of images in reduced searching time.

SI are stored in spatial databases [4] and require specialized metadata tags so that the images can be used meaningfully in different fields such as agriculture, geology, forestry, metrology, region planning, education, intelligence and warfare [5]. The processing of SI requires high performance data management and accurate IR mechanisms preferably based on metadata schemes [6], [7].

In this paper, we review and compare different IR models which can be used for SI. The next section of this paper reviews various aspects of IR models associated with SI and gives a comparison of these IR models. Section 3 summarizes our discussion while, in last section we give some future directions.

II. COMPARISON OF DIFFERENT IR MODELS FOR SI

A major problem with SI is that a huge amount of computations are required to process SI and it poses severe drawbacks in catastrophic situations. A proposed remedy to this problem is to adopt a grid computing technique [8]. Different metadata schemes like Dublin Core, Text Encoding Initiative, Metadata Encoding and Transmission Standard, and Metadata Object Description Schema are discussed in [9]. The authors also discussed the mapping of different metadata standards. Image processing and character extraction methods are used to acquire useful information from abundant information present in a SI [10]. The researchers used content based information retrieval model in an unconstrained environment. A new information retrieval model is introduced in [11]. The model is based on three types of models; linear, finite state and knowledge based. A semantic scheme is introduced in [12] to improve quality and reduce quantity of information in indexed SI. An IR scheme based on image content and metadata is proposed in [13] for cross collection search in distributed architectures.

The authors discuss the problems of UNOSAT when large quantities of computation are required to process SI during catastrophes [14]. Major problem is of storage space required to keep record of SI automatically. The UNOSAT project is a United Nations program that provides, among other services, satellite imagery to the international humanitarian community in case of natural disasters. It adopts a grid computing technique as a solution, which provides dynamic access to large amount of computing and storage resources. It defines the requirement, design and validation of metadata catalogue that is helpful to access SI metadata on grid. Metadata catalogue is designed for two types of prototype services; web pages and mobile devices. Metadata schema adopts geo network standard that is defined on XML schema. It is based on database schema to keep record of XML files. The implementation of catalogue is created by American Mountain Guides Association (AMGA) which gives help to query the catalogue from prototypes and web portal of UNOSAT. AMGA provides different API (Python, Java, C++ etc) and is an implementation of certificate based authentication for web service. AJAX service verifies SI metadata catalogue in web portal and J2ME uses in mobile phone programming for metadata catalogue. UNOSAT web portal provides user to assess high quality SI from remote areas in the world.

The authors describe a revision and expansion of metadata, structure of metadata, what does metadata do, schemes and element sets, how to create metadata and interoperability and exchange of metadata [15]. Interoperability helps the human and machines understand the metadata. It has ability to exchange data with minimal loss of content and functionality from different software platforms, data structures, hardware and interfaces. They discuss different metadata scheme like Dublin Core, The Text Encoding Initiative (TEI), Metadata Encoding and Transmission Standard (METS), Metadata Object Description Schema (MODS) etc. XML and SGML are tools which provide help to create metadata. Metadata quality control helps to control content of vocabularies, location, material of collection including clear statement of condition and terms of use of digital object, and support long terms management of objects in collections. Metadata crosswalk is the mapping of different metadata standards, which are based on the similarity of two schemes. It connects elements' semantics and syntax from one metadata scheme to another metadata scheme. The metadata registry helps to integrate resources for legacy data and also documents multiple schemes. Registries play a vital role in management of metadata. Therefore, metadata has very important role for improving retrieval of information from database.

The research describes how to process satellite cloud images for acquiring the useful information from the abundant information of cloud images [16]. Different image processing and character extractions methods are used for information retrieval of satellite cloud images. Traditional Information retrieval methods have some limitations for image retrieval, accurately and quickly. Whereas, content based information retrieval (CBIR) can retrieve image from unconstrained environment. They used pretreatment method for processing satellite cloud images to obtain the image character data. Pretreatment method matches the image and retrieval of images from database. Therefore, threshold processing and image filtering methods are used for better results before image processing and analysis. In image database, cloud image character plays a vital role based on colour character, texture character, shape character and spatial relation character. Traditional image management becomes slow in processing when there is an increase in the number of images for retrieval. Whereas, content based methods use the visual character for image retrieval. It is adaptive method but the main problem is how to use the information from database adequately.

Sheng et. al., describe new structure of information retrieval model where retrieval is based on three types of models: linear, finite state and knowledge based which covers the environmental epidemiology, oil / gas production, agriculture and forest [17]. In proposed model they define main issue of information retrieval from large archives is scalability and how to retain the scalability of image retrieval. The proposed model uses linear model which refers to locate top-K set of represent data whereas, linear regression derived the coefficient for model in term of maximized and minimized. Finite state model defines states of top-K data pattern with the help of finite state model machine which obtains flow of logical states of information retrieval of image and data. Bayesian Network and knowledge model is graphical model which refers to locate set of variables for top-K data patterns for defining the probabilistic and fuzzy rules within model. They define metrics for proposed model to check data representation and predict risky area and high interest area. Accuracy of model that predicts high and low risk locations depends upon number of occurrences of events. Efficiency of information retrieval of model is measured by complexity as well as size of data which derive complexity ratios on progressive executions of model and data representation.

The authors describe functionality of metadata in terms of technical and semantic level. They also describe how to index web service interchange of metadata and how to improve the quality and decrease the quantity of information which is required for searching on internet [18]. There are three major functionalities of metadata 1) description of resources, 2) production of metadata, and 3) use of metadata. Description of resources defines the set of element information which describes for particular object and identifies useful information. Production of metadata means how to maintain data which is produced by different levels of information retrieved through metadata.

The research describes functionality of information retrieval technology which is combination of experimentation and theory [19]. All search engines are based on it because experiments discover practical experience of retrieval of different things which are based on theory. Without theory experiment becomes trial and error so main challenge is how information retrieval modal is applicable to any theory. Information retrieval consists of three things: index documents, user information and matching. Index documents store the documents and images partially.

Paul H, Lewis et. al., describe a new approach of image retrieval that is combination of content and metadata based which provides cross collection search within distributed architecture [20].

Commonly used metadata tags with their possible values are given in Table 1.

Table 1: Metadata Tags with their Possible Values

Metadata tags	Possible Values
Region	Countries, Cities, Villages, Forests, Mountain ranges, Desert, Coastal area, Glaciers, Sea, Oceans, Fields, Farm land, Industrial area
Weather	Cloudy, Dust storm, Fog, Freezing rain/ice, Hailing, Hurricanes, Lighting, Raining, Sleet rain drop, Snow , Sun light, Thunderstorms, Tornado, Wind
Agriculture	Crop assessment, Crop health, Soil moisture, Climate change, Irrigation system
Natural Environment	Climate changes, Land degradation, Natural disaster, Eco system, Water resources, Pollution, Weather damages
Defense, Security and Intelligence	Space, Military, Special weapons monitoring, Air Forces, Army, Navel, Respond immediately to crisis, Target detection, Location and damage assessment
Risk and Disaster	Volcanoes, Tsunamis, Earthquake, Floods, Forest Fire, Storms
Building	Colonies, Hotel, AirPort, Dry port, Railway Stations, Hospitals, Shopping plaza, Historical building, Governmental building, T.V station, Sensitive and defensive building, Parks, Industrial building
Road	Motor way, Bridges, Street, Underpass, Railway lines, Runways, Road in cities and villages
Water Resources	Oceans, Canal, Lakes, Rivers, Sea, Tube well, Wells,
Land Resources	Forests, Animals, Coals, Fossils fuel, Petroleum, Mining
Energy	Solar energy, Lightning
Air	Water vapours, Gases, Ultraviolet solar radiations
Time	Day, Night, Evening, Morning, Dawn, Afternoon
Life Style	Wildlife, Urban, Rural
Regional Planning	Road maps, City map, Country map, Oceans map, Mountain ranges maps, Sea shores and Landscape mapping

Various IR models such as Linear Model (LM), Finite State Model (FSM), Bayesian Network Based Model (BNM), and

Rigorous Model (RM) are used for SI. These models are compared in Table 2 on the basis of metadata tags given in Table 1.

Table 2: Comparison of Various IR Models for SI

Metadata Tags	LM	FSM	BNM	RM
Region	Yes	Yes	Yes	Yes
Weather	No	Yes	No	No
Agriculture	Yes	Yes	Yes	No
Building	No	No	No	Yes
Road	No	No	No	Yes
Water	No	No	No	Yes
Land	Yes	Yes	Yes	Yes

LM derives linear regression and its variables, which depend on changing of time. Metadata tags for LM are region and agriculture. LM uses historical / background data to predict probability of changing in circumstances where regression techniques derive coefficient of model which show condition of probability.

FSM define the states of system or environmental condition with the help of finite state machines which has set of elements, structural language and logical synthesis that evaluate or match condition such as observations about region, weather and agriculture.

BNM is a graphical model that defines a set of variables which relate to probability of risk with knowledge maps, diagrams and statistical techniques for managing incomplete data. Metadata tags used by this model for IR are region and agriculture.

RM investigates on how to get accurate distortions on high resolution SI and extract cartographic feathers from these. The model consists of two layers; mathematical model and rational function. The mathematical model gets information from three dimensional SI whereas rational function derives the co-linearity equation independent of terrain. This model uses region, building, road, water and land metadata tags.

III. CONCLUSION

In this paper, we have discussed the inter-relation between metadata for SI and IR models. We reviewed various strategies proposed for IR from SI. A comparison of different IR models, Linear Model, Finite State Model, Bayesian Network Based Model, and Rigorous Model is presented in this paper. This comparison is based on different metadata tags like Region, Weather, Agriculture, Building, Road, Water and Land.

IV. FUTURE DIRECTIONS

In future, on the basis of discussion and comparison presented in this paper, a new IR model more suitable for SI may be proposed.

REFERENCES

[1]. R. Datta, D. Joshi, J. Li, and J. Z. Wang; Image Retrieval: Ideas, Influences, and Trends of the New Age, ACM Computing Surveys, 2008, vol. 40(2).

[2]. C. R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, K. Palaniappan; GeoIRIS: Geospatial Information Retrieval and Indexing System - Content Mining, Semantics Modeling, and Complex Queries, IEE Transactions on Geoscience and Remote Sensing, 2007, vol. 45(4), 839-852.

[3]. P. Enser; The Evolution of Visual Information Retrieval, Journal of Information Science, 2008, vol. 34(4), 531-546

[4]. R. D. S. Torres, C. B. Medeiros, M. A. Gonçcalves, E. A. Fox; A Digital Library Framework for Biodiversity Information Systems, International Journal on Digital Libraries, 2006, vol. 6(1), 3-17.

[5]. R. D. Holowczak, F. J. Artigas, S. A. Chun, J. S. Cho, H. S. Stone; An Experimental Study on Content-Based Image Classification for Satellite Image Databases, IEEE Transactions on Geoscience and Remote Sensing, 2002, vol. 40(6), 1338-1347.

[6]. P. G. B. Enser; Visual Image Retrieval, Annual Review of Information Science and Technology, 2008, 42(1), 1-42.

[7]. L. Gueguen, M. Datcu; A Similarity Metric for Retrieval of Compressed Objects: Application for Mining Satellite Image Time Series, IEEE Transaction on Knowledge and Data Engineering, 2008, vol. 20(4), 562-575.

[8]. W. Daniel, L. Sandoval; Access to Satellite Image Metadata on the Grid, Universite de Geneve, 2006.

[9]. A. Brand, F. Daly, B. Meyers; Metadata Demystified: A Guide for Publishers, NISO Press and The Sheridan Press, 2003.

[10]. Y. Hao, W. S. Guan, Y. Zhu, Y. H. Tang; Contented-Based Satellite Cloud Image Processing and Information Retrieval, Lecture Notes in Computer Science, 2007, vol. 4688, 767-776.

[11]. C. S. Li, Y. C. Chang, L. D. Bergman, J. R. Smith; Model-Based Multi-Modal Information Retrieval from Large Archives, ICDCS Workshop of Knowledge Discovery and Data Mining in the World-Wide Web, 2000.

[12]. R. Lannella, A. Waugh; Metadata: Enabling the Internet, DSTC Pty Ltd., 1997.

[13]. P. H. Lewis, K. Martinez, F. S. Abas, M. Faizal, A. Fauzi, S. C. Y. Chan, M. J. Addis, M. J. Boniface, P. Grimwood, A. Stevenson, C. Lahanier, J. Stevenson; An Integrated Content and Metadata Based Retrieval System for Art, IEEE Transactions on Image Processing, 2004, vol. 13, 302-313.

[14]. W. D. L. Sandoval, M. Lamanna, B. Koblitz. Access to satellite image metadata on the Grid. Université de Genève. 2006.

[15]. R. Guenther, J. Radebaugh. Understanding metadata. National Information Standard Organization (NISO) Press, Bethesda, USA. 2004.

[16]. Y. Hao, W. ShangGuan, Y. Zhu, Y. Tang. Contented-based satellite cloud image processing and information retrieval. In Bio-Inspired Computational Intelligence and Applications, Springer Berlin Heidelberg. 2007, 767-776.

[17]. C. S. Li, Y. C. Chang, L. D. Bergman, J. R. Smith. Model-based multi-modal information retrieval from large archives. IBM Thomas J. Watson Research Division. 2000.

[18]. R. Iannella, A. Waugh. Metadata: enabling the Internet. DSTC Pty Limited. 1997.

[19]. D. Hiemstra. Information retrieval models. Information Retrieval: searching in the 21st Century, 2009, 2-19.

[20]. P. H. Lewis, K. Martinez, F. S. Abas, M. F. A. Fauzi, S. C. Chan, M. J. Addis, J. Mike, ... J. Stevenson. An integrated content and metadata based retrieval system for art. Image Processing, IEEE Transactions on, 2004, 13(3), 302-313.