

AN ALGORITHM OF ONTOLOGICAL MODEL FOR INDEXING AND ACCESSING RESEARCH KNOWLEDGE

^{*1}Mabayoje M. A. and ²Olabiyisi S. O.

¹ Department of Computer Science, Faculty of Communication and Information
Sciences,
University of Ilorin, PMB1515, Ilorin, Kwara-Nigeria. **Tel.:** +2348063185885,
+2348054290856, Services100ng@yahoo.com.

²Department of Computer Science and Engineering, Ladoke Akintola University of
Technology, Ogbomosho, Oyo-Nigeria. **Tel:** +2348036669863
Tundeolabiyisi@hotmail.com

Abstract

An enormous amount of data is available in various departments in tertiary schools. Much of these relevant research materials in this massive and swiftly growing dataset will be of no relevance if there are no efficient tools to integrate and access them.

This paper presents a framework for integrating research works that cover similar fields of knowledge submitted by researchers to various departments in tertiary schools. The general aim of this effort is to present a model that can automatically classify the research reports according to keywords of topics and abstract. It will also manage metadata regarding individual researcher and supervisor of the research project. Ontological approach and dynamically acquired background domain knowledge will be used in this regard. This effort is aimed at providing a mean of assisting subsequent researchers to easily locate and reuse relevant documents through a structured front end.

The algorithm is implemented in Visual Basic.Net and tested on windows.

Keywords: Algorithm, Data, Domain, knowledge, Metadata, Model, Ontology, Reuse, System, Visual Basic.

INTRODUCTION

Information or data gathering, storage and retrieval have always been necessary for day-to-day running of establishments and organizations. In the academics, most of such data fall within academic research efforts. A practical approach to achieving good information storage and retrieval in term of academic research projects is through computer-based information systems [1].

Problem of information management in the academic arena has been related mostly to lack of practical systems that would take care of storage and extraction of academic information in respect of previous research efforts.

Going by modern related scientific trend of such information management, the search engine has been found to be one big solution to this problem. When incorporated into the information extraction system being proposed to solve this problem, academic information and data so embedded could be accessed and extracted through the net. This is because search engine is one of the most widely used methods for navigating in cyberspace today. Considering the amount of information that is available from a good search engine, it is similar to having the Yellow Pages or a guide book, all of which could serve other purposes like administrative purposes in addition to the academic ones.

This effort presents a framework for easy integration and accessing of academic research

project within a given domain for further research, for new ideas to emerge or to implement existing work. It aims at providing the tools that can assist in ontology building and to utilize the background ontology for document indexing.

It presents a model for automatic augmentation of various (abstract/summary) segments of previous undergraduate and post graduate research works with Metadata using Dynamically Acquired Background Knowledge, (DABK). This will assist researchers to easily locate information related to all research works in the departments through front-end structure format. The system will also provide access route for researchers outside the departments to locate relevant information on both graduate and undergraduate research works via the internet.

This research is aimed at providing source of material for postgraduate students as well, and it can be used as ready made software for accreditation purpose being the front-end structure format, while update and upgrade of data can also be done via the back-end format. It also provides an intelligent interface to allow for the retrieval of the stored information.

Related Concepts

Ontology

The field of computer and information science describes ontology as indicating an historical object that is designed for a purpose, which is to enable the modeling of knowledge

about some domain, real or imagined [2]. It is an arrangement of a conceptualization, that represents the objects, concepts, and other entities which are interpreted to exist in some area of interest and the relationships that hold among them, [3].

Ontological method is also used in the filtering, ranking and presentation of the results covering quality issues such as contradictions and related information, i.e. a different possible answer to the same query or an answer to a different but related query.

Ontology in computer science is specifically related to knowledge sharing and reuse, a specification of a conceptualization which has some properties for knowledge sharing among Artificial Intelligence (AI) software [4].

Information Retrieval

Information Retrieval is the scientific method of searching information either in documents or searching for documents themselves. It also includes searching within databases which could either be relational stand-alone databases or hyper textually-networked databases like the World Wide Web [5]. It is the science of locating, from a large document collection, those documents that fulfill a specified information need [6].

It also covers other areas of science like mathematics, library science, information science,

information architecture, cognitive psychology, linguistics, statistics and physics. However each of these areas of relevance has its application of skills, theory, literature, and technologies.

In each of these areas, information overload are reduced through the use of automated Information Retrieval System (IRS). IRS is used in places like universities and some other tertiary institutions. Access to books, journals, and other documents are easily achieved in libraries by the use of IRS. Programmers have succeeded in their effort at creating applications that have very well functioned in information retrieval. Examples of such applications are found within Web search engines like Google, Yahoo search, Live search (Formerly MSN search) etc. [7].

Areas where Information Retrieval techniques are employed are Adversarial information retrieval, Automatic summarization, Multi-document summarization, Cross-lingual retrieval, Document classification, Spam filtering, Open source information retrieval, Question answering, Structured document retrieval, Topic detection and tracking etc.

Measuring Retrieval Effectiveness

It is interesting to note that the technique of measuring retrieval effectiveness has been largely influenced by the particular retrieval strategy adopted and the form of its output. The two metrics that can be used to measure how well

an information-retrieval system is able to answer queries are: Precision; which measures what percentage of the retrieved documents are actually relevant to the query. The second is Recall. It measures what percentage of the relevant documents to the query was retrieved. Both Precision and Recall are measured in 100 percent [8].

In addition to the ones earlier proposed, i.e. Precision and Recall, two major measures of assessments were further proposed: Fall-out and F-measure. All of these measures assume a ground truth notion of relevancy, i.e. every document is known to be either relevant or non-relevant to a particular query.

Database and Database Management System

A Database is a large storage of data held in a computer and made easily accessible for needed purpose, [9]. A single organized collection of a structured data which are stored with a minimum duplication of data items in order to provide a consistent and controlled pool of data is referred to as database [10].

Databases are classified according to their approaches of organization. They include:

a) Relational database (model), which uses types of tables called relations that represent both data and relationships among those data. Each table has multiple columns, and each column has a unique name.

b) Network Model: Data in the network model are

represented by collections of records and relationship among data; and are represented by links, which can be viewed as pointers.

c) Hierarchical Model is similar to the network model in the sense that data and relationships among data are represented by records and links [8].

[8] explains database management system as consisting of a collection of interrelated data and a set of programs to access those data. It constructs, expands and maintains the database. It makes available the controlled interface between the user and the data in the database. Through its maintenance of indices the database management system allocates storage to data so that any required data can be retrieved and any separate items of data in the database can be cross referenced, [10, 11].

Algorithm

Algorithm is referred to as a set of rules or procedures that must be followed in solving a particular problem. [12]. [13] described an algorithm as a procedure or formula for solving a problem.

MATERIALS AND METHOD

Our major materials for this study are research reports from Computer Science Departments, University of Ilorin submitted as part of prerequisite for the award of first degree

within ten academic years.

In carrying out this study, we adopted the ontological approach to identify the various categories onto which gathered research topics and abstract could be mapped and classified as distinct entities (E.g. Networking, Database, Security etc). Adequate background knowledge pertaining to the research areas are also acquired and represented in such way that could facilitate the mapping. Then we tried to Segment various documents by employing background knowledge

to map each document section to its corresponding category. We created different identity for the data in a persistent data store. We tried to enable easy search across indexed documents by providing a user interface. This is achieved by linking each research work title to its corresponding abstract/full-text.

The figure below shows the interface to upload the research documents into database based on our methodology.

The screenshot displays a web application titled "PROJECT DATABASE MANAGER". At the top, there are navigation links for "Home", "About", and "Contact", along with a "Project Id" dropdown menu. The main interface is divided into several sections. On the left, a "Faculty" dropdown is set to "SCIENCE", and a "Click to Load Departments" section shows "COMPUTER SCIENCE" selected with a "toolbarBtn" button. The central area contains form fields for "Project Id" (01/55EC029), "Author" (Atolagbe-Olaoye, Kolawole), "Year" (2005), "Supervisor" (Prof. T. M. Adeniran), and "Category" (DATABASE MANAGEMENT). Below these fields is a text area labeled "Copy and Paste Abstract in the Box Below" containing a sample abstract. To the right of the form, there are "Browse for" buttons for "Title", "Abstract Location", and "Full Text Location". At the bottom, there are "Add New", "Edit", and "Delete" buttons, and a partially visible text "One or more of the".

FIGURE 1: The interface to upload the research documents into database

RETRIEVAL EXPERIMENTS

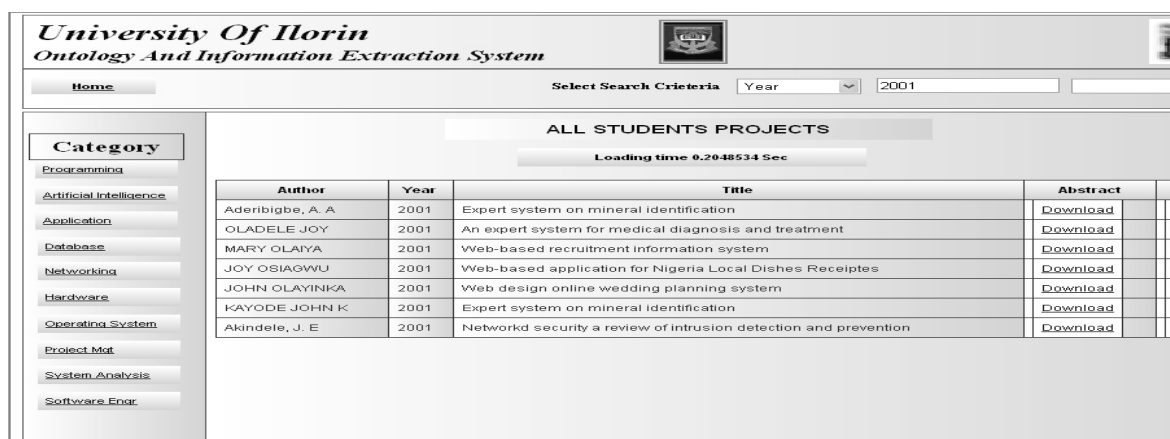
Ontological method adopted in the research provides easy retrieval of relevant research reports to the user's query based on the following steps.

1. Select appropriate Category of ontology.
2. Accept user's query (in English).
 - 2.1 Take stored research reports and search queries posed in English.
 - 2.2 Perform a fully automatic content

analysis of texts.

- 2.3 Match analysed search statements and contents of research reports.
3. Display the stored research reports based on their degree of relevance to the user's queries.

This Figure2 shows search results screen of relevant research reports by the system of above steps.



University Of Ilorin
Ontology And Information Extraction System

Home Select Search Criteria Year 2001

Category

- Programming
- Artificial Intelligence
- Application
- Database**
- Networking
- Hardware
- Operating System
- Project Mgt
- System Analysis
- Software Engr.

ALL STUDENTS PROJECTS

Loading time 0.2048534 Sec

Author	Year	Title	Abstract
Aderibigbe, A. A	2001	Expert system on mineral identification	Download
OLADELE JOY	2001	An expert system for medical diagnosis and treatment	Download
MARY OLAIYA	2001	Web-based recruitment information system	Download
JOY OSIAGWU	2001	Web-based application for Nigeria Local Dishes Recipes	Download
JOHN OLAYINKA	2001	Web design online wedding planning system	Download
KAYODE JOHN K	2001	Expert system on mineral identification	Download
Akindete, J. E	2001	Networkd security a review of intrusion detection and prevention	Download

FIGURE2: Search results screen of relevant research reports based on category of ontology.

CATEGORY MATCH MEASURE (CMM) IN ONTOLOGY

Category Match Measure (CMM) is used to evaluate the coverage of an ontology for the given search text. Categories that contain all search texts will obviously score higher than others. Exact matches have higher score than Partial matches.

For instance, using “Database” as search text, category of ontology with exact search terms will score higher in this measure than other categories which contain partially classes.

Definition 1: Let $C[o]$ be a set of categories in ontology o , and T is the set of search texts.

$$E(o, T) = \sum_{c \in C[o]} \sum_{t \in T} I(c, t) \quad (1)$$

$$I(c, t) = \begin{cases} 1: & \text{if } label(c) = t \\ 0: & \text{if } label(c) \neq t \end{cases} \quad (2)$$

$$P(o, T) = \sum_{c \in C[o]} \sum_{t \in T} J(c, t) \quad (3)$$

$$J(c, t) = \begin{cases} 1: & \text{if } label(c) \text{ contains } t \\ 0: & \text{if } label(c) \text{ not contain } t \end{cases} \quad (4)$$

where $E(o, T)$ and $P(o, T)$ are the number of classes of ontology o that have labels that match any of the searchtexts t exactly or partially, respectively.

$$CMM(o, \tau) = \alpha E(o, T) + \beta P(o, T) \quad (5)$$

where $CMM(o, \tau)$ is the Categories Match Measure for ontology o with respect to search text τ . α and β are the exact matching and partial matching weight factors respectively.

Exact matching is favored over partial matching if $\alpha > \beta$.

RESULT AND DISCUSSION.

The System is tested and evaluated with application of Precision and Recall measures. Hundred of undergraduate's projects with various research work topics are gathered. Firstly, a collection of five queries are tested on twelve ontologies. The relevant research document of various ontologies can be selected through an interface. The active ontological classifications

used in this system are Networking, Security, Database Management System, Information System, Programming and Artificial Intelligence.

A numbers of tests are performed in order to evaluate the effect of ranking model features on degree of relevance of retrieved documents. For example the following queries keyed-in provided the following results according to their degree of relevance:

User Query: "Database"

Results shown in Table 1 are documents that contain searched text “Database” and they are sorted based on the degree of relevance of each

document to the searched text. Also,

Figure 3 is graphical representation of the displayed results

<i>Doc</i>	<i>SORTED RANKS</i>
<i>A</i>	<i>0.00981</i>
<i>B</i>	<i>0.009621</i>
<i>C</i>	<i>0.00947</i>
<i>E</i>	<i>0.009362</i>
<i>F</i>	<i>0.009239</i>
<i>I</i>	<i>0.009197</i>
<i>G</i>	<i>0.008961</i>
<i>M</i>	<i>0.008813</i>
<i>K</i>	<i>0.008629</i>
<i>H</i>	<i>0.00831</i>
<i>J</i>	<i>0.008253</i>
<i>L</i>	<i>0.007581</i>
<i>P</i>	<i>0.005919</i>
<i>D</i>	<i>0.005762</i>
<i>O</i>	<i>0.005511</i>
<i>N</i>	<i>0.005181</i>
<i>Q</i>	<i>0.004096</i>
<i>R</i>	<i>0.002211</i>
<i>T</i>	<i>0.002211</i>
<i>X</i>	<i>0.002203</i>
<i>Y</i>	<i>0.001201</i>

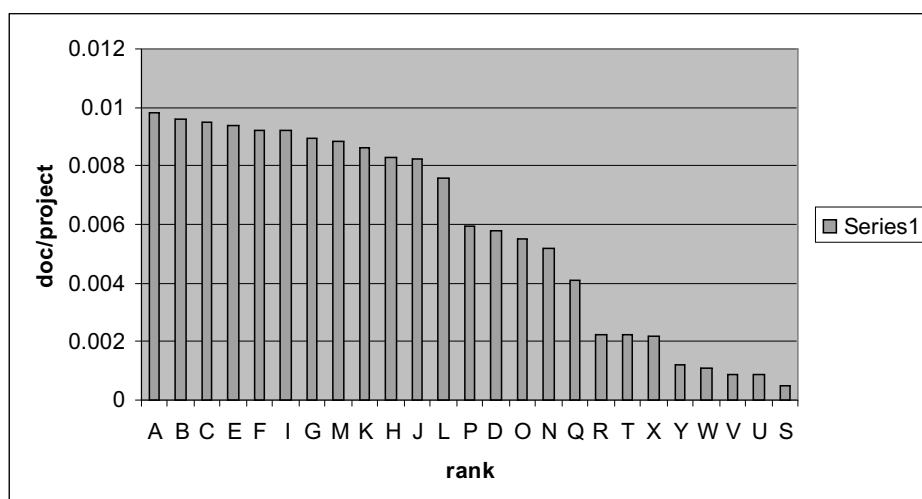


Figure 3: Chart Representation of Searched and Sorted documents

Table 1: Results of Searched and Sorted Documents

CONCLUSION

In this paper, we have attempted to develop a framework for integrating research works, so that subsequent researchers (graduates and undergraduates) can reuse such relevant information in their individual areas of research. We espoused Ontological methods, which have become popular in the domain of knowledge reuse as well as knowledge integration and management. To evaluate the coverage of the ontology of given query, this paper adopted Category Match Measure (CMM), which contain all search texts. The search texts obviously score higher than others. Exact matches have higher score than Partial matches. Also, we tested the effectiveness of our framework by adopting the two metrics- Precision and Recall to assert how well this model is able to answer queries.

With this framework, data of academic reports (projects) are uploaded from the back-end of the system. Users make use of uploaded data in term of needed relevant information at the front end. This is achieved by typing in required information in text format. The model is flexible so that updates and upgrading could be done. Document is retrieved by the system through the following frontend procedures: Parse the query. - Convert words into wordIDs.-Seek to the start of the doclist in the database for every word. - Scan through the doclists until there is a document that matches all the search terms. - Compute the rank

of that document for the query. - Sort the documents that have matched by rank and return the results to the user display interface for reuse.

The model could be upgraded by a user who has knowledge about the domain. This comes in two ways- uploading of new document and updating of existing ones. Uploading of new document is achieved through the following backend procedure: Enter document and necessary information through the administrative end. - Save document (automatically) into database based on different categories of ontology. For the update of existing data, i.e. background knowledge: Substitute background knowledge with current relevant information.

This work has created enhancement for academic administrative purposes as well as academic research purpose. Academic research data management in tertiary schools would be further enhanced by it, e.g. during accreditation as well as knowledge reuse for subsequent researchers.

Future effort on this model could be carried out so that management and integration of data on research reports in all Nigerian Universities could be achieved. Such effort could present model that would facilitate central database in that respect.

REFERENCES

- [1] Stephen D. (1996). *Information Systems for You*. Cheltenham GL50 1yw: Stanley Thornes (Publishers) Ltd.
- [2] Alani H. & Brewster C. (2005). *Ontology Ranking Based on the Analysis of Concept Structures*. In 3rd Int. Conf. Knowledge Capture (K-Cap), pages 5158, Canada. Banff.
- [3] Gruber, T.(1995). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies. Vol.43, Issues 5-6, Novemer 1995, p.907-928.
- [4] Gruber, T. (1992). *Ontolingua: A mechanism to support portable ontologies*. Technical report, Technical Report KSL91-66, Stanford University, Knowledge Systems Laboratory.
- [5] Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- [6] Frakes W.B. & Baeza-Yates R. (1992). *Information Retrieval: Data Structure and Algorithms*. NY: Prentice-Hall.
- [7] Behrouz A.F (2003). *Data Communications and Networking*. New York: McGraw-Hill.
- [8] Silberschatz A. et al. (2001). *Database System Concepts*. New York: McGraw-Hill.
- [9] Hornby, A.S. (2000). *Oxford Advanced Learner's Dictionary of Current English* (fifth edition). Oxford: Oxford University Press.
- [10] French C.S. (2001). *Data Processing and Information Technology*. London: Continuum.
- [11] Mabayoje M. A. (2009) *Ontology and Information Extraction System: A Model for Integrating Research Works*, M. Sc. Thesis, University of Ilorin, Ilorin, Nigeria.
- [12] Russell L.S. (1998). *Introduction to Computing and Algorithms*. Addison Wesley USA: Longman Inc.
- [13] Niklause W. (2005). *Algorithms + Data Structures = Programs*. New Delhi-110015: Prentice-Hall of India.