Online ISSN : 0975-4172
Print ISSN : 0975-4350

Global Journal of Computer Science and Technology : C

Software and Data Engineering

DISCOVERING THOUGHTS AND INVENTING FUTURE

HIGHLIGHTS

- Clinical Practices Guidelines
- Intelligent Information Retrieval
- Decision Tree Classifiers
- Social Network Data

Datacentre

Volume 12 | Issue 10 | Version 1.0 ENG

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 12 Issue 10 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

(A Delaware USA Incorporation with "Good Standing"; Reg. Number: 0423089)

Sponsors:Global Association of Research

Open Scientific Standards

Publisher's Headquarters office

Global Journals Inc., Headquarters Corporate Office, Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141 United States

USA Toll Free: +001-888-839-7392

USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Association of Research, Marsh Road, Rainham, Essex, London RM13 8EU United Kingdom.

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investers@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Editorial Board Members (HON.)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University Feinberg School of Medicine

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo,School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric SciencesPurdue University Director
National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, TaiwanUniversity Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan.Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

Contents of the Volume

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Knowledgebase Representation for Royal Bengal Tiger in the Context of Bangladesh

By Md.Sarwar Kamal & Sonia Farhana Nimmy
BGC Trust University Bangladesh, Chittagong

Abstract - Royal Bengal Tiger is one of the penetrating threaten animal in Bangladesh forest at Sundarbans. In this work we have had concentrate to establish a robust Knowledgebase for Royal Bengal Tiger. We improve our previous work to achieve efficiency on knowledgebase representation. We have categorized the tigers from others animal from collected data by using Support Vector Machines(SVM) .Manipulating our collected data in a structured way by XML parsing on JAVA platform. Our proposed system generates n-triple by considering parsed data. We proceed on an ontology is constructed by Protégé which containing information about names, places, awards. A straightforward approach of this work to make the knowledgebase representation of Royal Bengal Tiger more reliable on the web. Our experiments show the effectiveness of knowledgebase construction. Complete knowledgebase construction of Royal Bengal Tigers how the efficient out-put. The complete knowledgebase construction helps to integrate the raw data in a structured way. The outcome of our proposed system contains the complete knowledgebase. Our experimental results show the strength of our system by retrieving information from ontology in reliable way.

Keywords : *Ontology, Linked data, Web Semantics, XML parsing, N-triples, Royal Bengal Tiger.*

GJCST-C Classification: *I.2.4*

Strictly as per the compliance and regulations of:

Knowledgebase Representation for Royal Bengal Tiger in the Context of Bangladesh

Md.Sarwar Kamal[α] & Sonia Farhana Nimmy[σ]

Abstract - Royal Bengal Tiger is one of the penetrating threaten animal in Bangladesh forest at Sundarbans. In this work we have had concentrate to establish a robust Knowledgebase for Royal Bengal Tiger. We improve our previous work to achieve efficiency on knowledgebase representation. We have categorized the tigers from others animal from collected data by using Support Vector Machines(SVM) .Manipulating our collected data in a structured way by XML parsing on JAVA platform. Our proposed system generates n-triple by considering parsed data. We proceed on an ontology is constructed by Protégé which containing information about names, places, awards. A straightforward approach of this work to make the knowledgebase representation of Royal Bengal Tiger more reliable on the web. Our experiments show the effectiveness of knowledgebase construction. Complete knowledgebase construction of Royal Bengal Tigers how the efficient out-put. The complete knowledgebase construction helps to integrate the raw data in a structured way. The outcome of our proposed system contains the complete knowledgebase. Our experimental results show the strength of our system by retrieving information from ontology in reliable way.

IndexTerms : Ontology, Linked data, Web Semantics, XML parsing, N-triples, Royal Bengal Tiger.

I. Introduction

The sovereign Royal Bengal Tiger is drifting near the frontier of extinction. Once, the tiger cracked the whip over a supreme part of the globe ranging from the Pacific to the Black Sea and from Ural Mountains to the Mountain Agung. It is a paradox of fate that tiger is facing an assailment of poaching throughout its range. The main factor contributing in the decline of cat population is habitat degradation. But poaching has put them in a vulnerable condition to survive. The forest department sources said the big cat species are now disappearing fast from the world as the current population of tiger is only about 3700, down from around one lakh in 1900.There are only five sub-species of tigers surviving in the world which are Bengal tiger, Siberian tiger, Sumatran tiger, South- China tiger and Indo-China tiger. Balinese tigers, Javanese tigers and Caspian tigers have already vanished from the planet as the experts estimated that the remaining species of the big cat are likely to disappear immediately with the advent of next century. Official sources said at least 60 tigers were killed in the last three decades as the animals came to the nearby locality in search of food. According to review of the ministry, the big cats kill 25 to40 people annually while two to three tigers fall victim of mass-beating. According to a study conducted jointly by the United Nations, Bangladeshi government and Indian government in 2004, as many as 440 tigers have been found in the Bangladeshi part of the Sundarbans, the sources said. Right now tigers occupy only 7% of their historic range and they live in small islands of forests surrounded by a sea of human beings. Over the past few centuries tigers lost more than 80% of their natural habitats and what remain are only small fragments under heavy anthropogenic pressure.

This paper Organized as follows. In section II we have narrates Knowledgebase and Ontological basics and terminology which are essential for representation of Knowledgebase. In section III we described the General terminologies of Knowledgebase. In section IV we have described briefly Support Vector Machines (SVM) on the eve of categorized the Tiger from other animals. In section V we have elaborate INTRINSIC INFORMATION CONTENT METRIC and in next section we cited the Instance Matching Algorithm. last but not the least we have rape out by defining the challenges of the Ontology Instances Matching.

II. Knowledgebase and Ontology

Knowledge bases are playing an increasingly important role in enhancing the intelligence of Web and enterprise search and in supporting information integration. Today, most knowledge bases cover only specific domains, are created by relatively small groups of knowledge engineers, and are very cost intensive to keep up-to-date as domains change. At the same time, Wikipedia has grown intoone of the central knowledge sources of mankind, maintained by thousands of contributors Kobilarovetal. Collected data are organized to parsing and enable them to extract easily on the web. The complete knowledgebase contain information about Royal Bengal Tiger to enrich it. This knowledgebase helps to get informative knowledge about Royal Bengal Tiger who are an important part of our country as well as whole world. Our motivation is to provide a perfect representation of Royal Bengal Tiger on the web through Knowledgebase. The knowledge captured in the ontology can be used to parse and generate N-triples.

Author α : Lecturer, Computer science and engineering, BGC Trust University Bangladesh, Chittagong.E-mail : sarwar.bgctub@gmail.com
Author σ : Lecturer, Computer science and engineering, BGC Trust University Bangladesh, Chittagong. E-mail : nimmy_cu@yahoo.com

Structured data is easy to extract on the web which can be accessible for people to reach their goal. Our motive is to take the data in a structured way.

a) Ontology Alignment

Alignment A is defined as a set of correspondences with quadruples < e; f; r; l > where **e** and **f** are the two aligned entities across ontology's, **r** represents the relation holding between them, and **l** represents the level of confidence [0, 1] if there exists in the alignment statement. The notion **r** is a simple (one-to-one equivalent) relation or a complex (subsumption or one-to-many) relation Ehrig (2007). The correspondence between **e** and **f** is called aligned pair throughout the paper. Alignment is obtained by measuring similarity values between pairs of entities.

The main contribution of our Anchor-Flood algorithm is of attaining performance enhancement by solving the scalability problem in aligning large ontology's. Moreover, we obtain the segmented alignment for the first time in ontology alignment field of research. We achieve the best runtime in world-wide competitions organized by Ontology Alignment Evaluation Initiative (OAEI) 2008 (held in Karlsruhe, Germany) and 2009 (held in Chantilly, VA, USA).

b) Intrinsic Information Content

We propose a modified metric for Intrinsic Information Content (IIC) that achieves better semantic similarity among concepts of ontology. The IIC metric is integrated with our Anchor-Flood algorithm to obtain better results efficiently.

c) Ontology and Knowledge Base

According to Ehrig (2007), an ontology contains core ontology, logical mappings, a knowledge base, and a lexicon. A core ontology, S, is defined as a tuple of five sets: concepts, concept hierarchy or taxonomy, properties, property hierarchy, and concept to property function.

$$S = (C, \leq_c R, \sigma, \leq_R)$$

where **C** and **R** are two disjoint sets called concepts" and relations" respectively. A relation is also known as a property of a concept. A function represented by σ(r) =< dom(r); ran(r) > where r € R, domain is dom(r) and range is ran(r). A partial order≤R represents on R, called relation hierarchy, where r1≤R r2 iff dom (r1) ≤C dom (r2) and ran (r1) ≤C ran (r2). The notation ≤C represents a partial order on C, called concept hierarchy or taxonomy". In a taxonomy, if c1 <C c2 for c1; c2€C, then c1 is a sub concept of c2, and c2 is a super concept of c1. If c1 <C c2 and there is no c3€C with c1 <C c3 <C c2, then c1 is a direct sub concept of c2, and c2 is a direct super concept of c1 denoted by c1 ≺ c2. The core ontology formalizes the intentional aspects of a domain. The extensional aspects are provided by knowledge bases, which contain asserts about instances of the concepts and relations. A knowledge base is a structure KB = (C,R, I, C, , R) consisting of

_ two disjoint sets C and R as defined before,

_ a set I whose elements are called instance identifiers (or instance for short),

_ a function C : C→ Ə(I) called concept instantiation,

_ a function { R: R → Ə(I2) with (r) ⊆ ϡC (dom(r)) × ϡC (ran(r)), for all r ε R. The function R is called relation instantiation.

With data types being concepts as stated for core ontology, concrete values are analogously treated as instances.

III. General Terminology

This section introduces some basic definitions of terminologies of semantic web to familiarize the readers with the notions used throughout the paper. It includes the definitions of ontology and knowledgebase, linked data, Geonames, Geospatial data, and N-triples from semantic web to comprehend the essence of our paper.

a) N-Triples

N-Triples is a format for storing and transmitting data. It is a line-based, plain text serialization format for RDF (Resource Description Framework) graphs, and a subset of the Turtle (Terse RDF Triple Language) format.[1][2] N-Triples should not be confused with Notation 3 which is a superset of Turtle. N-Triples was primarily developed by Dave Beckett at the University of Bristol and Art Barstow at the W3C. N-Triples was designed to be a simpler format than Notation 3 and Turtle, and therefore easier for software to parse and generate. However, because it lacks some of the shortcuts provided by other RDF serializations (such as CURIEs and nested resources, which are provided by both RDF/XML and Turtle) it can be onerous to type out large amounts of data by hand, and difficult to read.

b) Geonames

Geonames is a geographical database available and accessible through various Web services, under a Creative Commons attribution license. Geonames is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. All lat/long coordinates are in WGS84 (World Geodetic System 1984). Users may manually edit, correct and add new names using a user friendly wiki interface.

c) Geospatial Data

Geospatial data is information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth, typically represented by points, lines, polygons, and or

complex geographic features. This includes original and interpreted geospatial data, such as those derived through remote sensing including, but not limited to, images and raster data sets, aerial photographs, and other forms of geospatial data or data sets in both digitized and non-digitized forms.

d) Neighbouring of Geospatial Data

At first, we find the neighbours of a division. In the same way we also find the neighbours of other six divisions. After that, we find the neighbours of all districts. At last, we find the neighbours of all sub districts one by one.

e) Linked Data

With the structures of ontology and ontology knowledge base, semantic web visionaries coined the term linked data, which uses Resource Description Framework (RDF) and RDF triples to connect related instances. The term refers to a style of publishing and interlinking structured data on the Web. The basic assumption behind Linked Data is that the value and usefulness of data increases the more it is interlinked with other data. In summary, Linked Data is simply about using the Web to create typed links between data from different sources. However, semantic knowledge base and linked data is used synonymously throughout this paper.

f) Semantic Web

The Semantic Web1 has received much attention recently. Its vision promises an extension of the current web in which all data is accompanied with machine understandable metadata allowing capabilities for a much higher degree of automation and more intelligent applications (Berners-Lee et al., 2001). To make this idea more concrete, consider the statement The University of Georgia is located in Athens, GA. To a human with knowledge of colleges and universities and the geography of the southeastern United States, the meaning of this statement is clear. In addition, upon seeing this statement, other related information comes to mind such as professors who work at the University. In a Semantic Geospatial Web context (Egenhofer, 2002), this related information would be GIS data and services, such as road network data and facility locations for the Athens area which could be combined with way finding services. The goal of the Semantic Web is to make the semantics of such data on the web equally clear to computer programs and also to exploit available background knowledge of related information. On the Semantic Web this statement would be accompanied with semantic metadata identifying an instance of the concept University with the name The University of Georgia. Similarly, the instance of City and State, Athens, GA, would unambiguously describe the university's geographic location. Note the distinction between semantic metadata describing high-level concepts and relationships and syntactic and structural metadata describing low level properties like file size and format. To create this semantic metadata, we must identify and mark occurrences of known entities and relationships in data sources. This tagging process is known as metadata extraction and semantic annotation. These annotations are especially important for multimedia data, as non textual data has a very opaque relationship with computers. Some examples of annotation of textual and multimedia data are presented in (Dill et al., 2003; Hammond et al. 2002), and (Jin et al., 2005) respectively. To provide ontological metadata in a machine process able form, a standard way to encode it is needed. The W3C has adopted Resource Description Framework (RDF) as the standard for representing semantic metadata. Metadata in RDF is encoded as statements about resources. A resource is anything that is identify able by a Uniform Resource Identifier (URI). Resources can be documents available on the web or entities which are not web-based, such as people and organizations.

IV. Support Vector Machines

Support Vector Machine (SVM) is one of the latest clustering techniques which enables machine learning concepts to amplify predictive accuracy in the case of axiomatically diverting data those are not fit properly. It uses inference space of linear functions in a high amplitude feature space, trained with a learning algorithm. It works by finding a hyperplane that linearly separates the training points, in a way such that each resulting subspace contains only points which are very similar. First and foremost idea behind Support Vector Machines (SVMs) is that it constituted by set of similar supervised learning. An unknown tuple is labeled with the group of the points that fall in the same subspace as the tuple. Earlier SVM was used for Natural Image processing System (NIPS) but now it becomes very popular is an active part of the machine learning research around the world. It is also being used for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by V.Vapnik.

Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. The flexibility of kernel functions allows the SVM to search a wide variety of hypothesis spaces. All hypothesis space help to identify the Maximum Margin Hyperplane(MMH) which enables to classify the best and almost correct data The following figure shows the process of SVMs selection from large amount of SVMs.

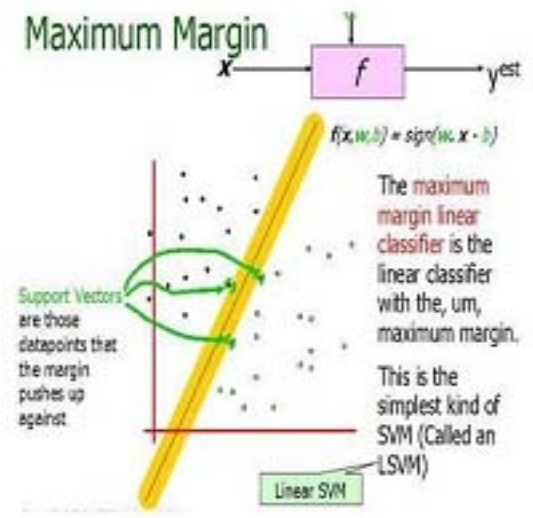


Fig 1: Maximum Margin Hyper Plane

Expression for Maximum margin is given as [4][8] (for more information visit [4]

$$\text{margin} \equiv \arg\min_{\mathbf{x}\in D} d(\mathbf{x}) = \arg\min_{\mathbf{x}\in D} \frac{|\mathbf{x}\cdot\mathbf{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

The above illustration is the maximum linear classifier with the maximum range. In this context it is an example of a simple linear SVM classifier. Another interesting question is why maximum margin? There are some good explanations which include better empirical performance. Another reason is that even if we've made a small error in the location of the boundary this gives us least chance of causing a misclassification. The other advantage would be avoiding local minima and better classification. Now we try to express the SVM mathematically and for this tutorial we try to present a linear SVM. The goals of SVM are separating the data with hyper plane and extend this to non-linear boundaries using kernel trick [8] [11]. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If Yi= +1;

[b] If Yi= -1; wxi + b ≤ 1

[c] For all i; yi (wi + b) ≥ 1

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training data is good and every test vector is located in radiusr from training vector. Now if the chosen hyper plane is located at the farthest possible from the data [12]. This desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets. Thus we have [a], [b] & [c].

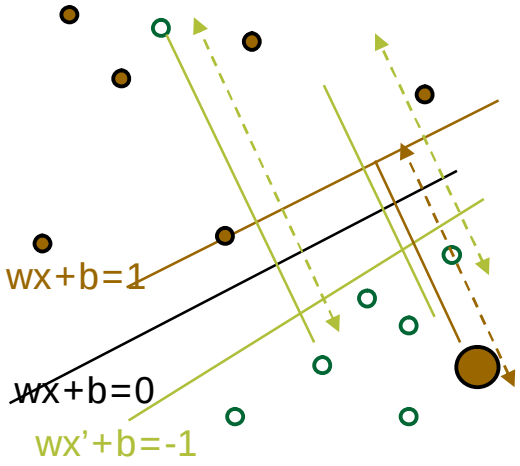


Figure 2 : Representation of Hyper planes

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyperplane to nearest points. Maximum Margin = M = 2 / ||w||

V. Related Work

Before this work we have had work to prepare ontology for medical document classification. We have reviewed 20 research journals on the eve of knowledgebase representation for Tigers but we got only a few that does not indicates the outcome for Tigers knowlegebase.

VI. Proposed Modification In Intrinsic Information Content Metric

To overcome the limitation of the state-of-art metrics of computing semantic similarity among concepts within domain ontology and to cope with the new ontologies with the introduced complex description logics, we propose a modified metric of computing intrinsic information content. The metric can be applied to a simple taxonomy and to a recent complex OWL ontology as well.

The primary source of IC in ontology is obviously concepts and concept hierarchy. However, OWL ontology also contains properties, restrictions and other logical assertions, often called as relations. Properties are used to define functionality of a concept explicitly to specify a meaning. They are related to concept by means of domain, range and restrictions.

According to Resnik, semantic similarity depends on the shared information. As Resnik introduces the IC which represents the expressiveness of a particular concept. Classical metric of IC are based on the available concepts in taxonomy or in a large text

corpora. However, as time passes on, the definition and the content of ontology becomes more and more complex. The expressiveness of a concept is not only rely on the concept taxonomy but also on the other relations like properties and property-restrictions.

We already have discussed about the probable sources of information content(IC) or the expressiveness of semantic similarity among the concepts of ontology. We find that the IC of a concept is negatively related to the probability of a concept in external large text corpora Resnik (1995). We also find that the IC of a concept is inversely related to the number of hyponyms or the concepts it subsumes Seco et al. (2004). Moreover, we observe that description logic (DL) based ontology of semantic technology is formal and explicit in its conceptualization with the help of relations. Every concept is defined with sufficient semantic embedding with the organization, property functions, property restrictions and other logical assertions. Current ontology of semantic technology is defined as an explicit specification of a conceptualization" Gruber (1995). Although the most domain ontologies are not as complete as Word Net in terms of concepts and concept organization, they have well support from logical assertions to define a concept concisely. Therefore, we can obtain sufficient IC of a concept without depending on the external large text corpora heavily, required that we use intrinsic information of the concept. One of the good sources of intrinsic information of a concept is its relations by means of property functions and property restrictions. Our relation based IC is defined as:
Icrel(c)

$$Ic_{rel}(c) = \frac{Log(rel(C)+1)}{Log(total\_rel+1)} \quad (1)$$

Where rel stands for the relation of properties, property function and restrictions, rel(c) denotes the number of relations of a concept c and total rel represents the total number of relations available in the ontology.

As long as the information content of a concept depends both on the hyponyms or sub sumption relations of a concept and the related properties of the concept, we need to integrate the icre(c) with the Seco's metric This integration introduces a coefficient factor ρ and the equation becomes as:

$$ic(c) = \rho.icrel(c) + (1-\rho).icseco(c) \quad (2)$$

Concepts	Number of relations	Number of Hypotenuse	IC_{scco}	IC_{rel}	$IC_{modified}$
Date	3	0	1.000	0.332	0.641
Page Range	2	0	1.000	0.263	0.603
Organization	0	3	0.613	0.000	0.283
Institution	3	1	0.693	0.322	0.257
Publisher	3	2	1.000	0.222	0.123
School	3	3	1.000	0.123	0.968
List	0	0	1.000	0.258	0.987
Person List	4	0	1.000	0.125	0.789
Journal	2	1	1.000	0.236	0.456
Address	3	0	1.000	0.125	0.489
Person	0	1	1.000	0.000	0.478
Conference	0	3	1.000	0.231	0.258
Reference	1	0	1.000	0.963	0.369
Academic	6	0	1.000	0.000	0.123
PhDThesis	5	1	1.000	0.217	0.147
MastersThesis	2	2	1.000	0.235	0.258
Misc	0	2	0.873	0.148	0.000
Motion	0	2	0.521	0.148	0.123
Picture Part	0	3	0.123	0.000	0.236
In Collection	0	0	1.000	0.789	0.214

Table 1: contains IC values measured by Saco's metric and our modified metric

Where the coefficient factor ρ is defined by the nature of ontology. While a small size of ontology is often incomplete by its concepts alone, the coefficient factor tends to increase to focus on relations. On the contrary, when relations are inadequate to define a concept and there are a large number of concepts in the taxonomy, ρ tends to decrease its value. However, we definitely need a trade-off to select the coefficient factor and we define it as:

$$\rho = \frac{Log(total\_rel+1)}{Log(total\_rel)+Log(total\_concept)}$$

Where total_rel is the maximum number of relations while total_ concepts is the maximum number of concepts available in an ontology.

From the experiments, we also observe that the deeper concepts have more expressiveness or larger IC values. Therefore, it guarantees that our modified IC metric takes the depth of a concept implicitly and the children of a concept explicitly.

However, we do not take the link type and local concept density into account unlike expressed in Jiang & Conrath (1997). As we consider thyponyms by incorporating the Saco's IC metric, it considers the edges between sub sumption concepts implicitly Furthermore, we also compute semantic similarity for every possible pair of concepts of the ontology.

e1	e2	Sim_{seco}	$Sim_{proposed}$
Reference	PhD Thesis	0.113	0.782
Reference	Master's Thesis	0.113	0.782
Reference	In Collection	0.113	0.782
Reference	In Proceedings	0.113	0.782
Reference	Article	0.113	0.790
Reference	Chapter	0.113	0.784
Reference	In Book	0.113	0.784
Reference	TechReport	0.113	0.777
Reference	Deliverable	0.113	0.784
Reference	Manual	0.113	0.790
Reference	Unpublished	0.113	0.790
Reference	Booklet	0.113	0.777
Reference	Lecture Notes	0.113	0.788
Reference	Collection	0.113	0.782
Reference	Monograph	0.113	0.782
Reference	Proceeding	0.113	0.782

Table 2: contains semantic similarity between Reference to each of its leaves considering Saco's metric and our proposed metric

VII. Instance Matching Algorithm

The operational block of the instance matching integrates ontology alignment, retrieves semantic link clouds of an instance in ontology and measures the terminological and structural similarities to produce matched instance pairs. Pseudo code of the Instance Matching algorithm:

Algo. InstanceMatch (ABox ab1, ABox ab2, Alignment A)
for each insi element of ab1
cloudi=makeCloud(insi,ab1)
for each insj element of ab2
cloudj=makeCloud(insj,ab2)
if $\forall$a(c1; c2) elements of A|c1 elements of Block(ins1:type) ^
c2 elements of Block(ins2:type)
if Simstruct(cloudi; cloudj) $\geq \delta$
imatch=imatch $\bigcup$ makeAlign(insi; insj)

VIII. Ontology Instance Matching Challenges

The ontology schema, which includes concepts, properties and other relations, is relatively stable part of an ontology. However, concepts and properties of ontology are instantiated very often by deferent users in deferent styles. Thus, ontology instances are dynamic in nature and are challenging to be matched. Structural variants compose of the most challenging variations in defining instances. To define an instance of a concept, ontology users usually take support from the properties, either object properties or data properties. Properties always behave like functions having domains and ranges. There might be a great variation of using property functions in their range values. The range of an Object Property is an instance while the range of a Data type Property is an absolute value. There is always a chance of defining an Object Property of ontology as a Data type Property in ontology and vice versa. The cases of defining aproperty by another instance in one ABox and defining the property by a value in other ABox yield a great challenge in instance matching.

a) Approach to Solve the Challenges

We resolve typographical variation by the methods of data cleansing. The task of data cleansing comprises the detection and resolution of errors and inconsistencies from a data collection. Typical tasks are syntax check, normalization, and error correction. First of all, our syntax check and normalization process check the data type of an instance and classify on three important information types: time data (using regular expression), location data (using Geo Names Web service) and personal data. In our current realization, we use a couple of manually defined normalization rules for each information type. We implemented the module in a modular way, so that the used algorithm and rules of normalization can be extended and substituted. In instance matching, we need to look up the type (concept as a type of an instance) match of instances first. To cope with the logical variation, we first look up a block of concepts that includes the original type of an instance against another block of concepts which includes the type of another instance to be compared with instead of comparing two types alone. A relational block is defined as follows:

Definition 1: As concepts are organized in a hierarchical structure called a taxonomy, we consider a relational block of a concept c as a set of concepts and simply referred to block throughout this paper, and defined as:

$$\text{block}(c) = \{\text{children}(c) \bigcup \text{siblings}(c) \bigcup \text{parents}(c) \bigcup \text{grandparents}(c)g\}$$

where children(c) and parents(c) represent the children and the parents of a particular concept c, respectively within a taxonomy, whereas siblings(c) is defined as children (parents(c)-c and grandparents(c) is defined as parents (parents(c)) In an ontology, neither a concept nor an instance comprises its full specification in its name or URI (Uniform Resource Identifier) alone. Therefore we consider the other semantically linked information that includes other concepts, properties and their values and other instances as well. They all together make an information cloud to specify the meaning of that particular instance. The degree of certainty is proportional to the number of semantic links associated to a particular instance by means of property values and other instances. We refer the collective information of association as a Semantic Link cloud (SLC), which is defined as below:

Definition 2: A Semantic Link Cloud (SLC) of an instance is defined as a part of knowledge base Ehrig

(2007) that includes all linked concepts, properties and their instantiations which are related to specify the instance sufficiently.

IX. Conclusions

In this dissertation, we described the Anchor-Flood algorithm that can align ontologies of arbitrary size effectively, and that makes it possible to achieve high performance and scalability over previous alignment algorithms. To achieve these goals, the algorithm took advantage of the notion of segmentation and allowed segmented output of aligned ontologies. Specifically, owing to the segmentation, our algorithm concentrates on aligning only small sets of the entire ontology data iteratively, by considering\locality of reference". This brings us a by-product of collecting more alignments in general, since similar concepts are usually more densely populated in segments. Although we need some further refinement in segmentation, we have an advantage over traditional ontology alignment systems, in that the algorithm finds aligned pairs within the segments across ontologies and it has more usability in different discipline of specific modelling patterns. When the anchor represents correct aligned pair of concepts across ontologies, our Anchor-Flood algorithm finds segmented alignment within conceptually closely connected segments across ontologies efficiently. Even if the input anchor is not correctly defined, our algorithm is also capable of handling the situation of re- porting misalignment error. The complexity analysis and a different set of experiments demonstrate that our proposed algorithm outperforms in some aspect to other alignment systems. The size of ontologies does not affect theefficiency of Anchor-Flood algorithm. The average complexity of our algorithm is ON log(N), where N is the average number of concepts of ontologies.

References Références Referencias

1. Benjamin's, V Contreras, J., Corcho, O. & G_omez-P_erez, A. (2004).Six Challenges for the Semantic Web. AIS SIGSEMIS Bulletin,
2. Berners-Lee, T., Fischetti, M. & Dertouzos, M. (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper San Francisco.
3. Caracciolo, C., Euzenat, J., Hollink, L., Ichise, R., Isaac, A.,Malais_e, V., Meilicke, C., Pane, J., Shvaiko, P., Stuckenschmidt, H., _Sv_ab-Zamazal, O. & Sv_atek, V. (2008). Results of the Ontology Alignment Evaluation Initiative 2008. Proceedings of Ontology Matching Work-shop of the 7th International Semantic Web Conference, Karlsruhe, Germany. Collins, A. & Quillian, M. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior.
4. Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
5. Ehrig, M. (2007). Ontology Alignment: Bridging the Semantic Gap. Springer, New York.
6. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. The MIT Press, Cambridge
7. Euzenat, J., Ferrara, A., Hollink, L., Joslyn, C., Malais_e, V.,Meilicke, C., Nikolov, A., Pane, J., Scharffe, F., Shvaiko, P.,Spiliopoulos, V., Stuckenschmidt, H., _Sv_ab-Zamazal, O., Sv_atek, V., Santos, C.T. & Vouros, G. (2009). Preliminary results of the Ontology Alignment Evaluation Initiative 2009. Proceedings of Ontology MatchingWorkshop of the 8th International Semantic Web Conference, Chantilly, VA,USA.
8. Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000
9. Image found on the web search for learning and generalization in svm following links given in the book above. Engineering Review.
10. Grau, B., Parsia, B., Sirin, E. & Kalyanpur, A. (2005). Modularizing OWL ontologies. Proc. KCAP-2005 Workshop on Ontology Management, Banff , Canada.
11. Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997.
12. J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.
13. Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database.
14. Hu, W., Cheng, G., Zheng, D., Zhong, X. & Qu, Y. (2006a). The Results of Falcon-AO in the OAEI 2006 Campaign. Proceedings of Ontology Matching (OM-2006), Athens, Georgia, USA.
15. Hu, W., Zhao, Y. & Qu, Y. (2006b). Partition-based Block Matching of Large Class Hierarchies. Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Beijing, China.
16. Hu, W., Qu, Y. & Cheng, G. (2008). Matching Large Ontologies: A Divideand- Conquer Approach. Data and Knowledge Engineering
17. Jiang, J. & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings on International Conference on Research in Computational Linguistics, Taiwan.
18. Knappe, R., Bulskov, H. & Andreasen, T. (2007). Perspectives on ontology-based querying. International Journal of Intelligent Systems.
19. Kobilarov, G., Bizer, C., Auer, S. & Lehmann, J. Dbpedia-a linked data hub and data source for web and enterprise applications. Programme Chairs.

20. Lin, D. (1998). An information-theoretic de_nition of similarity.
21. Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity Flooding: AVersatile Graph Matching Algorithm and its Application to Schema Matching.Proceedings of the 18th International Conference on Data Engineering (ICDE2002), San Jose, CA, USA.
22. Miller, G. & Charles, W. (1991). Contextual Correlates of Semantic Similarity. Language and cognitive processes.
23. Mitschick, A., Nagel, R. & Mei_ner, K. (2008). Semantic Metadata Instantiation and Consolidation within an Ontologybased Multimedia Document Management System. In Proceedings of the 5th European Semantic Web Conference ESWC.
24. Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE transactions on systems, man and cybernetics.
25. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.
26. Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.Journal of arti_cial intelligence.
27. Rogers, J. (2005). OpenGALEN: Making the Impossible Very Difficult.http://www.opengalen.org.
28. Rosse, C. & Mejino, J. (2003). A Reference Ontology for Biomedical Informatics: the Foundation Model of Anatomy. Journal of Biomedical Informatics.
29. Seco, N., Veale, T. & Hayes, J. (2004). An intrinsic information contentmetric for semantic similarity in WordNet. Proceedings of ECAIf2004, the 16th European Conference on Arti_cial Intelligence.
30. Seddiqui, M. & Aono, M. (2008). Alignment Results of Anchor-Flood Algorithm for OAEI-2008. Proceedings of Ontology Matching Workshop of the 7th International Semantic Web Conference, Karlsruhe, Germany.
31. Seddiqui, M.H. & Aono, M. (2009a). An E_cient and Scalable Algorithm for Segmented Alignment of Ontologies of Arbitrary Size. Journal of Web Semantics: Science, Services and Agents on the World Wide Web.
32. Seddiqui, M.H. & Aono, M. (2009b). Anchor-Flood: Results for OAEI-2009.Proceedings of Ontology Matching Workshop of the 8th International Semantic Web Conference, Chantilly, VA, USA.
33. Seidenberg, J. & Rector, A. (2006). Web Ontology Segmentation: Analysis, Classification and Use. Proceedings of the 15th International Conference on World Wide Web (WWW2006), Edinburgh, Scotland.
34. Shannon, C. & Weaver, W. (1948). A mathematical theory of communication. Bell System Technical Journal.
35. Shvaiko, P. & Euzenat, J. (2008). Ten challenges for ontology matching.Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), Monterrey, Mexico.
36. Stoilos, G., Stamou, G. & Kollias, S. (2005). A String Metric for Ontology Alignment. Proceedings of the 4th International Semantic Web Conference (ISWC2005), Galway, Ireland.
37. Stuckenschmidt, H. & Klein, M. (2004). Structure-based Partitioning of Large Concept Hierarchies.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Understanding Rule Behavior through Apriori Algorithm over Social Network Data

By S.S.Phulari, P.U.Bhalchandra, Dr.S.D.Khamitkar & S.N.Lokhande
S.R.T.M.University, Nanded, MS, India

Abstract - APRIORI algorithm is a popular data mining technique used for extracting hidden patterns from data. This paper highlights practical demonstration of this algorithm for association rule mining over a survey data set of students related to social network usage. We concluded with discussions on the number of research observations including new rules generated during the process.

GJCST-C Classification: *E.m*

Strictly as per the compliance and regulations of:

Understanding Rule Behavior through Apriori Algorithm over Social Network Data

S.S.Phulari[α], P.U.Bhalchandra[α], Dr.S.D.Khamitkar[α] & S.N.Lokhande[α]

Abstract - APRIORI algorithm is a popular data mining technique used for extracting hidden patterns from data. This paper highlights practical demonstration of this algorithm for association rule mining over a survey data set of students related to social network usage. We concluded with discussions on the number of research observations including new rules generated during the process.

I. Introduction

Data mining is a technique that helps to extract important data from a large database. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. The myth possible with data mining includes automated prediction of trends and behaviors and automated discovery of previously unknown patterns. The most commonly used techniques in data mining are:

1. Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
3. Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.
6. Apriori is a classic algorithm used in data mining for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

Author α : School of Computational Sciences, S.R.T.M.University , Nanded, MS, India. E-mail : Santoshphulari@gmail.com , E-mail : srtmun.parag@gmail.com, E-mail : s.khamitkar@gmail.com

II. Analysis of Apriori Algorithm

Apriori was proposed by Agrawal and Srikant in 1994. The algorithm finds the frequent set L in the database D. It makes use of the downward closure property. The algorithm is a bottom search, moving upward level; it prunes many of the sets which are unlikely to be frequent sets, thus saving any extra efforts. Apriori algorithm is an algorithm of association rule mining. It is an important data mining model studied extensively by the database and data mining community. It Assume all data are categorical. It is initially used for Market Basket Analysis to find how items purchased by customers are related.

The problem of finding association rules can be stated as : Given a database of sales transactions, it is desirable to discover the important associations among different items such the presence of some items in a transaction will imply the presence of other items in the same transaction. As example of an association rule is:

Contains (T, “baby food”) → Contains (T, “diapers”) [Support= 4%, Confidence=40%]

The interpretation of such rule is as follows:

- 40% of transactions that contains baby food also contains diapers;
- 4% of all transactions contain both of these items.

The calculations of the Support(S) and Confidence(C) are very simple:

- CONF (A → B) = SUPP(AUB)
- SUPP(A)
- S (A) = (Number of transactions containing item A) /(Total number of transactions in the database)

- S (A → B) =(Number of transactions containing items A and B) / (Total number of transactions in the database)

The above association rule is called single-dimension because it involves a single attribute or predicate (Contains). The main problem is to find all association rules that satisfy minimum support and minimum confidence thresholds, which are provided by user and/or domain experts. A rule is frequent if its support is greater than the minimum support threshold and strong if its confidence is more than the minimum confidence threshold.

Discovering all association rules is considered as two phase process where we find all frequent item sets having minimum support. The search space to enumeration all frequent item sets is on the magnitude of 2 * n. In second step, we generate strong rules. Any association that satisfies the threshold will be used to generate an association rule. The first phase in discovering all association rules is considered to be the most important one because it is time consuming due to the huge search space (the power set of the set of all items) and the second phase can be accomplished in a straightforward manner.

III. Algorithm for Apriori

The pseudo code for the algorithm is given below. For a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed; though note that T is a multi set. C_k is the candidate set for level k. Generate () algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma.

$count[c]$ accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

$$
\begin{aligned}
&\text{Apriori}(T,\epsilon)\\
&\quad L_1 \leftarrow \{\text{ large 1-itemsets }\}\\
&\quad k \leftarrow 2\\
&\quad \textbf{while } L_{k-1} \neq \emptyset\\
&\qquad C_k \leftarrow \{c \in a \cup \{b\} | a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}\\
&\qquad \textbf{for transactions } t \in T\\
&\qquad\quad C_t \leftarrow \{c | c \in C_k \wedge c \subseteq t\}\\
&\qquad\quad \textbf{for candidates } c \in C_t\\
&\qquad\qquad count[c] \leftarrow count[c] + 1\\
&\qquad L_k \leftarrow \{c \in C_k | count[c] \geq \epsilon\}\\
&\qquad k \leftarrow k + 1\\
&\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

IV. Steps in Finding the Association Rules Using Apriori

A large supermarket tracks sales data by stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, and {2,4}. Each number corresponds to a product such as "butter" or "bread". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately: Table 1 explains the working of Apriori algorithm. We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 3. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, priori *prunes* the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent and generate a list of all 3-triples of the frequent items (by connecting frequent pairs with frequent single items). In the example, there are no frequent 3-triples. Most common 3-triples are {1,2,4} and {2,3,4}, but their support is equal to 2 which is smaller than our min support. Table 2 explains these items.

Item	Support
1	3
2	6
3	4
4	5

Table 1:

Item	Support
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

Table 2 :

V. Implementing Apriori Algorithm in Weka

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General Public License.

VI. Data Set Features

A closed questionnaire of 56 questions, labeled A,B,C…… BB was prepared and circulated among 56 students. Maximum questions were having four options to answer. Theses answers were caught as a1, a2, a3, a4 (a means answer) . These questionnaire were circulated randomly to avoid mass copying of the answers and then collected after one hour. Out of 56 , only 43 questionnaire were correct in all respects. Remaining 13 needed interactions with the corresponding students as few questions on them were not answered by them. Since 13 students refused to re answer these, we have rejected them out. Microsoft Excel was used to tabulate the data in the questionnaire and 43 rows were created . A CSV(Comma Separated Values) sheet was made from it which has been fed as input to the WEKA Algorithm.

E ***Talking about Face book, how frequently do you log in?***

e1 Several times a day ☐ e2 At least once a day ☐

e3 At least once a week ☐ e4 At least once a month ☐

F ***When you access Face book, on average how much time do you spend looking at the wall and photos of your contacts?***

f1 Less than 15 min ☐ f2 from 15 to 30 min ☐

f3 From 30 min to 1 h ☐ f4 More than 1 h ☐

G. ***Have you joined any Face book groups?***

g1Yes ☐ g2 No ☐

H. ***Indicate how many social networking site you are registered with apart from Face book***

h1 None ☐ h2 1 ☐

h3 Less than 5 ☐ h4 More than 5 ☐

VII. Rules Generated

1. G=g1 34 ==> K=k1 34 conf:(1) [Those who join face book groups also have knowledge of Security Settings, Accuracy -34 %]
2. Ae=ae1 33 ==> Af=af1 33 conf:(1) [Those who use internet for preparing projects also use internet for preparing seminars , Accuracy -33 %]
3. D=d1 Ac=ac1 33 ==> Af=af1 33 conf:(1) [Those who have active account on facebook , and download lecture notes , also use internet for preparing seminars, Accuracy -33%]
4. G=g1 Af=af1 33 ==> K=k1 33 conf:(1) [Those who joined groups in facebook , and download seminar from internet , also have knowledge of security settings of facebook Accuracy -33%]
5. K=k1 Ae=ae1 32 ==> Af=af1 32 conf:(1) [Those who download lecture notes , also use internet for preparing projects and seminars, Accuracy -32%]
6. D=d1 G=g1 31 ==> K=k1 31 conf:(1) [Those who have active account on facebook and joins groups on facebook , can have knowledge of security settings, Accuracy-31%]

VIII. Conclusions

In this paper, we have studied association rule mining over survey dataset. Our study shows that mining multiple-level association rules from databases has wide applications and efficient algorithms can be developed for discovery of interesting and strong such rules in the database The larger the set of frequent item sets the more the number of rules presented to the user, many of which are redundant.

References Références Referencias

1. Arun K. Pujari, " Data Mining Techniques", 14th impression, 2008
2. R. Agrawal, T. Imielinski, A. Swami,"Mining Association Rules Between Sets of Items in Large Databases", Proc. SIGMOD Conference, 1993.
3. Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules in large

databases", In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

4. Klementtinen M., et al "Finding interesting rules from large sets of discovered association rules." Proceedings of the CIKM 1994.
5. http://en.wikipedia.org/wiki/Apriori_algorithm

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Semantic Clustering of Genomic Documents Using Go Terms as Feature Set

By Dr.B.L.Shivakumar & V.Bhuvaneswari
Bharathiar University

Abstract - The biological databases generate huge volume of genomics and proteomics data. The sequence information is used by researches to find similarity of genes, proteins and to find other related information. The genomic sequence database consists of large number of attributes as annotations, represented for defining the sequences in Xml format. It is necessary to have proper mechanism to group the documents for information retrieval. Data mining techniques like clustering and classification methods can be used to group the documents. The objective of the paper is to analyze the set of keywords which can be represented as features for grouping the documents semantically. This paper focuses on clustering genomic documents based on both structural and content similarity .The structural similarity is found using structural path between the documents. The semantic similarity is found for the structurally similar documents. We have proposed a methodology to cluster the genomic documents using sequence attributes without using the sequence data. The sequence attributes for genomic documents are analyzed using Filter based feature selection methods to find the relevant feature set for grouping the similar documents. Based on the attribute ranking we have clustered the similar documents using All Keyword approach (KBA) and GO Terms based approach (GOTA). The experimental results of the clusters are validated for two approaches by inferring biological meaning using Gene Ontology. From the results it was inferred that all keywords based approach grouped documents based on the semantic meaning of Gene Ontology terms. The GO terms based approach grouped larger number of documents without considering any other keywords, which is semantically relevant which results in reducing the complexity of the attributes considered. We claim that using GO terms can alone be used as features set to group genomic documents with high similarity.

Keywords : *Semantic Clustering, Go Terms, Attributes, Feature Set, Xml.*

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Semantic Clustering of Genomic Documents Using Go Terms as Feature Set

Dr.B.L.Shivakumar [α] & V.Bhuvaneswari [σ]

Abstract - The biological databases generate huge volume of genomics and proteomics data. The sequence information is used by researches to find similarity of genes, proteins and to find other related information. The genomic sequence database consists of large number of attributes as annotations, represented for defining the sequences in Xml format. It is necessary to have proper mechanism to group the documents for information retrieval. Data mining techniques like clustering and classification methods can be used to group the documents. The objective of the paper is to analyze the set of keywords which can be represented as features for grouping the documents semantically. This paper focuses on clustering genomic documents based on both structural and content similarity .The structural similarity is found using structural path between the documents. The semantic similarity is found for the structurally similar documents. We have proposed a methodology to cluster the genomic documents using sequence attributes without using the sequence data. The sequence attributes for genomic documents are analyzed using Filter based feature selection methods to find the relevant feature set for grouping the similar documents. Based on the attribute ranking we have clustered the similar documents using All Keyword approach (KBA) and GO Terms based approach (GOTA). The experimental results of the clusters are validated for two approaches by inferring biological meaning using Gene Ontology. From the results it was inferred that all keywords based approach grouped documents based on the semantic meaning of Gene Ontology terms. The GO terms based approach grouped larger number of documents without considering any other keywords, which is semantically relevant which results in reducing the complexity of the attributes considered. We claim that using GO terms can alone be used as features set to group genomic documents with high similarity.

Keywords : Semantic Clustering, Go Terms, Attributes, Feature Set, Xml.

I. Introduction

Biological data sources are characterized by a very high degree of heterogeneity in terms of the type of data model used, the schema design within a given data model, as well as incompatible formats and nomenclature of values. The biological databases generate huge volumes of genomics and proteomics data after the draft of human genome sequences in 2001. The researchers use the existing sequence information to find similar patterns of genes, proteins and derive other sequence information. Each data source has custom text formats, and these formats change occasionally. Furthermore, an entire data source may be retired or completely restructured using a new schema. Some data sources are inconsistent at the semantic level, and frequently, there is inadequate use of controlled vocabularies and common data elements to specify the metadata.

The National Center for Biotechnology Information (NCBI) is one major resource that maintains public biomedical annotation databases, which are represented in different useful formats that includes XML format. The XML format of databases is very useful, because it is one of the powerful languages for representing the biological data in semi structured form and also the extraction of biological entities from XML format of databases are very easy at any extent. The content similarity measure needs distances that estimate similarity in terms of the textual content inside elements, while the structure dimension needs distances that estimate similarity in terms of the structural relationships of the elements [9].

The Genomic sequence data are stored in public databases like NCBI, Uniport in various formats. The genomic sequence data consist of large number of attributes for describing the sequences. Finding the important attributes for comparing the genomic sequence data based on annotation, becomes the challenging task. Feature selection methods can be used to analyze and study the best features used for representing sequence information for association and clustering of documents.. The complexity of clustering the documents based on the description without considering the sequence data depends on the features selected for clustering. We have analyzed and ranked the features using Filter Based Approach by using CHIR and x^2 statistics.

The Gene Ontology (GO) is one of the most important ontologies in the bioinformatics community and is developed by GO consortium. It is specifically intended in annotating gene products with consistent, controlled and structured vocabulary. The semantic similarity between the documents is determined based on its contents. Many approaches has been used to cluster biological documents based on contents. We have proposed an idea using Gene Ontology terms as a filter to group documents to get meaningful clusters and

Author α : Professor and Head, Department of Computer Applications, SNR Sons College, Coimbatore-6. E-mail : blshiva@yahoo.com
Author σ : Assistant Professor, Department of Computer Applications, Bharathair University, Coimbatore–46.
E-mail : bhuvanes_v@yahoo.com

compared the same by considering other attributes as keywords leaving GO terms using. In this paper the grouping of biological documents in Xml is done based on structural similarity followed by semantic similarity.

The paper is organized as follows: Section 2 provides the literature review of the study for clustering XML documents and Filter based Feature Selection methods. Section 3 explains the proposed methodology in detail. Section 4 discussed the experimental results of the proposed work followed by conclusion in Section 5.

II. Related Work

The background study related to the work is discussed in detail in the following section. The various approaches to find the similarity between the documents are syntactic similarity and semantic similarity. The related work based on structural and semantic similarity to cluster the documents is as follows. The structural similarity between xml documents is found using graph edit distance measure by Nierman and Jagadish. Edit distance is operations performed on a graph to transform form one form to other[12]. Raffaele has proposed an XML based approach for automatic musicological analysis[14]. Joachim and Paul have presented the use of XQuery with illustrations fop retrieving musical features in music Xml [8].

Tagarelli A and Grew has addressed the problem of clustering xml data based on structure and contents [15]. Ma & Chbili have studied the method for using same schema for finding similarity of XML data based on structure and content [10]. Thedoore and Cheng have proposed a method for clustering XML documents based on structure using tree representation [16] . Docuet A has proposed an approach for clustering homogenous xml documents based on Kmeans algorithm [6]. Panagiotis and Christos has proposed a clustering algorithm for Heterogeneous and homogenous XML using Edge summaries [130]. Nayak R has discussed clustering of heterogeneous Xml documents [11]. Bertino has given an matching algorithm for measuring structural similarities between Xml documents and DTD applications [3] . Yu-Chih and Jia has proposed an approach for extraction and clustering structural features for Music XML [19]. Wang [18] proposed a hierarchical algorithm for structural similarity which reduces the join cost for querying XML documents which is stored in relational tables..

The contents in the biological databases are represented as xml tags. Inferring information from the xml tags which have biological semantic meaning is very complex. The bioinformatics community used various ontologies to infer meaningful biological similarities across documents. The various work proposed by researchers using Go ontology for clustering are as follows: Meeta Mistry and Paul Pavlidas has proposed various content similarity measure using GO and also represented GO as flat matrix representation [9]. Catia Pesquita [5] has evaluated GO based semantic similarity measure using the relationship with sequence similarity as a means to measure based on the presence and absence of these annotations. Brendan and Sheehan [2] has proposed and idea to measure the semantic similarity based on set based and vector based approaches using GO based on conceptual level and structure level. Julie Chabalier ad Jean Mosser [7] has used vector space model for computing semantic similarity between genes using a traversal approach. Andreas Schlicker [1], Francisco has presented a new method for comparing set of GO terms and assessing the functional similarity of gene products. Gene products are said to be functionally similar if they have comparable molecular functions and are involved in a similar biological process.

Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. Bassam Al-Salemi Used Feature Selection techniques such as Mutual Information (MI), Chi-Square Statistic (CHI), Information Gain (IG), GSS Coefficient (GSS) and Odds Ratio (OR) to reduce the dimensionality of feature space by eliminating the features that are considered irrelevant for the category [4].

III. Methodology

The proposed methodology shows in Figure. 1 consists of two phases for clustering genomic sequence documents using the sequence descriptions. The first phase is the structural similarity phase where the original documents are analyzed based on structure. The filtered structurally similar documents are passed for measuring the content similarity. The second phase the features of the sequence documents are analyzed based on supervised statistical techniques and semantic grouping of documents are done. Two approaches are proposed to group similar documents based on the features. The first approach All Keyword based approach the clusters are analyzed using all the keywords. The second approach GO Terms based clusters analyze the similarity among documents using the GO as keywords. The proposed clustered approaches are validated

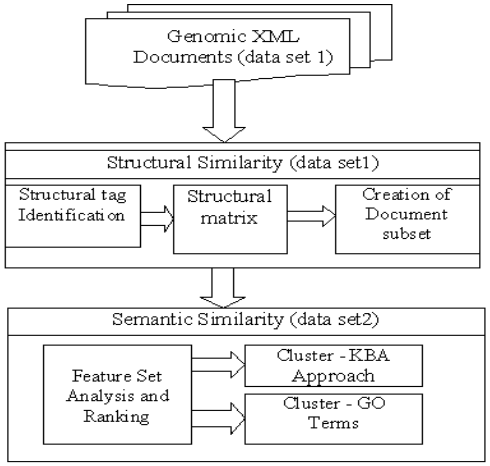


Fig. 1: Methodology

a) Phase – I Structural Similarity

The structural similarity of XML documents is based on the path of the elements given in the document. The structure of XML document is represented as a tree structure in which it is broken down into collection of distinct paths. The structural similarity is measured using the distinct paths. The sequence database maintains the sequence information as tags in Xml documents. The genomic data in XML format has more than 3500 tags to represent the functional descriptions about the sequences like accession no, taxonomy, organism, lineage, sequence title, sequence descriptions, alternate name, gene name, author details, and identifiers related to other databases like GO, KEGG, PUBMED. To measure the structural similarity between the documents the structural matrix is constructed, in which each document is checked for the below said tags, where there is possibility of more than one occurrence of a particular tag. The total count of occurrence of each tag is entered into the matrix, in absence of a tag value zero is entered into the corresponding place. The content similarities of documents are analyzed only for the documents that are structurally similar. The proposed work the dataset contains sequence attributes for both *E.Coli* and *Human* organism.

b) Phase Ii Semantic Similarity

i. *Dataset*

In our experiment the public database downloaded from NCBI for *E.Coli, human sequencec* in xml format is used. The NCBI dataset is the integrated, text-based search and retrieval system used at the major databases, including PubMed, Nucleotide and Protein Sequences, ProteinStructures, Complete Genomes, Taxanomy, and others. The GO Ontology recent download 2010 was used to verify the clusters generated based on the functionality of genes described in the second approach. We have extracted 150 documents from the databases for the organisms and stored in db2 for further extraction and querying. The work is implemented using two softwares. The Xml preprocessing and extraction is carried out using DB2 an IBM Product using XQuery language and we have linked with.NET Framework using COM.

ii. *Feature Set Identification*

Filter Based approaches supervised methods like X^2 statistics, CHIR statistics are used for analyzing the features of the xml document for the proposed work . The X^2 Statistics is used to measure the independence between the keyword and the category[4]. This can be done by comparing observed frequency in the 2-way contingency table with the expected frequency when they are assumed to be independent. CHIR is a supervised learning algorithm based on X^2statistics, which determines the dependency between a keyword and a category and also the type of dependency [17]. Type of dependency indicates whether the feature is a positive or negative dependency for the category. The Features are Ranked based on X^2_{max}, X^2_{avg} and rx^2statistics. The highly ranked Features are used for analyzing the term relevance. The Feature sets are identified based on ranking

The documents are initially clustered for analyzing the features using hierarchical clustering algorithm for assigning class labels. The proposed work we have considered 150 documents with 358 extracted keywords. On clustering the 150 documents 30 clusters are generated. From the generated cluster it is found that single document is found in many clusters and maximum documents are found in 9 clusters. So we have taken the cluster which contains highest number of documents to analyze the feature attributes and find the term relevance using filter based approach. Among 358 keywords retrieved we have identified three feature set based on the ranking with 156, 77 and 58 keywords respectively.. The feature set identified are considered for grouping the documents based on its contents.

iii. *Semantic Similarity*

The content similarity is the main task involved in document clustering, in which the important terms from the documents that differentiate the documents are identified. The term matrix (vector space model) is constructed for the documents which are structurally similar. Consider there are n number of documents in a data set D, that are denoted by d_1, d_2, d_3, …., d_n and the distinct terms from the above document are denoted by t_1, t_2, t_3, ….., t_m. Then the term matrix of size $n \times m$, where n is the number of documents in dataset and m is the number of distinct terms appeared in the data set D, is constructed..

Two different clustering approaches namely All keywords based approach and GO Terms based approach are proposed for clustering similar documents based on the sequence annotations. All keyword based

approach the feature set extracted from the documents are represented as keywords and the term matrix is generated. The documents are clustered using the existing similarity measures like Euclidean, jaccard and cosine . In GO Term based cluster approach the GO terms alone from the feature set are extracted and the mapping is done to find the corresponding genes for the GO terms and viceversa using GO ontology. The term matrix is constructed for the genes and GO terms and the documents are clustered.

c) All Keywords Based Approach -Kba

The feature set with 156 keywords and 77 keywords is used as dataset and the feature matrix is constructed. Some biological keywords like *Alternate name, Go terms, Gene name, Sp_block keywords and Ecnumber* are ranked high in the feature set identification , which has a positive dependency for the clusters generated. The selected feature set attributes were analyzed with respect to the document by varying the no of attributes and clusters were generated. In order to get high degree of cohesion in documents in each clusters we used kernel approach [10], in which the documents in each clusters have high degree of similarity. Kernel is the count of the individual unique keywords from the term matrix greater than a particular threshold. The kernels were created for values starting from 30 and varying it up to 55.The clusters were generated by varying the kernel to find the similarity among attributes. The clusters generated for the kernel values are shown in Figure 3.

d) Go Terms Based Approach - Gota

The proposed idea of our work choosing GO terms as keywords for clustering documents is based on the idea that Documents are said to be semantically based on the gene products. Gene products are said to be functionally similar if they have comparable molecular functions and are involved in similar biological process. GO annotation capture the available information of genes and used as a basis for defining a measure of functional similarity between genes which is used in our second approach to group documents based on semantic similarity. Each gene is related with more than one GO terms.

A Vector Space Model(VSM) is used to compute similarity between pair of gene products .VSM are essentially used in information retrieval for computing the similarity between documents described as vectors of Keywords[2007]. We have used the same model for our second approach to find associative relations between the terms in the GO. To compute the similarity between documents the gene products are described as vectors of GO terms. A gene product is represented by a specific vector g as follows: $g=(t_1,t_2....t_n)$ where t_i is the numeric value that the term takes on for the gene product and n is the number of go terms associated with the gene products. A value $t_i = 0$ means when there exists no association between GO terms and genes. The existing similarity measures are used to cluster the documents.

All the go terms and gene names are extracted from the feature set and a mapping is done with existing go terms and genes using the bioinformatics famous Gene Ontology recently downloaded. The term matrix is constructed representing genes as rows and GO terms as columns for the clustering phase .In the proposed approach we have included only the Go terms as features for clustering leaving other attributes from the documents. We have compared the two approaches and results are discussed in section 4.

IV. Results and Discussion

The dataset with 150 documents is given as input for the first phase of clustering to extract documents that are structurally similar which is heterogeneous containing information for two organism *E.Coli* and *human*. The structural path is used to analyze the structural similarity. Structurally similar 107 documents were retrieved based on the approach, which is given as input for the content similarity phase.

The various feature sets identified using Filter based on the r anking of rx^{2v} statistics is shown in Figure 2. We have considered the feature set with count of keywords with high , low and average 156, 77, and 56 respectively for the proposed study.. The identified feature set is passed for finding the semantic similarity using All keyword based and GO Term based approach.

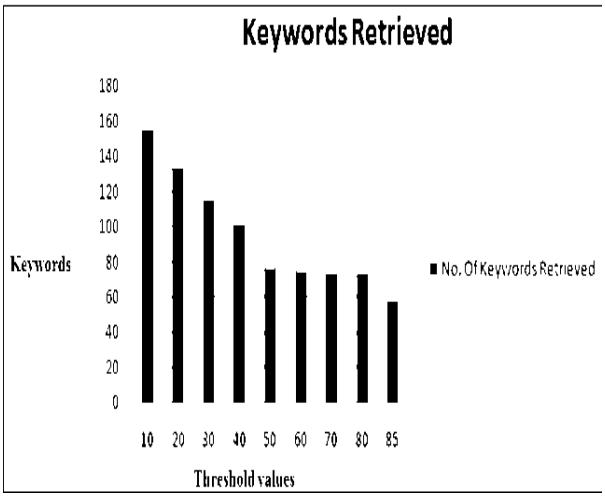


Figure 2: Identified Feature Set

a) Keyword Based Approach - Kba

The clustering result of all KBA for kernel values 30, 45, and 55 for 30 clusters is shown in Figure 3.

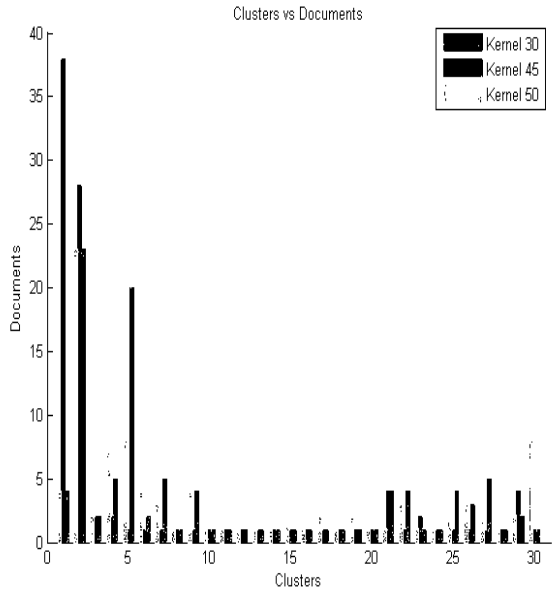


Figure 3: Clsuter using Keyword Based Approach

The grouping of documents is based on the 156 keywords which are functionally related with each other. The snapshot of the document grouped in some clusters for a set of keywords for the above kernals is shown in Table 1 and Table 2.

Table 1: Documents clustered for Kernels

Cluster No	Kernal 50	Kernal 45	Kernal 30
	Document ID		
1	1,36,46,54	1,24,25,34,35,3,38,39,44,46,47 ,49,50,53,54,57,58,59,60,63,94	1,36,54,70
2	2,3,4,7,9,13,14,15,16,17,18,19,20,78, 80,92,93,95,96,97,99,100,101	2,3,4,5,6,7,8,9,10,12,13,14,15, 16,17,18,19,20,90,91,92,93,95, 96,97,98,99,100,101,102,103	2,3,4,7,9,13,14,1516,17,18,19,20,7 8,80,92,93,95,96,97,99,100,101
5	5,10,77,79,81,90,91,98	45	5,6,8,10,12,77,79,81,82,83,85,87,8 8,89,90,91,98,102,103
10	43	74	43
15	52,66,67,74,75	78	48
18	57	81	55
19	65,69	82	56

Table 2: Key Terms

Kernal 50	Kernal 45	Kernal 30
ATP-binding	Complete proteome	Oxidoreductase
1.1.1.-	GO:0055114	
Cytoplasm	Kinase	
GO:0005524	Nucleotide-binding	
GO:0005737	Allosteric enzyme	
GO:0055114	Transferase	
Metal-binding		
NAD		
Nucleotide-binding		
1.6.5.3		
Polymorphism		
Transferase		

It is found that same documents are found in clusters for the kernel values 30 and 50. It is also found the grouping of documents for clusters {10, 15, 19, 19} are different and contains only one document. In order to asses the semantic meaning of the clusters formed biologically we have analyzed the terms related to the clustered documents. We have inferred from our results that the terms that grouped the documents are biologically associated with each other. The terms responsible for one functionality had other related associated terms. The document with keyword cytoplasm had its associated terms like nucleus, cytoskeleton which are called as cellular components which is associated with a gene name, and Go number. The documents with term oxidation reduction had related terms like fatty acid metabolism, biosynthetic receptor which is responsible for biological activity. The terms like Aledhydde dehydrogenase is associated with keywords like lipid binding, protein binding etc. The above inference of our results motivated us to go for the second approach proposed to cluster documents based on the Go terms and genes which is used by many researchers for gene clustering. The results of our second approach are briefed below

b) Go Term Based Approach -Gota

Clustering documents based on functionality of the genes using GO terms is proposed in the second approach with the same dataset. The gene names and the corresponding go terms are extracted for the documents which are structurally similar. A total of 71 gene names and 238 go terms were extracted from the dataset and stored in structure for further processing. On implementation of the clustering algorithm the documents were grouped into 30 clusters. The number of documents grouped in each cluster is given in Figure 4.

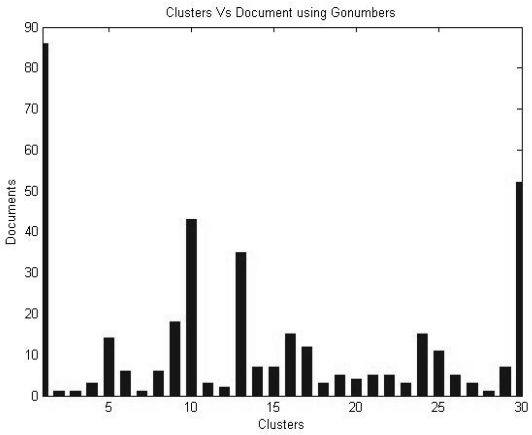


Figure 4: Clusters using GO Terms

Cluster 1 grouped maximum of 86 documents with 178 GO Terms and 55 related genes. Some of the documents were overlapped in other clusters to.

c) Validating Proposed Approaches

The two approaches Keyword Based and GO Terms based cluster results are shown in Figure 5. The go term 55114 grouped 54 documents which is responsible for the biological activity the oxidation reduction and GO Term 5737 is responsible for cellular activity in cytoplasm. The term 5488 is responsible for Molecular function for binding. The clusters with Go Terms are highly semantically relevant based on functionality than keyword based approach. Some of the documents were found to be overlapped because the functionality of one process inhibits the other. The goterms and its associated genes are functionally related to a process which can be found in the Go Taxanomny. From the results we state that the GO annotations is remarkably useful for grouping documents based on the functionality rather than using the conventional methods

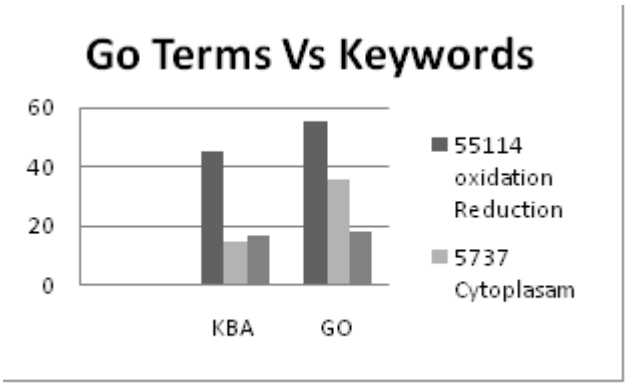


Figure 5: Validating Clusters for KBO and GOTA

The GO Terms and its associated genes are functionally related to a process which can be found in the Go Taxanomy. From the results we state that the GO annotations is remarkably useful for grouping documents based on the functionality rather than Considering other features. The experimental results it is found that the GO terms 55114 grouped larger documents in the second approach which is responsible for oxidation reduction. The same keyword grouped documents for kernel 45 in all KBA. The documents also found distributed in the remaining clusters of the first approach , based on the specific keywords , However the biological inference of both approaches are similar , based on the literature.

V. Conclusion

This paper presents an approach to cluster xml genomic documents using both syntactic and semantic approaches. The structural similarity of documents is done based on the path similarity as in xml documents all information is maintained in tags. The dataset used in the work contains heterogeneous documents with different structural tags for different taxonomies. The structurally similar documents filtered are analysed for Features set Identification using Filter Based approach. The attributes were statistically analyzed and identified three best feature sets. The feature set is used for grouping documents using two proposed approaches Keyword Based Approach and Go Term based approach. The two approaches are compared for their biological relevance. The experimental results it was found that GO Term based clusters documents based on functionality and the terms are related with keywords. Finally, we conclude that using the GO annotations as feature set is efficient to cluster documents which also reduce the dimension of the datasets.

Acknowledgment

This work was performed as part of the Minor Research Project, which is supported and funded by University Grants Commission, New Delhi, India.

References Références Referencias

1. Andreas Schlicker, Francisco S Domingues et.al ,"A new measure for functional similarit of gene based on Gene Ontology", *BMC Bioinformatics*, June 2006
2. Brendan Sheehan ,Aaron Quigley et.al , "A relational based measure of semantic similarity for gene ontology", *BMC Bioinformatics*, Nov. 2008
3. Bertino, Elisa, Giovanna Guerrini, and Marco Mesoti, "A Matching Algorithm for Measuring the Structural Similarity between An XML Document and A DTD and its Applications," *Information Systems*, Vol. 29, No. 1, March 2004.
4. Bassam Al-Salemi, Mohd. Juzaiddin Ab Aziz, "Statistical Bayesian Learning for Automatic Arabic Text Categorization", In Journal of Computer Science 7 (1): 39-45, 2011.

5. Catia Pesquita , Daniel Faria, "Metrics for Go based protein semantic similarity: a systematic evaluation", *BMC Bioinformatics*, April 2008.
6. Doucet, A. and Ahonen-Myka, H. "Naïve Clustering of a large XML Document Collection". In P*roceedings of the 2002 Initiative for the Evaluation of XML Retrieval Workshop (INEX '02)*, 2002, pp. 81-87.
7. Julie Chabalier, Jean Mosser et.al,"A transversal approach to predict gene product networks from ontology-based similarity", *BMC Bioinformatics*, July 2007.
8. Joachim Ganseman , Paul Scheunders and Wim D'haes "Using XQuery on Musical Databases for Musicological Analysis" *In Proceedings of ISMIR 2008 – Data Exhange , Archiving and Evaluation.*
9. Meeta Mistry, Paul Pavlidis 'Gene Ontology term overlap as a measure of gene functional similarity', *BMC Bioinformatics*, Aug 2008.
10. Ma, Y. and Chbeir, R. Content and Structure Based Approach for XML Similarity. In *Proceedings of the 2005 Conference on Instructional Technologies (CIT '05)* (Binghamton, Canada, 2005), 2005, pp. 136-140
11. Nayak, R. and Xu, S. XCLS: A Fast and Effective Clustering Algorithm for Heterogeneous XML Documents. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '06)* (The Singapore, April 9-12, 2006). 2006, pp. 292-302
12. Nierman A and Jagadasih H.V "Evaluating structural similarity in XML documents", *In proceedings of the WebDB Workshop,*Madison, June 2002.
13. Panagiotis Antonellis , Christos Makris , Nikos Tsirakis , " XEdge : Clustering Homogenous and Heterogeneous XML Documents using Edge Summaries", *ACM* , March 16-20, 2008,Fortaleza, Brazil.
14. Raffaele Viglianti, "MusicXML: An XML based approach to automatic musicological analysis," in *Conference Abstracts of the Digital Humanities 2007 conference*, Urbana-Champaign, Illinois, USA, Jun. 4-8 2007, pp. 235–237.'
15. Tagarelli, A. and Greco, S. toward Semantic XML Clustering. In *Proceedings of the 2006 Siam Conference on Data Mining (SDM '06)* (Maryland, USA, 2006). 2006, pp. 188-199.
16. Theodore Dalamagas , Tao Cheng et.al " Clustering XML Documents by Structure"
17. Tao Liu, Shengping Liu, Zheng Chen, "An Evaluation on Feature Selection for Text Clustering", Proceedings of the Twentieth International Conference On Machine Learning(ICML-2003)."
18. Wang, Lian, David Wai-lok Cheung, Nikos Mamoulis, and Siu-Ming Yiu., "An Efficient and Scalable Algorithm for Clustering XML Documents by Structure," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 1, January 2004
19. Yu-Chih Shen , Jia-Lein Hsu , Shuk-Chun Chung, "MF Tree: Extracting and Clustering the Structural Features from Music Object ib MusicXML".

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Application of Information Technology in Consumer Indexing through Geographical Information System for Power Utilities

By Jalpesh Solanki & Utkarsh Seetha
Jodhpur National University

Abstract - DC, DR, MSSQL, SOA, Microsoft BizTalk, CRM, VRF, NLD, MPLS.

GJCST-C Classification: *H.3.1*

Strictly as per the compliance and regulations of:

Application of Information Technology in Consumer Indexing through Geographical Information System for Power Utilities

Jalpesh Solanki [α] & Utkarsh Seetha [σ]

Abstract - DC, DR, MSSQL, SOA, Microsoft BizTalk, CRM, VRF, NLD, MPLS.

I. Introduction

For operations as complex as described above, the system must be designed in a way that it stands to the rigorous needs of Rajasthan Distribution Companies. As part of our approach this research takes into consideration certain design aspects of the system with reference to which the system is conceptualized and developed to cater the need for GIS and its uses.

Design considerations are the parameters that have been used as basic philosophy to begin the thought process for evolving a solutions approach. There are four basic considerations based on which any IT system is designed. Based on these considerations GIS and other applications will work upon.

- Reliability
- Availability
- Scalability
- Security

Taking these considerations as fundamental, this research has assured their presence in each of the system building block.

- Application Software
- DC
- DR
- Network
- Hardware
- Services

Based on the design considerations mainly SOA need to have a robust integration engine, Microsoft BizTalk Server 2009 is being proposed as the integration middleware and MSSQL Server 2008 as database layer.

Moreover the utilities want the solution to be designed in such a way that in future it is possible to segregate one or more of the entities at the application and database level. Moreover the segregation mechanism should be flexible and at no cost to the DISCOMs. This therefore calls for an innovative approach on part of the ITIA to deliver such a solution at a competitive price.

Author α : Research Scholar Jodhpur National University (Faculty of Computer Application).
Author σ : Restructured Power Development and Reforms Programme.

The research specifies that no Separate instances are acceptable for each individual DISCOMs as per PFC guidelines but it is possible to create Logical partitioning for the individual DISCOMs in the same set of common servers. Therefore system design research would use a single set of application & a single set of database servers and will create logical partitions for different DISCOMs so as to fulfill the stipulation.

The global competition and swiftness of changes emphasize the importance of human capital within organizations, as well as the swiftness and ways of knowledge gaining of that capital. In the economy where uncertainty is the only certainty, knowledge is becoming a reliable source of sustained competitive advantage. Knowledge is becoming basic capital and the trigger of development. Previously built on foundations of possessing specific resources and low costs, present day competition is based on knowledge possessing and efficient knowledge management. Modern organizations therefore use their resources (money, time, energy, information, etc.) for permanent training and advancement of their employees. Organizations which are constantly creating new knowledge, extending it through the entire organization and implementing it quickly inside the new technologies, develop good products and excellent services.

II. Geographical Information System (GIS)

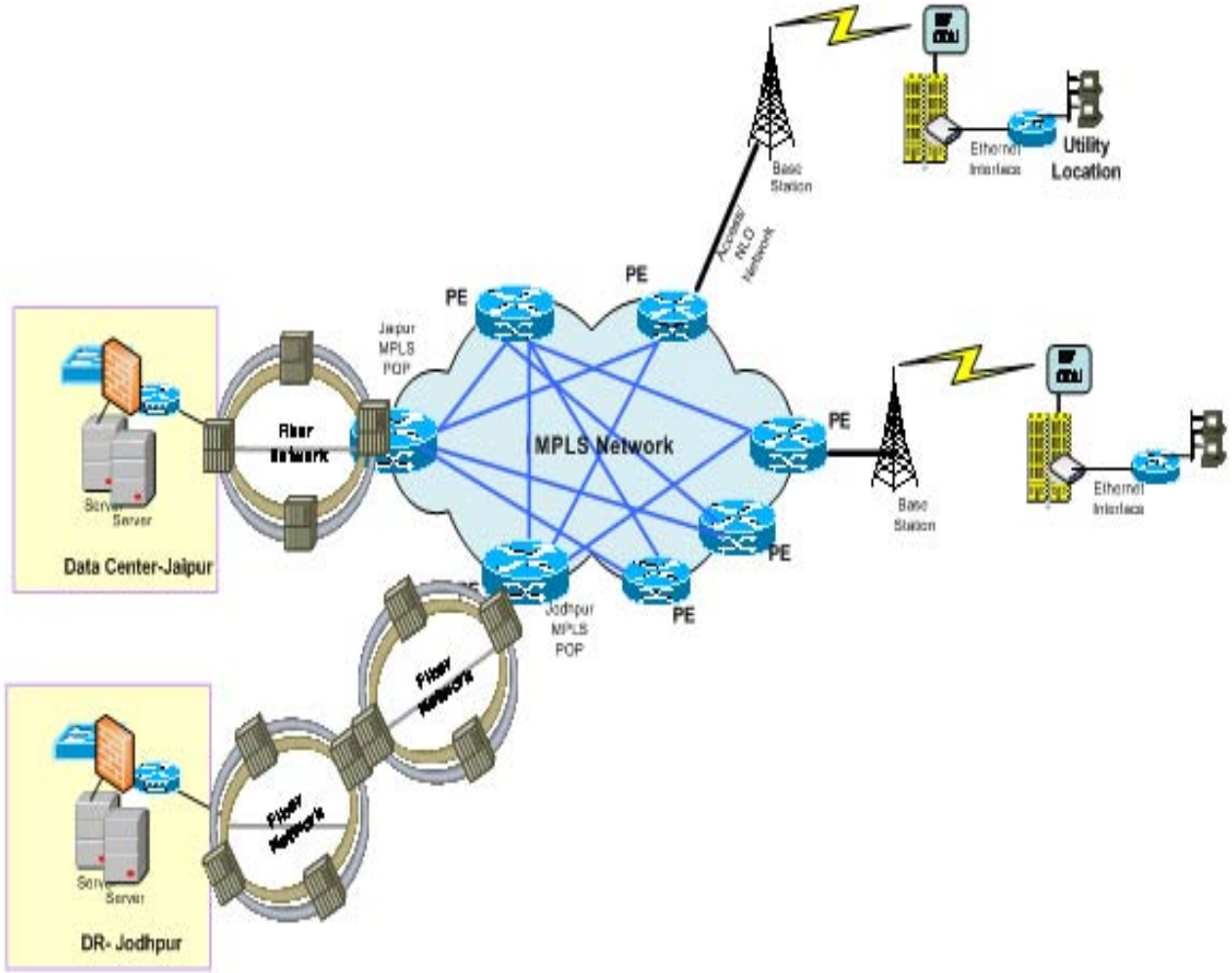
In order to obtain a new connection or other CRM workflows, the business workflow user needs to visit GIS link in systems Integrated Home Page Or CRM application where he/she wish to provide site verification. Following different roles can open a GIS map.

i. Business Work Flow User (To create a new connection.)
ii. Technical User
iii. General Viewer
iv. CCC Viewer

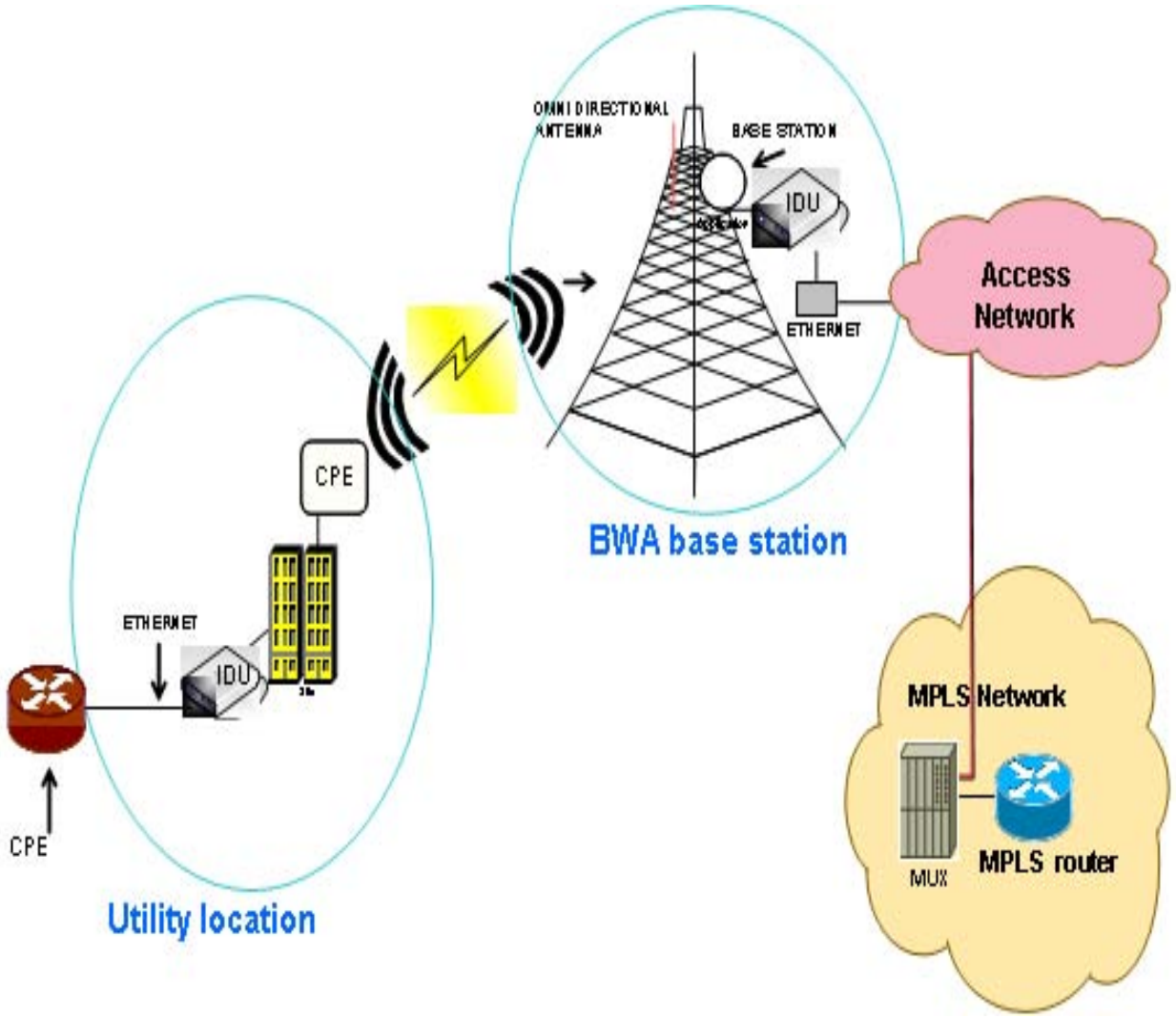
a) Remote (Utility location) Connectivity to use GIS applications at site

➢ Service Provider(s) selected by SI will connect all the 834 locations on RF last mile or any other feasible technology.

- Service Provider(s) selected for this project already have a presence in terms of Base Stations from where last mile will be extended.
- All the RF or equivalent last miles will be backhauled on Service Provider's Access or NLD network to the nearest MPLS POP in Rajasthan
- All the locations will become part of specific VRF that will have route towards Data center and also to DR in case of failure of DC.
- Service Provider will provide the RF ODU and IDU at all the locations
- Interface of IDU will be Ethernet and which will terminate to router.



Considering the geographical constraints maximum locations in Rajasthan are proposed on RF connectivity. As shown in the diagram, Service Provider can reach out to any site via RF medium. Using the existing BTS coverage Service Provider shall deploy the P2P/ P2MP radio at the locations to terminate the circuit at the nearest BTS. The BTS is connected with BCS, who in turn is connected with MSC. To carry the intercity traffic, MSC is connected with SDH backbone.



b) Secondary Link Connectivity plan to use GIS and other related applications at site

Option – 1:- VSAT Connectivity

- Service Provider will provide secondary link on VSAT last mile which is a different technology.
- Service Provider will implement VSAT at all the 834 locations.
- VSAT will terminate on Ethernet Interface of Bidder's router.
- On utility router a dynamic (BGP) protocol will be configured for Primary Path (MPLS on RF last mile) and static protocol with 100 metric will be configured for secondary path
- Primary link will be on Priority and traffic will be always routed to Primary link.
- On failure of Primary link router will route the traffic to secondary VSAT link.
- When Primary link will be restored, traffic will again be routed to Primary link.

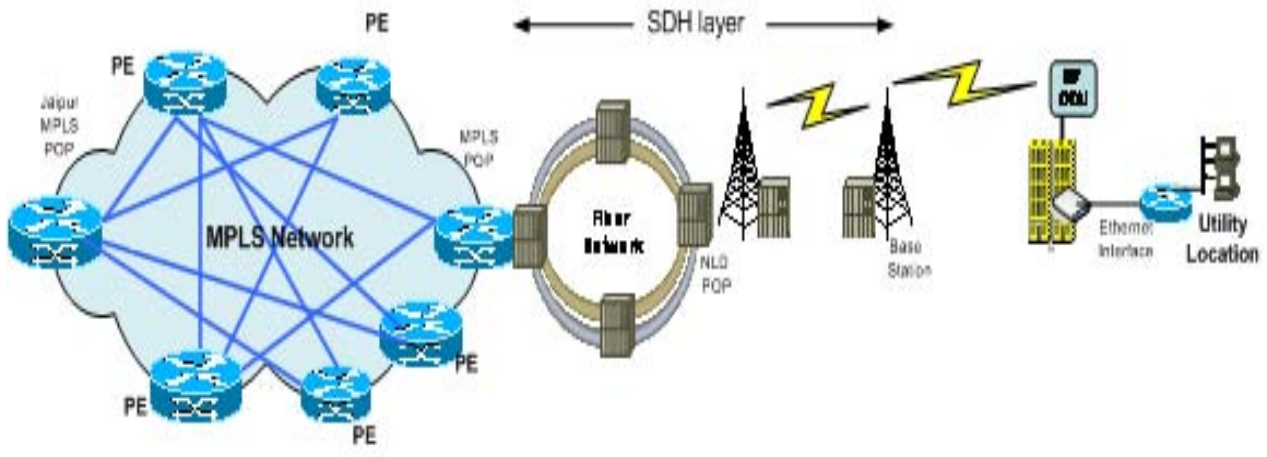
The network will be configured on Ku Band shared hub of Service Provider. The network is proposed on MFTDMA star technology. The network will cover a total of 834 locations for secondary connectivity. Jaipur will be configured as the central site for the network. This site will have backhaul connectivity to Service Provider's VSAT NoC. Disaster recovery site for the network will be setup at Jodhpur. The topology for the network is shown in the figure below

i. *Transmission Security on WAN Network*

MPLS VPN provides layer 3 connectivity throughout the network in a secure manner. A single circuit provides the needed connectivity for all sites. Each customer's routing information is kept securely separate from every other customer's routing information through the use of a route distinguisher (RD) that is unique to a particular customer. The use of the RD allows the provider to give each customer a logically separate PE router. PE routers will remain a shared resource unless otherwise negotiated. The customer routing information is maintained by a specific routing protocol instance tied to its RD. The routing table assembled by this routing protocol instance is known as a virtual routing and forwarding (VRF) table. In essence, it is simply an extension of the customer's routing table,

because it includes all of the customer's advertised prefixes and hence it is inherently secure.

Service Provider will ensure the security of the traffic from CE Router to PE Router even if they not directly connected as in case of non availability of MPLS POP. In case of non availability of MPLS POP traffic will be on SDH layer and which is highly secure.

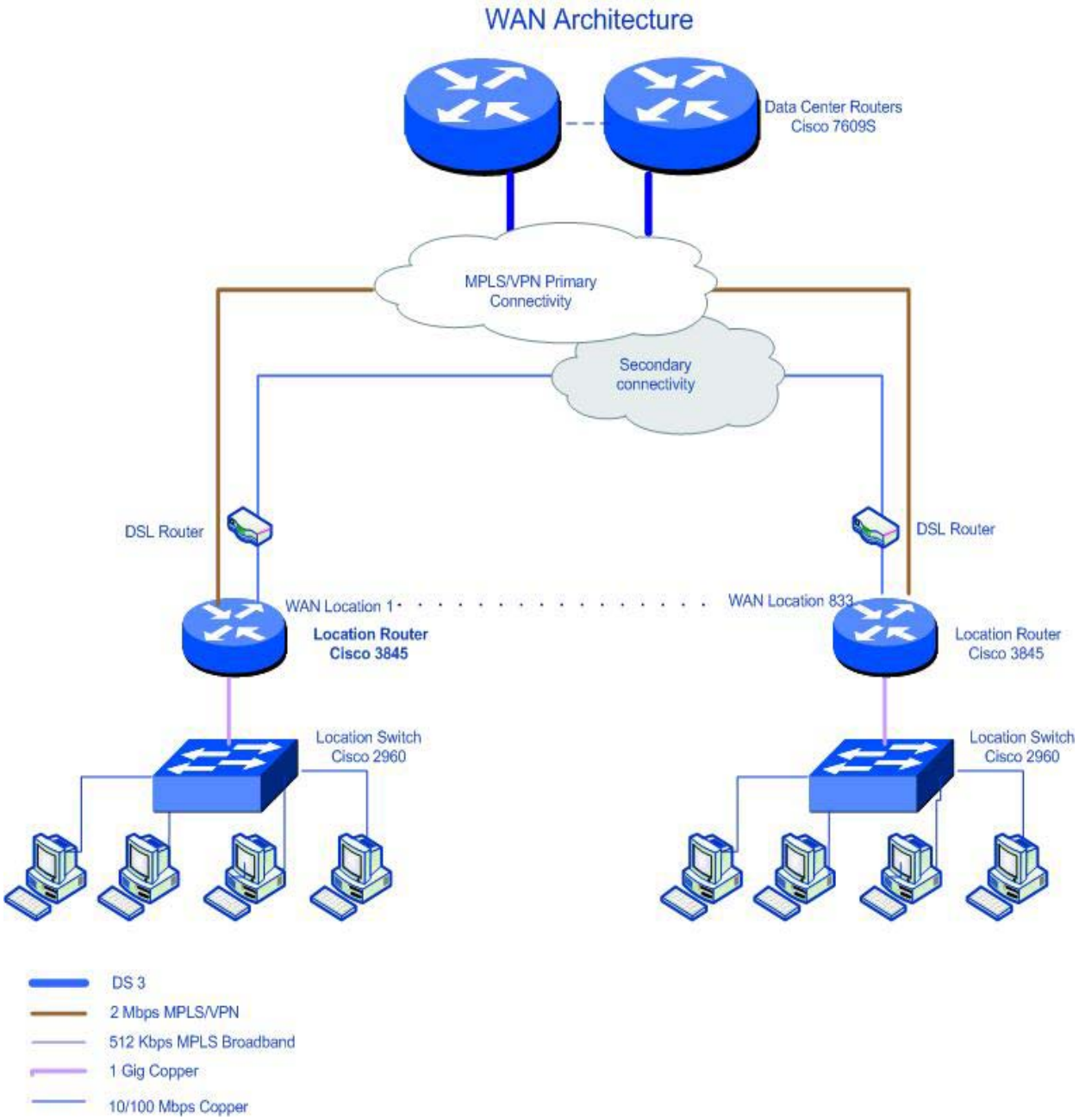


WAN Architecture Diagram to use GIS application at site

c) WAN Connectivity Requirements

The WAN Connectivity requirements as specified in the SRS and RFP are as stated below –

- Provide MPLS VPN connectivity for 834 of locations to Data Centre in Hub and Spoke topology
- All the applications will be hosted in Centralized Data Centre at State Level and which is required to be connected to all the locations.
- High Availability connectivity solution for Data Centre and DR
- The Data Centre shall have facilities for connecting to Utility HQ, all the remote utility offices in Circles, divisions, Sub divisions etc. as per the requirement of utility and all the Customer care Centers
- The Sub divisional offices and Other Offices would be connected to the Data centre through minimum 2 Mbps VPN connectivity
- The Customer care centers would also be connected to the Data centers through a minimum 2 Mbps VPN connectivity
- The Internet connectivity at the Datacenter will be 20Mbps primary and 5Mbps Backup from different Service Providers. At the DR Site, 2Mbps Internet connectivity will be provided. It will be terminated on separate Internet router.
- WAN link of 5 Mbps is required between DC and DR for replication of Data.



References Références Referencias

1. R-APDRP project, Power Finance Corporation of India 2009-2010
2. Ministry of Power, Government of India 2009-2010
3. Jaipur Vidyut Vitiran Nigam Limited, Jaipur 2009 – 2010
4. Ajmer Vidyut Vitiran Nigam Limited, Ajmer 2009 – 2010
5. Jodhpur Vidyut Vitiran Nigam Limited, Jodhpur 2009 – 2010
6. ITIA (Information Technology Implementation Agency), Restructured – Accelerated Power Development and Reform Programme (R-APDRP), Rajasthan 2009 – 2010.

Global Journals Inc. (US) Guidelines Handbook 2012

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

- 'FARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FARSC" can be added to name in the following manner. eg. **Dr. John E. Hall, Ph.D., FARSC or William Walldroff Ph. D., M.S., FARSC**
- Being FARSC is a respectful honor. It authenticates your research activities. After becoming FARSC, you can use 'FARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 60% Discount will be provided to FARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- FARSC will be given a renowned, secure, free professional email address with 100 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- FARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 15% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- Eg. If we had taken 420 USD from author, we can send 63 USD to your account.
- FARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- After you are FARSC. You can send us scanned copy of all of your documents. We will verify, grade and certify them within a month. It will be based on your academic records, quality of research papers published by you, and 50 more criteria. This is beneficial for your job interviews as recruiting organization need not just rely on you for authenticity and your unknown qualities, you would have authentic ranks of all of your documents. Our scale is unique worldwide.
- FARSC member can proceed to get benefits of free research podcasting in Global Research Radio with their research documents, slides and online movies.
- After your publication anywhere in the world, you can upload you research paper with your recorded voice or you can use our professional RJs to record your paper their voice. We can also stream your conference videos and display your slides online.
- FARSC will be eligible for free application of Standardization of their Researches by Open Scientific Standards. Standardization is next step and level after publishing in a journal. A team of research and professional will work with you to take your research to its next level, which is worldwide open standardization.

- FARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), FARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 80% of its earning by Global Journals Inc. (US) will be transferred to FARSC member's bank account after certain threshold balance. There is no time limit for collection. FARSC member can decide its price and we can help in decision.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

- 'MARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'MARSC" can be added to name in the following manner. eg. Dr. John E. Hall, Ph.D., MARSC or William Walldroff Ph. D., M.S., MARSC
- Being MARSC is a respectful honor. It authenticates your research activities. After becoming MARSC, you can use 'MARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 40% Discount will be provided to MARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- MARSC will be given a renowned, secure, free professional email address with 30 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- MARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 10% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- MARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- MARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), MARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 40% of its earning by Global Journals Inc. (US) will be transferred to MARSC member's bank account after certain threshold balance. There is no time limit for collection. MARSC member can decide its price and we can help in decision.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJMBR for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5.STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. **Pick a good study spot:** To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. **Think technically:** Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be

sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to

shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic

principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

T

U

V

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350