Improving the Performance of Semantic Search by Categorizing the Semantic Web by integrating different Processes for tourism domain

P. Jayaprabha[1], Dr.N.K.Kuppuchamy[2], P. J. Shriidhar[3], Dr. A. SenthilKumar[4]

[1]HOD/MCA, [2]HOD/CSE, [3]Research Scholar, [4]ASP/CS Department
[1,2]Vidyaa Vikas College of Engineering & Technology, Tiruchengode, Tamilnadu, India – 673601
[3]Bharathiyar University, [4]Arignar Anna government arts college namakkal.

***Abstract*: Semantic Web is the dream of the Tim Burner's Lee who invented World Wide Web. This has been achieved in today's world to some extent. Now-a-days, most of the web sites are developed based on semantics. The advent of semantic web overcomes the difficulty of the search engine, it provides better result due to the information in the web are structured, machine readable and it is very easy to integrate. Semantic web without categorization provides poor result when compare to semantic web with categorization. In order to improve the performance of search engine, Categorization of semantic web is done. This Semantic web categorization enhances the result of user query. The proposed approach is to categorize the semantic web based on four processes and promoting the efficiency of information retrieval. The processes can extract both the explicit relations extracted directly from the ontologies in a traditional way and the potential relations inferred from existing ontologies by extracting important class names, using WordNet relations and detecting the methods of describing the Web resources. It can achieve a notable improvement in categorizing the valuable Web resources based on incomplete ontologies for a Semantic Web search system.**

***Keywords*: Ontology, RDF, Relations, Category, Classnames, WordNet, Pattern analysis.**

1.0 Introduction

Recently, the Semantic Web increases in size and contains variety of resources. The web users can make efficient use of web resources that are machine readable and understandable in Semantic Web due to semantics. Providing categorization to the semantic web improves the performance of the search engine. Many works have been done for categorizing semantic web resources. But all these works are done for well defined (or) well structured ontology. As the World Wide Web changes dynamically, it's impossible to structure a perfect ontology including the relations of all the web resources. In Semantic Web, a class is a type of web resources. A class has some properties and it can be a subclass of another class. Thus, it cannot be matched by the ontology-based search engines for a web resource which is not defined as a class in the structured ontologies. To use web resources efficiently, it becomes an important and emerging issue to detect the potential relations on existing ontologies. In order to overcome the above difficulty, the proposed approach extracts potential relations among the web resources by using the knowledge such as the class hierarchy in ontologies, the relations of word in dictionaries and the description of web resources. It can break through the limitation of incomplete ontology. In other words, the web resources related to the categories can be categorized by using both the explicit relation and the detected relations.

2.0 Related Work:

Keyword used in search queries often has multiple meanings. Consequently, search results corresponding to different meanings may be retrieved, making identifying relevant results inconvenient and time-consuming. Grouping of the search results based on the different meanings of the query by utilizing the semantic dictionary WordNet to determine the basic meanings or senses of each query term and similar senses are merged to improve grouping quality. Grouping algorithm employs a combination of categorization and clustering techniques and achieves high grouping accuracy [1].

The next generation of Tourism Information System is expected to be: enable semantics-based information processing, exhibit natural language capabilities, facilitate inter-organization exchange of information in a seamless way and evolve proactively in compete with dynamic user requirements [3], [4].

The OntoWebber system [4] is an ontology-based approach to website management. It facilitates the design, creation, generation and maintenance of Web sites using a set of software tools. It also enables the personalization of Web site views based on individual users. Another notable approach is the Hera project, which is a methodology that supports the design and engineering of Web Information Systems (WIS) using Semantic Web technology. The main focus of the Hera project is to support Web design and implementation particularly hypermedia aspects.

Different processes for categorizing semantic web resources. The processes can extract both the explicit relations extracted directly from the ontologies in a traditional way and the potential relations inferred from existing ontologies by focusing on some new challenges such as extracting important class names, using WordNet relations and detecting the methods of describing the Web resources [9].

3.0 Technologies used

Semantic web

The Semantic Web is used for representing information in the World Wide Web in a machine-readable fashion: such that it can be used by machines not just for display purposes, but for automation, integration and reuse across applications.

These machine-interpretable descriptions allow more intelligent software systems to be written, automating the analysis and exploitation of web-based information. Software agents will be able to create automatically new services from already published services, with potentially huge implications for models of e-Business.

The Semantic Web functions because of a highly specialized type of data and information tagging that can be implanted within Web pages.

Fundamentally, each Semantic Web tag, known as a triplet, links together a subject, verb, and object, creating a relationship between them. Three simple examples of a semantic tag might be: <Boston> <is in the state of> <Massachusetts>; <Beacon Hill> <is a neighborhood of> <Boston>; <I> <like> <Boston>. These two statements each consist of two nouns separated by a verb and are interlinked with one another through Boston.

Semantic web uses many ontology languages to describe semantic data. Some of the ontology languages are follows

- RDF (Resource Description Framework)
- OWL (Web Ontology Language)
- DAML (DARPA Agent Markup Language)
- SPARQL (Simple Protocol and RDF Query Language)
- GRDLL (Gleaning Resource Descriptions from Dialects of Languages)
- OIL (Ontology Inference Layer)

Ontology

Information integration from different sources needs to be a shared by understanding of the relevant domain. Knowledge representation formalisms provide structures for organizing this knowledge, but provide no mechanisms for sharing it.

Ontologies provide a common vocabulary to support sharing and reuse of knowledge. Ontology is a fundamental component for achieving the Semantic Web. Ontology has the capability to solve a number of problems in tourism. This includes: 1) enabling interoperability of heterogeneous platforms; 2) standardization of business models, business processes, and knowledge architectures; and 3) serving as a model of knowledge representation for the generation of knowledge-based information services [2].

The Resource Description Framework

RDF provides a means for adding semantics to a document without making any assumptions about the structure of the document. It is an XML application customized for adding Meta information to Web documents.

The Resource Description Framework attempts to address XML's semantic limitations. It presents a simple model that can be used to represent any kind of data. This data model consists of nodes connected by labeled arcs, where the nodes represent web resources and the arcs represent properties of these resources. It should be noted that this model is essentially a semantic network, although unlike many semantic networks, it does not provide inheritance.

The data model of RDF provides three object types: resources, property types, and statements.

- **A resource** is an entity that can be referred to by a address at the WWW (i.e., by an URI). Resources are the elements that are described by RDF statements.
- **A property** defines a binary relation between resources and/or atomic values provided by primitive data type definitions in XML.
- **A statement** specifies for a resource a value for a property. That is, statements provide the actual characterizations of the Web documents.

The Semantic Web is a web of data. There is lots of data, all use every day, and it's not part of the web.

WordNet

WordNet is a heavily-used lexical resource in natural-language processing and information retrieval. More recently, it has been adopted in Semantic Web research community. It is mainly used for annotation and retrieval in different domains.

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called synsets and each expressing different concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The meaningfully related words and concepts can be navigated with the internet.

WordNet resembles a thesaurus, in that group words together based on their meaning. WordNet interlinks not just word forms, strings of letters, but specific sense of words. Words that are found close to another in the network are semantically disambiguated. WordNet labels the semantic relations among the words and provides explicit pattern other than meaning similarity.

Apriori algorithm

The Apriori algorithm is used for generating sequence of patterns. It is used to filter the non-frequent elements. This algorithm is used to count no. of singleton elements, no. of 2- sequences, no. of 3-sequences, and this process is repeated until no more frequent sequences are found in the database.

There are two main steps in the algorithm.

- Candidate Generation. Given the set of frequent (k-1)-frequent sequences F(k-1), the candidates for the next pass are generated by joining F(k-1) with itself. A pruning phase eliminates any sequence, at least one of whose subsequences is not frequent.
- Support Counting. Normally, a hash tree–based search is employed for efficient support counting. Finally non-maximal frequent sequences are removed.

4.0 About Tourism

Tourism plays major role in entertaining all sorts of people from young age to elder, poor person to rich one's. In order to plan a trip, a user must visit a series of WebPages to get correct information related to his trip. A trip usually includes many parts such accommodation, transportation, insurance, visa services, guide services, excursions in the destination. Due to the heterogeneity of the travel product, travel agency consultant or a person who is planning the trip itself must have access to different sources of information. The above difficulties are overcome by developing semantic web to the tourism domain. It provides inter-operatability between the web data.

Tourism product is also immaterial, meaning that traveler cannot see or touch the tourism product before the trip. That is why reliable information about destination, accommodation options and other parts of the tourism product is extremely important for both people working in tourism industry and tourists themselves.

Tourism product cannot be stored in storage. If a hotel room or a seat in an airplane remains empty today, this is lost revenue for the tourism company. This is a reason why effective distribution and inventory management are key factors in the tourism business.

To overcome the above difficulties, tourism related unstructured, semi-structured heterogeneous web content are converted into semantic web format RDF and again this web content are categorized using different algorithms to obtain more relevant output for user query and this in turn improves the tourism business by effective distribution and enhanced inventory management of tourism product.

5.0 Semantic Web Categorization algorithm

Domain selected for proposed work is tourism. The popularity of the WWW resulted a flurry of websites covering tourism related information covering almost everything in the universe.

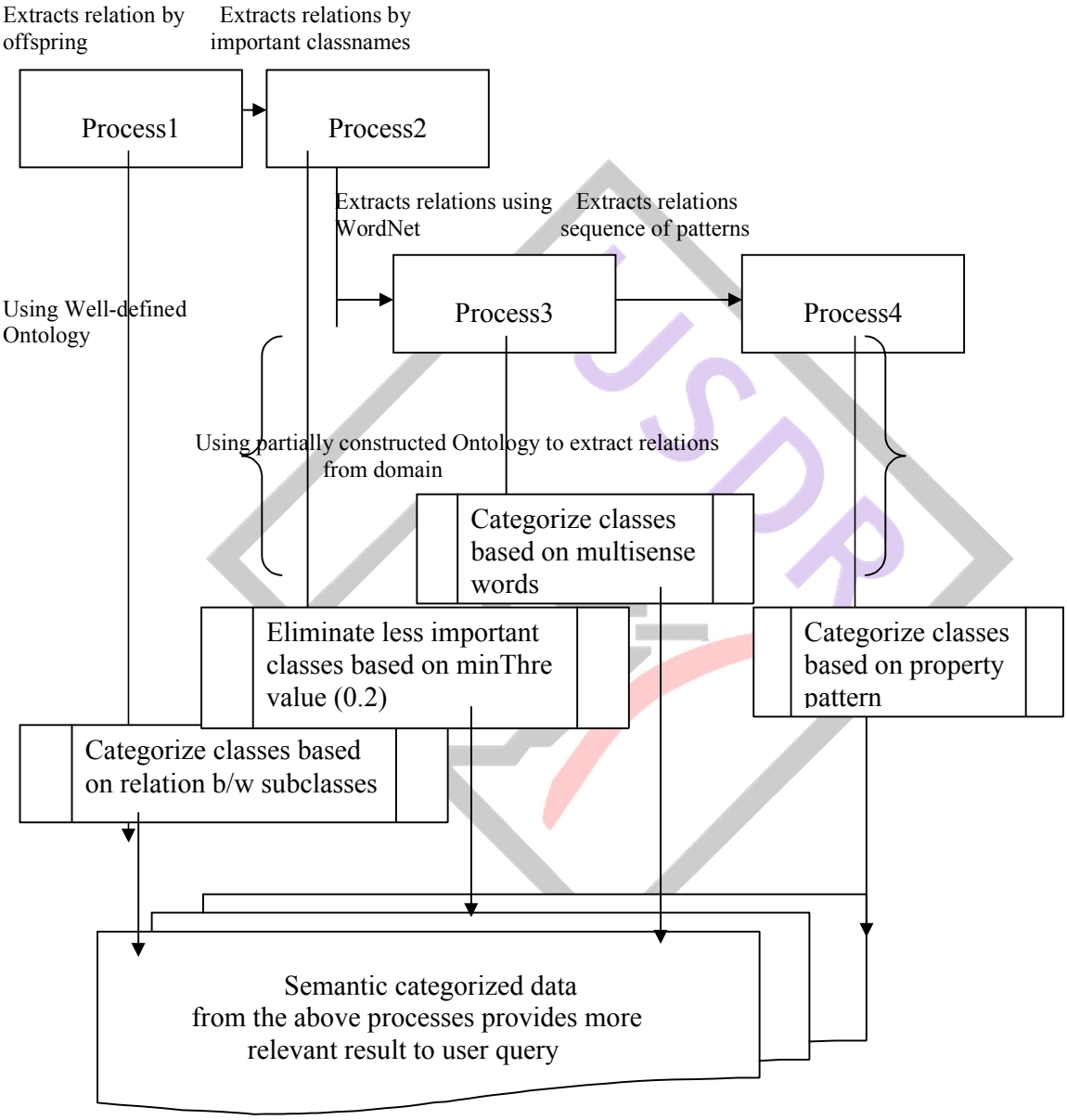
Tourism is one of important domain referring to many factors and has plenty of domain knowledge, which is the essential base of travel information systems.

The proposed work is to categorize the semantic web related tourism domain using four different processes.

1. Process1
2. Process2
3. Process3
4. Process4
5. Search

This is shown in the fig.1.

Fig.1 Semantic web categorization by four different processes



Semantic web data will be available for semantic search. In order to obtain more relevant result for user query, again this semantic web data is categorized using four algorithms namely process1 used for categorizing document by offspring, process2 used for categorizing document by important classname, process3 used for categorizing document by detecting potential relation from dictionary using WordNet, process4 used for categorizing document by detect relation in terms of property pattern analysis. Each process will be elaborated in the following section.

Recently, the Semantic Web increases in size and contains variety of resources. The web users can make efficient use of web resources that are machine readable and understandable in Semantic Web due to semantics. Providing categorization to the semantic web improves the performance of the search engine. As the World Wide Web changes dynamically, it's impossible to

structure a perfect ontology including the relations of all the web resources. In Semantic web, a class is a type of web resources. A class has some properties and it can be a subclass of another class. Thus, it cannot be matched by the ontology-based search engines for a web resource which is not defined as a class in the structured ontologies. To use web resources efficiently, it becomes an important and emerging issue to detect the potential relations on existing ontologies.

Since problems such as lack of relations and definitions always exist in real world Semantic Web data, all these four processes are used to categorize the semantic web resource types by using not only well-defined relations but also some possible semantic relations for incomplete ontologies.

Central idea of these methods is used to extract potential relations among the semantic web resources.

5.1 Process1 – Categorize offspring

P**rocess1** module extracts the relation by using the knowledge such as the class hierarchy in ontologies. Relations among classes defined in Web ontologies. All extended
classes based on relations such as "rdf:subClassOf," "owl:equivalentClass" and "owl:sameAs" should be in the same category.

Proces1 is intending to categorize the classes, well defined in ontologies. If a class has a relation to a category, its subclasses ("rdf:subClassOf") could have a weaker relation to the category. The equivalent classes ("owl:equivalentClass" and "owl:sameAs") have equivalent relation strength.

Let **A** denotes a set of these categories specified for categorization and **a** denotes a category in **A**. This denotation is used in each process. A measure of the relation between class **c** and category **a** is defined as relation strength and denoted as **rc,a**. Defined some class names to get base classes for each category, and the relation strength of initial classes to the category equals 1.

Process1 is performed by using a recursive function **RER(c0, a, rc0,a)**. Let **c0** be a class in category **a** with the relation strength as **rc0,a**. If class **c**, which is an equivalent class or a subclass of class **c0** in category **a**, has not been categorized to category a, RER(c0, a, rc0,a) can categorize class c to category a with relation strength as rc,a. By recursively calling itself, RER(c0, a, rc0,a) categorizes all the offspring classes of c0. Here, given a coefficient k for the degressive strength of relation in the hierarchical structure. While checking the number of categorized classes for a set of values of k, found that the number increase much while k is not larger than 0.95, and increase very little when k is larger than 0.95. Since too high value of k decreases the precision of categorization result, specify k as 1 for a same level and as 0.95 for a lower level class. The relation strength of class c to category a is calculated by Eq. (1).

$$rc,a = k \times rc0,a \qquad \text{Eq.(1)}$$

$$\left.\begin{array}{l} k = 1 \text{ equivalentClass} - \text{relation} \\ k = 0.95 \text{ subClass} - \text{relation.} \end{array}\right\} \qquad \text{Eq.(2)}$$

It takes input from well defined ontology i.e download from the web.

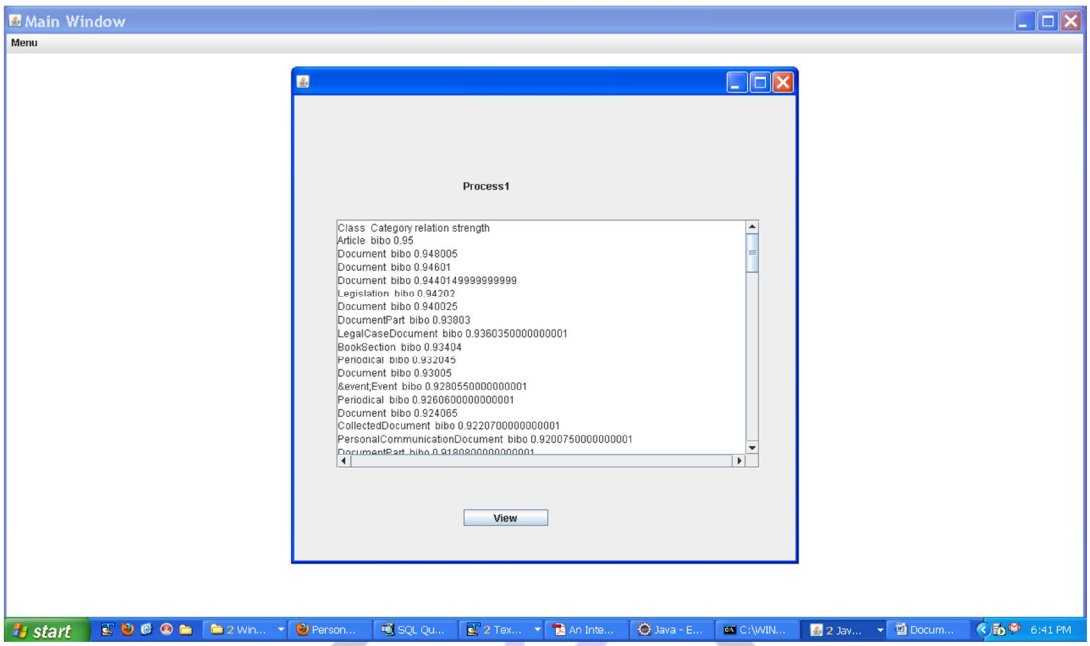
Algorithm for Process1

```
Foreach (category a ∈ A)
    Foreach (class c0 ∈ a)
        RER(c0, a, rc0,a);

Function RER(c0, a, rc0,a)
   Foreach (class c ∈ c0.Children)
       rc,a = k × rc0,a;
       If (c.Contain(a) = false)
          a.Add(c, rc,a);
          RER(c, a, rc,a);
  Foreach (class c ∈ c.Equivalents)
          rc,a = rc0,a;
       If (c.Contain(a) = false)
          a.Add(c, rc,a);
          RER(c, a, rc,a);
```

Using above algorithm the semantic web are categorized for well defined ontologies. It is shown in the fig.2.

Fig.2 Process1 – Categorize offspring



5.2 Process2 – Categorize by important class name

By performing **Process1**, some classes categorized for each category. In Process2, extract the class names that represent the characters of category a, then categorize the classes having the important class names to category a.

In **Process2**, categorize the classes not existing in the ontology based on the important class names. For each category a, extract all the class names by a function getName(a). Then compute a coefficient l to evaluate how much a class name cName can represent the character of a by function getW(cName, a) based on tfidf [7]. The weight of class name ci in category aj is denoted by Wi,j , which is calculated by Eq. (3).

$W_{i,j} = tfi_{,j} \times \log(Ndf)$ Eq.(3)

$tfi_{,j}$ = number of occurrences of i in j,
dfi = number of categories containing i,
N = total number of categories in A.
Since Wi,j is largely affected by the number of classes in a category, normalize Wi,j in [0, 1], and W'i,j denotes the normalized value which can be calculated by Eq. (4).

$W\_i,j = Wi,j/\max(Wi,j)$ Eq.(4)

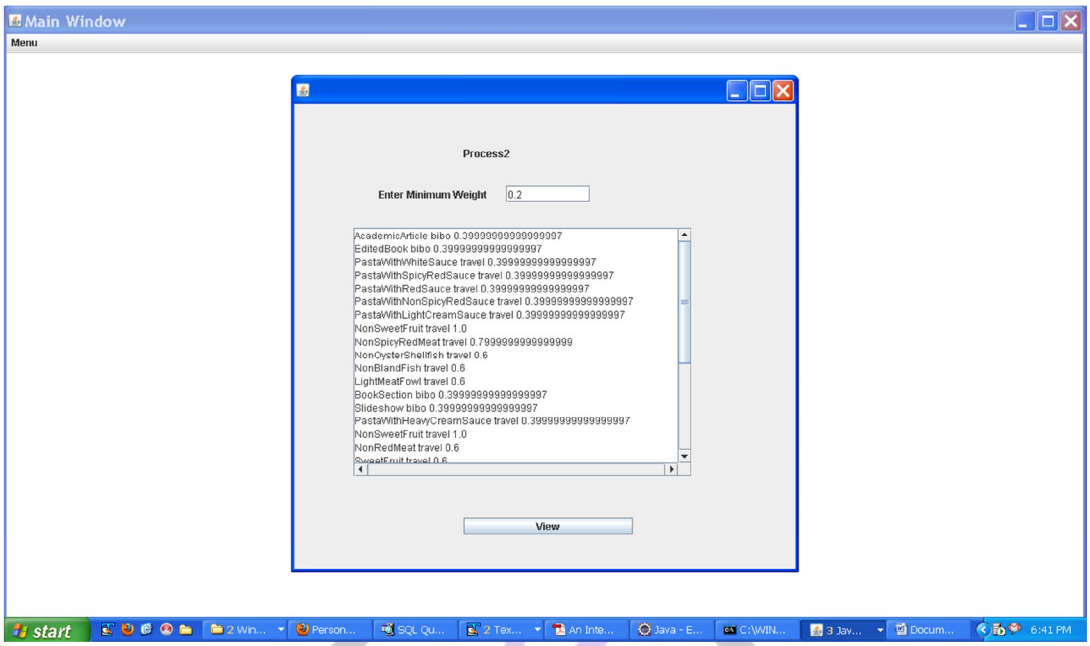
Consider the class names having W'i,j larger than a threshold minW as the important class names for the category. In the preliminary experiments, obtained good result of the categorization while giving a threshold minW ranging from 0.02 and 0.04. Here, specified the range as 0.02 by considering both effectiveness and efficiency. The algorithm is described concretely as follows. For each important class name cName, extract all the classes having class name cName in category a from all the classes by function getClass(cName). From these classes, extend category a with each class c which denotes a class having a class name of cName but not existing in category a. For a set of classes having class name cName and existing in category a, getR(cName, a) is a function to get the relation strengths of the classes in the set to category a, and avg(getR(cName, a)) denotes the average. Then, the relation strength of c to a can be calculated by Eq. (5).

$rc,a = avg(getR(cName, a)) \times W\_i,j$ Eq.(5)

It takes input form

In the **Process2**, extend the categories with all the offspring classes of the newly categorized classes by executing **Process1** again.

Fig.3 Process2 – Categorize by important class name



5.3 Process3: Detecting potential relation from dictionary

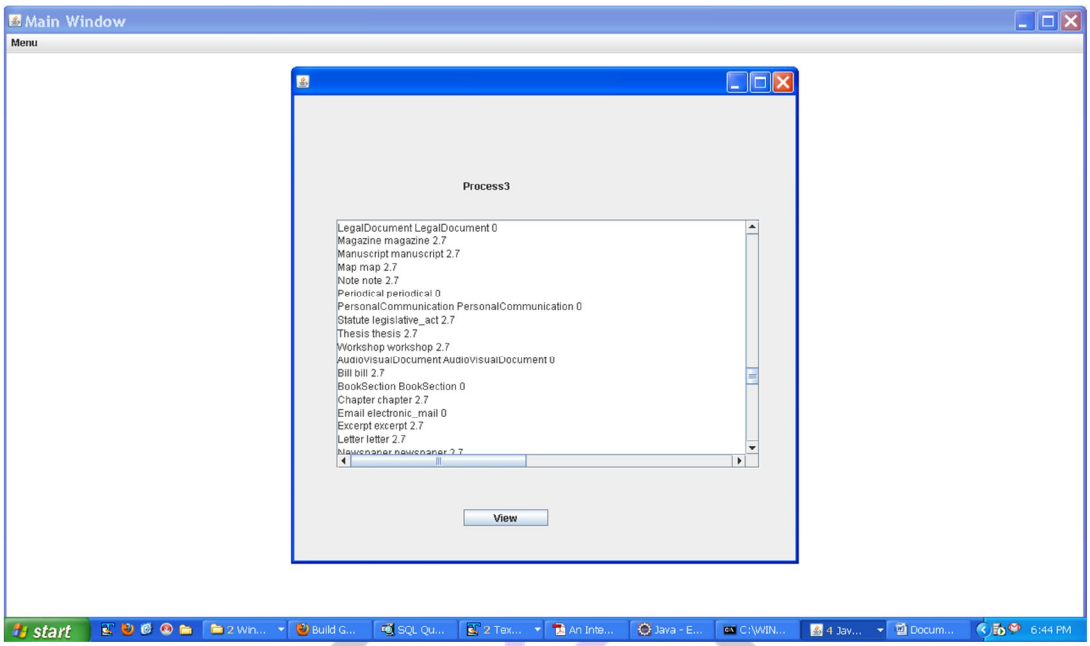
Process3 categorizes the classes using relations among class names based on the superior-inferior relations defined in dictionaries (WordNet). The algorithm is described concretely as follows. Get all the class names in category a by function getName(a) where pruned the class names that exist in more than one category. For each class name wd0 ∈ getName(a), extract all the offspring class names from WordNet and store them in B. For each class name wd in B, Get all the classes whose name is wd. If these classes do not exist in category a, they are categorized into a. There exist many multisense words that are not in common use. For example, word "fish" in WordNet has four senses such as a kind of aquatic vertebrates, a kind of food, a person who is born while the sun is in Pisces, and a sign of the zodiac.

In general, the first sense denotes the most common meaning. Use the relation between the first word senses of nouns in WordNet to simplify the experiment. Since an "is-a" relationship can be considered as an extension to a lower level, compute relation strength based on that of classes having a superior class name. rc,a denotes the relation strength of class c having class name wd to category a. wd0 denotes the superior of wd, and getR(wd0, a) computes the relation strength between classes having class name as the superior of wd to category a. avg(getR(wd0, a)) denotes their average.

Then, rc,a can be calculated by Eq. (6).

$$r_{c,a} = \text{avg}(\text{getR}(wd0, a)) \times 0.95 \quad . (6)$$

Fig.4 Process3 – Categorize by detecting potential relation from dictionary



Then, execute Process2 again to extend categories using all the offspring classes with the root names having high tfidf values.

5.4 Process4: Detecting relation by property pattern analysis

In **Process4**, perform a deeper categorization for the classes not existing in the ontology based on the description of the Web resources. In the Semantic Web, one resource is usually described by several properties. For example, a Web resource "Tour:India" is described by properties "has-tourist places in India" "has-web-address," "tourist-agency". It is considerable that if these properties were used together only for describing tourism, then the other resources described by these properties would have high possibility to be a similar class to tourism.

Here, define property set p as a set of properties that describes about a same resource. sup of property set p in category a is defined as the number of resources described by p in a divided by the number of all the categorized resources (Eq. (7)), which denotes the frequency of p occurred in a. conf of property set p in category a is defined as the number of resources described by p in a divided by the number of all the resources described by p (Eq. (8)), which denotes the possibility of p related to a. Then, property pattern in category a is defined as a property set having sup and conf larger than the threshold minsup and minconf respectively. The number of properties in a property pattern is defined as length of the pattern.

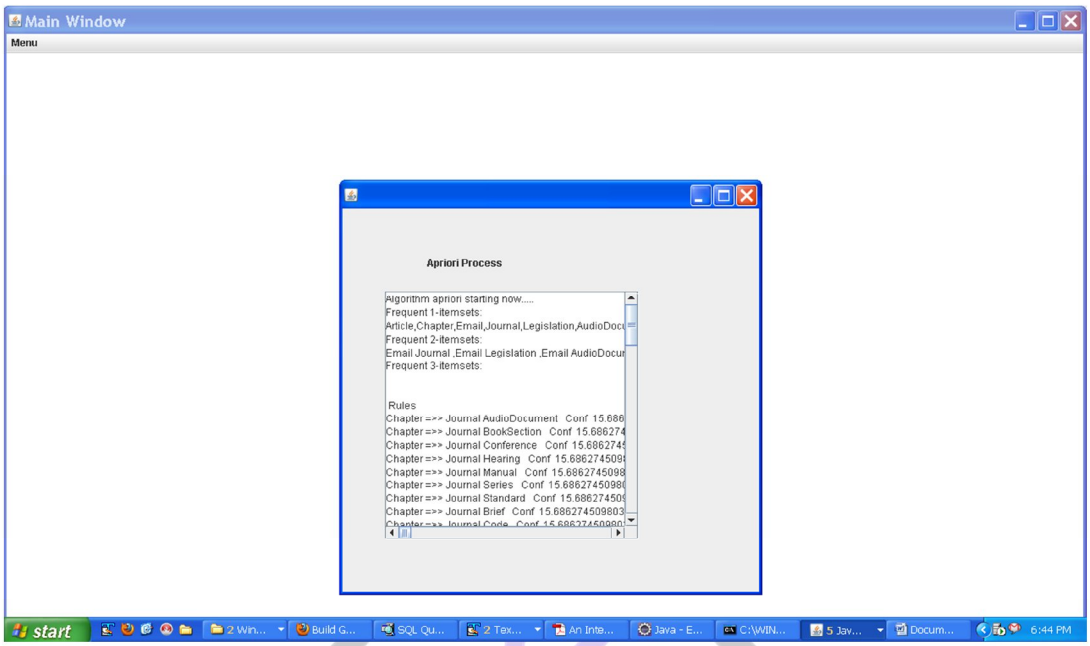
Sup p,a = |resources described by p in a|/|all the categorized resources|, (7)
conf p,a = |resources described by p in a|/ |categorized resources described by p|. (8)

Calculate the relations by considering the frequency of the property pattern in categorized Web resources. In category a, N is the count of the resources in a, and Nr is the count of resources described by property pattern p. Get the relation strengths of classes related to property pattern p to category a by using function getRp(p, a). avg(getRp(p, a)) denotes the average. Then, the relation strength of the detected class c to category a can be calculated
by following Eq. (9).

$$rc,a = avg(getRp(p, a)) \times ((Nr/ N0 + 1)/2 \quad . (9)$$

Since some classes are newly categorized from the resource type, execute Process2 again to extend the category using those newly categorized classes.

Fig.5 Process4 – Categorize by detecting relation by property pattern analysis

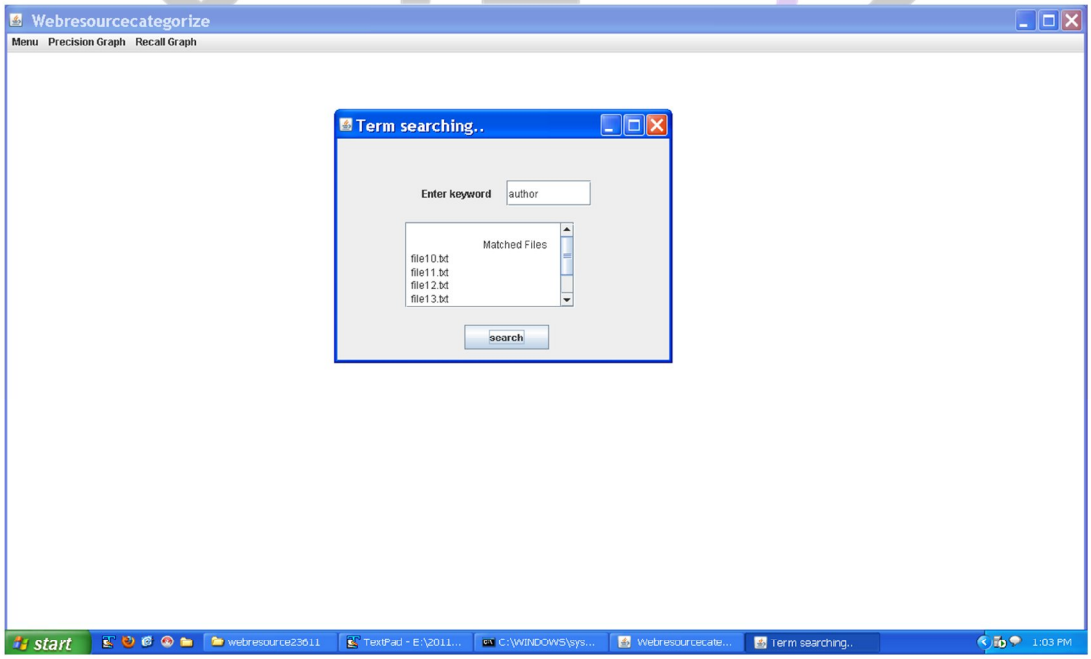


In the above process, the semantic web resources are categorized for efficient search of data for user query.

5.5 Search:

This module is used to find particular keyword that occurred in all the above four process. That is relation strength of a category and the classes.

Fig.6 Search keyword matched in all the four processes



6. Evaluation of processes using performance measure

The process flows as process1-process2-process3-process4, because this flow yields more categorization than other flow of data. Two performance measures such as Precision and Recall are used for the evaluation.

For a given user and query, let *S*1 denote a set of the relevant resources marked by the users, *S*2 denote a set of results output by system.

Precision is the ratio of the intersection of *S*1 and *S*2 to the results *S*2.

Recall is the ratio of the intersection of *S*1 and *S*2 to the relevant resources *S*1.

7. Conclusion and Future Enhancements

The semantic web resources are categorized using different processes namely categorize by offspring's, important classnames, WordNet, Pattern analysis for not only well defined ontology, but also for incomplete ontology too. That is the main advantage of this proposed work. After performing semantic web categorization using above four processes, it yields a very good semantic web repository. Search engine now provides very useful information for search query based on above semantic web categorization related to tourism. In future it will be extended to other domains.

References:

[1] Bhogal, J., Macfarlane, A. & Smith, P. (2007) 'A review of ontology based query expansion', Inf. Process. Manage. Vol. 43, No. 4, pp. 866-886.

[2] Bry, F., Koch, C., Furche, T., Schaffert, S., Badea, L. & Berger, S. (2010) 'Querying the Web Reconsidered: Design Principles for Versatile Web Query Languages', Int. J. Semantic Web Inf. Syst., Vol. 1, No. 2, pp. 1-21.

[3] Carmel, D., Yom-Tov, E., Darlow, A. & Pelleg, D. (2006) 'What makes a query difficult?', paper presented to the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA.

[4] Daramola, O., Adigun, M., Ayo, C., Building an Ontology-based Framework for tourism Recommendation Services, ENTER 2009, pp. 135-147, Amsterdam, Netherlands (2009).

[5] Horrocks, I. (2007) 'Semantic web: the story so far', paper presented to the Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A), Banff, Canada.

[6] Moscato, F., Martino, B.D., Venticinque, S. & Martone, A. (2009) 'OVerFA: a collaborative framework for the semantic annotation of documents and websites', Int. J. Web Grid Services, Vol. 5, No. 1, pp. 30-45.

[7] Noah, S.A., Alhadi, A.C. & Zakaria, L.Q. (2005) 'A semantic retrieval of web documents using domain ontology', Int. J. Web Grid Services, Vol. 1, No. 2, pp. 151-164.

[8] Panagis, Y., Sakkopoulos, E., Garofalakis, J. & Tsakalidis, A. (2009) 'Optimisation mechanism for web search results using topic knowledge', International Journal of Knowledge and Learning, Vol. 2, pp. 140-153.

[9] Reza Hemayati, Weiyi Meng, Clement Yu, "Semantic-based grouping of search engine results using WordNet", published in the proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management conference on Advances in data and web management.

[10] Staab, S., Werthner, H., Ricci, F., Zipf, A., Gretzel, U., Fesenmaier, D.R., Paris, C., and Knoblock, C.: Intelligent systems for tourism, IEEE Intelligent Systems, Volume 17, Issue 6, Nov/Dec, 53-66. (2002).

[11] Werthner, H. and Klein, S.: Information Technology and Tourism—A Challenging Relationship, Springer-Verlag, New York, 23. (2004).