Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ODMM - AN ONTOLOGY BASED DEEP MINING METHOD TO CLUSTER THE CONTENT FROM WEB SERVERS

[1] S.GANESH KUMAR, [2] Dr.K.VIVEKANANDAN
[1] Asstt.Prof., Department of Computer Science and Engineering, SRM University
[2] Professor, Pondichery Engg College.
E-mail: [1]sganeshkumar78@gmail.com, [2] kvivek27@yahoo.com

ABSTRACT

ODMM aims to present a novel ontology-based content-mining approach to cluster research proposals deep web search based on their similarities. The used method is efficient and effective for clustering research proposals with English texts. Text-mining methods have been proposed to solve the problem by automatically classifying text documents. Current search methods for grouping proposals are based on manual matching of similar search discipline keywords. The advantages of this method are that it can extract three types of data records, namely, synonyms data records, hypernymy data records, and hyponyms data records, and also provides options for aligning iterative and disjunctive data items. The proposed ODMM is used together with statistical method and optimization models and consists of reference to the ontology; the new proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm. The SOM algorithm is a typical unsupervised learning neural network model that clusters input data with similarities. Our new techniques used in data extraction from deep webs needs to be improved to achieve the efficiency and finally the result is given like the result comes for user query on multi view point like web links, news contents and the synonym hyponym and hypernym for the input term specified.

Key terms : *Semantic Web Services, ontology, mapping, ontology search engine*

1. SEMANTIC WEB

The Semantic Web introduces the idea to help computers read and use the content of the Web. The proposed concept ODMM is simple i.e. semantics (meaning) added to Web pages can make the existing World Wide Web content machine readable and machine understandable. We are not introducing artificial intelligence, but just preprocessing the input terms before providing the user with the output. This new idea of ODMMwould be an extension to the current one and not a replacement for the present World Wide Web (Web 2.0).Semantic Web helps in finding, extraction, representing, interpreting, and maintaining the information. Semantic Web will provide an important platform for improving search strategies and enhance the probability ofof making user satisfied by doing the refinement to the user query i.e. preprocessing it,thereby making information as machine process-able and machine understandable semantically and providing the output which the user desired for.

2. ONTOLOGY

Using Ontology we can annotate semantics and thereby obtain a foundation for Semantic Web. Ontology provides a common language vocabulary, an approach for grammatically publishing data, and providing a semantic description of data that could be used to conserve the Ontologies and make data ready for inference.An ontology is the used to represents data as a collection of concepts into various domain, using vocabulary to cluster the types, properties and interrelationships prevailing among them. Ontologies describe individuals (instances), classes (concepts), attributes, and relations. [1]

Common components of ontologyinclude :

- **Individuals**: instances or objects
- **Classes**: sets, collections, concepts, classes in programming, types of objects, or kinds of things
- **Attributes**: aspects, properties, features, characteristics, or parameters that objects (and classes) can have
- **Relations**: ways in which classes and individuals can be related to one another
- **Function terms**: complex structures formed from certain relations that can be used in place of an individual term in a statement
- **Restrictions**: formally stated descriptions of what must be true in order for some assertion to be accepted as input
- **Rules**: statements in the form of an if-then (antecedent-consequent) sentence that describe

the logical inferences that can be drawn from an assertion in a particular form

- **Axioms**: assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative

3 Linking Open Data

Linking Open Data (LOD) provides real meaning to Semantic Web. Integration of data on larger scale, and linking data on the Web would make the data more discoverable, accessible and ready to perform speedy operation. The W3C also maintains a collection of Semantic Web Case Studies and Use Cases that show how Semantic Web technologies, including Linked Data, are used in practice. In the proposed system linkage of data would be done on the basis of the semantics and ontologies. Linked Data is about using the Web to connect related data

grammar and formal logic. In those disciplines, axioms include only statements asserted as a priori knowledge. As used here, "axioms" also include the theory derived from axiomatic statements

Events: the changing of attributes or relation

Web to lower the barriers to linking data that wasn't previously linked, or using the currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."

The followingfig.1and fig. 2shows the example of status of linkage of data in the consecutive years which is continuously growing.

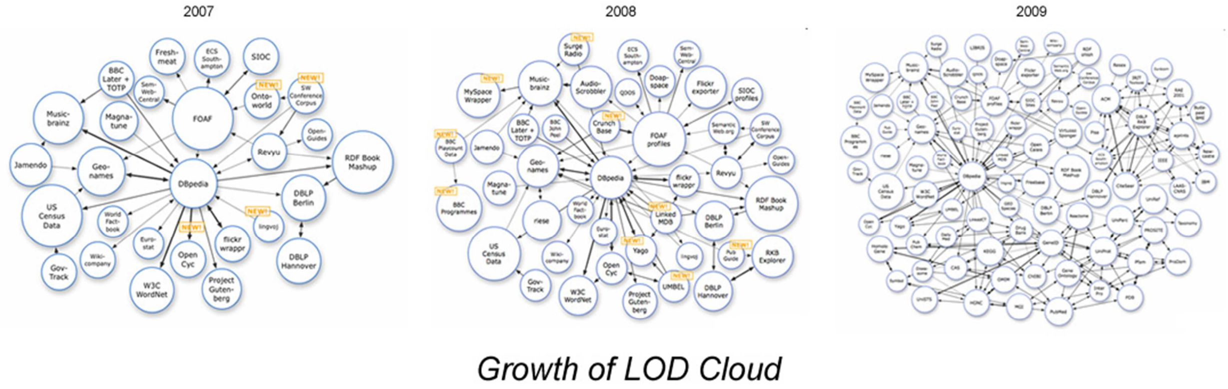


Fig. 1 The Growth of Linkage Data through years

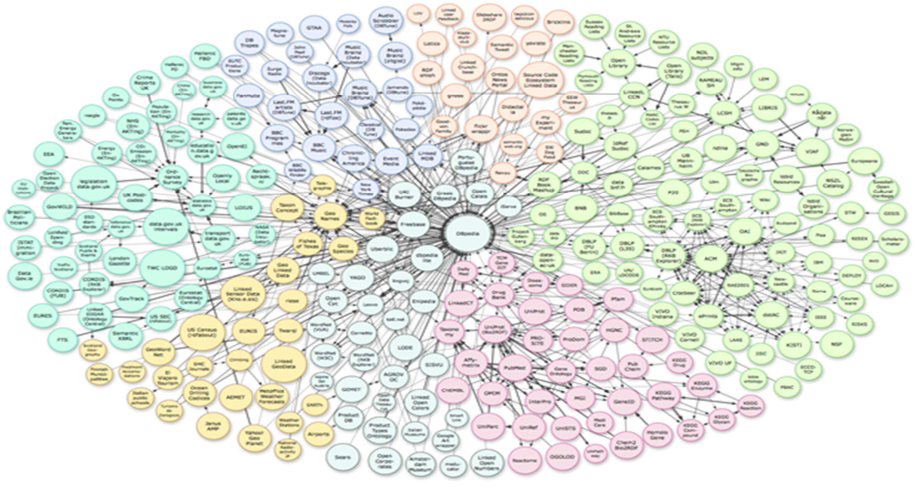


Fig. 2 ked Data At Present

4 WEB 2.0 AND WEB 3.0

The term "Web 2.0" (2004–present) is commonly associated with web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web. [3]
The current Web has its limitations when it comes to:

- finding relevant information
- extracting relevant information
- combining and reusing information

Hence the need arouse to bring the Web 2.0 to the next level by doing following modification :

- Increasing automatic linking among data
- Increasing recall and precision in search
- Increasing automation in data integration
- Increasing automation in the service life cycle
- Adding semantics to data and services is the solution

The following table would clearly differentiate between Web 2.0 and Web 3.0

Table 1. Comparing Web2.0 and Web 3.0

Web 2.0	Web 3.O
Syntax	Semantics
Objective	Subjective
Analytical	Synthesizing
Looks at parts	Looks at whole
Rational	Intuitive

5 RESEARCH WORK

In this section we will see about the contribution by various approaches in encouraging the development of this proposal. Text Mining has emerged as a definitive technique for extracting the unknown information from large text document. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. Ontology's make the task of searching similar pattern of text that to be more effective, efficient and interactive. The current method for grouping proposals for research project selection is proposed using ontology based text mining approach to cluster research proposals based on their similarities in research area. [4]Ontology mapping in the context of Question Answering can provide more correct results if the mapping process can deal with uncertainty effectively that is caused by the incomplete and inconsistent information used and produced by the mapping process. [5] service capabilities need to be manually analyzed, which lead to the development of the Semantic Web for automatic service discovery and retrieval of relevant services and resources. This work proposes the incorporation of Semantic matching methodology in Semantic Web for improving the efficiency and accuracy of the discovery mechanism. [6] A key goal of the Semantic Web is to shift social interaction patterns from a producer-centric paradigm to a consumer-centric one. Treating customers as the most valuable assets and making the business models work better for them are at the core of building successful consumer-centric business models. It follows that customizing business processes constitutes a major concern in therealm of a knowledge-pull-based human semantic web. [7]

6 PROPOSAL OF ODMM

Semantic Web Services (SWS) mainly works on the following five step process :

1. "Advertisement" by service provider in which the benefits about the service is explained
2. "Discovery" is used to find the list of services that can satisfy the user needs
3. "Selection" refers to picking the best option depending on QoS
4. "Composition" is combining the selected Web Services into a compound one
5. "Invocation" if for calling the web services for execution

As shown in following fig. the entire concept would be dependent on the backbone that would be created as a two stage process :

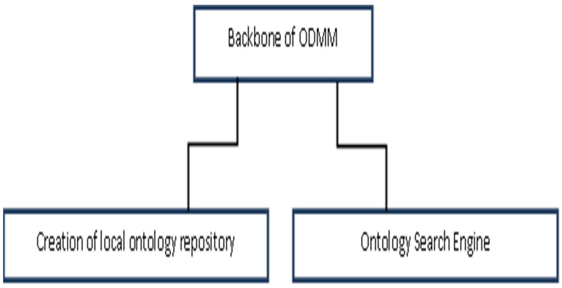


Fig 3. Backbone of ODMM

As shown in the above fig. 3 the primary job would be to create an ontology repository on basis of which we would be clustering the contents. In order to convert the present day data into the semantic based data we need to:

Stage 1 : Creation of local Ontology repository

1. Map form WSDL to OWL-S

   As the semantic web needs all the data to be clustered on the basis of the meaning, it becomes our primary job to map from WSDL TO OWL-S.

2. Register the unclassified data into the database
   This phase includes the registration 0f WSs and SWSs according to the semantic cluster they would belong too.
3. Classify all possible data into clusters
   Here we would be actually clustering our content according to the meaning.

Step 2 : Ontology Search Engine

1. Get the input from the user
   The input which the user would provide will be preprocessed and the required output would be provided by clustering the content on the basis of the preprocessing.
2. Using ontology cluster the output in terms of synonym, hyponym and hypernym
   The output generated would be classified into following categories :
   1. Synonym – equivalent meaning words to the input term
   2. Hyponym – equivalent meaning word to the synonym of the input term
   3. Hypernym – Technical equivalent word of the input term

The OWL-S provides the following components for the development of ODMM :

1. Service Profile – (provides) what does the service do
2. Service Model – (describes) how does the service work
3. Service Grounding – (supports) how to access the service

The following figure provides the architecture of the ODMM

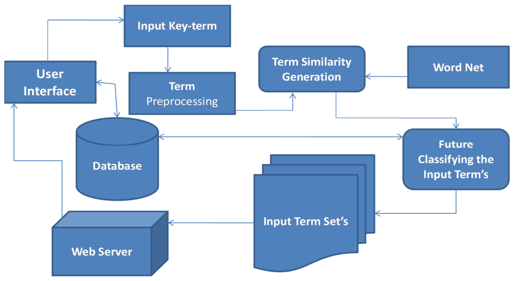


Fig. 4 : Architecture of ODMM

The working of the ODMM would be as follows :

- At first the user would be entering the input term through the user interface
- The input term would be first preprocessed, i.e. the stop words would be removed
- At the same time the input term would be saved into the database in the search list
- Depending on the preprocessing the key terms would be extracted and the similarity generation would be performed using the semantics.
- For this generation, we would be use Word Net,a tool that would provide semantic support
- The term generation would be further classifying the input term into synonym,hyponym and hypernym
- Different sets would be formed for the user to select from
- The classified set would be given to the web server, which would forward this to the user interface
- Finally the user has the option to select and get his desired results.

Use case diagram of ODMM would clear the respective roles of the users of this proposal

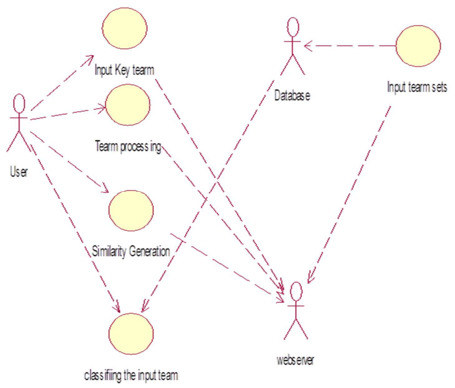


Fig. 5 : Use Case Dagram of ODMM

The above diagram form fig. 4clearly specifies how each component of the architecture would be divided as per the user.The main user in ODMM would be classified on the basis on the methods they would be performing. As each user have their specified tasks to be performed the entire implementation is predefined.

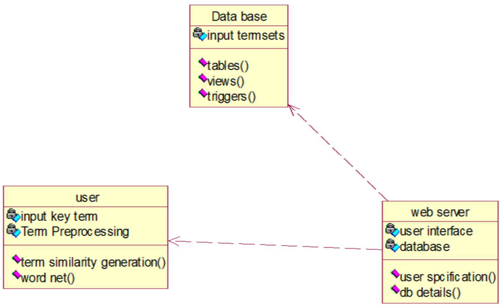


Fig. 6 Class Diagram og ODMM

The entire process flow of proposed ODMM is as follows :

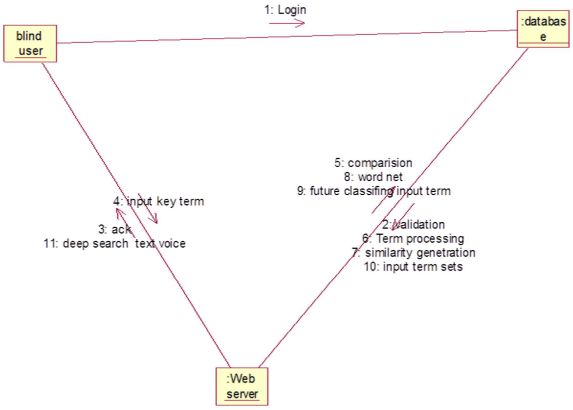


Fig 7. Process flow of ODMM

7. MODULE DEVELOPMENT

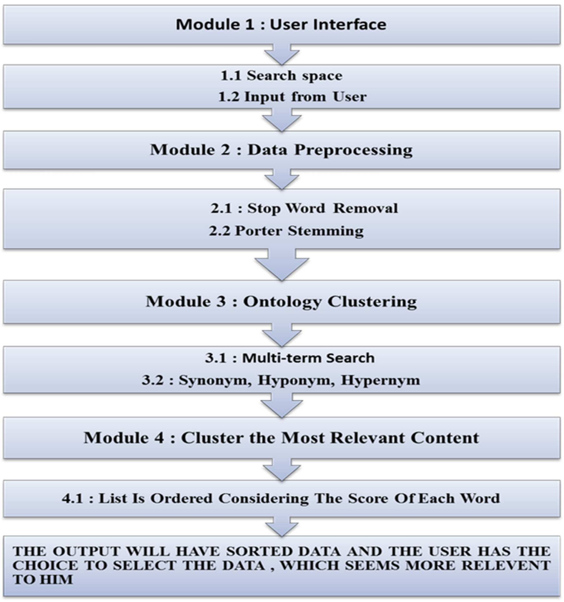

Module 1 : User Interface
1.1 Search space
1.2 Input from User
Module 2 : Data Preprocessing
2.1 : Stop Word Removal
2.2 Porter Stemming
Module 3 : Ontology Clustering
3.1 : Multi-term Search
3.2 : Synonym, Hyponym, Hypernym
Module 4 : Cluster the Most Relevant Content
4.1 : List Is Ordered Considering The Score Of Each Word
THE OUTPUT WILL HAVE SORTED DATA AND THE USER HAS THE CHOICE TO SELECT THE DATA , WHICH SEEMS MORE RELEVENT TO HIM

Fig. 8 : Module formation of ODMM

The entire process will be divided into 4 stages.

1. The User interface
   It provides with the search space and the space for the output.
2. Data processing
   The primary task of data preprocessing is stop word removal. Another innovative introduction is the introduction of removal of stop words. Stop words are those additive words that are entered by user as input term to make his/her query meaningful. The following table below provides a list of certain stop words used in general.

Table 2 : List Of Common Stop Words

a	can't	he	it's	over	they	when
about	cannot	he'd	its	own	they'd	when's
above	could	he'll	itself	same	they'll	where
after	couldn't	he's	let's	shan't	they're	where's
again	did	her	me	she	they've	which
against	didn't	here	more	she'd	this	while
all	do	here's	most	she'll	those	who
am	does	hers	mustn't	she's	through	who's
an	doesn't	herself	my	should	to	whom
and	doing	him	myself	shouldn't	too	why
any	don't	himself	no	so	under	why's
are	down	his	nor	some	until	with
aren't	during	how	not	such	up	won't
as	each	how's	of	than	very	would
at	few	i	off	that	was	wouldn't
be	for	i'd	on	that's	wasn't	you
because	from	i'll	once	the	we	you'd
been	further	i'm	only	their	we'd	you'll
before	had	i've	or	theirs	we'll	you're
being	hadn't	if	other	them	we're	you've
below	has	in	ought	themselves	we've	your
between	hasn't	into	our	then	were	yours
both	have	is	ours	there	weren't	yourself
but	haven't	isn't	ourselves	there's	what	yourselves
by	having	it	out	these	what's	

Porter Stemming Algorithm

The next important step in the development would be using a certain algorithm to cluster the content in terms of the relevance. The help of porter stemming algorithm would be taken for this purpose. Our proposed text clustering has a frequent concept to cluster the text documents. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. Our Proposed algorithm utilizes the semantic relationship between words to create concepts. The Relationship between words like synonyms, hypernymy, also be identified &hypernymy is most effective for Text clustering. As shown below in the figure Porter stemmer is divided into five steps, step 1 is divided further into steps 1a, 1b and 1c, and step 5 into steps 5a and 5b. Step 1 removes the i-suffixes, and steps 2 to 4 the d-suffixes. Composite d-suffixes are reduced to single d-suffixes one at a time. So for example if a word ends **icational**, step 2 reduces it to **icate** and step 3 to **ic**. Three steps are sufficient for this process in English. Step 5 does some tidying up.

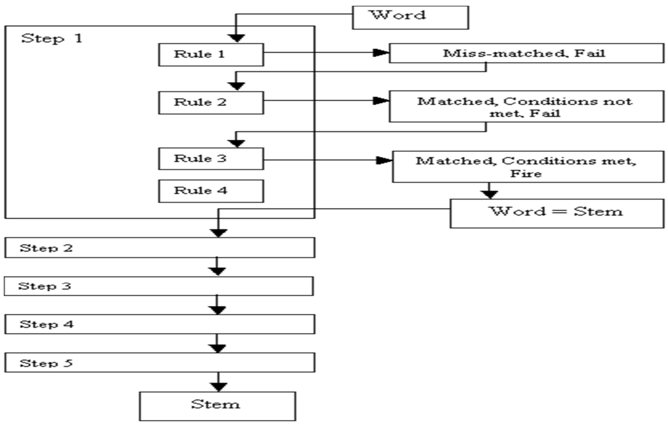


Fig. 9 : Porter Stemming Process Flow

Self Organizing Map Algorithm

The SOM algorithm is a typical unsupervised learning neural network model that clusters input data with similarities. Text-mining methods have been proposed to solve the problem by automatically classifying text documents. We propose the semantic web based search engine which is also called as Intelligent Semantic Web Search Engines. We use the power of xml meta-tags deployed on the web page to search the queried information. The xml page will be consisted of built-in and user defined tags. Here propose the intelligent semantic web based search engine. We use the power of xml meta-tags deployed on the web page to search the queried information. The xml page will be consisted of built-in and user defined tags. The metadata information of the pages is extracted from this xml into rdf. our practical results showing that proposed approach taking very lesstime to answer the queries while providing more accurate information.

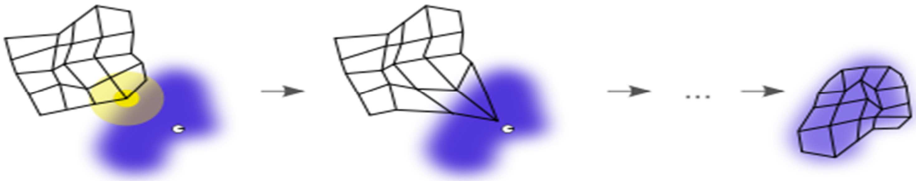


Fig. 10

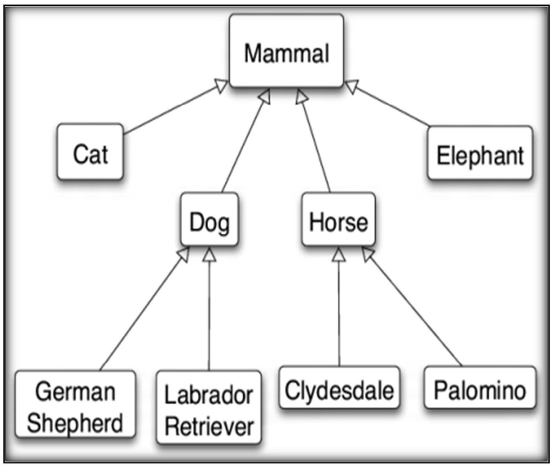
0 :Example of Self Organizing Map

An illustration of a self-organizing map is shown in the above figure. The blue blob is the distribution of the training data, and the small white disc is the current training datum drawn from that distribution. At first (left) the SOM nodes are arbitrarily positioned in the data space. The node (highlighted in yellow) which is nearest to the training datum is selected. It is moved towards the training datum, as (to a lesser extent) are its neighbors on the grid. After many iterations the grid tends to approximate the data distribution (right).

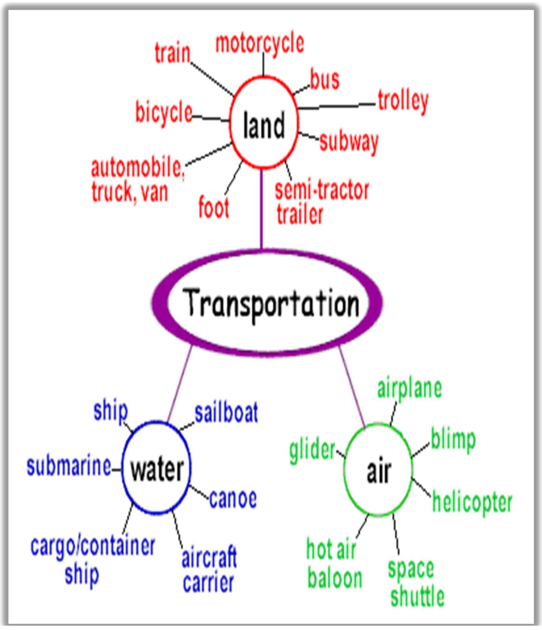
8 HOW WILL IT WORK

The input term given by the user would be preprocessed and cluster and the output would be given through the web server.

Example 1 : if the user input term is mammal, all the animals fall into this category. Hence the output would be clustered in the following fashion.



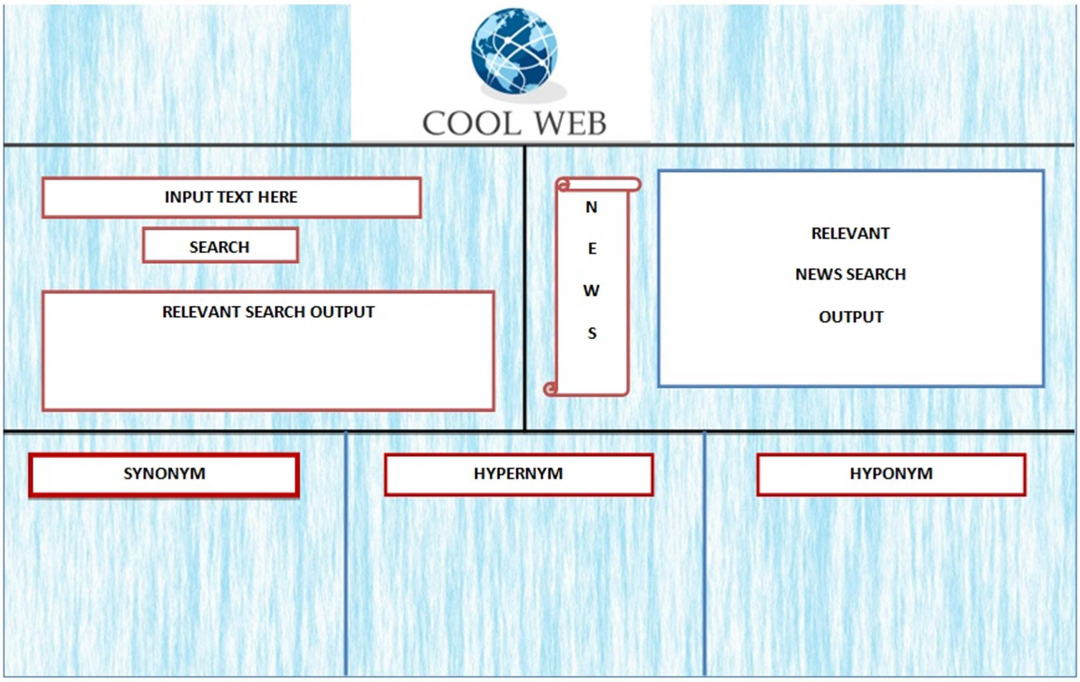
Example 2: if the user input term is transportation, all the animals fall into this category. Hence the output would be clustered in the following fashion.



As shown above we have clustered the term according to the semantics, thereby introducing intelligent search and preprocessing of data, the following section will give you an outlook of how the system would look like on completer implementation of above said algorithms and concepts.

9 PROPOSED

Appearance OfODMM



The output would consist of multiple frames for the following :

1. Input Search Field
2. Relevant search output for non-semantic words
3. Output frames for synonym, hyponym and hyperrnym
4. Frame for News related to the input term

10 CONCLUSION

ODMM is designed to be a scalable search engine. The primary goal is to provide high quality search results over a rapidly growing World Wide Web. ODMM employs a number of techniques to improve search quality including page rank, anchor text, and proximity information. The proposed System is a complete architecture for gathering web pages, indexing them, and performing search queries over them and all these works fine for a random collection of data which has no connection between them.

11 FUTURE WORK

Since the proposed system is for text clustering only at present, once the proposed system is implemented, ODMM intends to extend its scope for images, videos and other non-textual contents also.

REFERENCES

[1] Cai M., Zhang W.Y., and Zhang K., "ManuHub: *A semantic web system for ontology-based service management in distributed manufacturing environments*," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 3, pp. 574–582, May 2011.

[2] Liang Q., Wu X., Park E. K, Khoshgoftaar T. M, and C. H. Chi, *"Ontology-based business process customization for composite web services*," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 717–729, Jul. 2011.

[3] Liu O. and Ma J., "*A multilingual ontology framework for R&D project management systems*," Expert Syst. Appl., vol. 37, no. 6, pp. 4626–4631, Jun. 2010.

[4] Nagy M. and Vargas-Vera M., *"Multiagent ontology mapping framework for the semantic web,"* IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 693–704, Jul. 2011.

[5] http://en.wikipedia.org/wiki/Ontologycomponents

[6] http://blog.semantic-web.at/2009/05/

[7] http://en.wikipedia.org/wiki/Web_2.0