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

S.GANESH KUMAR, Dr.K.VIVEKANANDAN
1 Asstt.Prof., Department of Computer Science and Engineering, SRM University
2 Professor, Pondicherry Engg College.
E-mail: sganeshkumar78@gmail.com, 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, hyponymy data records, and hypernyms 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 ODMM would 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 of making user satisfied by doing the refinement to the user query i.e. preprocessing it, thereby making information as machine processable 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 ontology include:

- **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 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

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.
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 arose 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>
<thead>
<tr>
<th>Web 2.0</th>
<th>Web 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>Semantics</td>
</tr>
<tr>
<td>Objective</td>
<td>Subjective</td>
</tr>
<tr>
<td>Analytical</td>
<td>Synthesizing</td>
</tr>
<tr>
<td>Looks at parts</td>
<td>Looks at whole</td>
</tr>
<tr>
<td>Rational</td>
<td>Intuitive</td>
</tr>
</tbody>
</table>

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

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 of 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

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 user interface, 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

The above diagram form fig. 4 clearly 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.
The entire process flow of proposed ODMM is as follows:

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

**Fig. 6 Class Diagram of ODMM**

**Fig. 7. Process flow of ODMM**

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

### 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, hyponymy, 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 *ational*, 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.
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 less time to answer the queries while providing more accurate information.

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

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 hypernym
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


