A Unified Framework for Ontology Driven Web Mining and Semantic Clustering of Web Documents

Tulika Narang
University of Allahabad

Abstract—The objective of the paper is to propose a framework for web mining. An architectural pattern for multilayered architecture that implements ontology driven web mining and semantic clustering of web documents is proposed. The paper discusses a multilayered architecture for semantic clustering of web documents. The multilayered architecture maps to an architectural pattern. The mapping from pattern to implementation is also discussed in the paper.

Semantic clustering is supported by including ontology as background knowledge. The architecture provides a unified and universal framework for ontology driven web mining and semantic clustering of web documents.

Keywords- Architecture pattern, Ontology, Web mining, Clustering.

I. INTRODUCTION
Web mining is the discovery and analysis of useful information from World Wide Web. It is the discovery of knowledge from Web data. Mining web data includes web page content, hyperlink structure and web log data [1,17,18].

Mining with ontology yields better results as compared to traditional Bag of Words approach. The inclusion of Wordnet ontology enhances the semantic content and yields improved clustering results [7,8,12].

There is no proposed architecture for web mining. The domain of web mining lacks a unified and universal framework. The framework can be provided by an architecture that maps to a pattern. The architecture pattern provides a solution to frequently occurring same or similar problem domains. It represents a structural organization schema for a domain. It defines the various modules and roles of each module.

II. BACKGROUND & RELATED WORK
A. Web Mining
Web mining is the application of data mining techniques for extracting knowledge from web data. Data on the web includes content of Web pages, information available via hyperlinks and Web log data. Web mining is the application of data mining techniques for extracting knowledge from web data. It includes search and retrieval of information from resources such as Web documents. It is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. It is further categorized as [17,18,19]:

- Web content mining
- Web structure mining
- Web usage mining

Web content mining is the process of extracting useful information from the content of web page. The technologies used are Natural Language Processing and Information retrieval. Information retrieval is the process of organizing, storing, retrieval and evaluation of information relevant to user’s query [1,23,24].

Web Content Mining includes several concepts of traditional text mining techniques. Clustering is one such method. Web document clustering puts similar documents together in a group or cluster. This makes information retrieval process more effective. It finds inherent groupings of pages so that a set of clusters is produced. The clusters contain relevant pages to a specific topic. The irrelevant pages are outliers and are not included in any group. The clustering methods group the documents into clusters [25]. Each cluster or group represents a particular topic that is different than those topic represented by the other groups. It is unsupervised learning in which similar documents are grouped to form classes or groups. A good clustering aims at high inter-cluster dissimilarity and high intra-cluster similarity.

The objective of web page clustering is to find similar groups of web documents. The resultant groups have similar Web pages while Web pages from different resultant groups are dissimilar [1,13,17,18].
The inclusion of ontology enhances the semantic and context and improves the retrieval process on the World Wide Web. The emphasis is to retrieve the most useful and relevant. Traditional keyword based search has limitations of overload and mismatch associated. The challenge is to overcome these limitations [14,22].

B. Ontology
An ontology is a formal description of concepts and relationships that can exist for a community of human and/or machine agents.[1,4,2,15,16,26]
Wordnet ontology provides an explicit lexicon for disambiguating the meaning of a word [5,21]. The inclusion of ontology enhances the semantic and context and improves the retrieval process on the World Wide Web [16].

C. Clustering and Semantic clustering
Clustering is the process of grouping Web sources into sets or clusters so that similar objects are in the same cluster and dissimilar objects are in different clusters. It creates new groupings such that each cluster is similar in some ways to each other and dissimilar to those in other clusters. Clustering on the Web identifies homogeneous groups of web documents. The objective is to organize large document collections into meaningful groups of document clusters [11,12,13].

Clustering can be broadly classified as Flat and Hierarchical. Flat clustering results in a flat set of clusters without any explicit structure that would relate clusters to each other. Hierarchical clustering creates a hierarchy of clusters. Hierarchical clustering can further be classified as agglomerative or divisive. An agglomerative method starts with each document representing a single cluster. This is a bottom up approach to clustering. In divisive approach all observations start in one cluster and continue until each object is in a separate group. It is a top down approach to clustering [3].
Semantic clustering includes the semantic relationships between words. This enhances the meaning associated and results in better clustering of web documents. The traditional clustering methods rely only on Bag of Words approach and do not consider the semantics associated. This limitation of traditional approaches poses the challenges in effective clustering and mining relevant documents from document corpus. Semantic similarity imposed by semantic clustering methods overcomes the limitation of traditional methods [8,9,10,20].

III. PROBLEM STATEMENT
A wide range of semantic clustering methods have been proposed for effective web mining. It has been proved that inclusion of semantics and context improves the clustering result. The inclusion of ontology for background knowledge improves the results obtained in clustering web documents [7].
There is no unified and universal framework proposed for the domain of semantic clustering of web documents. The paper is an attempt to provide the framework. The framework includes a multilayered architecture and proposes an architectural pattern. The architecture maps to the design pattern. The proposed design pattern is an architecture pattern. Architecture patterns define the basic characteristics and behavior of an application.

IV. SOLUTION
The solution to the problem is provided by the following:
1. Multilayered architecture
2. Proposed Architecture Pattern
3. Implementation
A. Multilayered Architecture
The proposed multilayer architecture is represented as[1]:

![Fig 1. Multilayered architecture for web mining[1,2]](imageurl)

The multilayer architecture comprises of:
Layer 0: It represents the Web itself. Due to the implicit characteristics of WWW it poses challenges for effective and useful knowledge discovery[1].
Layer1: It is the Web page descriptor layer. It defines descriptive information related to Web documents such as address of the page, keywords, timestamp; access frequency. The clustering process focuses on keywords for web content mining[1].
Layer2: The layer represents the clustering process. [1]
B. Proposed Architecture Pattern

The layered architecture provides a solution to the stated limitation. Layered architecture lays emphasis on grouping of related functionality within an application into distinct layers. The layers are stacked vertically on top of each other. The layers are related by a common role or responsibility. The layered pattern is an inverted pyramid of reuse. A layer aggregates the functionalities and abstractions of the layer directly beneath it [27,28,29].

The layering can be of following types:

- **STRICT**
  The components of a layer interact with components of the same layer or with components that are directly beneath it.

- **RELAXED**
  The components of a layer interact with components of the same layer or with components of any layer that are beneath it.

A layered architectural style focuses on the essential principles of Abstraction, Encapsulation, High cohesion. Reusability and Loose coupling. [27,28].

The objective of the principles is to view the system as a whole with emphasis on individual responsibility and role of each layer. The separation of functionality ensures that layer performs a task related to the layer.

C. Implementation

1. Mapping from architectural pattern to implementation method:
   1) Layer 0: The document corpus that is candidate for clustering is set of web documents
   2) Layer 1: The attribute keywords are extracted for representing the documents in the candidate corpus
   3) Layer2: Clustering using SOM and wordnet ontology is performed and results evaluated.

2. Steps to generate vector of documents and clustering using the layered approach[2]:
   - Identification and extraction of terms from documents.
   - Mapping the domain lexicon to ontology structure(Wordnet ontology)
   - Indexing of documents
   - Generating vector of documents(VSM model via TMG tool)[3]
   - Clustering using Self Organising Map in Layer 3[2]

Self Organising Map(SOM) based clustering with ontology yields better results[2,6]. The inclusion of Wordnet lexical categories improves the clustering process. The clustering process takes the advantage of the semantics associated and results in better and semantically enhanced clusters of documents.

**Figure2.SOM clustering on Reuters dataset [6]**

The clustering results show that the clustering quality for input data set is much better than the traditional Bag of Words.

REFERENCES


[14] Yuefeng Li and Ning Zhong, “Mining Ontology for Automatically Acquiring Web User Information needs”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 4, April 2006


[17] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, ISBN 81-8147-049-4


