International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 2, No. 3
August 2013

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 2, No. 3, August 2013

Research Paper

SOM BASED CLUSTERING OF WEB DOCUMENTS USING AN ONTOLOGY

Tulika Narang[1]* and R R Tewari[1]

Corresponding Author:* ***Tulika Narang,* ✉ *n.tulika@gmail.com*

The paper discusses SOM based clustering of web documents. The clustering method uses ontology for document representation. The approach is an attempt to overcome the drawbacks of overload and mismatch in information retrieval from World Wide Web. The proposed solution is a neural network based clustering technique, Self Organizing Maps (SOM). Clustering on the web has been proposed based on the idea of identifying homogeneous groups of web documents. The clustering results in groups of relevant web documents. The relevance of a web document is computed using a relevance function.

Keywords: SOM, Ontology, Clustering, Information retrieval

INTRODUCTION

World Wide Web (WWW) is a large-scale document collection. It is a huge, widely distributed, global information repository. Information Retrieval (IR) on WWW considers the combination of textual content of documents, structure of the Web, and the search behavior of users. The information available on the Web is very different from the information contained in either libraries or classical IR collections. A large amount of information on the Web is duplicated, and content is often mirrored across many different sites. Also many documents can be inaccurate and inappropriate according to users' request. Only a small subset of the information is relevant or useful to a user. It is a challenge to find high-quality web pages on a specified topic. IR on WWW suffers from the drawbacks of overload and mismatch. It is not easy to obtain right information for a particular user. Overload and mismatch are essential issues regarding extraction of information form WWW. It difficult for users to search and retrieve documents that is relevant to their particular needs. Users browse through a large hierarchy of concepts to find the relevant information. The query submitted to a search engine has to wade through irrelevant documents. The problem of overload occurs when a large number of irrelevant documents may be considered to be relevant. Mismatch

[1] Centre of Computer Education, Institute of Professional Studies, University of Allahabad, Allahabad, India.

occurs when retrieved information is not according to users' expectations.

The growth of Web has enhanced the problems with its usage. In particular, the quality of Web search and corresponding interpretation of search results are not according to users' expectations. The retrieved information has drawbacks of mismatch and overload (Yuefeng Li and Ning Zhong, 2006; Li and Zhong, 2004).

The focus is to retrieve the most useful and relevant. The challenge has promoted to find methods for effective and efficient searching on the web. The challenge can be overcome by improved information retrieval methods. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs such as search strings in web search engines. In information retrieval several objects match the query with different degrees of relevancy.

One solution is to construct meaningful classifications of objects (Hector Oscar, 2007 Jiawei Han and Micheline Kamber, 2005; Margaret H Dunham and Sridha, 2007). The essential application is to group similar objects into classes. Classes or clusters are collections of objects whose intra-class similarity is high and inter-class similarity is low.

Cluster analysis is a technique for multivariate analysis that assigns items to automatically created groups based on a calculation of the degree of association between items and groups (Jiawei Han and Micheline Kamber, 2005; Margaret H. Dunham and Sridha, 2007; Richard J Roiger and Michael, 2005). It deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters. The groups or clusters which are formed should have a high degree of association between members of the same group and low degree between members of different groups. It is a tool of information discovery. It has the potential to reveal previously undetected relationships between data. The goal of clustering is to create classes or clusters such that the objects within a group are similar and related. The greater the similarity the better and distinct is cluster analysis. In context of web mining cluster analysis aims at grouping Web pages on the basis of document content. It is useful for organizing documents to improve retrieval and support browsing.

The proposed methodology performs clustering of Web documents using ontology (Ernesto William De Luca, 2006). Clustering is performed using a neural network technique, SOM. It represents each cluster as an exemplar. An exemplar acts as a prototype of the cluster. New objects can be distributed to the cluster whose exemplar is the most similar. Similarity is computed using a similarity function.

The word "ontology" has a long history in philosophy. It refers to the subject of existence. In the context of knowledge sharing Gruber defined ontology, "a specification of a conceptualization". That is, ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. Ontology provides a shared vocabulary, which can be used to model a domain that is, the type of objects, and/or concepts that exist, and their properties and relations. It defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members).

BACKGROUND AND RELATED WORK

Information Retrieval (IR) is concerned with searching for documents, for information within documents, and for metadata about documents. It also deals with searching structured storage, relational databases, and the WWW. It is an essential task in web content mining. Applying data mining techniques to web page content is referred to as web content mining (Jiawei Han and Micheline Kamber, 2005; Margaret H Dunham and Sridha, 2007).

It is a sub-area of web mining, partially built upon the established field of information retrieval. IR domain is concerned with searching for documents, for information within documents, and for metadata about the documents. It involves the development of sophisticated systems that can act autonomously or semiautonomously on behalf of a particular user to discover and organize Web-based information. It aims at retrieving the useful and leaving the rest. It is concerned with finding the relevant subset from the available set of documents (Li and Zhong, 2004, Baeza Yates and Ribeiro Neto, 1999).

The Web consists of a variety of Web sources. In order to facilitate data availability and accessing, and to meet user preferences, the Web sources are clustered with respect to a certain parameter or characteristic such as their popularity, structure, or content. The vast amount of content on the Internet has made difficult for users to find and utilize information. It is difficult to classify and catalog documents. Traditional web search engines often return hundreds or thousands of results for a search. There is a requirement for efficient and automated information retrieval methods. The process of extracting useful information from WWW is termed as Web mining (Li and.Zhong, 2004). It includes the data present in Web pages and data related to Web activity. It is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. An appropriate clustering algorithm is a useful first step in extracting relevant information from data sources such as WWW. It groups similar documents together to make information retrieval more effective. Web document clustering methods identify inherent groupings of pages so that a set of clusters is produced in which clusters contain relevant pages to a specific topic. The grouping of documents is done on the basis of related content. This information is useful in various applications, for example, in Web search engines. This improves the information retrieval process (i.e., clustering Web queries). In addition, the clustering of Web documents increases Web information accessibility and improves content delivery on the Web.

The irrelevant pages are outliers and are not included in any group. The clustering methods group the documents into clusters. Each cluster or group represents some topic that is different than those topic represented by the other groups. It is helpful for discrimination, summarization, organization, and navigation for unstructured Web pages. It aims at improving the traditional informational retrieval performed by search engines (Liu and Chang, 2004).

The proposed solution is an artificial neural network based clustering technique, Self Organizing Maps (SOM). SOM based clustering initially proposed by Kohonen (Tenvo Kohonen, 1998). It performs clustering by means of competitive learning. It is particularly suited for dimensionality reduction and exploratory data

analysis. The goal is to identify a finite and discrete set of groupings in the patterns. The similarity between objects in a group is required to be larger than the similarity between objects belonging to different clusters. In the SOM the neurons are usually arranged in a two dimensional lattice or feature map. Each neuron receives inputs from the input layer and from the other neurons in the map. The input samples are described with real vectors. Each neuron contains a model vector that can be regarded as a prototype of the patterns in the cluster.

During the learning, the network performs clustering and the model vectors are changed so as to reflect the similarity of neighboring clusters. The goal of the mapping is to represent the points in the source space by corresponding points in a lower dimensional target space. In particular, the training is aimed at preserving as much as possible the distance and proximity relationships among input samples. The basic SOM training algorithm relies on the Euclidean distance to compare the patterns and the model vectors. The Euclidean distance between two patterns is in general very sensitive to small transformations of the patterns.

PROPOSED METHODOLOGY

The proposed solution emphasizes ontology driven clustering of web documents (Yuefeng Li and Ning Zhong, 2004). SOM based clustering is selected to find some structure in a set of patterns without predefining the classes (Tenvo Kohonen, 1998; Juha Vesanto and Esa Alhoniemi, 2000). Given a corpus of documents and a user's query, the task is to identify and retrieve the most relevant document. The documents relevant to a class are grouped together on the basis of relevance function. Each class forms a cluster of similar web documents. The similarity measure is the relevance factor (Yuefeng Li and Ning Zhong, 2006; Xiaohui tao, Yuefeng Li and Ning Zhong, 2011). Relevance denotes how well a retrieved document or set of documents meets the information need of the user. A document belongs to a class only if the relevance value is greater than or equal to a specified threshold.

The proposed methodology comprises of following two essential steps:

1. Preprocessing
2. Clustering

Preprocessing

Preprocessing involves two essential tasks:

1. Representation of the domain
2. Calculating the importance of semantic element.

Representation of Domain

The domain is represented as ontology built using text corpus. Then documents are indexed. The documents are numerically represented by vectors whose dimensions correspond to indexing units. The vectors store the weight of the indexing unit and are input to the clustering algorithm.

Ontology construction results in a hierarchical structure comprising of concepts and the relationships between them. This structure includes the inter-relationships between concepts.

An ontology is represented as a tuple, T (Hector Oscar Nigro, 2007, Wanlong LI *et al.,* 2006).

T= (L, F, C, H, Root)

where,

L: Domain lexicon that contains set of terms

C: Set of concepts of the ontology.

F: Reference function that links a set of terms to the set of concepts they refer to.

H: Hierarchy that contains relationship between concepts, generally related via the subsume relationship.

ROOT: Top concept and belongs to set C

The following steps are performed to generate vector of documents from ontology structure:

- The input is a set of Web documents. The input maps to domain lexicon. Terms are identified and extracted from documents. This represents the domain lexicon.
- The lexicon is mapped to ontology structure. The ontology structure is a hierarchical structure comprising of set of concepts representing the domain.
- When ontology is available documents are indexed.
- Indexed documents are used to generate vectors of documents. These are input to the clustering method. This map to vector space model for domain representation. Each document is represented as a feature vector, whose length is equal to the number of unique document attributes in the collection. Each component on that vector has a weight indicating the importance of each attribute in the characterization of the document. Generally, these attributes are terms extracted from the document.

Computing the Importance of Semantic Element

The importance of a semantic element is based on its frequency of its occurrence in the document. A semantic element, i.e., a term is important to a document only if it is present in the document. A weight value 0 associated with a term implies that the term is not at all important to the particular document. In other words, a document is relevant to a particular domain only if it contains important semantic elements.

The significance is computed using a relevance function. Relevance ranking allows the salience of each term to be assessed relative to the document collection as a whole.

Relevance function R(d) for web documents is defined as (Yuefeng Li and Ning Zhong, 2006; Xiaohui tao *et al.*, 2011):

$R(d) = \Sigma pr_{\beta}(t)\ \rho(t, d)$

where $\rho(t, d) = 1$ if $t \in d$, otherwise 0

The relevance function is based on the following two dimensions:

- Exhausitivity: It describes the extent to which pattern or topic discusses what users want.
- Specificity: It describes the extent to which the pattern or topic focuses on what users want.

A pattern (P) is a set of term frequency pairs. Each term or keyword is an individual semantic element. A measure support (P) is used to describe the extent to which the pattern is discussed in the set of documents under consideration (Yuefeng Li and Ning Zhong, 2004). The greater the support is, the more important the pattern is.

$pr_{\beta}(t)$ is directly propositional to support (P).

β is a mapping that explicitly describes the relationship between patterns and the common hypothesis space.

Clustering Web Documents

Clustering of Web documents precede with preprocessing textual documents. It is the process of collecting Web sources into groups so that similar objects are in the same group and dissimilar objects are in different groups. It has been proposed based on the idea of identifying homogeneous groups of web documents. A neural network approach to clustering is proposed (Hector Oscar Nigro, 2007, Margaret H Dunham and Sridhar, 2007).

It represents each cluster as an exemplar. An exemplar acts as a prototype of the cluster. New objects can be distributed to the cluster whose exemplar is the most similar. Similarity can be computed using a similarity function.

It is unsupervised learning and deals with instances which have not been pre-classified in any way. It organizes a set of objects in a multidimensional space into cohesive groups, called clusters. The scope of applying clustering algorithms is to discover useful but unknown classes of items. It is an approach of learning where instances are automatically placed into meaningful groups based on their similarity.

Clustering is similar to classification. But unlike classification groups or clusters are not predefined (Jiawei Han and Micheline Kamber, 2005). In classification data items are assigned to a predefined category based on a model that is created from preclassified training data (supervised learning).

The goal of clustering is to separate a given group of data items (the data set) into groups called clusters. It emphasizes that items in the same cluster are similar to each other and dissimilar to the items in other cluster. In clustering methods no labeled examples are provided in advance for training. It is also called unsupervised learning).

The proposed approach is based on the vector space model. The similarity measure is the relevance factor. The clustering algorithm applied is SOM based clustering based on neural network approach of clustering.

To represent text and web document content for clustering a vector-space model is used. The process of clustering documents begins with selecting the type of the characteristics or attributes (e.g., words, phrases or links) of the documents on which the clustering will be based and their representation. Clustering is then performed using as input the vectors that represent the documents.

In vector space model a document is represented as a vector of the terms that appear in all the document set. Each term in a document becomes a feature dimension. Each feature vector contains term weights of the terms appearing in that document. The term weighting scheme is usually based on TF-IDF method in IR (Hector Oscar Nigro, 2007).

The value assigned to each dimension of a document may indicate the number of times the corresponding term appears on it. It is a weight that takes into account other frequency information, such as the number of documents upon which the terms appear. This model is simple and allows the use of traditional machine learning methods that deal with numerical feature vectors in a Euclidean feature space.

The paper presents a SOM based clustering technique (Hector Oscar Nigro, 2007), Tenvo Kohonen, 1998). It is a neural network clustering

method. A vector representing the neuron is called neural vector. This vector has same number of dimensions as the input vectors. These vectors are inputs to the clustering process. To generate vectors from documents an indexer is required. Indexing is dependent on term extraction. Term extraction includes the process of stemming. A term represents a semantic element having meaning in the particular domain. A term can be a keyword in the document. The importance of a term is measured using Term Frequency (TF) or Term Frequency-Inverse Document frequency (TF.IDF) (Yuefeng Li and Ning Zhong, 2006; Hector Oscar Nigro, 2007). TF is the number of occurrences of the term in the document. The greater the value of TF the more important is the term. TF.IDF is an improvement over TF. An extension to TF.IDF measures is CF.IDF measure (Hector Oscar Nigro, 2007). CF is the sum of the TF for all terms representing the concept. A concept is retrieved from the web document by locating terms or keywords present in the document.

CONCLUSION AND FUTURE WORK

The paper is an attempt to overcome the limitations of overload and mismatch on WWW. The objective is to retrieve the most useful and relevant web documents from the available huge repository.

The proposed solution focuses SOM based clustering of web documents using ontology. It clusters set of web documents into classes or groups. Clustering web documents separates unrelated pages and clusters related pages (to a specific topic) into semantically meaningful groups. It is useful for summarization, organization and navigation of unstructured Web pages. Each cluster is semantically similar on the basis of terms. A relevance function computes similarity measure. A document is relevant to a particular domain if it contains terms or keywords essential in describing the domain. The formal implementation of the proposed algorithm and its performance is in process. We will implement the approach for designing a domain specific web search engine. It will contrast general search engines that index large portions of WWW.

REFERENCES

1. Baeza Yates R and Ribeiro Neto B (1999), *Modern Information retrieval,* Addison Wesley.
2. Ernesto William De Luca (2006), "*Andreas Nürnberger, "Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation*", *International Journal Of Intelligent Systems*, Vol. 21.
3. Hector Oscar Nigro (2007), "Sandra Gonzalez Cisaro, and Daniel Hugo Xodo, Data Mining with Ontologies-Implementations, Findings, and Frameworks", IGI Global, ISBN 978-1-59904-618-1
4. Jiawei Han and Micheline Kamber (2005), Data Mining Concepts and Techniques, ISBN 81-8147-049-4
5. Juha Vesanto and Esa Alhoniemi (2000), "Clustering of the Self-organizing Map", *IEEE Transactions on Neural Network*, Vol. 11, No. 3.
6. Li Y and Zhong N (2004), "Web Mining Model and Its Application on Information gathering," *Knowledge Based Systems*, Vol. 7.
7. Liu and K Chang (2004), *SIGKDD*

Explorations, Special Issue on Web Content Mining, Vol. 6, No. 2.

8. Margaret H Dunham and Sridhar S (2007), "Data Mining–Introductory and Advanced Topics, Second Impression", ISBN 81-7758-785-4
9. Richard J Roiger and Michael W (2005), "Geatz, Data Mining-A Tutotrial-based Primer, First Indian Reprint", ISBN 81-297-1089-7.
10. Tenvo Kohonen (1998), *The self organizing map*, Elsevier.
11. Wanlong LI, Dayou Liu, Shanhong Zheng and Suyun Jiao (2006), "A Novel Computational Approach to Concept Semantic Similarity", International Conference on Computer, Mechatronics, Control and Electronic Engineering.
12. Xiaohui tao, Yuefeng Li and Ning Zhong (2011), "A Personalized Ontology Model for Web Information Gathering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 4.
13. Yuefeng Li and Ning Zhong (2004), "Capturing Evolving patterns for Ontology based Web mining", International Conference on Web Intelligence.
14. Yuefeng Li and Ning Zhong (2006), "Mining Ontology for Automatically Acquiring Web User Information needs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 4.

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com