Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Perusal for Information Resurgance Using Semantic Web

[1]Rubavathy. P, [2]Shilpa.S
[1]Assistant Professor, [2]Student,
[1,2]Department of Computer Science, St. Joseph's College of Arts and Science for Women, Hosur, Tamilnadu, India

Abstract - The Information retrieval systems based entirely on syntactic contents have serious limitation. In order to achieve high precision and recall on information systems, the incorporation of natural language processing technique that provide semantic information is needed. For this reason by determining the semantic for the constitutes of documents a clustering method is presented in this paper. The goal is to find the conjoined point which can combine the both textual part and visual part and to use for information system. It can help to well extract the meaning of a term.

Keywords *: Semantic, NLP, Clustering, Ontologies*

I. INTRODUCTION

The goal of Information Retrieval system is to search the documents which user wants to obtain in a fast and efficient way. If there is an information retrieval system that comprehends the semantic content of documents and reflects the preferences of user, it can be very helpful to search the information on the internet and to improve the performance of the existing system.

Because the Information retrieval system must interpret the contents of information items or documents in a collection and rank them according to the degree of relevance to the user query, the representation of document contents and the user preference are important factors in the retrieval process. Semantic web process and clustering provides a common framework that allows data to be shared and reused across applications, enterprise etc.

II. SEMANTIC WEB MINING CONCEPTS

The Semantic Web can be defined as an extension of the current web. Here the information is presented in a well-defined manner, better enabling computers and people to work in cooperation. Data in the Semantic Web is defined and linked in a way that can be used for more effective discovery, automation, integration and reuse across applications. This data can be shared and processed by automated tools as well as people.

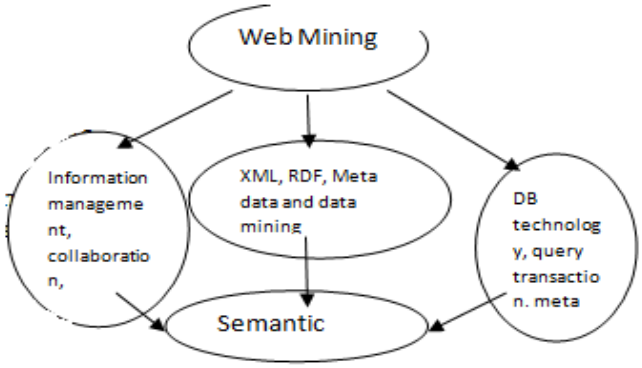

Figure 1: Semantic Web Mining Technologies

The Semantic Web will provide an infrastructure that enables not just web pages, but databases, services, programs, sensors, personal devices, and even household appliances to both consume and produce data on the web. Semantic web mining is essentially mining the information pertaining to the semantic Web. This means mining Web pages so that the machine can better understand the information. It also means mining the data sources to develop an effective semantic Web.

III. INTEGRATING DOMAIN KNOWLEDGE INTO WEB MINING

Web mining is the process of discovering and extracting useful knowledge from the content, usage, and structure of one or more Web sites. Semantic Web mining involves the integration of domain knowledge into the Web mining process. Domain knowledge can be integrated into the Web usage mining, and personalization. On the other hand, data mining tasks can also help to enhance the process of domain knowledge discovery. Domain knowledge can improve the accuracy of document clustering and classification and induce more powerful content patterns. For example, in domain ontologies are employed in selecting textual features. The selection is based on lexical analysis tools that map terms into concepts within the ontology. The approach also aggregates concepts by merging the concepts that have low support in the documents. After preprocessing, only necessary concepts are selected for the content clustering step.

Traditional approaches to content mining and information retrieval treat every document as a set or a bag of terms. Without domain semantics, we would treat "human" and "mankind" as different terms, or, "brake" and "car" as unrelated terms. In a concept is defined as a group of terms that are semantically relevant, for example, as synonyms.

With such concept definitions, concept distribution among documents is analyzed to find interesting concept patterns. For example, one can discover dominant themes in a document collection or in a single document; or find associations among concepts.

IV. CLUSTERING IN RETRIEVING

Clustering discover grouping of domain objects. Clustering in this task of grouping a set of objects in such a way that objects in the same group are more similar(in some sense or another) to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

V. MULTIMEDIA RETRIEVAL USING CLUSTERING AND SEMANTIC WEB

A. Text Retrieval

The traditional keyword-based search uses annotations of content to find the relevance's. The classic technique Vector-Space Model (VSM) is a combination of annotation weighting and ranking algorithm. It uses tf-idf model, term frequency (frequency of a term in document) and inverse document frequency (frequency of a term across the collection) for giving weight to the terms. Based on the relevance of information to the user queries, the results are ranked. The problem with KeyWord based search is complexity between words and concepts due to synonymy (different words denote same content) or homonymy (same word denote different concepts). But in semantic web the search is not only based on the keyword and also based on the meaning of keywords (semantics).

The additional information such as type, structure, relations, classification, and rules about the concepts which are referenced in the documents, are represented in ontology based knowledge base which returns the instances against the user queries. The documents are semi automatically annotated with the instances. They assigned to a weight that reflects how relevant the instance is considered for the document semantic. Predicate describes the relationship between subject and object. For retrieving the data related to user queries, the relevance between document metadata and RDF triples measured by concepts similarity, relations similarity and RDF triples similarity.

The SPARQL query language for RDF used to retrieve data according to the relations of synonym, inheritance and part-of relations between the concepts from the large data sets. A new structure of ontology which has 6-tuples (atomic term, complex terms, instances, instances description, attribute assignment and axioms) has been proposed for semantic information retrieval.

B. Image Retrieval

Image retrieval goes through three kinds of retrieval techniques including text-based retrieval, content-based retrieval and semantic based retrieval. Most of the text-based image retrieval system requires the images to be annotated manually. In text-based retrieval, images retrieved through matching the user queries and annotated keywords. The content-based image retrieval system uses low-level visual features like colour, texture, shape and location, etc. for retrieving the images. These low level features are extracted from images automatically. The semantic-based image retrieval technique is the combination of low level features and semantic of the images.

Initially proposed image retrieval system based on Natural language Processing (NLP) approach. This processing model integrates WordNet, online lexical information system and low level visual features. A hybrid model which comprises ontology reasoner and Bayesian Network (BN). The ontology reasoner retrieves all the neighboring information of the input keyword and the BN model used to calculate the relevance between input keyword and neighboring information. So this model not only retrieves the target image but also retrieves the images having neighboring information about the keyword. Three sub ontologies such as Domain Ontology, Textual Description Ontology and Visual Description Ontology constitute multi-modality ontology. Later, the new ranking mechanism based on semantic similarity between the concepts used to improve the results. Chang and Huang proposed the method of partitioning each image in the different major categories into suitable regions with the help of novel fuzzy segmentation algorithm.

Based on the Discrete Cosine Transform (DCT) coefficients images classified into different major categories. The regions derived from images in each individual major category are classified into different sub-categories based on low level features. Then the image search carried through these sub-categories. Fuzzy Domain Ontology (FDO) introduced in. FDO has been constructed by adding fuzzy membership to the concepts and relationship between them in domain ontology, which is used to describe the high-level semantic features. In addition, the Classifier Ontology maps the low-level visual features to high-level semantic features and semantic ontology describes semantics of images in the image database. Liua, Shaob and Liua presented an image retrieval technique which accepts image as input and will construct the image ontology with the help of SIFT (Scale Invariant Feature Transform) features. SIFT used to extract the low level features of the image. The retrieval technique will calculate the feature similarities between input image and images in ontology. It also considers the similarity of semantic features and concepts.

C. Video Retrieval

Mostly video files annotated by a large number of tags in archives. So, in text-based video retrieval users may face difficulties to find the exact video. Content-based video retrieval is used to find the videos efficiently for users. In this method, low-level visual feature extraction, shot detection and object recognition are the most important phases. Pictorially enriched ontologies are based both on visual and linguistic concepts as described in , where visual features extracted from MPEG videos.

Annotation is performed automatically on selected video clips associating occurrences of events or entities to high level concepts, by checking their similarity to visual concepts that are hierarchically linked to higher level semantics. Based on four dimensions such as Action, Location, Time and Shooting technique events are modelled as in. Combination of low-level visual features called as semantic patterns, which is used to characterize concepts in the above four dimensions. Video is transformed into a 20-Dimensional multi stream by sequentially aggregating low-level visual features from all shots. The video ontology further refined by extracting new semantic patterns from subspaces of videos which cannot be retrieved by previously extracted patterns. The authors also noted complex events can be retrieved by combining semantic patterns in different dimensions.

In this model, each image/video is represented by its semantics and low level features. The semantic tolerance relation model defines the semantic categories of image/key-frame which represent the segment of video, together with the tolerance degrees between them. In this model, the semantics of each image/key-frame is represented by the class weighted vector. These weights assigned based on classification result to the defined classes. The Bayesian Classifier is used to classify the images in this model.

Using WordNet, the semantic linguistic relations between the concepts is determined for defining the ontology schema. Then the concept detectors are linked to the corresponding concepts in ontology. The rules defined in Semantic Web Rules

Language (SWRL) automatically, learn the knowledge embedded in the ontology.

D. Audio Retrieval

In this digital media era, audio plays significant role among the users because of its expressiveness. Most of the audio information retrieval system compares the audio domain with the features extracted from the audio signals. This model consists of several tasks such as Segmentation, Meta data acquisition, Selection, Scheduling, etc. The segmentation phase used to identify the boundaries of audio objects, which is defined by five tuple such as Identifier, Start time, End time, Description and Audio data. The accuracy of audio information retrieval is based on the generation of metadata. Both fully automated content extraction and selected content extraction approaches have been used for metadata acquisition. Metadata has the description of audio with boundary. Based on the user's query the audio information will be retrieved. Query-by-example audio retrieval system based on semantic similarity proposed in. The user provides an audio example instead of semantic description and the system returns audio content that is similar to the query. Unordered set of feature vectors are extracted from the audio signals to represent the audio tracks. There are two kinds of approaches used in this audio retrieval system. First one is query-by-acoustic example (QBAE) which retrieves the audio acoustically similar to the query. Second approach is query-by-semantic example which retrieves the audio based on the semantic information. An integrated system introduced in , which can be used for text based retrieval, content-based query-by-example and automatic annotation. It uses hybrid network that linking the sounds through a measure of perceptual similarity and linking the semantic tags through user provided weights or lexical libraries which helpful for improving the retrieval of audio information.

CONCLUSION

This paper discussed about various retrieval techniques used for multimedia using ontology for the semantic web. It is focused mainly towards text, image, audio and video retrieval. It is observed that semantic-based search improves the performance of the retrieval of multimedia in all the ways. This paper includes important techniques of multimedia retrieval among more techniques. However, it provides retrieval mechanisms from the past to present which are crucial. the approach of semantic retrieving for web documents in certain domain, extracting relevant information based on the semantic web.

References

[1] D. Manas, C. Hasan, S. Debakar, A. Khandakar. (2010). Focused Web Crawling: A Framework for Crawling of Country Based Financial Data. 978-1-4244-6928-4/10, IEEE.

[2] H. Debashis, S. Biswajit, K. Amritesh. (2010). Adaptive Focused Crawling Based on Link Analysis. 201O 2nd International Conference on Education Technology and Computer (ICETC). 978-1-4244-6370-1, IEEE.

[3] H. Rui, L. Fen, S. Zhongzhi. (2008). Focused Crawling with Heterogeneous Semantic Information. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-3496-1/08, IEEE.

[4] J.F. Allen, "Maintaining Knowledge About Temporal Intervals," Comm. ACM, vol. 26, no. 11, Nov. 1983, pp. 832-843.

[5] M. Bouet et al., "Powerful Image Organization in a Visual Retrieval System," Proc. 6th ACM Int'l Multimedia Conf. (ACM Multimedia 98), ACM Press, New York, 1998, pp. 315-322.