ISSN No. 0976-5697

Volume 7, No. 5, September-October 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Clustering and Labeling of Images under Web Content Mining

C.Menaka M.C.A., M.Phil,
Research Scholar,
Bharathiar University, Coimbatore,
Tamilnadu,India

Dr. N. Nagadeepa., Principal,
Karur Velalar College of arts and science for women,
Karur, Tamilnadu,India

Abstract: In internet images plays vital role for users. Everyday numbers of users are trying to access and manipulating the images very efficiently. From web we can see images in variety of formats. Depends upon the user requirements the request may fulfilled. This paper focuses the overview of clustering of images and giving annotation for specific images which makes more benefit to users who are really in need.

Keywords: clustering, labeling, mining

I. INTRODUCTION

The term internet is more essential for humans nowadays. In earlier days books are the only source to obtain knowledge and also for enhancing it. Today internet replaces it for acquiring knowledge. Also it helps to enhance our skills at the maximum. The basic thing behind this is retrieving [1] information from web. We can say web is extracting information from which comes under the category of web content mining [2] which is a type of web mining. It is necessary to know what it is. Web mining is the process of extracting knowledge by applying data mining algorithms. Again this can be divided into three categories: Web usage mining, Web structure mining and web content mining. The term web usage mining refers the details of log who is logging into the site for retrieving information [1][2].

Web structure mining gives the study of format what they have used to create such web site using scripting languages. Web content mining is more important than the other since it holds variety of structured, semi-structured and unstructured data.

II. WEB MINING

The term Web-Mining refers to computational processes that aim to discover useful information or knowledge from data on the Web. Based on the primary kinds of data used in the mining process, web mining can be subdivided into the three non-disjoint types: 1) Web structure mining, 2) Web content mining and 3) Web usage mining. Web structure mining discovers knowledge from the hyperlink structure of the Web. Web content Mining extracts useful information from page contents from
page contents and web usage mining extract knowledge from the usage patterns that people leave behind as they interact with the web .Based upon its use mining can be decomposed into various following subtasks.

1. Resource finding
2. Information selection and preprocessing
3. Generalization
4. Analysis

III. WEB CONTENT MINING

Web content mining is somehow different from data mining and text mining. It is related to data mining the reason is because many data mining techniques can be applied in Web content mining. Moreover it is related to text mining because most of the web contents are texts. Web data may be semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires some of various creative applications of data mining and/or text mining techniques. Traditionally, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems. Web content mining problems and discuss existing techniques for solving these problems. Some other emerging problems will also be surveyed.

Data/information extraction: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

Web information integration and schema matching: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

Opinion extraction from online sources: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.

Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain..

Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

IV. CLUSTERING

Here particularly, Web image classification is a most challenging task in image processing. There are number of techniques used in the existing systems to classify the images in web.

Clustering analysis is an important aspect in web mining research. A widely adopted definition of optimal clustering is a partitioning that minimizes distances within a cluster [4] and maximizes distances between clusters. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the data to be clustered, and the data are subsequently assigned to those clusters. It made huge progress in such aspects as image texture segmentation, remote-sensing image segmentation and edge detection etc.,

A cluster is a list of data objects. Algorithms [5] for clustering helps to categorize data organize data. Cluster analysis can be used as a standalone data mining tool to gain insight into the distribution of data. Various clustering techniques:

A. Text based Clustering:

The text-based web document clustering approaches characterize each document according to its content, i.e. the words contained in it (or phrases or snippets). The basic idea is that if two documents contain many common words then it is very possible that the two documents are very similar. The approaches in this category can be further categorized according to the clustering method used into the following categories: partitional, hierarchical, graph based, neural network- based and probabilistic algorithms [6].

B. Partitional Clustering:

Partitional clustering algorithms are divided into iterative or reallocation methods and single pass methods. The most common partitional clustering algorithm is k-means, which relies on the idea that the center of the cluster, called centroid, can be a good representation of the cluster. The algorithm starts by selecting k cluster centroids. Then the cosine distance between each document in the collection and the centroids is calculated and the document is assigned to the cluster with the nearest centroid. Then the new cluster centroids are recalculated and the procedure runs iteratively until some criterion is reached. Many variations of the k-means algorithm are also proposed, e.g. ISODATA and bisecting k-means [8].

C. *Hierarchical Clustering*:

Hierarchical clustering algorithms produce a sequence of nested partitions. Usually the similarity between each pair of documents is stored in a n x n similarity matrix. At each stage, the algorithm either merges two clusters (agglomerative methods) or splits a cluster in two (divisive methods). The result of the clustering can be displayed in a tree-like structure, called a dendrogram, with one cluster on the top containing all the documents of the collection and many cluster on the bottom with one document each.. Almost all the hierarchical algorithms used for document clustering are agglomerative (HAC).[2][4]

D. *Graph Based Clustering*:

The documents to be clustered can be viewed as a set of nodes and the edges between the nodes represent the relationship between them. The edges bare a weight, which denotes the degree of that relationship. Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge-cut, i.e. the sum of the weights of the edges that are cut, is minimized. Since each edge in the graph represents the similarity between the documents, by cutting the edges with the minimum sum of weights the algorithm minimizes the similarity between documents in different clusters[7][8].

E. *Neural Network based Clustering*:

The Kohonen‟s Self-Organizing feature Maps (SOM) [10] is a widely used unsupervised neural network model. It consists of two layers: the input layer with n input nodes, which correspond to the n documents, and an output layer with k output nodes, which correspond to k decision regions. The input units receive the input data and propagate them onto the output units. Each of the k output units is assigned a weight vector. During each learning step, a document from the collection is associated with the output node, which has the most similar weight vector. The weight vector of that „winner‟ node is then adapted in such a way that it will become even more similar to the vector that represents that document [7][8].

F. *Fuzzy Clustering*:

All the above approaches produce clusters in such a way that each document is assigned to one and only one cluster. Fuzzy clustering approaches, on the other hand, are non-exclusive, in the sense that each document can belong to more than one cluster. Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that a document can belong to more than one cluster is described by a membership function. The membership function calculates for each document a membership vector, in which the *i*-th element [7][8] indicates the degree of membership of the document in the *i*-th cluster.

G. Probabilistic Clustering:

Another way of dealing with uncertainty is to use probabilistic clustering algorithms. These algorithms use statistical models to calculate the similarity between the data instead of some predefined measures. The basic idea is the assignment of probabilities for the membership of a document in a cluster. Each document can belong to more than one cluster according to the probability of belonging to each cluster.

V. TYPES OF CLUSTERING

A. k-means clustering :

It identifies the mutual exclusive clusters of spherical shape. It uses statistical methods to assign rank values to the cluster. It organizes objects into k-partitions, each represents a cluster.

B. Hierarchical clustering :

It decomposes the given set of data objects. in this tree of clusters called dendogram. Compute the distance between the new cluster and each of old clusters. It can be further divided into two types') Agglomerative and 2) Divisive [9][10]

C. *DBSCAN* – Density Based Spatial Clustering of Applications with Noise.

Based on the density of nearby objects it grows. Here number of point parameter impacts detection of outliers. It uses DENCLUE algorithm for low-dimensional data.

D. OPTICS – Ordering Points to Identify Clustering Structure. It is used to detect the meaningful clusters on data of varying density. OPTICS abstracts from DBSCAN by removing this each point is assigned as core distance.

E. STING – Statistical Information Grid

It is a grid based multi resolution technique, Attributes regarding grid cell like mean, maximum and minimum values stored as statistical parameters. Quality depends on the granularity of the lowest level of grid structure.

VI. METHODOLOGY PROPOSED :

The following procedure is used for downloading and the creation of clusters and then process of retrieving images.

Algorithm 1:

- Enter the URL.
- Remove noisy data in web page
- Display the images.
- Create a cluster and label the images.
- Move the retrieved images into cluster.
- Content extraction techniques are applied here.
- Do the image search by entering image as input.
- Check with various clusters
- Display the result with label.
- Repeat the process for n number of searches.

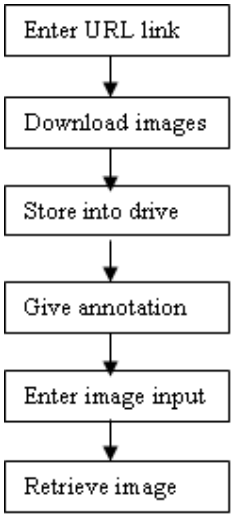

Fig .1: Annotating images

Labeling images:

The reason for labeling the images is to identify the images soon and on category wise. It helps the user for efficient manipulation. Also it avoids repetition of images and its information.

The main benefit is name of the label is also used for search category.

Procedure:

1. Input:

The label set $X = [X_1, X_2, X_3 \ldots X_n]$ of training data is considered. Matrix can be formed Y and Query image z.

2. Formula for solving Co-efficient vector taken is pq of the query image with z over Y can be obtained by solving the equation

3. Image annotation Obtain the label vector $X^{,}$ of the query image z.

Proposed method consists of the following modules:

Module 1:

1. Enter the web link as user input.
2. Web link related to cartoon images
3. Removing noisy information like banners, ads etc.,
4. Display the images

Module 2:

1. Create various numbers of clusters.
2. Depends upon the cartoon characters number of clusters can be created.
3. Store the image of cartoons into the respective clusters.

Module 3:

1. Develop a tool for searching the image which is stored.
2. Implementation Technology used here is ASP.NET.

Module 4:

1. Enter the input as image
2. Search over the clusters
3. All the images which is related to the entered query are retrieved and displayed.

VII. CONCLUSION

Users are fond of manipulating images from web. Various techniques have been proposed traditionally. In this paper proposed method exhibits the purpose of using specific image category like cartoon images with its label and its description. It helps to reduce the search time and processing speed can be increased. To experiment this only limited set of mostly used pictures can be considered as training set data.

VIII. REFERENCES

[1] Robert Cooley, Bamshad mobasher, jaideep srivatsasva, web mining information and pattern discovery on the "WWW".

[2] Mary carvin,"Data mining and the web : What they can do together.

[3] Ji Zhang Wynne Hsu Mong Li Lee "Image mining: Issues, Frameworks and Techniques.

[4] Aura conci., Everest mathias M.M castro "Image mining by color content"

[5] B. Xu, J. Lu, &G. Huang, A constrained non-negative matrix factorization in information retrieval, Proc. 2003 IEEE Int. Conf. on Information Reuse and Integration, Las Vegas, NV, 2003, 273-277.

[6] Y. Wang, Y.Jia, C.Hu, & M. Turk, Fisher nonnegative matrix factorization for learning local features, Proc. 6th Asian Conf. on Computer Vision, Jeju Island, Korea, 2004.
[7] D. Guillamet, J. Vitri`a & B. Schiele, Introducing a weighted non-negative matrix factorization for image classification, Pattern Recognition Letters, 24(14), 2003, 2447-2454.
[8] S.Z. Li, X.W. Hou, H.J. Zhang, & Q.S. Cheng, Learning spatially localized, parts-based representation, Proc. 2001 IEEE Conf. on Computer Vision and Pattern Recognition, Kauai, HW, 2001, 207-212.
[9] V. Vapnik, The nature of statistical learning theory (Berlin: Springer-Verlag, 1995).
[10] K. Yu, L. Ji, & X. Zhang, Kernel nearest-neighbor algorithm, Neural Processing Letters, 15(2), 2002, 147156.