Lexical Tag Parsing, Contour Filter Refine and Multilevel Annotation Techniques for Extracting Relevant Cartoon Images

C. Menaka M.C.A, N. Nagadeepa

***Abstract:** Many number of techniques are used in the existing systems to classify the images in the process of web image classification. In this work, proposed technique considers two HTML tags namely alt and src. In a group of web pages these tags are taken into account to download the images. Mainly this approach considers the cartoon image category web link then images can be extracted and stored. LTP techniques is applied here to parse the given tags. Images are clustered and stored in their respective folders as per the category after clustering process. CFR algorithm is used here to refine the images for storing. MIA technique is applied here to give annotation for all images which is in the cluster for best retrieval. Finally based upon the given input as image resultant image can be searched from various available clusters and return to the user along with its detailed description.*

***Keywords:** Image clustering, LTP, MIA, CFR, Image annotation, SIC.*

I. INTRODUCTION

Internet is a global network. It is providing a group of information and communication facilities to the Internet users. Web consists of huge heterogeneous and less structured data. Mining this data is one of the challenging tasks [1]. Content extraction of web is an important issue. Since it is composed of noisy data such as advertisements, navigation links, TOC, copyright statements, service catalogs etc., Most of the applications can get benefit through the extracted content such as document classification indexers, crawlers and IR. This research aims to focus the images retrieved from web page [2]. Various cartoon images are used here for the research. set of cartoon images are retrieved from web page by removing the noisy data and non-relevant information of web techniques. Clusters are created to store and retrieve images [2][3].

II. OBJECTIVE OF THE RESEARCH

1. To ensure downloading size of the images can get reduced.
2. To obtain the less data size can be less by considering only object from the origin.
3. To acquire the considerable bandwidth.
4. To avoid the irrelevant data.

Manuscript published on 30 August 2017.
***Correspondence Author(s)**

C. Menaka M.C.A., M.Phil., Research Scholar, Bharathiar University, Coimbatore (Tamil Nadu), India.
Dr. N. Nagadeepa, Principal, Karur Velalar College of Arts and Science for women, Karur (Tamil Nadu), India.

III. STATEMENT OF PROBLEM

Web based image classification technique is to classify cartoon characters for famous cartoon websites.

Image classification itself a sensitive job in image processing whereas online image processing double the complication. Basically text based search is used for retrieving images based on filename and description [3].

Moreover while downloading images the web page may give irrelevant data, noisy information, and redundant data. If this information are downloaded, data size may be greater since it consumes more downloading capacity. Also size of bytes may acquire more and search time is more. In general, CBIR (content based image retrieval) approach is used for image search. It can be done either by text based or content based. In the form of text based search manual annotation must be given [4]. Once it is known particular image can be easily searched, but always it is not possible to get accurate images. In case of content based search it is more effective for feature extraction and also it is less expensive. Related images can be retrieved in one search. It find difficult in identifying the human face and some specified characters [6].

Moreover, to achieve accurate output continuous filtering process should be done. Once the image is downloaded, the object can be stored along with its boundary it is not necessary to use the boundary region for future use where as only the object is considered and going to use. So here it implies that the specific object with its region consumes more memory size, increased bandwidth and speed. As said, earlier CBIR is useful for feature extraction which includes color, texture properties and shape feature but it is inefficient to find the desired outcome [3][4].

IV. EXISTING SYSTEM

Clustering is being considered as a technique for image analysis. MSA clustering a non-parametric method to find arbitrary shape clusters in the feature space of an image. Main use of this technique is improving the classification accuracy. Based on MSA, re-clustering technique is used. It follows two step processes 1. Original image into homogeneous image segments. 2. Cluster the image segments to obtain result [1]. Most of the clustering algorithms based on 1.Hierarchical (Either aggregates or divides) 2.Partitional (clusters) splits into subsets. Moreover, in K-means k-points are generated randomly for k-clusters.

Here data points are assigned to the closest cluster. Centers of the clusters are recomputed .Each step points are re-assigned. If there is no more movements the process get stopped.

Issue here is to find the number of clusters k. Here the same MSA is applied to improve the classification accuracy [2][3]. Even it produces more accuracy compared to traditional methods still it does not bother about the irrelevant, noisy and redundant data.

Commonly used search technique is CBIR and ISRC.1. CBIR is one of the query based search method so that the search analyzes the contents of the image apart from metadata such as keywords, tags or descriptions which is associated with the image [4]. Content refers colors, shapes, textures or any other information that can be derived from the image itself. Annotation is given manually here to retrieve images from large set of databases. Even this search is subjective and has not been well defined [5].

2. ISRC – Image on web has become one of the most important information for browsers. The large number of results retrieved from image search engine increases the difficulty in finding intended images. ISRC may produce solution for this approach. There is no annotation based search even it consider both textual and visual features.

Disadvantages:

1. Noisy information cannot be eliminated.
2. Also extract Irrelevant data
3. Redundant data
4. Continuous filtering need to acquire accuracy
5. Storage space may be more since the image contains more resolution

V. PROPOSED SYSTEM

A. Lexical Tag Parsing:

Algorithm 1: Lexical Tag Parsing

Image tag analysis is to parse the src tag and the plain text analysis detects the alt tag to detect the labels of the image. It can be defined by

LTP(x) where x is the webpage
T – Number of Tokens in webpage
k – Number of tags in webpage
s – <src> tag
a - <alt> tag
C[n]- Cluster array
Img_{src} - cluster named folders
img_k = {} , a set to contain image in it.

```
Parse each word in x and to a Bag of words set $W_b$.
initially $W_b$ ={}
for n ← 1 to T
do for i ← 1 to k
if (i==s)
$Img_k$ <- add(s)  // add image into the set
If (i==a)
C[n] ← 1            // keep track of active cluster
        Create folder $Img_{src}$ named with $Img_k$
  for each image in Imgk
    for j  1 to N
do $Img_{src}$ ← SIM($Img_{src}$, $Img_k$)
        $Img_k$ ← add(k)
        C[n] ← 0 (deactivate cluster)
```

Here the SIM () function is used to find the similarity of the folder name and the image name. If both are similar then the image will be stored in that folder.

B. Contour Filter Refine Method:

Contour tracing is a preprocessing technique performed on digital images to extract image properties like shape and structure. Once the properties are extracted, it's different characteristics can be obviously examined and used as features which can be later used in pattern classification techniques. So accurate tracing of the contour will produce more accurate features which will increase the chances of classifying a given pattern more accurately. So we analyze the edge histogram, color layout, Texture properties of the images in order to find the similarity between two images [4]. These standard properties of images can be used to cluster images by performing step by step approach. First the filtering of images on color properties of a particular cluster will be made so that irrelevant images to that cluster will be pruned out. After that refine method will be done so that the most matching images will be grouped with that particular cluster. Traditional image based search and clustering techniques mainly focused on content based as a whole. So it will be little time consuming as individual image will be compared with the entire image to find its similar image [6].

Initially pruning of image will be made using

Algorithm 2. Filter by Refine

Input: $Img_{src,,}$ Img_k
Output: Img_{src}^* (containing clusters of similar images)

```
Description:
$Img_{src}^*$ = 0
for each image in $Img_k$ do
$M_{MS}$ = MaximumSimilar($Img_k$);
if |$M_{MS}$| > $\beta$ then
add $Img_k$ to $Img_{src}^*$
else
discard $Img_k$
end if
end for
return $Img_{src}^*$;
```

C. Multilevel Image Annotation:

The main objective of this technique is to divide the query images into clusters and assign few labels to each cluster. That is, we give each cluster with a few representative labels and make the labels unique across clusters.

There are many possibilities for one type of images to be grouped into two different clusters. In order to address that issue we are making this factorization so that detailed information can be acquired by giving more than one label description for a cluster. So on comparing the labels defined in each clusters and the label of the image that is being downloaded we can get an idea that the image is well suited to be placed in this cluster itself [3][5].

D. Synergic Image Clustering (SIC)

Synergic image clustering is the synergistic final delivery of image [8], its cluster and the image's description. This technique is used to make an effective and efficient image search result clustering. SIC as an efficient technique to organize Web image search results into very accurate semantic clusters.

Different from all the existing web algorithms that can only cluster the top images using either visual or link features, our proposed technique first identifies several semantic clusters[11][12] related to the given image, then assigns all the resulting images to its corresponding clusters.

Our algorithm has three advantages over existing ISRC (Image Search Result Clustering) algorithms [3][4]. First the most important image groups can be found more accurately. Second, entire group of images will be taken into consideration in the clustering process instead of only a smaller part. And finally, our algorithm is efficient enough to be implemented in practical systems.

Given the cluster names, merging and pruning technique is utilized to obtain the final cluster names. First, we merged the same or very similar candidates from different set of sources [3][8]. Second, the description of the images is utilized to prune out the candidate cluster names of possibly unhelpful clusters. Finally, the resulting cluster names are utilized as queries to search a description of that cluster.

The cluster names with many or too few resulting images are first considered for analysis. The reduced thumbnails of top ranked images are used as representation images of the clusters. The byproducts, two problems of the existing web image search engines are solved to some extent by our algorithm. One is that with the existing image search engines, one kind of images tends to dominant the search results. For example, for query images on "Tom", this character can be in any color and structure and some other characters in cartoon can also be similar with this each other like color, layout or texture. So we need to synergize all the properties of the images to cluster them into exact classification. While with SIC as we are considering our MIA and CFR techniques, other cartoon characters, e.g. Jerry or Ben 10 could be easily rationalized and discarded by our application. The other limitation addressed is that for some general queries, especially those game related queries, e.g. Power puff, Sally and so on that are similar to cartoon will also be ranked very high. With SIC, the most related key phrases using our multilevel annotation could be found for each image could be filtered out and more relevant images could be grouped with the appropriate e classification.

VI. CONCLUSION

The proposed methodology has the benefit by using LTP, CFR, and MIA to acquire the better results to reduce the downloading size, to find accurate images while searching. Then it can produce the relevant images and it can find the exact object from its boundary region. So that bandwidth can be reduced. When compared to the common search techniques the developed methods provide maximum accuracy in search since the recent search engine still uses CBIR approach. Even it does not provide the desired outcome.

VII. RESULT AND DISCUSSION

Table 1: Data SET Comparison of Existing System and Proposed System

S. No		No. of images	Relevant	Irrelevant	ads	Banners					
	Name of the websites	Exis	Pro	Exis	Pro	Exis	Pro	Exis	Pro	Exis	Pro
1	www.nick.com	60	60	25	48	23	12	7	-	5	-
2	www.cartoonindia.com	54	54	24	46	25	8	3	-	2	-
3	www.chottabheem.com	48	48	11	41	28	7	5	-	4	-
4	www.disney.com	67	67	24	55	42	12	1	-	-	-
5	www.doraemon.com	28	28	22	22	6	6	-	-	-	-
6	www.cartoonnetworkasia.com	54	54	30	43	21	11	2	-	1	-
7	www.powerrangers.com	56	56	15	46	41	10	7	-	6	-
8	www.cartoonnetwork.co.uk	58	58	12	51	32	7	6	-	8	-
9	www.pokemon.com	59	59	13	53	32	6	8	-	6	-
10	www.disneyjunior.disney.com	54	54	5	47	34	7	7	-	8	-
11	Character.disney.in	46	46	13	35	23	11	-	-	-	-

12	www.disneyclips.com	38	38	11	32	27	6	-	-	-	-
13	www.kiddle.com	43	43	12	40	31	3	-	-		-

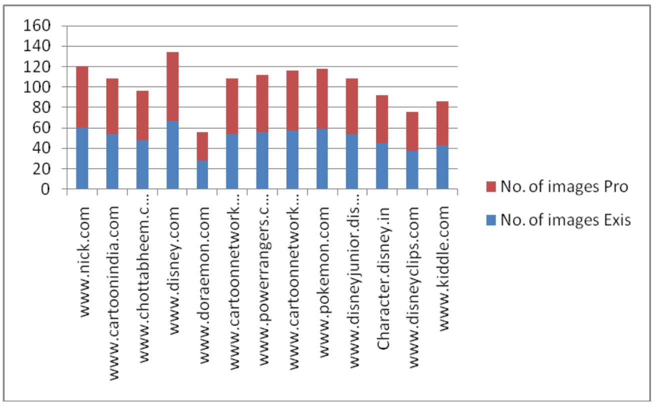

Fig.1 No. of Images Obtained in Proposed and Existing System

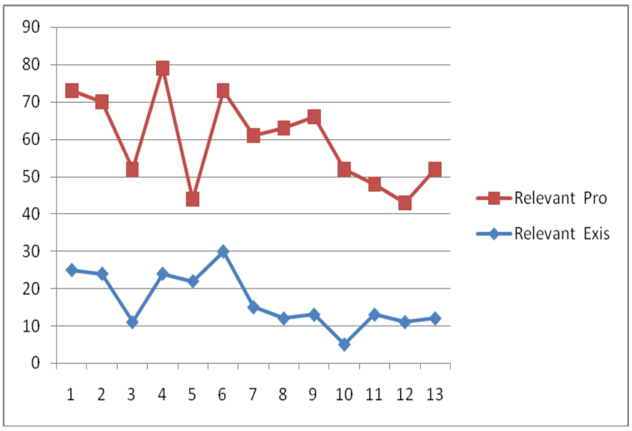

Fig.2 No. of Relevant Images Obtained in Proposed and Existing System

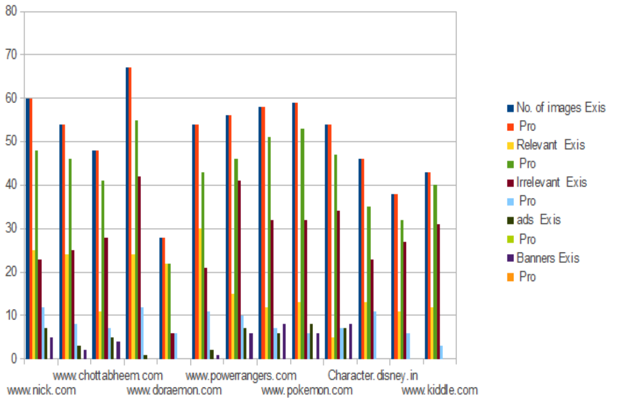

Fig.3 Overall Comparison Obtained from Proposed and Existing System

Table 2: Factors Comparison of Existing System and Proposed System

S. No	Factors	Existing system (%)	Proposed system (%)
1	Speed	65	95
2	Accuracy	57	93
3	Bandwidth	75	90
4	Storage capacity	34	90

Table 3: Algorithms

S. No	Existing system	Proposed system
1	Text – based approach	LTP
2	CBIR – content based approach	CFR
3	ISRC	MIA

REFERENCES

1. S. Xia, Z. Q. Xiang, Y. X. Zou* ADSPLAB/EL LIP, "Integrating Visual and Textual Features for Web Image Clustering", IEEE International Conference on Multimedia Big Data.
2. Shukui Bo, Yongju Jing ," Image Clustering Using Mean Shift Algorithm" Fourth International Conference on Computational Intelligence and Communication Networks.
3. A.Kannan, Dr.V.Mohan, Dr. Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques" IEEE International Conference on Computational Intelligence and Computing Research.
4. Felci Rajam and S.Valli, A survey on content based image retrieval, Life science journal 2013; 10(2) June 2013.
5. Allan hanbury, "A survey of methods for Image annotation", International journal of computer applications, Volume 37 No.6,January 2012.
6. C.Wang, F.Jing , L.Zhang, H.Zhang, "content based image annotation refinement ", proc. of CVPR,2007.