A NOVEL APPROACH FOR ANNOTATING IMAGES BY SEMANTIC SIMILARITY & KEYWORD BASED ASSOCIATION ANALYSIS

Payal Gulati,
Dept. of Computer Engineering, YMCA University of Science & Technology, Faridabad, Haryana, India
gulatipayal@yahoo.co.in
A.K. Sharma
Dept. of Computer Engineering, YMCA University of Science & Technology, Faridabad, Haryana, India

ABSTRACT

Tremendous growth of multimedia data has resulted in enhancing the number of images over the web. As a consequence of this, search engines face challenges in finding relevant images. Therefore, images must be properly annotated and effectively indexed for increasing relevancy. Image annotation has been a topic of ongoing research for more than a decade leading to several interesting techniques. However, retrieval results are not satisfactory because there is a semantic gap between low-level visual features and high-level semantic concepts. This motivates our work. In this work, a novel approach for finding relevant annotation based on semantic similarity and subsequently enriching the annotation by using keyword based association analysis is proposed.

Index Terms—Image Annotation, Page Manager, Association Rule Mining, Knowledge Base, Wordnet, Query Processor, Image Repository.

I. INTRODUCTION

The size of the web has already surpassed 13.86 billion pages [1] and yet there is no sign of leveling off; causing a challenge for image search engines in searching images over these pages. Image annotation or automatically annotating images with keywords is a solution to this problem. Images are annotated by extracting relevant keyword from page title, surrounding text, metadata, image caption etc. Proper annotations of images increase relevancy and efficiency of retrieval. Despite the continuous efforts in inventing new annotation algorithms, the annotation performance is usually not satisfactory due to limited vocabulary. Moreover, images over the web are not annotated with semantic descriptors, raising difficulty in searching for relevant images. In this paper, a novel approach for annotating images both by semantic similarity and keyword based association analysis is being proposed. This paper is organized in the following way. Section II discusses the related work done in this domain. Section III presents the proposed architecture for annotating images by semantic similarity and keyword based association analysis. Algorithm and flow diagram are discussed in section IV and V respectively. Implementation and comparison of the proposed and existing systems is presented in section VI. Finally section VII comprises of the conclusion.

II. RELATED WORK

Image annotation is to assign metadata with images based on visual features of the images and as well as the associated text [2]. It is an active research topic of recent years for understanding images and simplifying the web image retrieval process [3]. The metadata added to images allows effective searches. Manually adding metadata is an extremely laborious and time-consuming task [4]. To overcome this, many researchers have developed an approach for automatically annotating images [5, 6, 7] by using text from the image tag contained in image file and the text surrounding the image. Content descriptive and content dependent information can be associated with the images.
In Content-descriptive annotation, images are annotated by extracting relevant keyword from page title, surrounding text, metadata, image caption, alt tags etc. However indexing embedded image by image caption can be quite confusing as image caption may not always relate with image semantics. For instance an apple image can be captioned as image1.gif on the web page. Also, extraction of information from the surrounding text is somewhat heuristics as it is difficult to determine which area is more relevant. Despite the continuous efforts in inventing new annotation algorithms, the annotation performance is usually not satisfactory due to limited vocabulary leading to inaccurate image annotations.

Content-dependent metadata consists of color histogram features, texture features calculated by different algorithms etc. However, retrieval results of CBIR systems are not satisfactory because there is always a semantic gap between low-level visual features and high-level semantic concepts. For instance, the user queries in natural language and seldom specifies the visual features of the image while submitting the query. Targeting this problem, Li and Sun presented an approach that incorporates lexical semantics from WordNet into the image annotation process. Jin et al. proposed a novel approach that prunes irrelevant keywords by the usage of WordNet. Paek et al. combined visual and textual features for image annotation. Some researchers have annotated images using ontologies. Hanbury reviewed three image annotation approaches: free text annotation, keyword annotation and annotation based on ontologies.

In this work, the keywords are extracted by text based annotation. The extracted keywords are further enriched by semantics using WordNet and keywords inferred from association knowledge base.

III. PROPOSED ARCHITECTURE

This section describes the architecture of the proposed system shown in Fig. 1.

Following subsections describe various components of the proposed system.

A. Page Manager and Valid Image Checker

Page manager collects web pages from the web, downloads HTML source codes, checks for the presence of image tags and hands the page to the next phase: valid image checker. In this phase, web page containing images are analyzed for checking image's validity. Image is considered as valid if its height is more than 50 pixels. Thus, it will ignore the unwanted images such as backgrounds, icons, advertisements etc.

B. Parser

It crawls the downloaded HTML code of web page, does link extraction, collects images and extracts associated text from page title, surrounding text, image caption and alt tag. The extracted text is then buffered, from which frequent keywords are determined by the annotator.

C. Annotator

Apart from determining frequent keywords, annotator helps in finding relevant annotation according to the semantic similarity and enriching the final annotation by using keyword

Fig.1. Annotation on the basis of Semantic Similarity and Keyword Association Analysis
based association analysis, leading to accurate image annotations.

**D. Association Rule Mining**

Association rule mining is the data mining technique for extracting patterns from the database or web page's HTML source code. The major statistics computed for the association rules, support and confidence are given in equation (1) and (2). In the proposed work, it is used for extracting patterns [17] from the downloaded source code, which are then stored in association knowledge base.

**Support:** \( (X=>Y)=P(X\cup\bar{Y}) \)........... (1)

**Confidence:** \( (X=>Y) = P (Y | X) \) ........ (2)

**A. Association Knowledge Base**

The knowledge base is a repository of extracted rules which have been derived using association rule mining [17]. A sample knowledge base containing rules is shown in Fig.2. The inference derived from these association rules help in annotation process.

**F. Wordnet**

WordNet is a lexical database for the English language [18] with synsets (set of synonyms) as its building block. A synset denotes concepts paired with description (a.k.a. gloss) of the synset.

Apart from this, they are interconnected with relational links, such as hypernymy, meronymy, antonymy and others. In this work the noun-taxonomy of WordNet, is used by annotator for enriching the semantics.

**G. Query Processor**

Query pre-processor processes user queries. It returns images from the web image repository corresponding to the queries fired by the user.

**I. ALGORITHM**

In this section an algorithm has been proposed whose step by step sequence is shown below.

**Algorithm Semantic_KeywordAsso_Annotation (p0)**

begin
p0 is a valid web URL hyperlink
Q is a queue of valid hyperlinks
P is a set of web pages
H is a set of hyperlinks
B is a Buffer for temporary storing information from tags and surrounding text
Q ← p0 // insert p0 into the queue Q
while |Q| ≠ Ø do
p ← Q // get head of queue Q
P ← P U p // appending p by inserting page p
D ← Download HTML Source Code of web page, p
ExtractedImageSearch for image tag in D
If (HeightOf(Extracted_Image) > 50 pixels)
// i.e. not an unwanted image
//Extract image caption, page title, metadata, alt tag in buffer
B ← [Info(ImageCaption) U Info(Page_Title) U Info(Alt_Text) U Info(Surrounding_Text)]
Frequent_Keywords Frequent Keywords Analyzer(B)
//extracting frequent keywords from buffered information
Inferred_Keywords AssociationAnalysis (B)
//extracting keywords from inference rules from //Knowledge Base
Semantic_KeywordsEnrichedSemantics(B)
//extracting noun taxonomy from WordNet for all //keywords in buffer
empty B
B ← [Frequent_Keywords U Inferred_Keywords U Semantic_Keywords] // update buffer B
//index image details of the form
<ImageCaption, B,URL, date, size>
end if
extract URL hyperlinks h contained in p into H
for all h ∈ H, h ∉ Q do
Q ← h
end for
empty B
end while
end
The flow diagram of the algorithm discussed in section IV is shown in Fig. 3.
Let p0, Q and P be seed URL, URL queue and set of pages respectively. Seed URL is the starting URL from which crawling for web pages starts. For instance, http://en.wikipedia.org/wiki/Jawaharlal_Nehru [19] be the seed URL i.e. p0, then on the basis of algorithm discussed Q and P are as follows.
Page Manager downloads HTML source code for the page p0. Fig.4. and Fig.5. shows the contents of web page p0 and its HTML code respectively.
Now, the valid image checker checks the validity of the image on p0. It analyzes the width and height of the image. The width and height for image (Bundesarchiv_Bild_183618490_0001_Indien_Ott_Grotewohl_bei_Ministerpr%c3%A4sidnt_Nehru_cropped.jpg) on page p0 is 220 and 273 pixels respectively, hence image is valid. Fig.6. shows the valid image on page p0.
The information extracted from various tags (page title, image caption, Image alt tag) and text surrounding the image (Fig. 6.) is shown in Fig.7. The frequent keywords from extracted information are buffered.

The enriched annotation using association rules...
and ontology is shown in Fig. 8. At last all the hyperlinks on p0 are inserted in queue H. The resultant queue H is as follows.

![Figure 8. Enriching Annotation](image)

### I. IMPLEMENTATION AND SIMULATED RESULTS

The prototype for the proposed system has been developed in Java. Implementation Results for URL

http://en.wikipedia.org/wiki/Jawaharlal_Nehru are shown in Fig. 9. and Fig. 10. below.

![Figure 9. Extracting Keywords from Tags](image)

![Figure 10. Enriching keywords by semantics and Association Knowledge base](image)

Extensive experiments have been conducted over certain web pages containing valid images. Table 1. shows the annotation process done by the proposed system, and it is observed that the proposed system does better annotation as compared to original annotation, followed by existing systems.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 1   Proposed System Annotations
### VII. Conclusion

The size of image collections is increasing rapidly and as a consequence of this; images must be effectively indexed and properly annotated for increasing relevancy. This work proposes a novel approach for finding relevant annotation, expanding it by using keyword based association analysis and enriching key terms using semantics. The proposed approach proves to be more efficient than existing approaches because the annotator annotates by referring semantics from WordNet and by inferring rules from association knowledge base; leading to accurate image annotations. Further the simulated results show the correctness of the proposed work.

**REFERENCES.**


