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

[8]. In Content-descriptive annotation, images are annotated by extracting relevant keyword [9] 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 [10] 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 [11, 12]. 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 [13]. Jin et al. proposed a novel approach that prunes irrelevant keywords by the usage of WordNet [14]. Paek et al. combined visual and textual features for image annotation [15]. Some researchers have annotated images using ontologies. Hanbury reviewed three image annotation approaches: free text annotation, keyword annotation and annotation based on ontologies [16].

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.

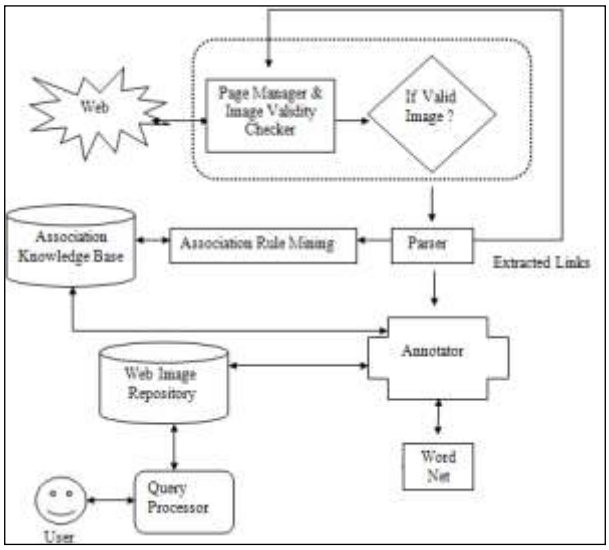


Fig.1. Annotation on the basis of Semantic Similarity and Keyword Association Analysis

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

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 U□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

R1 : star + galaxy sky
R2 : star + cloud sky
R3 : star + cloud moon
R4 : ……..
R5 : …….

Fig.2. Knowledge Base

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 nriching 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.
I. FLOW DIAGRAM

```
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
Extracted_ImageSearch 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_KeywordsAssociationAnalysis (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 Є 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 with 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.

Qhttp://en.wikipedia.org/wiki/Jawaharlal_Nehru (inserting p0 in Q) P P U http://en.wikipedia.org/wiki/Jawaharlal_Nehru (updating P by inserting page p0)

Page Manager downloads HTML source code

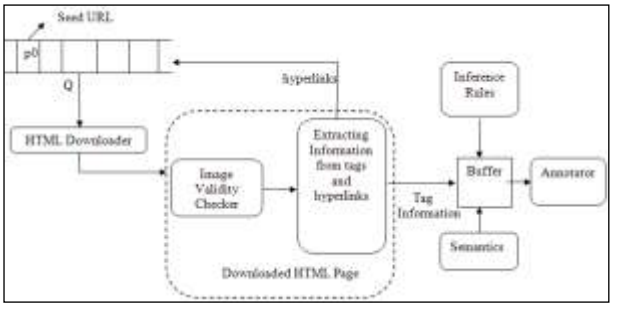


Fig.3. Flow Diagram

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

Fig.4. Web Page p0 containing image

Fig.5. p0 in HTML format

validity of the image on p0. It analyzes the width and height of the image. The width and height for image (Bundesarchiv_Bild_183618490001,_Indien,_Otto_Grotewohl_bei_Ministerpr%c3%A4sident_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

Fig.6. Valid Image

(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

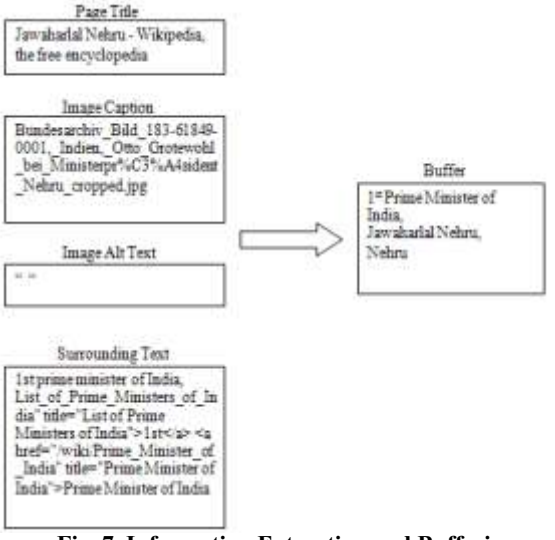


Fig. 7. Information Extraction and Buffering

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

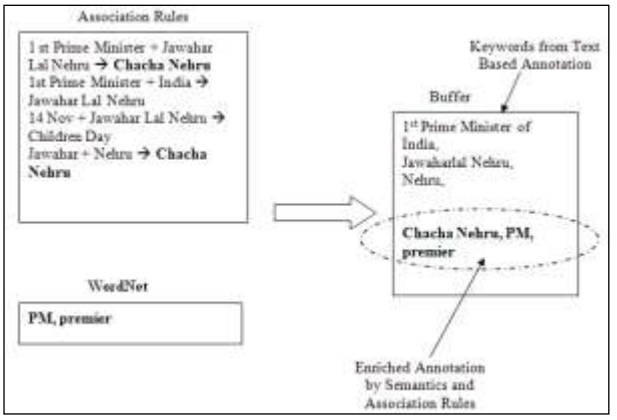


Fig.8. .Enriching Annotation

I. IMPLEMENTATION AND SIMULATED RESULTS

H H U http://ibnlive.in.com/news/gandhi-jinnah-both-failed-jaswant/99323-37.html U http://nuclearweaponarchive.org/India/Bhabha.html U http://www.jnmf.in/history.html U http://www.india-intro.com/remembering-nehru-and-others.html U http://www.indohistory.com/jawaharlalnehru.html U http://www.harappa.com/sounds/nehru.html U http://www.india-today.com/itoday/millennium/100people/nehru.htmlU http://indrus.in/articles/2011/08/02/12_indians_who_are_famous_in_russia_12823.html

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.

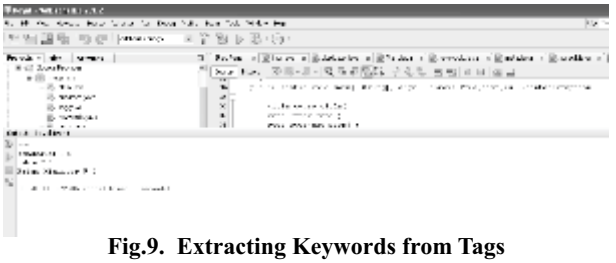


Fig.9. Extracting Keywords from Tags

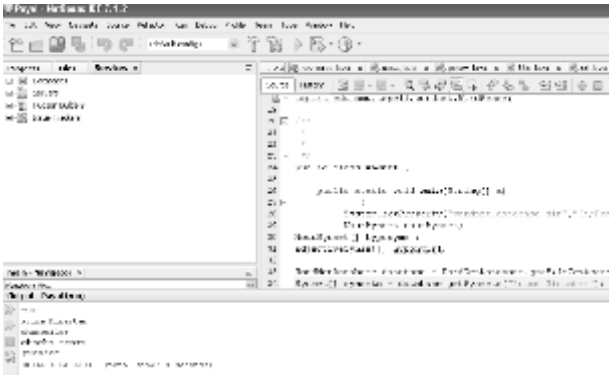
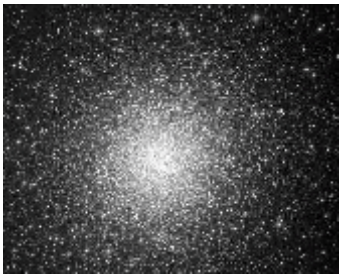


Fig. 10. Enriching keywords by semantics and Association Knowledge base

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 1 Proposed System Annotations

URL	**http://en.wikipedia.org/ wiki/Star_cluster [20]**	**http://www.culturalindia.net/ national-symbols/ national-bird.html [21]**
image		

Parsing Phase	**Page title:** Star cluster, **Image caption:** M92_arp_750pix.jpg, **Surrounding Text :** Star clusters or star clouds are groups of stars. Two types of star clusters can be distinguished: globular clusters are tight groups of hundreds of thousands of very old stars which are gravitationally bound, while open clusters, a more loosely clustered group of stars, generally contain less than a few hundred members, and are often very young. Open clusters become disrupted over time by the gravitational influence of giant molecular clouds as they move through the, but cluster members will continue to move in broadly the same direction through space even though they are no longer gravitationally bound; they are then known as a stellar association, sometimes also referred to as a moving group. **Image alt tag:** img alt = "" (MISSING)	**Page title:** National Bird Of India - National Bird India - India National Bird - Indian National Bird **Image caption:** national-peacock.jpg **Surrounding Text :** Peacock (Pavo cristatus), which is a symbol of grace, joy, beauty and love is the national bird of India. Peacock occupies a respectable position in Indian culture and is protected not only by religious sentiments but also by parliamentary statute. The Indian peacock is a colorful, swan-sized bird with a fan-shaped crest of feathers on its head, a white patch under the eye and a long-slender neck. The male peacock is more colorful than the female one. **Image alt tag:** National Bird of India
Original Annotation	star, cluster, cloud	National, bird, Indian National bird, peacock
Expanded Annotation using Association rule mining	star + galaxy sky star + cloud sky star + cloud moon	National + bird peacock Symbol+ peacock joy + grace + beauty + loveRain peacock
Enriched Annotation using Semantics	celestial body, galaxy, constellation, heavenly body, extragalactic nebula	Peafowl, Pavo cristatus
Proposed System Annotation	star, cluster, cloud, galaxy, constellation, heavenly body, extragalactic nebula, sky	Peafowl, Pavo cristatus, national bird of India, joy grace beauty love

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.

[1] World Wide Web Size.[online] http://www.worldwidewebsize.com (Accessed 21 January 2013).

[2] Sumathi, T., Hemalatha, M.(2011) 'A combined hierarchical model for automatic image annotation and retrieval', Advanced Computing (ICoAC), Third International Conference on , vol., no., pp.135-139, 14-16 Dec.

[3] Pecenovic,Z.,Do,M.,Ayer, S. and Vetterli,M.(1998) 'New methods for image retrieval', In Proc. of ICPS'98 Congress on Exploring New Tracks in Imaging, pp. 242-246, Antwerp, Belgium.

[4] Inoue,M.(2004) 'On the need for annotation-based image retrieval', Workshop on Information Retrieval in Context.

[5] Sumathi T., Lakshmi C. D., and Hemalatha M., "An Overview of Automated Image Annotation Approaches", International Journal of Research and Reviews in Information Sciences, Vol. 1, No. 1, 2011.

[6] Li,X.,Chen,L., Zhang,L.,Lin,L. and Ma, W.-Y. (2006) 'Image annotation by large-scale content-based image retrieval', In ACM Multimedia, pages 607-610.

[7] Tsai, Chih-Fong., Hung, Chihli.(2008) 'Automatically Annotating Images with Keywords: A Review of Image Annotation Systems', Recent Patents on Computer Science, 1, 55-68.

[8] Shah,K., and Sheth, A (1998). Logical Information Modeling of Web- accessible Heterogeneous Digital Assets. Proc. of the Forum on Research and Technology Advances in Digital Libraries, (ADL'98), Santa Barbara, CA. April, pp 266-275.

[9] Jung, K., Kim, K. and Jain, A. (2004) 'Text Information Extraction in Images and Video: A Survey', Pattern Recognition, vol. 37, no. 5,pp. 977-997.

[10] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., Jain, R.(2000) 'Content-based image retrieval at the end of the early years', IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12)1349–1380.

[11] Wang, X.-J., Zhang, L., Li, X. and Ma, W.-Y. (2008) 'Annotating images by mining image search results', IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1919–1932.

[12] Zhang, Dengsheng, Islam, Md. Monirul, Lu, Guojun.(2012) 'A review on automatic image annotation techniques', Pattern Recogntion, Volume 45, Issue 1, January, Pages 346-362.

[13] Li, Wei., Sun, Maosong.(2006) 'Automatic Image Annotation Based on WordNet and Hierarchical Ensembles', Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science Volume 3878,pp 417-428.

[14] Jin, Y., Khan, L., Wang, L. and Awad, M. (2005) 'Image annotations by combining multiple evidence & Wordnet', In ACM Multimedia, pages 706-715.

[15] Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang, S.-F., McKeown, K.R.(1999) 'Integration of visual and text- based approaches for the content labeling and classification of photographs', In: Proc. ACM SIGIR workshop on multimedia indexing and retrieval.

[16] Hanbury, A. (2008) 'A Survey of Methods for Image Annotation', Journal of Visual Languages & Computing, vol. 19, no. 5, pp. 617-627.

[17] Zhao, Q., and Bhowmick, S. S.(2003) 'Association Rule Mining: A Survey', Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116.

[18] Miller, G. A. (1995) 'WORDNET: A Lexical Database for English', Communications of ACM(11): 39-41.

[19] Wikipedia Jawaharlal Nehru.[online] http://en.wikipedia.org/wiki/Jawaharlal_Nehru (Accessed 14 November 2012)

[20] Wikipedia Star Cluster. [online] http://en.wikipedia.org/wiki/Star_cluster (Accessed 14 November 2012).

[21] Cultural India.[online] http://www.culturalindia.net/national-symbols/national-bird.html (Accessed 14 November 2012)