Usage Based Tag Enhancement of Images

Balaji Vasan Srinivasan
Big data Experience Lab,
Adobe Research, Bangalore
balsrini@adobe.com

Noman Ahmed Sheikh
Indian Institute of Technology,
Delhi
nomanahmedsheikh11@gmail.com

Roshan Kumar
Indian Institute of Technology,
Kanpur
roshankr1995@gmail.com

Saurabh Verma
Indian Institute of Technology,
Roorkee
saurv4u@gmail.com

Niloy Ganguly
Indian Institute of Technology,
Kharagpur
ganguly.niloy@gmail.com

ABSTRACT

Appropriate tagging of images is at the heart of efficient recommendation and retrieval and is used for indexing the image content. Existing technologies in image tagging either focus on what the image contains based on a visual analysis or utilize the textual tags from accompanying the images as the image tags. While the former is insufficient to get a complete understanding of how the image is perceived and used in various context, the latter results in a lot of irrelevant tags when the accompanying text is large. To address these, we propose an algorithm that analyzes the usage of an image from its accompanying textual content, extracts image-relevant tags in the accompanying text to and enhances the tags around the image. Evaluation based on human annotators and existing metrics from baselines shows the viability of the proposed approach.

CCS Concepts

•Information systems → Web searching and information discovery;

Keywords

image tagging, usage content, YAGO

1. INTRODUCTION

A popular English idiom says "An image is worth a thousand words". Content writers always look out for good visual supplements to enrich their content and make it more appealing to the target audience. In the era of data explosion, it is necessary to annotate content (images, video, etc.) with appropriate tags for efficient organization that can be leveraged for the retrieval and recommendation needs. However, the size of online visual data clearly calls for an automatic approach to tag these visual data.

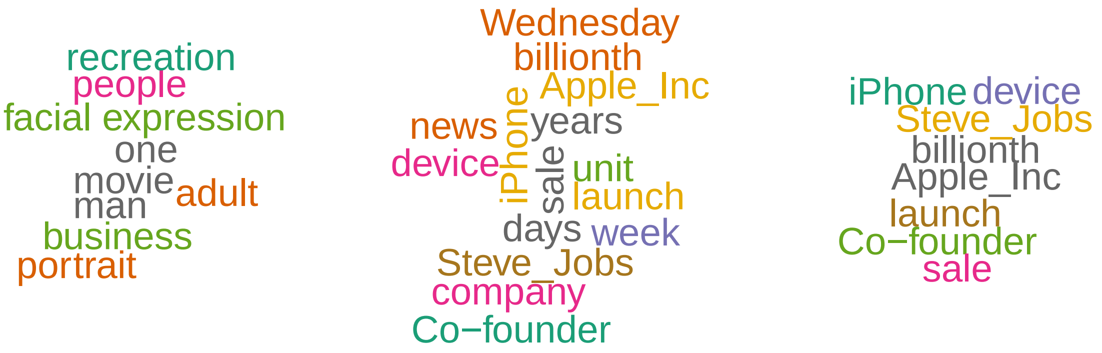
Apple sells its 1 billionth iPhone

Apple on Wednesday announced that it sold its one billionth iPhone last week. The news comes about two years after the company sold the 500 millionth unit of its handheld device. The iPhone was first introduced in 2007 by late Co-founder Steve Jobs and had registered its one millionth sale after 74 days of the launch.

Table 1: Example: An image of Apple co-founder Steve Jobs along with the text from an article using a similar image in InShorts[1], an online news aggregator.

Existing tagging systems work towards capturing the denotational aspects of the image, viz. what the image denotes/contains. This includes tags capturing the various aspects present in the image. These details are either captured via the visual features of the images or via human added tags. However, the former tags are often generic and do not capture the entire information that is contained in the image. Let us consider an example in Table 1 which shows an

International Conference on World Wide Web 2017
© 2016 ACM. ISBN 978-1-4503-2138-9. . . $15.00
DOI: 10.1145/1235



(a) Visual Tags

(b) Text based Tags

(c) Tags from the proposed system

Figure 1: Tags for the image in Fig. 1 based on the visual tagger from [21], textual parsing and our algorithm.

image of the Apple co-founder Steve Jobs from the web. Fig 1(a) shows the set of tags for the image based on the visual tagging system in [21]. It can be seen that the tags thus obtained are generic in nature e.g. ‘person’, ‘business’ and do not capture any deeper information about the image e.g. Steve Jobs, Apple Inc., etc.

Often such images are used in different illustrations which contain valuable information about the image. To address the short comings of the visual tags, the accompanying content of the images can be analyzed to extract the tags e.g. in search engines. Such information can enhance both the denotational and connotational (how the image is perceived) understanding of the image. To test this hypothesis, we conducted a survey among 31 participants to rate the relevance of the text around an image in several articles on the web and its usefulness to enhance the understanding of the image. It was observed that in 91.23% of annotations, the participants found the text relevant to the image. Survey respondents further opined that while the original image tags were very appropriate, the image had a different connotation when appeared along with the text, thus calling for a need to incorporate these into the image tags.

We identified an article from InShorts[1], an online news aggregator. Table. 1 includes the text from this article. A simple text based tagging can add a lot of noise to the tags as seen in Fig. 1(b), where the text in Fig. 1 was parsed to extract the textual tags. The level of noise will increase with the size of the accompanying content. This calls for an automated tagging system that optimally combines the tags from accompanying text with the image tags capturing the right denotational and connotational information around the images while still minimizing the noise in the resulting tags.

In this work, we propose a novel framework to combine the tags derived from accompanying content with the image tags based on the visual features, thus combining the information from content and usage cues. We thus achieve a balance between connotational and denotational aspects of an image. We show that such a combination beats the state-of-the-art (visual and textual) tagging engines in our subjective and objective evaluations.

The paper is organized as follows. In Section 2, we describe the existing state of image tagging and position our framework with respect to existing systems. Section 3 introduces the key components of the proposed system. In Section 4 evaluates the different parameters of the proposed system to arrive at the right system configuration. We then compare the performance of the proposed system against existing works via subjective and objective evaluations. Section 5 concludes the paper.

2. RELATED WORK

Tagging and understanding textual content has been widely studied. The first step in textual tagging is extracting and detecting named entities; the popular one here is the Stanford NLP parser [14]. Once the named entities are identified, they are disambiguated and resolved into various categories [12]. Finally, the inter relationships in the content or hierarchies are identified by a semantic understanding of the text. In these works, the entities in the textual content are typically processed into a rich semantic representation (e.g. [2]) which is utilized to gain a deeper understanding of their inter-relationships.

Yang et al. [26] extract the textual tags based on a nearest-neighbor based approach and utilize the neighbors to extract the relationships between entities. Nallapatti et al. [17] use “event threading” to join different pieces of text and identify the undercurrent events in the textual topics. Shahaf et al. estimate the importance and “jitteriness” of the entities in the text and use it to infer the connections between different parts of the textual content.

With the advent of knowledge bases like DBpedia [1], Freebase [3] and YAGO [22], relationships from these sources are used to further enhance the understanding of the textual content. Kuzey et al. [9] resolve temponym based on a YAGO based entity resolution to understand textual content with temporal scopes. They develop an Integer Linear Program that jointly optimizes the mappings to knowledge base for a rounded document representation. Tandon et al. [23] mine activity knowledge from Hollywood narratives to answer questions around these activities. They capture the spatio-temporal context of the topics by constructing multiple graphs to capture relationships among activity frames which is leveraged for effective understanding. However, none of these works aim at understanding content based

[1] https://www.inshorts.com/news/apple-sells-its-1-billionth-iphone-1469693675991

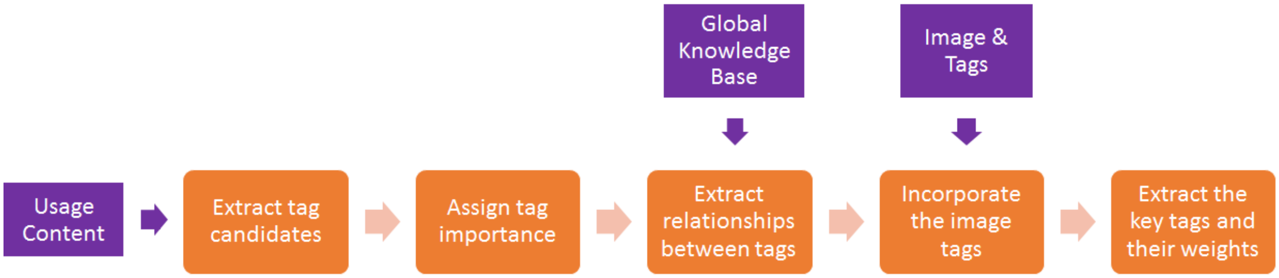

Global Knowledge Base
Image & Tags
Usage Content
Extract tag candidates
Assign tag importance
Extract relationships between tags
Incorporate the image tags
Extract the key tags and their weights

Figure 2: Framework for tag extraction and enhancement in the tagging engine

on multiple cues which is the key challenge in our problem, where we have to combine the content and usage cues in tagging.

There also exists a large body of literature in the space of image tagging. Li et al.[11] propose methods for assignment of tags from visual aspects and use them for effective retrieval of images. Once the image is tagged, the relationships with other images have also been used for further enhancing the tags [18] or propagating the tags to neighboring images [5]. Like in the textual tagging, the tags can be enhanced and disambiguated with the help of knowledge base and conceptnets [25]. With the successful emergence of deep learning for image understanding, convolution neural networks have been used to find an intermediary representation *Visual Word2Vec* [8] in order to generate the image tags from this latent space. However, all these works focus on tagging the image from their visual cues/content. In our problem, we capture the usage of the images along with the visual content in the image tags to have a rounded understanding of the image.

One work that is close to the proposed solution framework is by Leong et al. [10], which relies exclusively on accompanying content for mining information relevant to the image. They construct relationships among entities based on multiple factors to arrive at the final set of tags. However, they do not use the visual tags of the images to align the accompanying content to the image and therefore have the same pitfall that we illustrated in our example in Fig. 1.

3. USAGE BASED TAGGING

The proposed solution enriches the tags around an image which may not be initially contained in the set of image based tags based on the visual features. Our approach takes as input the image tags (author given and the auto tags) along with the content for which the image is used as an illustration. The content is processed to extract key tag candidates which are then pruned with the help of a knowledge base to come up with a set of enriched candidate tags and their relationships. The relationships are leveraged to extract top tags for the image along with the confidence around the tag. Fig. 2 shows a schematic of the proposed solution framework.

We score the importance of the tags (Section. 3.2) based on the local significance of the candidates in the image/textual content thus capturing the prominent tags from the content.

Scoring the tags just based on the local context can lead to a myopic understanding of the image (and its context). We therefore represent these tag candidates in a graph which allows capturing not only the local significance of the tags but also their relationships with other tags. We capture the inter-tag relationships (Section 3.3) as weights on the edges based on a global corpus to understand the global relationship between the tags.

The final set of tags should not only be relevant to the image but also be diverse capturing multiple aspects of the image. After incorporating the information from both the text and image to the tag graph, we perform a random walk based node ranking starting from the image tags to ensure relevance to the visual tags. Starting from multiple image tags via the random walk yields diversity in our tags as well.

3.1 Candidate Tag Extraction

The proposed approach starts with disambiguating the accompanying content for ambiguous entities via Ambiverse [6]. Ambiverse provide a technology to automatically analyze a textual data and disambiguate named entities. It relies on the knowledge base YAGO [22], which is derived from Wikipedia, can be thought of as a very large collection of entities. YAGO also contains accurate characterizations of all entities. These multiple characteristics of the entities are used along with the context of every entity in the text to disambiguate them into formal entries in YAGO. We replace each occurrence of the entity with their disambiguated version. This helps in reducing the ambiguities that can get into the candidate tags. The disambiguated content is extensively parsed to identify all named entities using the Stanford NLP Parser [14].

We then establish the relationships between these entities across the entire accompanying content. Note that the image may/may not be relevant to the entirety of the entities in the accompanying text and we address this in Section 3.4. At the end of this step, we have a set of all candidate tags which we shall be considering for our final tags.

3.2 Scoring Tag Importance

For each tag candidate extracted in the previous step, a score is assigned based on their importance in the local context. We first calculate the total frequency count of the candidate in the usage content accounting for the co-reference of the candidates via proper nouns by an appropriate co-reference parsing. Thus, not just the direct mentions, the indirect mentions of the entities are also accounted to calculate the local importance.

For every tag candidate we also compute the average distance of the entity from the root node of the corresponding dependency tree (obtained by passing the accompanying content through a dependency parser[4]). A candidate tag at the root is the central topic of discussion in a sentence and hence is more important indicating the local relevance of the entity in the discussed subject. Note that, the smaller this metric, more relevant is the tag. We normalize the measure to be in the range $0 - -1$ and subtract the normalized score from 1.

The average of the two measures yields the final tag importance (n_i).

3.3 Inter-tag Relationship Extraction

We build the relationships between each tag candidates based on a global knowledge base from two independent measures.

In the first measure, we used the Word2Vec [15] model trained on a corpus of Google News dataset with 100 billion words resulting in a final corpus of about 3 million word representations. Word2Vec yields a 300 dimensional vector for every tag candidate that represents the word in the space of the trained deep neural network. To measure the relationship between a pair of words, we compute the cosine-similarity between the vectors in this space which captures the semantic closeness of the words as described in [15].

In our second measure, we calculate the point-wise mutual information [24] between two entities based on their co-occurrences in the Wikipedia articles as the "co-occurrence" score. This yields a similarity score based on how coherent the two tags are with respect to the entire Wikipedia corpus (English).

The first measure based on Word2Vec captures the semantic similarity between the tags because the Word2Vec space groups similarly meaning entities together. Therefore, entities closer in this space can often be interchangeably used in several context. On the other hand, the second measure based on Wikipedia captures topical closeness - since entities that occur together in the several article are closer in this space. Our final edge weight (e_{ij}) is computed as the average of the two measures. We will show in Section 4.2.2 that the combination performs better than the individual measures based on several evaluation metrics.

3.4 Incorporating Image Tags

The edge weights (from Section 3.3) along with the node importance (from Section 3.2) yield a graphical representation of the candidate tags with the edge weights capturing the global relationship between the tags and the node weights indicating their local importance in the usage content. To extract the usage-specific tags from the accompanying content, it is important to understand how these tag candidates relate to the visual tags.

Images are often accompanied with author tags and/or tags from the visual features capturing the information contained in the images. Starting with these tags and retaining the visual tagger's confidence as their importance, we calculate the similarity score between the visual tags and every tag candidate in the graph based on Section 3.3. We use this score to either merge the visual tags with a pre-existing tag (if the the computed similarity is greater than a threshold) or add as independent nodes to the graph.

First, the tag pairs with similarity greater than a threshold (0.95 in our experiment) are merged into a single node. We then propagate the importance of the merged node to the adjacent nodes (at a distance of 2 edges) using an exponential decay. This ensures the propagation of the strength of the merged nodes to its neighbors and thus emphasizing the relevant pieces of the tag graphs with respect to the visual tags.

For the tag pairs less than the matching threshold, an edge is added between every tag candidate whose similarity with the visual tag is significant (> 0.1 in our experiments). This ensures that the non-merged visual tags are connected to the relevant parts of the tag-graph.

The series of steps is summarized in Algorithm 1.

Algorithm 1 Tag Unifier

```
procedure UNIFY(tagsFromImage, TagGraph)
    for tag ∈ tagsFromImage do
        tag ← normalize(tag)
        for node ∈ TagGraph do
            val ← similarity(tag,node) (from Sec.3.3)
            if $val > \sigma_1$ then
                MergeNodes(tag,node)
                node.weight ← MergedWeight()
                PropagateWeight(node)
            else if $val > \sigma_2$ then
                edge ← createNewEdge(tag,node)
                edge.weight ← val
            else
                continue
            end if
        end for
    end for
end procedure
```

3.5 Tag Extraction

With the graphical representation of the tags, the problem of identifying tags that capture the context around the image boils down to identifying the top nodes in the tag graph that are closely connected to the image tags. For this we use a random walk based algorithm [19], starting the random walk from the visual tags, thus ensuring the node ranking relevant to the tag images and avoiding irrelevant tags from the accompanying text.

We define the probability of the random walk moving from a node i to another node j as,

$$P(tr_{i \to j}) = e_{ij} \times n_j \tag{1}$$

where, e_{ij} is the weight of the edge (from Section 3.3) between tags i and j and n_j is the node importance of tag j from Section 3.2. The probability of the node staying in the same node is defined as

$$P(tr_{i \to i}) = n_i \tag{2}$$

The probabilities above are normalized to conform to the requirements of a probability distribution. The final set of tags is then extracted by performing a random walk starting from the visual/author tag nodes. This ensures that the tags selected are not just based on their importance from the accompanying text but also emphasizes on a strong relationship with the visual tags. The set of tags with weights above a certain threshold is output as the final set of tags for the images.

3.6 Survey-based evaluation

We conducted a survey among 45 participants to rate the overall relevance and diversity of the tags on a scale of $0-10$

for the outputs from the proposed approach as well as those provided by the visual tagger, Clarifai [21] on a subset of 20 images. Fig 3 shows the relevance and diversity of the two tags based on the annotation. On a scale of 10 for tag relevance to the image, usage tags were rated at 8.08 on an average against the score of 5.73 for the visual tags. For diversity, usage tags received a rating of 6.79, whereas, the visual tags received 5.93. This indicates that the proposed approach increases the overall relevance of the tags to the image and also performs better in terms of the diversity of the tags indicating the viability of the proposed approach.

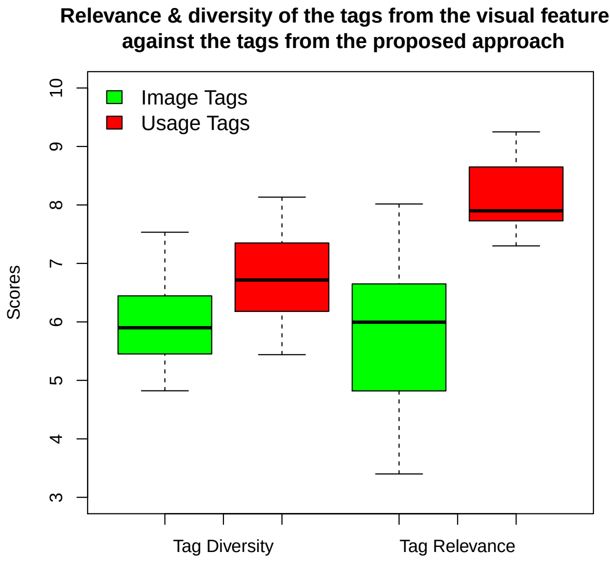


Figure 3: Relevance and diversity of tags based on annotations from 45 participants

4. EXPERIMENTAL EVALUATION

We utilized the dataset[2] curated by Leong et al. [10] which contains 300 image-text pairs collected by issuing a query to Google Image API and processing one of the query results that has a significant text around the images. Leong et al. [10] have also created a gold standard tags based on manual annotations from 5 annotators via Amazon Mechanical Turk accepting annotations from annotators with approval rating $> 98\%$. We used the datasets along with the gold standard tags for our evaluations. We used the Clarifai API [21] to generate the visual tags for all our experiments.

4.1 Metrics for evaluation

Human annotations cannot be extended for a comprehensive evaluation of the tags. We therefore extend several existing metrics to measure different aspects of the tags which are described below.

4.1.1 Term Significance [13]

The term-significance [13] is calculated as the significance of the tags to the textual content and is calculated by computing the Normalized Discounted Cumulative Gain(NDCG)[7] over the term frequency of the tags from the usage content normalized based on the tag's inverse document frequency in a global corpus. The intuition here is to compute how important a tag is to the given context (usage) and normalize it with its "commonness" across a bigger corpus (as computed by the idf). We use Wikipedia as the bigger corpus similar to Leong et al. [10].

4.1.2 Tag Relevance

The term significance metric purely tests the relevance of the tags to the usage content and is biased towards the system proposed in [10]. It fails to capture the relevance of the tags to the gold standard tags or its overall diversity. We therefore propose two metrics to capture the tag relevance to the image and its overall diversity. To determines how relevant our tags are to the gold standard tags, we compute a weighted cosine similarity between the Word2Vec [15] representation of the extracted tags and the gold tags as given by,

$$sim = \frac{1}{N}\sum_{i} \frac{\sum_{a_j \in TopK(G_i,I_i)} cos(a_j, I_i)\gamma^j}{\sum_j \gamma^j}, \tag{3}$$

where N is the number of tags generated for the images, I_i is the vector representation of the i^{th} image tag and G_i is the set of all vector representations of the gold standard tags. The inner sum above computes a weighted average of the similarity between the generated tag and the top gold-standard tags. The parameter γ, $(0 \leq \gamma \leq 1)$, penalizes the generation of image tags that is similar to only a small set of the gold standard tags via a decayed-weighted-average. The outer summation calculates the average weighted similarity between the generated image tags and the gold standard tags.

4.1.3 Tag Diversity

Finally, for measuring the diversity in the tags, we use the **cophenet correlation coefficient** [20] (which is a measure of how faithfully a dendogram preserves the pairwise distances between the original un-modeled data points). We perform a hierarchical clustering on the tags based on their Word2Vec representation [16] and compute the cophenet correlation coefficient as the diversity score. Cophenet correlation coefficient is then given by,

$$c = \frac{\sum_{i<j}(x(i,j)-\bar{x})(t(i,j)-\bar{t})}{\sqrt{[\sum_{i<j}(x(i,j)-\bar{x})^2][\sum_{i<j}(t(i,j)-\bar{t})^2]}} \tag{4}$$

where, $x(i,j)$ is the distance between the i^{th} and j^{th} tag. $t(i,j)$ is the height of the node at which the clusters corresponding to ith and jth cluster are first joined together. A higher value of the cophenet correlation coefficient indicates the presence of more significant clusters and hence more tag diversity.

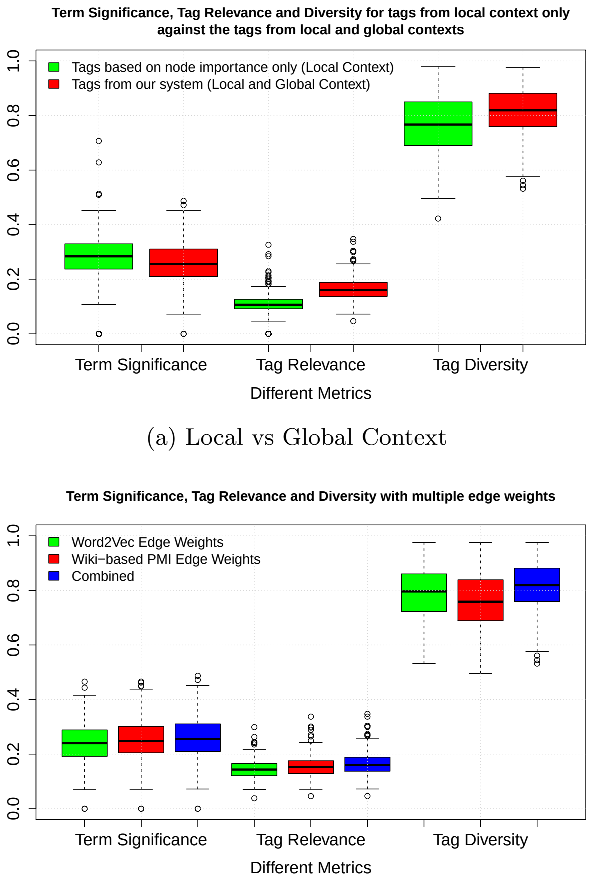
4.2 Evaluation of Algorithmic Parameters

Before comparing the performance of the proposed system, we independently evaluate the different parts of the algorithm and their importance in extracting relevant and diverse tags capturing the usage of the image.

[2] http://lit.csci.unt.edu/index.php/index.php?P=research/downloads#TEXT_MINIG"

4.2.1 Local vs Global Context

In this experiment, we compare the local context captured by the node importance (Sec. 3.2) against the combined context captured in our approach. We extract the top tags based on their node importance score and compare it against the top tags from our approach.



(a) Local vs Global Context

(b) Effect of Edge Weights

Figure 4: Term Significance, Tag Relevance (Eq. 3) and Diversity (Eq. 4) for tags for different algorithmic parameters. Fig. 4(a) compares the tags extracted solely based on Node Importance against the tags from the entire system (where the local and global context of the tags are jointly accounted for). Fig. 4(b) compares the effects of different edge weighting mechanisms on the tagging performance

Fig. 4(a) compares the Term Significance, Tag Relevance (Eq. 3) and Tag Diversity for the two cases. The term significance of the tags based on the local context with an average of 0.275 is marginally better than the term significance of the proposed system (average at 0.26). Since the term significance captures the local importance of the tags in the accompanying text, hence the tags from local context is expected to be better here.

However, the overall tag relevance (average of 0.1090 for local context against 0.1654 for the combined context) and tag diversity (average of 0.7554 for local context against 0.0.8155 for the combined context) is better with the combined approach since it accounts for the global relationship between the tags and hence better tags without compromising much on the term significance (since the difference between the two methods is not significant).

4.2.2 Effect of edge weights

In the next experiment, we compare the term significance, tag relevance and tag diversity between the edge weighting mechanisms based on Word2Vec, Wikipedia and the combined metric defined in Section 3.3.

From Fig. 4(b), it can be seen that while Word2Vec performs marginally better than the Wikipedia based relationship on the scales of term significance (average of 0.2516 for Word2Vec based metric against the 0.2398 average for the Wikipedia based metric) and tag relevance (average of 0.1567 for Word2Vec based metric against the 0.1451 average for the Wikipedia based metric).

In terms of overall tag diversity, Wikipedia based metric is marginally better than Word2Vec (average of 0.7897 for Wikipedia based metric against the 0.7617 average for the Word2Vec based metric). This could perhaps be because Wikipedia include more entities than the Google News Corpus on which the Word2Vec were trained, and hence aid in the extraction of diverse tags.

It can also be seen that the combined approach yields the best tags across all metrics and was used for our further comparisons.

4.2.3 Effect of visual tag quality

We finally compare the correlation between the quality of the visual tags and the tags from the proposed system. Fig. 5 shows the correlation between the two set of tags on the scales of Term Significance, Tag Relevance and Tag Diversity.

It can be seen that there is a strong dependence of the term significance and relevance of our tags with the visual tags as indicated by the slopes of 0.95 and 0.89 respectively of the corresponding line fits. This is expected since the algorithm performs the random walk starting from the visual tags and hence the output tag quality is pivoted on the quality of visual tags.

However the tag diversity is less dependent on the visual tags, since the diversity of the output tags is obtained more from the accompanying text than from the visual tags indicated by a lower slope of the corresponding line (0.36).

4.3 Tagging Performance

To evaluate our overall tagging system, we compare it against the baseline algorithm in [10]. Leong et. al [10] propose 3 independent algorithms based on “Wikipedia Salience”, “Flickr Picturability” and “Topic Modeling” to extract tags for an image its accompanying textual content. In their experiments, the Wikipedia Salience based tagger was best performing in terms of the precision and recall. We used this algorithm as the baseline for our evaluations.

We also compare the performance of the proposed tagging system against the visual tagger in [21] and the text based tagger in [10]. Fig. 6 shows the Term Significance, Tag Relevance and Tag Diversity for the tags from the system in [10] compared against our proposed system.

The term significance checks significance of the tags with respect to the accompanying text and hence text based taggers are expected to perform better in this measure. Along

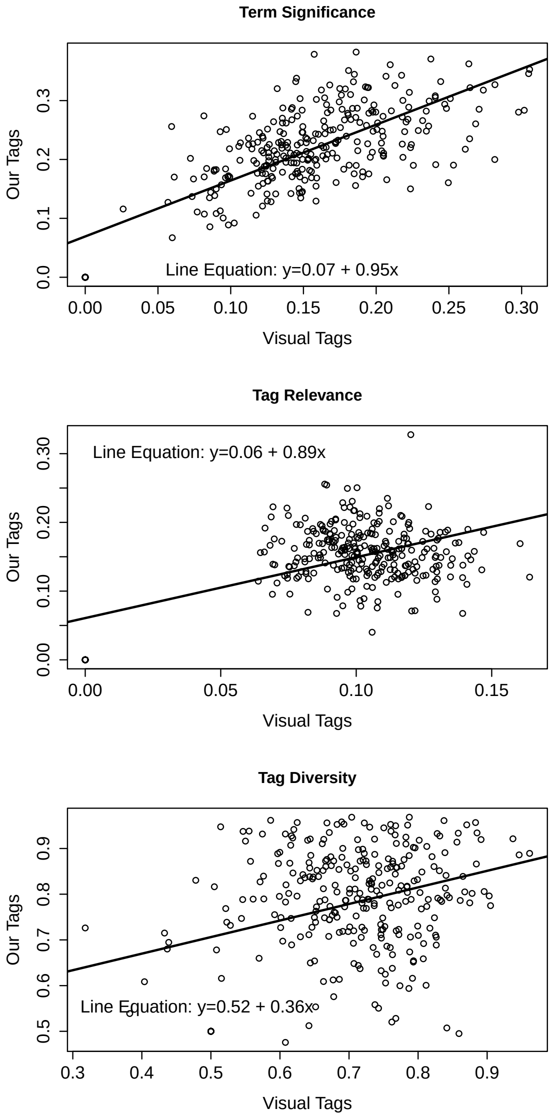


Figure 5: Correlation between the quality of visual tags and the tags from the proposed system

the expected lines, both the proposed tagger and the tagger by Leong et al. [10] perform better than the visual tagger. Between the text based taggers, the term significance is the best for the proposed method indicating the superiority of the tags in capturing the local context.

The tags from the proposed system are also more relevant/close to the human annotated tags based on the tag relevance (Eq. 3). The visual tags from Clarifai [21] is observed to have better relevance than the tags from Leong et al. [10] indicating that the human annotators focused on denotational aspects over the connotational aspects. A superior performance here indicate that the proposed system captures the denotational aspects as well as the connotational aspects.

Capturing the connotational aspects of the images yield more diversity as indicated by the superior performances of both the text-based taggers on the scales of diversity. Here again, the tags from the proposed system are more diverse than the tags from Leong et al. [10].

5. CONCLUSION

In this paper, we have proposed a novel graph based approach to enhance the tags of an image by capturing its usage. The proposed approach extract candidate tags and captures their local importance by a semantic parsing of the accompanying text. The relationship between the candidate tags are extracted based on a global corpus, thus accounting for their global context. The tags from the visual tagger are combined with the candidate tags, thus merging the information from two different cues (image and text), to extract the final set of tags.

The tags thus obtained are compared against existing baselines based on text based tagging and visual tagging. The proposed system outperforms the other systems in the lights of several quality metrics capturing the relevance and diversity of the tags. Such a tagging system will serve well to improve the image retrieval systems by effectively serving the user's context.

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *DBpedia: A Nucleus for a Web of Open Data*. 2007.

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation (amr) 1.0 specification. In *ACL Conference on Empirical Methods in Natural Language Processing*. ACL, 2012.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of data*. ACM, 2008.

[4] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

[5] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE International Conference on Computer Vision,*, 2009.

[6] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[7] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[8] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. *arXiv preprint arXiv:1511.07067*, 2015.

[9] E. Kuzey, V. Setty, J. Strötgen, and G. Weikum. As time goes by: comprehensive tagging of textual phrases with temporal scopes. In *Proceedings of the*

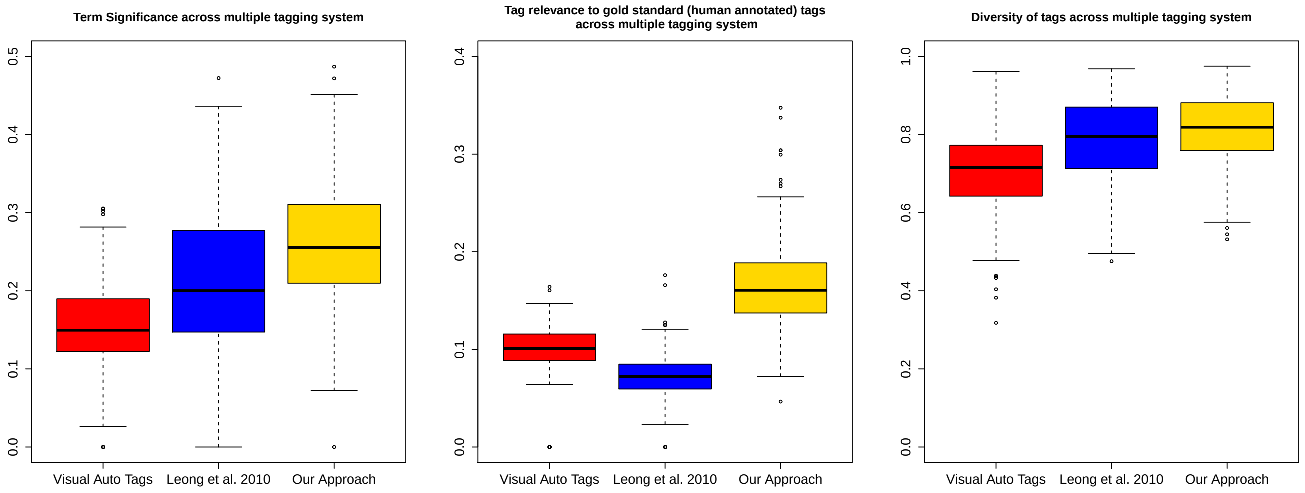


Figure 6: Term Significance, Tag Relevance (Eq. 3) and Diversity (Eq. 4) for tags based on Clarifai [21], Wikipedia Salience [10] and the proposed approach

25th International Conference on World Wide Web, pages 915–925, 2016.

[10] C. W. Leong, R. Mihalcea, and S. Hassan. Text mining for automatic image tagging. In *International Conference on Computational Linguistics: Posters*, 2010.

[11] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. Del Bimbo. Image tag assignment, refinement and retrieval. In *21st ACM International Conference on Multimedia*, 2015.

[12] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *35th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 731–740. ACM, 2012.

[13] Y.-T. Lu, S.-I. Yu, T.-C. Chang, and J. Y.-j. Hsu. A content-based method to enhance tag recommendation. In *IJCAI*, volume 9, pages 2064–2069, 2009.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. 2013.

[16] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, May 2013.

[17] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *13th ACM International Conference on Information and Knowledge Management*. ACM, 2004.

[18] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[19] Sanjay and D. Kumar. A review paper on page ranking algorithms. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, Volume 4(6):pp. 2806–2811, 2015.

[20] R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, pages 33–40, 1962.

[21] G. Sood. *clarifai: R Client for the Clarifai API*, 2016. R package version 0.4.0.

[22] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *16th International Conference on World Wide Web*, pages 697–706. ACM, 2007.

[23] N. Tandon, G. de Melo, A. De, and G. Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 223–232. ACM, 2015.

[24] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[25] L. Xie and X. He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *21st ACM International Conference on Multimedia*, pages 967–976. ACM, 2013.

[26] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer. Improving text categorization methods for event tracking. In *23rd ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.