A Semantic Approach to Recommending Text Advertisements for Images

Weinan Zhang, Li Tian, Xinruo Sun, Haofen Wang, Yong Yu
Dept. of Computer Science and Engineering
Shanghai Jiao Tong University
No. 800, Dongchuan Road, Shanghai, China 200240
{wnzhang, tianli, xrsun, whfcarter, yyu}@apex.sjtu.edu.cn

ABSTRACT

In recent years, more and more images have been uploaded and published on the Web. Along with text Web pages, images have been becoming important media to place relevant advertisements. Visual contextual advertising, a young research area, refers to finding relevant text advertisements for a target image without any textual information (e.g., tags). There are two existing approaches, advertisement search based on image annotation, and more recently, advertisement matching based on feature translation between images and texts. However, the state of the art fails to achieve satisfactory results due to the fact that recommended advertisements are syntactically matched but semantically mismatched. In this paper, we propose a semantic approach to improving the performance of visual contextual advertising. More specifically, we exploit a large high-quality image knowledge base (ImageNet) and a widely-used text knowledge base (Wikipedia) to build a bridge between target images and advertisements. The image-advertisement match is built by mapping images and advertisements into the respective knowledge bases and then finding semantic matches between the two knowledge bases. The experimental results show that semantic match outperforms syntactic match significantly using test images from Flickr. We also show that our approach gives a large improvement of 16.4% on the precision of the top 10 matches over previous work, with more semantically relevant advertisements recommended.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Experimentation

Keywords

Visual Contextual Advertising, Semantic Matching, Cross-media Mining

RecSys'12, September 9–13, 2012, Dublin, Ireland.
Copyright 2012 ACM 978-1-4503-1270-7/12/09 ...$15.00.

1. INTRODUCTION

Nowadays, Web pages no longer contain just textual information. Instead, more and more images have been uploaded and published on the web. For instances, social Web sites like Facebook[1] and Flickr[2] have billions of photo album pages with little text. Compared with the traditional textual Web pages, images become the main contents of these Web pages. Thus, traditional contextual advertising approaches cannot be directly applied to Web pages dominated by images because of the lack of textual information. Therefore, understanding the contents or topics of images and then recommending relevant advertisements based on these images becomes a challenging problem interesting to both academia and industry.

Visual contextual advertising (see Figure 1) refers to finding the most relevant advertisements for a target image without textual information such as tags. It can be regarded as a special case of contextual advertising where images become the context for recommending advertisements. While it is a young branch of contextual advertising, it is more challenging than advertising on textual Web pages because it requires techniques such as computer vision and cross-media transfer learning . In other words, visual contextual advertising aims at semantic matching between two heterogeneous features spaces (i.e., image feature space and text feature space).

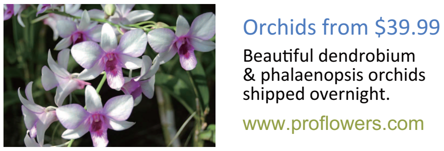

Orchids from $39.99
Beautiful dendrobium
& phalaenopsis orchids
shipped overnight.
www.proflowers.com

Figure 1: An example of visual contextual advertising.

Image annotation [4, 11] is one approach to visual contextual advertising. Intuitively, given a target image, text annotations are extracted based on a model trained by labeled images. Then these annotations are used to search for relevant advertisements, similar to keyword search in traditional contextual advertising. However, since it is time consuming and error prone to obtain high-quality labeled images, the quality of annotations cannot be guaranteed, which leads to poor recommendation performance. On the other hand, since the match process is performed between two heterogeneous feature spaces (i.e., images and text), het-

[1] http://www.facebook.com
[2] http://www.flickr.com

erogeneous transfer learning [30, 9] can be adapted to the image-advertisement match. The state-of-the-art algorithm for visual contextual advertising is ViCAD [8]. It first builds a bridge between the image feature space and text feature space through a feature translation model. Then it uses a method based on a language model to estimate the relevance of each candidate advertisement to the target image. While ViCAD is reported to outperform annotation-based approaches, the advertising precision is still not satisfactory as to be used in real world applications.

With a careful investigation of the performance of previous work, we find that the major weakness of ViCAD as well as the annotation-based approaches comes from mismatches between image tags and text advertisements due to their shortness, ambiguity, and variety. Figure 2 presents some examples which indicate the syntactic mismatch in these approaches. Detailed explanations are as follows.

- **Different term distributions in image tags and text advertisements.** Both ViCAD and annotation-based approaches make use of image-tag co-occurrence data. However, in the image-tag co-occurrence data, if the tag terms (also called text features) translated from target image features are very rarely used or have a different meaning in the advertisement contents (or bid keywords), no advertisement or irrelevant advertisements will be matched.

- **Semantic mismatch between text features and advertisements.** Even if the translated text features are accurate and can syntactically match some advertisements, these advertisements may be semantically irrelevant. This is because current approaches use a syntactic match to retrieve advertisements. Therefore, though the retrieved advertisements contain the image tags, they are irrelevant to the target images.

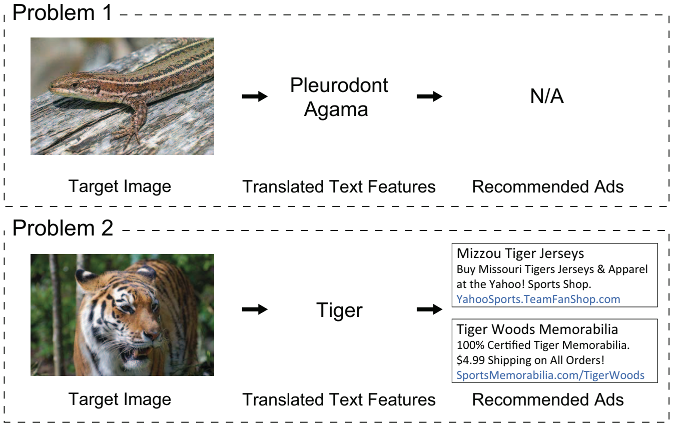


Figure 2: Two main problems in current approaches.

These two problems also occur frequently in traditional Web page contextual advertising. To overcome the syntactic mismatch problem, people use a broad match [12, 15] which finds the semantic relation between different keywords or phrases. In order to optimize the semantic relevance, several semantic approaches to contextual advertising have been proposed [5, 23]. In this paper, we follow this promising direction and propose a semantic approach to tackle the problems of visual contextual advertising. More specifically, we map the target image to some nodes of interlinked knowledge bases instead of to pure text features. Compared with pure text features, knowledge base nodes have their context and relationships with other relevant nodes, which helps solve the two problems. Using the precision@10 measure on the Flickr test dataset with 230 images, our approach outperforms the syntactic matching approaches by up to 16.4 percent.

To sum up, the contributions of this paper are threefold.

- We identify the problems of syntactic mismatch in existing approaches to visual contextual advertising.

- We propose a knowledge-driven cross-media semantic matching framework to solve these problems. To the best of our knowledge, this is the first work that studies the semantic match between images and text advertisements.

- In the experiment, our approach provides a substantial improvement over the existing approaches, making visual contextual advertising more applicable.

The rest of this paper is organized as follows. In section 2, we discuss related work. In section 3, we present our semantic approach to visual contextual advertising. In section 4, we describe the experiments and analyze the results. Finally, in section 5, we conclude this paper and discuss future work.

2. RELATED WORK

2.1 Contextual Advertising

Contextual advertising refers to placing relevant advertisements on third-party Web pages. The publisher and search engine will share the revenue once any advertisement on their pages is clicked. Studies [28] have shown that the relevance of the advertisements to the content of target pages makes a large difference at the click-through rate. Therefore, the work of matching target Web pages and advertisements is the key point of contextual advertising [23, 6].

Keyword-based approaches [29, 31] are widely used in contextual advertising. They first extract the keywords from a target Web page and then use these keywords to retrieve relevant advertisements just like sponsored search (another kind of Web advertising). However, due to the vagaries of keyword extraction and the lack of content in advertisements, keyword-based approaches always lead to irrelevant advertisements. Besides keyword-based approaches, the authors in [5] imported semantic information to enhance the matching work. They classified both pages and advertisements into a common taxonomy and combined the keyword-based approach with taxonomy matching to rank the advertisements. Moreover, Pak et al. [23] proposed an ESA [14] based approach which makes use of Wikipedia as the knowledge base to improve the performance of contextual advertising. However, they only chose one thousand entities and no link information was used. This work has much room to improve. On the side of efficiency, since analyzing the entire page content is costly and new or dynamically created Web pages cannot be processed to match the advertisements ahead of time, the authors in [2] proposed a summary-based approach to enhance the efficiency of contextual advertising with an ignorable decrease on effectiveness.

2.2 Cross-media Mining

Besides textual content, there are more and more multimedia elements such as images, audio, and video on Web pages. These elements, as pieces of information, are often important and illustrate the topic of a Web page. Mining on these multimedia elements has got considerable attention from both academia and industry. In particular, data mining across different media has become a promising research direction. IJCAI 2009 held a workshop focusing on cross-media information access and mining [1]. Recently, some applications using cross-media mining technologies were developed. Chao et al. [7] proposed TuneSensor, a semantic-driven service to recommend background music for Web photo albums. In contextual video advertising, the systems VideoSense [22] and vADeo [24] have been built based on video content analysis.

Regarding contextual advertising on Web images, image annotation approaches [4, 11] can be leveraged. However, image annotation is not specifically designed for recommending advertisements. The authors of ImageSense [21] first proposed to match advertising with images. But ImageSense mainly used surrounding text for advertisement match while visual relevance acted as a complement to that information. To the best of our knowledge, ViCAD [8] is the only work trying to match advertisements for a target image without any textual information. In ViCAD, the authors built an image-text feature mapping using a graphical model and a language model. Then, the conditional probability of any advertisement for a target image was determined. ViCAD is a very relevant work and will be compared in our experiments.

3. A SEMANTIC APPROACH TO VISUAL CONTEXTUAL ADVERTISING

In the field of contextual advertising, besides the direct syntactic `page-ad` matching, there are two major frameworks for matching the target Web page and advertisement.

```
page-keyword-ad
page-taxonomy-ad
```

In the `page-keyword-ad` framework [31], advertising keywords are extracted from the target Web page and then advertisements are matched with the keywords. In the `page-taxonomy-ad` framework [5], pages and advertisements are mapped to the same taxonomic structure and the semantic similarities are calculated using the mapping on the taxonomic hierarchy of pages and advertisements. Besides, the traditional syntactic matching is also combined into this framework.

For visual contextual advertising, the traditional image annotation approach is just like the `page-keyword-ad` approach and ViCAD corresponds to the syntactic matching. To the best of our knowledge, there is no previous work in visual contextual advertising using any semantic approach. In this section, we propose a semantic approach to visual contextual advertising, with the goal of improving the performance of the advertisement precision.

3.1 Problem Definition

First we formally define the problem of visual contextual advertising. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ be the text feature space, where t_i is a text feature and m is the size of the text feature space. Let $\mathcal{A}$ be the advertisement space and each advertisement $a \in \mathcal{A}$ is represented by a text feature vector $(t_a^1, t_a^2, \ldots, t_a^m)$, where t_a^k is the frequency of text feature t_k in a. Similarly, we denote image feature space $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, where v_i is an image feature and n is the size of image feature space. The image space is denoted as $\mathcal{I}$. And each image $\iota \in \mathcal{I}$ is represented by an image feature vector $(v_\iota^1, v_\iota^2, \ldots, v_\iota^n)$, where v_ι^k means the frequency of image feature v_k in ι. In addition, the text knowledge base is denoted as $\mathcal{O}_t = \{O_t, E_t\}$, where $O_t = \{ot_1, ot_2, \ldots, ot_\mu\}$ is the node set and $E_t = \{(ot_i, ot_j)\}$ is the edge set. Also the image knowledge base is defined as $\mathcal{O}_v = \{O_v, E_v\}$, where $O_v = \{ov_1, ov_2, \ldots, ov_\nu\}$ and $E_v = \{(ov_i, ov_j)\}$. For a given image $\iota \in \mathcal{I}$, the objective is to find the function $r(\iota, a) : \mathcal{I} \times \mathcal{A} \mapsto \mathbb{R}$ that accurately estimates the relevance of any candidate advertisement a to ι.

3.2 Semantic Visual Contextual Advertising Framework

In this subsection, we discuss the framework of semantic visual contextual advertising. As mentioned in Section 1, we first map the image and advertisement onto some nodes of interlinked knowledge bases. Since the feature spaces of image and text are heterogeneous, the image and text knowledge bases are always different. However, just like the image-text occurrence data, we can still find a way to match the nodes on the two knowledge bases[3]. We propose a framework for semantic matching of images and advertisements by building links between nodes of the image and text knowledge bases.

$$\texttt{image-knowledge}_{\texttt{image}}\texttt{-knowledge}_{\texttt{text}}\texttt{-ad}$$

In this framework, first, images and advertisements are mapped to nodes in the image and text knowledge bases, respectively. Then the matching between the nodes of interlinked image and text knowledge bases is processed. With the help of semantic link information in the knowledge bases, syntactic mismatches between the image features and text features can be reduced. Therefore, given a target image ι, the task of finding the best match advertisement can be written as

$$\arg\max_{a \in \mathcal{A}} \mathcal{M}(\psi(\iota), \phi(a)), \tag{1}$$

where

$$\psi(\iota) = \{(ov, \omega_{ov})\}_{ov \in O_v \; and \; \omega_{ov} > 0}, \tag{2}$$

$$\phi(a) = \{(ot, \omega_{ot})\}_{ot \in O_t \; and \; \omega_{ot} > 0}, \tag{3}$$

with

$$\sum_{ov \in O_v} \omega_{ov} = 1 \; and \; \sum_{ot \in O_t} \omega_{ot} = 1.$$

Here ψ and ϕ are the functions mapping text instances to nodes in the text knowledge base and image instances to nodes in the image knowledge base, respectively. Each mapped node is assigned a weight to express its relevance to the image or advertisement. $\mathcal{M}$ is a cross-knowledge base matching function for the two sets of weighted nodes on the combined structure of image and text knowledge bases. To sum up, our framework can be depicted as Figure 3.

[3] For ontology engineering, one of the most important processes is to find the match between two ontologies.

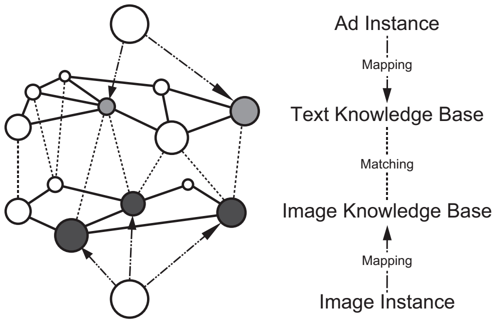


Figure 3: Framework of semantic matching between the image and advertisement.

The specific implementation of the functions in Equation 1 depends on the knowledge used, which will be discussed in detail later.

3.3 Knowledge Bases

In this section, we introduce the specific knowledge bases we use in this framework and the bridging knowledge data between the nodes of two knowledge bases.

Text Knowledge Base: Wikipedia. We use *Wikipedia* as the text knowledge base $\mathcal{O}_t$ in our framework. Wikipedia is a user-contributed online encyclopedia. It contains numerous entities with a formatted article description and interlinks to other relevant entities. Each entity article is written or revised by Web users so as to lead a comprehensive description of the entity. In addition, the interlinks between each two entities serve as auxiliary information and further explanation, which indicates their semantic relatedness. In sum, Wikipedia is a large-scale qualified knowledge base: so far in March 2012, it has more than 3.9 million articles written in English, with 19.67 edits for each article[4].

Image Knowledge Base: ImageNet/WordNet. We choose *ImageNet* [10] as the image knowledge base $\mathcal{O}_v$ in our framework. ImageNet is an image database organized according to *WordNet* [13]. WordNet is composed of synsets, each of which is described by several synonyms. The edges linking two synsets provides the semantic relation between them. The kinds of edges include: `antonym`, `hypernym`, `instance hyponym`, `part meronym`, `derivationally related form`, `member of this domain`, and so on. We regard synsets as concept nodes. The `hypernym` edges are used to construct a node hierarchy. Currently, there are 14.2 million images and 21.8 thousand nodes indexed in ImageNet[5]. Each node is assigned 1000 images on average. Images of each concept are human-annotated and have high quality. Therefore, using ImageNet, each node $ov_i \in \mathcal{O}_v$ is represented as a set of images.

Bridging Knowledge: YAGO. We connect Wikipedia nodes and ImageNet/WordNet nodes using YAGO [27]. Each Wikipedia node is labeled with types in YAGO's taxonomy, which is built on the topology of WordNet. Thus we can obtain a list of WordNet nodes for each Wikipedia node. For example, `Wikipeida::Aristotle` has the type of `WordNet::Person`, `WordNet::Scientist`, etc.

3.4 Mapping and Matching Functions

In this section we introduce the definition of the functions in our framework, based on the image and text knowledge bases discussed in Section 3.3.

3.4.1 *Text Mapping Function ϕ*

Given a candidate advertisement in bag-of-words form, ϕ maps a to the relevant nodes on $\mathcal{O}_t$. Here $\mathcal{O}_t$ represents the set of Wikipedia entities each with an article description. Because advertisement content is in short-text and diversely written, it is usually difficult to directly find Wikipedia entity names in advertisement content. For this reason, we make use of explicit semantic analysis (ESA) [14] to find the most relevant Wikipedia entities for each candidate advertisement.

Here each mapped node weight ω_{ot} in Equation 3 is defined by the ESA association strength. A widely used choice [26] is to select tfidf weighting

$$\omega_{ot} = \sum_{t \in ot} tfidf_{ot}(t) \cdot tf_a(t), \quad (4)$$

where $tfidf_{ot}(t)$ is the product of the frequency and inverse document frequency of t in the article of ot, $tf_a(t)$ is the frequency of a word or phrase t in the advertisement dataset. Particularly, top-3 weighted nodes are selected.

3.4.2 *Image Mapping Function ψ*

Given a target image $\iota = \{v_\iota^1, v_\iota^2, \ldots, v_\iota^n\}$, ψ maps ι to the relevant nodes on $\mathcal{O}_v$. Different from mapping advertisement content to nodes of Wikipedia as ϕ, the image mapping function ψ is closer to multi-label classification. Each node in ImageNet has about 1000 image instances; these are used as the training data and the target image is regarded as test data. Specifically, we use a node-level centroid based similarity function $\theta(\iota, ov)$ to obtain the closest k node set C_k to the target image ι. Specifically, θ can be implemented as cosine similarity after the process of principle component analysis (PCA) [18]. Moreover, since ImageNet has a hierarchical structure, we can implement a hierarchical centroid algorithm which leverages the ancestor information in the similarity calculation. Finally, the weight ω_{ov} for the mapped node ov is defined by the (normalized) similarity between ι and ov, calculated as

$$\omega_{ov} = \left(\theta(\iota, ov) \prod_{ov' \in A(ov)} \theta(\iota, ov')^{\omega} \right)^{\frac{1}{|A(ov)|\omega+1}}, \quad (5)$$

where $A(ov)$ denotes the set of ancestors of ov and ω is the weight assigned to each ancestor node; these are combined with a geometric mean. With leveraged ancestor information, ψ is less likely to map ι to irrelevant nodes. Particularly, we set $k = 7$ and $\omega = 0.6$ in our experiment, after preliminary parameter tuning.

3.4.3 *Cross-Knowledge Base Matching Function $\mathcal{M}$*

Above we have elaborated the text and image mapping functions which map the advertisements and images to Wikipedia and ImageNet/WordNet. In addition, these two knowledge bases could be bridged via YAGO (Section 3.3). Thus we can regard them as a combined knowledge base. Now we introduce the matching function $\mathcal{M}$ between the two disjoint sets of weighted nodes on the combined knowledge

[4] `http://en.wikipedia.org/wiki/Special:Statistics`

[5] `http://image-net.org/about-stats`

base. Here we uniformly use o_i to represent ov_i and ot_i since the two knowledge bases have been combined. We also define the mapped node sets of image and advertisement as O_ι and O_a respectively. We discuss two implementations of the cross-knowledge base matching function $\mathcal{M}$.

LOD Description Overlap (LODDO). This is an approach proposed in [32] for evaluating named entity semantic relatedness on linked open data (LOD). The authors propose to regard the neighborhood of an entity o in LOD as its description $\delta(o)$, defined as the set of entities linked to o. And the similarity between entity o_i and o_j is defined as the description overlap between $\delta(o_i)$ and $\delta(o_j)$.

$$LODDO(o_i, o_j) = \frac{|\delta(o_i) \cap \delta(o_j)|}{\min(|\delta(o_i)|, |\delta(o_j)|)} \tag{6}$$

Since Wikipedia and ImageNet/WordNet are also members of LOD, this approach can seamlessly be adapted to our matching function. The matching function between the target image ι and a candidate advertisement a can be calculated as the weighted average of the similarity of each image-advertisement entity pair.

$$\mathcal{M}(\psi(\iota), \phi(a)) = \sum_{o_i \in O_\iota} \sum_{o_j \in O_a} \omega_{o_i} \omega_{o_j} LODDO(o_i, o_j) \tag{7}$$

Hierarchy-based Matching. Taxonomy-based semantic matching has been used in contextual advertising [5]. As has been mentioned in Section 3.3, YAGO does provide a shared taxonomy between Wikipedia and WordNet. Thus we can map the nodes in both knowledge bases to a common taxonomy hierarchy, where we can implement hierarchy-based matching. Matching function $\mathcal{M}$ can be written as

$$\mathcal{M}(\psi(\iota), \phi(a)) = \left(\sum_{o_i \in O_\iota} \sum_{o_j \in O_a} \omega_{o_i} \omega_{o_j} LCA(o_i, o_j) \right)^{-1}, \tag{8}$$

where $LCA(o_i, o_j)$ means the maximal path length from o_i and o_j to their least common ancestor [5].

In the experiment, we will compare the above two cross-knowledge base matching functions to explore how to provide cross-media semantic matching appropriately.

3.5 Algorithm Chart

So far we have introduced our framework of semantic matching between an image ι and an advertisement a. Now the practical task is to retrieve and rank the relevant advertisements for a given image ι,. Since expansion and matching of graph structures are involved in our matching algorithm, it is very inefficient to traverse the advertisement dataset to perform a match between each advertisement and the target image. Here we propose the algorithm flow to efficiently solve the problem (see Figure 4).

In an offline process, we pre-calculate a set of relevant Wikipedia nodes for each advertisement a using ESA. Thus we can build an inverse advertisement index for each node, like the document index to each keyword in a search engine. For the online process, with a target image ι as input, first we use image mapping function ψ to get k ImageNet nodes $\psi(\iota)$. Then we link mapped ImageNet nodes to Wikipedia nodes via YAGO. With the advertisement index above, we can retrieve the indexed advertisements for each linked Wikipedia node, which lead to the candidate advertisement list $\mathcal{L}$. For each advertisement a in $\mathcal{L}$, calculate the similarity with ι via ontology matching function $\mathcal{M}$. Finally, rank the candidate advertisement list in descending order by the similarity score, get the top N advertisements as the output.

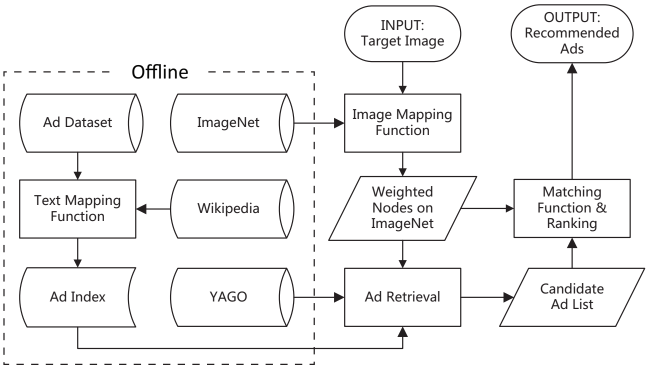


Figure 4: Algorithm Flow Chart.

3.6 Complexity

For the image mapping process, each image-centroid similarity function θ takes $O(|\mathcal{V}|)$. Thus the image mapping function takes $O(|O_v| \cdot |\mathcal{V}| + |O_v| \log |O_v|)$, where the notation is as in Section 3.1. In practice, the image feature number $|\mathcal{V}|$ is much larger than $\log |O_v|$. Thus the complexity of the image mapping function is $O(|O_v| \cdot |\mathcal{V}|)$.

For the matching process, let n_a be the maximum number of advertisements one Wikipedia node could retrieve, D be the maximum out-degree of WordNet nodes in YAGO. Thus the maximum number of candidate advertisements is $k_\iota \cdot n_a \cdot D$. For the LODDO matching function, the complexity is $O(k_\iota \cdot k_a \cdot N_\delta \log N_\delta)$, where N_δ is the description size. For the hierarchy-based matching function, the complexity is $O(k_\iota \cdot k_a \cdot d)$, where d is the depth of the YAGO taxonomy hierarchy, k_ι and k_a are the maximum number of mapped nodes for images and advertisements respectively. Thus the complexities of the matching processes are $O(n_a \cdot k_\iota^2 \cdot k_a \cdot D \cdot N_\delta \log N_\delta)$ and $O(n_a \cdot k_\iota^2 \cdot k_a \cdot d)$ respectively. In practice, $D \cdot N_\delta \log N_\delta)$ and d are not large numbers ($d < 15$ and $N_\delta < 30$). Uniformly, we use c to denote the upper bound of these two numbers and the matching complexity is $O(n_a \cdot k_\iota^2 \cdot k_a \cdot c)$.

To sum up, the overall complexity of the online algorithm is $O(|O_v| \cdot |\mathcal{V}| + n_a \cdot k_\iota^2 \cdot k_a \cdot c)$.

In our experiment, the average real run time for each test case is 0.751 seconds on a machine with an Intel(R) Core-2(TM) Quad Q8400 CPU with 2 cores at 2.6GHz and 2GB memory. Furthermore, the efficiency can be further improved with the optimization such as parallelization in the image ontology match process and advertisement index pruning.

4. EXPERIMENT

In this section, we introduce the datasets, compare algorithms and evaluation measures, and finally report and discuss our experimental results.

4.1 Datasets

4.1.1 Advertisement Dataset

The textual advertisements can be crawled from a mainstream commercial search engine. Specifically, we use AOL query log [16] as query set and then crawl the delivered advertisements on the search engine result page (SERP) for each query during March 2011. Specifically, there are 9,954,130 queries in the AOL dataset, where 1,118,729 queries attract at least one advertisement. As a result, we collect 1,607,688 unique advertisements.

For each advertisement, we crawl its title, creative, and display URL, as has been shown in Figure 1.

4.1.2 *Knowledge Bases*

As has been discussed in Section 3.2, there are text and image knowledge bases (Wikipedia and ImageNet/WordNet) and bridging data (YAGO).

Wikipedia - We obtained the Wikipedia dump of Jan.5, 2012. We selected the Wikipedia articles representing concrete concepts using heuristics similar to [14], resulting in a collection of 1,521,080 concept nodes. We use Lucene[6] to build the ESA index from articles describing the concepts.

ImageNet/WordNet - For WordNet structure, we download WordNet 3.0[7] and remove the edges with negative semantics (`Antonym`). The knowledge base contains 117,659 nodes and 377,592 edges, where 97,666 are `Hypernym` edges.

For the image data, we take the 1,000 ImageNet synsets released on April 30, 2010 which contain 2,522,812 images. Each image in this dataset has SIFT features extracted and 1000-clustered bag of words. To investigate whether the size of the image knowledge base is large enough to provide relevant advertisements, we will drive an experiment about the performance against the number of ImageNet nodes in Section 4.4.2.

YAGO - To connect Wikipedia and ImageNet/WordNet, we take YAGO dataset of `type_star` in version `yago2core-20120109`. On average, each Wikipedia concept is mapped to 25.2 WordNet concepts. In all, 4,564 WordNet concepts have at least one corresponding Wikipedia concept[8].

4.1.3 *Target Image Dataset*

In our experiment, we use a Flickr image set as our target dataset. This dataset contains 521 thousand images crawled from Flickr during 2010. Considering the large effort of human judgement, we randomly selected 230 images as the target images for testing[9].

The data preprocessing is the same as ImageNet. First we detect the interesting points for each image using SIFT descriptors [19]. Then we cluster 1,000 categories (same as [25]) for all interesting points to obtain a codebook, which turns out to be the image feature space and each image can be represented by image-bag-of-words. These image features are used in the similarity function $\theta(\iota, ov)$.

4.2 Compared Algorithms

Since there are few methods for visual contextual advertising except ViCAD, we compared all the methods that work [8]. The algorithms are listed below.

Annotation + Search (AS). First, the target image is annotated [20]. Then advertisements are retrieved and ranked by a search process using the annotations as query. The search engine is built based on Lucene. This work is just like the keyword-based methods used in traditional contextual advertising.

Annotation + Expansion + Search (ASEx). One intuitive approach to adding semantic matching into the traditional AS approach is to expand the extracted annotations using a semantic knowledge base and then search the advertisements with the expanded query set. Specifically, we implement ASEx similar to the work [17].

ViCAD. The heterogeneous transfer learning based ViCAD proposed in [8] has been discussed in Section 2.2.

ImageAdSense. This is our approach and the algorithm has been discussed in Section 3. In order to compare different matching functions, we implement LODDO and the hierarchy-based matching function mentioned in Section 3.4.3, denoted as iAdSense-LODDO and iAdSense-Tree. In order to investigate the impact of a cross-knowledge-base matching function, here we add an algorithm iAdSense-OneLayer, which only has ImageNet/WordNet. The mapping of advertisements to WordNet nodes is based on syntactic match.

4.3 Evaluation Measure

The input of the experiment is a target image ι and the output is k advertisements for ι. As the basis of the evaluation work, we invited six college students to judge the relevance of each image-advertisement pair as below.

- *Relevant.* The advertisement is relevant to the content of the target image, scored as 1.
- *Irrelevant.* The advertisement is not considered relevant to the content of the target image, scored as 0.

Each image-advertisement pair has at least two human judges. Then, we averaged the scores for each image-advertisement pair. Then we evaluated the performance of the algorithms using $P@n$ as the evaluation measure. Precision at position n ($P@n$) is defined to be the fraction of the top-n retrieved advertisements that are relevant [3].

$$P@n = \frac{\sum_{i=1}^{n} \pi_i}{n} \tag{9}$$

In Equation 9, π_i denotes the average rate score for the pair of the target Web page and the ith recommended advertisements. Since we cannot evaluate every image-advertisement pair, there is no good measure to evaluate the recall of each approach.

4.4 Experimental Results

4.4.1 *Overall Performance Analysis*

In the first part of the experiment, we judge the overall recommendation performance of the compared algorithms on test dataset. For the 230 test images, each algorithm recommends 10 top ranked advertisements. We use the evaluation measure $P@n$ (see Section 4.3) for the recommendation performance. The result for six algorithms is provided in Figure 5.

From Figure 5 we can have the following observation. (i) Three iAdSense-algorithms provide much better performance than AS, ASEx and ViCAD. The absolute improvement of P@10 of iAdSense-LODDO is 16.4% and 20.7%, compared with ASEx and ViCAD respectively, which verifies the

[6] `http://lucene.apache.org/`

[7] `http://wordnet.princeton.edu/wordnet/download/`

[8] Although the ratio of involved WordNet concepts is not high, these concepts are usually the representative category labels, which have links to most of WordNet Concepts.

[9] As a reference, 200 test images were selected in the experiment of previous work [8].

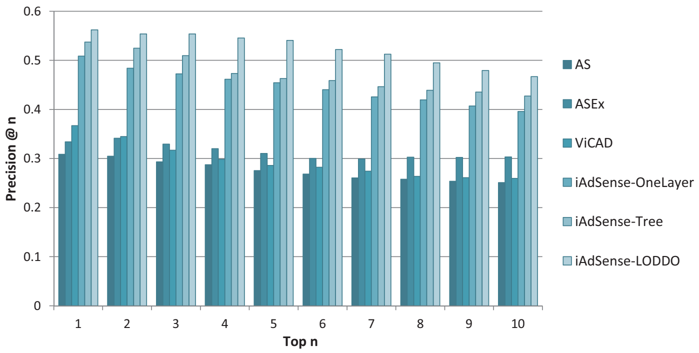


Figure 5: Figure representation of $P@n$ results of all compared algorithms on test dataset.

impact of semantic matching. (ii) In the comparison among these three iAdSense-algorithms, iAdSense-LODDO performs the best. This indicates that the semantic relatedness approach LODDO is well adapted to our framework. iAdSense-Tree has a little lower precision. This is because only hierarchy edges are used in iAdSense-Tree, while iAdSense-LODDO makes use of all edges of each node to provide a more comprehensive semantic description. iAdSense-OneLayer is not as good as others with two layers. This indicates the necessity of semantic text mapping. Syntactically mapping advertisement content to its words in WordNet will import much ambiguity since each word always occurs in several WordNet synsets. (iii) ViCAD outperforms AS but is not as good as iAdSense algorithms. The reason ViCAD is not as good as iAdSense-algorithms is the frequent noise in the tags of training images, which reduces the accuracy of cross-domain feature transferring. In addition, ViCAD is also a syntactic match approach and has the same problems as AS.

To sum up, the above comparison shows that iAdSense is more effective than previous approaches.

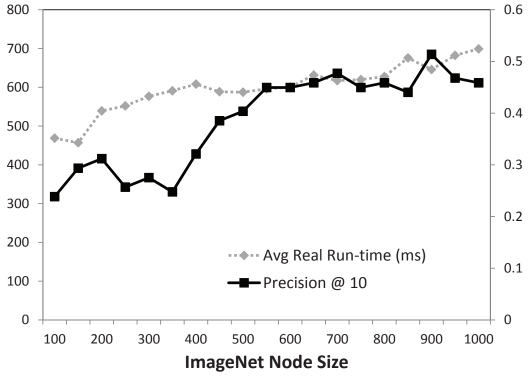
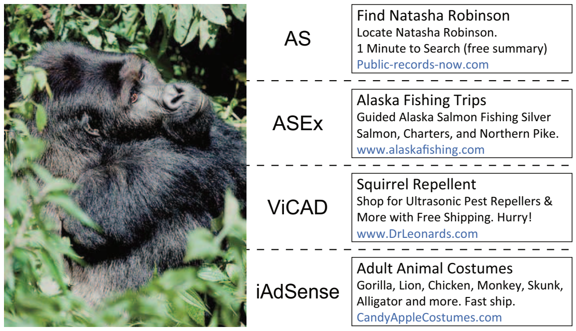


Figure 6: $P@10$ of iAdSense against the number of randomly selected ImageNet nodes.

4.4.2 *ImageNet Scale Analysis*

As mentioned in Section 4.1.2, we should investigate the recommendation performance of iAdSense against the number of ImageNet nodes and check whether it is enough to take the 1,000 ImageNet synsets released with SIFT features. Specifically, we vary the number of **randomly selected** ImageNet nodes from 50 to 1,000 with a step of 50. Then we evaluate the performance of iAdSense-LODDO in the same way as above[10]. The result is shown in Figure 6.



Find Natasha Robinson
Locate Natasha Robinson.
1 Minute to Search (free summary)
Public-records-now.com

Alaska Fishing Trips
Guided Alaska Salmon Fishing Silver Salmon, Charters, and Northern Pike.
www.alaskafishing.com

Squirrel Repellent
Shop for Ultrasonic Pest Repellers & More with Free Shipping. Hurry!
www.DrLeonards.com

Adult Animal Costumes
Gorilla, Lion, Chicken, Monkey, Skunk, Alligator and more. Fast ship.
CandyAppleCostumes.com

Figure 7: Advertisements recommended by the compared algorithms on one case.

From Figure 6 we can see that (i) as the number of nodes increases, the $P@10$ performance of iAdSense improves and its real run-time increases. (ii) The precision curve has a sigmoid-shaped trend: $P@10$ fluctuates without an obvious increase when number of nodes varies from 50 to 350; in the range of [350, 700], $P@10$ increases rapidly; after 700, $P@10$ fluctuates around 0.475. (iii) The real run-time curve has a stable increase rate against the number of ImageNet nodes. This is because the image mapping process is implemented as a hierarchical centroid algorithm, a memory-based approach, so more ImageNet synsets will surely bring an efficiency decrease. To sum up, 1,000 ImageNet size is a suitable scale for iAdSense considering both effectiveness and efficiency.

4.4.3 *Case Study*

Here we demonstrate a case that makes a difference among the compared algorithms. Figure 7 provides some advertisements recommended by the four algorithms for a target image about a gorilla. From the results we can find AS recommends an irrelevant advertisement. For ASEx, there is a topic drift between the target image and advertisement, which is caused by annotation expansion. ViCAD recommends a syntactic match advertisement. However, *squirrel* here refers to a brand name instead of a kind of animal, which is a case of semantic mismatch. iAdSense-LODDO recommends a suitable advertisement, where *Gorilla* in the advertisement refers to the animal in the target image.

Finally, we provide more cases of the results of semantic visual contextual advertising with respect to the test dataset. In Figure 8, there are two advertisements listed on the right of each target image. These advertisements are recommended by algorithm iAdSense-LODDO. More demonstrations are presented on the Web site of ApexLab[11].

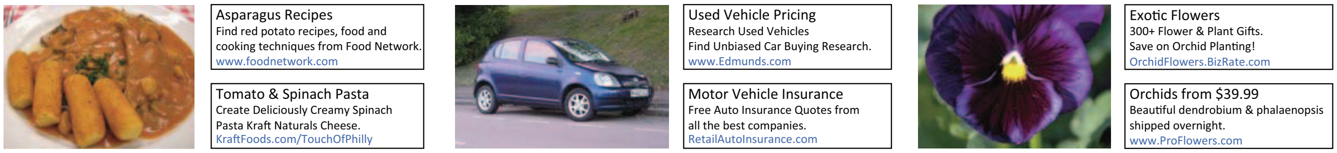
5. CONCLUSION AND FUTURE WORK

We investigate the current work on visual contextual advertising and point out the problems of semantic mismatch despite a syntactic match between image and advertisement content. In order to solve these problems, we proposed a semantic approach named iAdSense with the help of text and image knowledge bases. In the experiment, iAdSense provides an improvement of 16.4% over the previous approaches, with more semantically relevant advertisements recommended.

In future work, we will explore other knowledge bases to help in this framework. For example, we can use a more com-

[10] Due to the huge human labeling effort, the test set here is a subset of the test dataset.

[11] Online demo. `http://iadsense.apexlab.org`



Asparagus Recipes
Find red potato recipes, food and cooking techniques from Food Network.
www.foodnetwork.com

Tomato & Spinach Pasta
Create Deliciously Creamy Spinach Pasta Kraft Naturals Cheese.
KraftFoods.com/TouchOfPhilly

Used Vehicle Pricing
Research Used Vehicles
Find Unbiased Car Buying Research.
www.Edmunds.com

Motor Vehicle Insurance
Free Auto Insurance Quotes from all the best companies.
RetailAutoInsurance.com

Exotic Flowers
300+ Flower & Plant Gifts.
Save on Orchid Planting!
OrchidFlowers.BizRate.com

Orchids from $39.99
Beautiful dendrobium & phalaenopsis shipped overnight.
www.ProFlowers.com

Figure 8: Case study of iAdSense-LODDO results.

mercially relevant knowledge base to explore a better advertisement mapping. Moreover, we will work on the application of visual contextual advertising to E-commerce such as Taobao[12]. The input will be a product image and some relevant products will be recommended. In this topic, more specific image features will be selected and more information can be obtained from the product pages.

6. REFERENCES

[1] Ijcai-09 workshop on cross-media information access and mining.

[2] A. Anagnostopoulos, A. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel. Just-in-time contextual advertising. *CIKM*, 2007.

[3] Baeza-Yated, R., Ribeiro-Neto, and B. *Modem Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, 2008.

[4] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, M. I. Jordan, J. K, T. Hofmann, T. Poggio, and J. Shawe-taylor. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.

[6] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *Proceeding of the 17th international conference on World Wide Web*, pages 417–426. ACM, 2008.

[7] J. Chao, H. Wang, W. Zhou, W. Zhang, and Y. Yu. Tunesensor: A semantic-driven music recommendation service for digital photo albums. In *Proceedings of 16th International Semantic Web Conference*, 2011.

[8] Y. Chen, O. Jin, G. rong Xue, J. Chen, and Q. Yang. Visual contextual advertising: Bringing textual advertisements to images. In *Proceedings of the 24th AAAI Conference*, AAAI'10, 2010.

[9] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS'08*, pages 353–360, 2008.

[10] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.

[11] P. Duygulu, K. Barnard, N. de Freitas, P. Duygulu, K. Barnard, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. 2002.

[12] E. Even Dar, V. S. Mirrokni, S. Muthukrishnan, Y. Mansour, and U. Nadav. Bid optimization for broad match ad auctions. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 231–240, New York, NY, USA, 2009. ACM.

[13] C. Fellbaum. Wordnet: An electronic lexical database. In *Proceedings of 11th Eurographics Workshop on Rendering*. MIT Press, 1998.

[14] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007.

[15] S. Gupta, M. Bilenko, and M. Richardson. Catching the drift: learning broad matches from clickthrough data. In *Proceedings of the 15th international conference on Knowledge discovery and data mining*, 2009.

[16] IAB and PricewaterhouseCoopers. Iab internet advertising revenue report, 2010 full year results, april 2011.

[17] N. James and C. Hudelot. Towards semantic image annotation with keyword disambiguation using semantic and visual knowledge. In *Proceedings of IJCAI-09*.

[18] I. Jolliffe. Principal component analysis. In *Encyclopedia of Statistics in Behavioral Science*, 2005.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[20] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the International Conference on Computer Vision*, pages 316–329, 2008.

[21] T. Mei, X. Hua, and S. Li. Contextual in-image advertising. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 439–448. ACM, 2008.

[22] T. Mei, X. Hua, L. Yang, and S. Li. Videosense: towards effective online video advertising. In *Proceedings of the 15th international conference on Multimedia*, pages 1075–1084. ACM, 2007.

[23] A. Pak and C. Chung. A wikipedia matching approach to contextual advertising. *World Wide Web*, 13(3):251–274, 2010.

[24] S. Sengamedu, N. Sawant, and S. Wadhwa. vadeo: video advertising system. In *Proceedings of the 15th international conference on Multimedia*, pages 455–456. ACM, 2007.

[25] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proceedings of the International Conference on Computer Vision*, 2005.

[26] P. Sorg and P. Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. *Natural Language Processing and Information Systems*, pages 36–48, 2010.

[27] F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.

[28] C. Wang, P. Zhang, R. Choi, and M. Eredita. Understanding consumers attitude toward advertising. In *Eighth Americas Conference on Informatino System*, pages 1143–1148, 2002.

[29] X. Wu and A. Bolivar. Keyword extraction for contextual advertisement. In *Proceedings of the 18th World Wide Web Conference*, pages 1195–1196, 2008.

[30] Q. Yang, Y. Chen, G. rong Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the social web. In *In Proc. of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 2009.

[31] W.-t. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 213–222, New York, NY, USA, 2006. ACM.

[32] W. Zhou, H. Wang, C. Jiansong, W. Zhang, and Y. Yu. Loddo: Using linked open data description overlap to measure semantic relatedness between named entities. In *Proceedings of Joint International Semantic Technology Conference*, 2011.

[12] `http://www.taobao.com`