Mining Text Snippets for Images on the Web

Anitha Kannan
Microsoft
ankannan@microsoft.com

Simon Baker
Microsoft
sbaker@microsoft.com

Krishnan Ramnath
Microsoft
kramnath@microsoft.com

Juliet Fiss[*]
University of Washington
juliet@cs.washington.edu

Dahua Lin[†]
TTI Chicago
dhlin@ttic.com

Lucy Vanderwende
Microsoft
lucyv@microsoft.com

ABSTRACT

Images are often used to convey many different concepts or illustrate many different stories. We propose an algorithm to mine *multiple* diverse, relevant, and interesting text snippets for images on the web. Our algorithm scales to *all* images on the web. For each image, *all* webpages that contain it are considered. The top-K text snippet selection problem is posed as combinatorial subset selection with the goal of choosing an optimal set of snippets that maximizes a combination of relevancy, interestingness, and diversity. The relevancy and interestingness are scored by machine learned models. Our algorithm is run at scale on the entire image index of a major search engine resulting in the construction of a database of images with their corresponding text snippets. We validate the quality of the database through a large-scale comparative study. We showcase the utility of the database through two web-scale applications: (a) augmentation of images on the web as webpages are browsed and (b) an image browsing experience (similar in spirit to web browsing) that is enabled by interconnecting semantically related images (which may not be visually related) through shared concepts in their corresponding text snippets.

Categories and Subject Descriptors

H.2.8g [**Information Systems**]: Database Management: Database Applications: Image Databases

General Terms

Data Mining Algorithms, Applications

Keywords

Text mining for images; Text snippets; Interestingness; Relevance; Diversity; Browsing; Semantic image browsing; Web image augmentation

1. INTRODUCTION

The popular adage “a picture is worth a thousand words” reflects the effectiveness of an image in conveying many different ideas or “stories.” An image is typically used to illustrate one of these stories in any given webpage. However, the same image can be used in different webpages to illustrate different stories at different levels of descriptiveness and interestingness. By collecting the stories associated with an image together, we can create a host of new applications that seamlessly integrate the image and text modalities.

The focus of our paper is exactly this: We propose a mining algorithm to obtain the most relevant and interesting text snippets for images on the web. Using this database of images and their associated text snippets, we present two new applications that we have implemented at web scale. We believe, however, that these applications represent the “tip of the iceberg”: many more applications are possible.

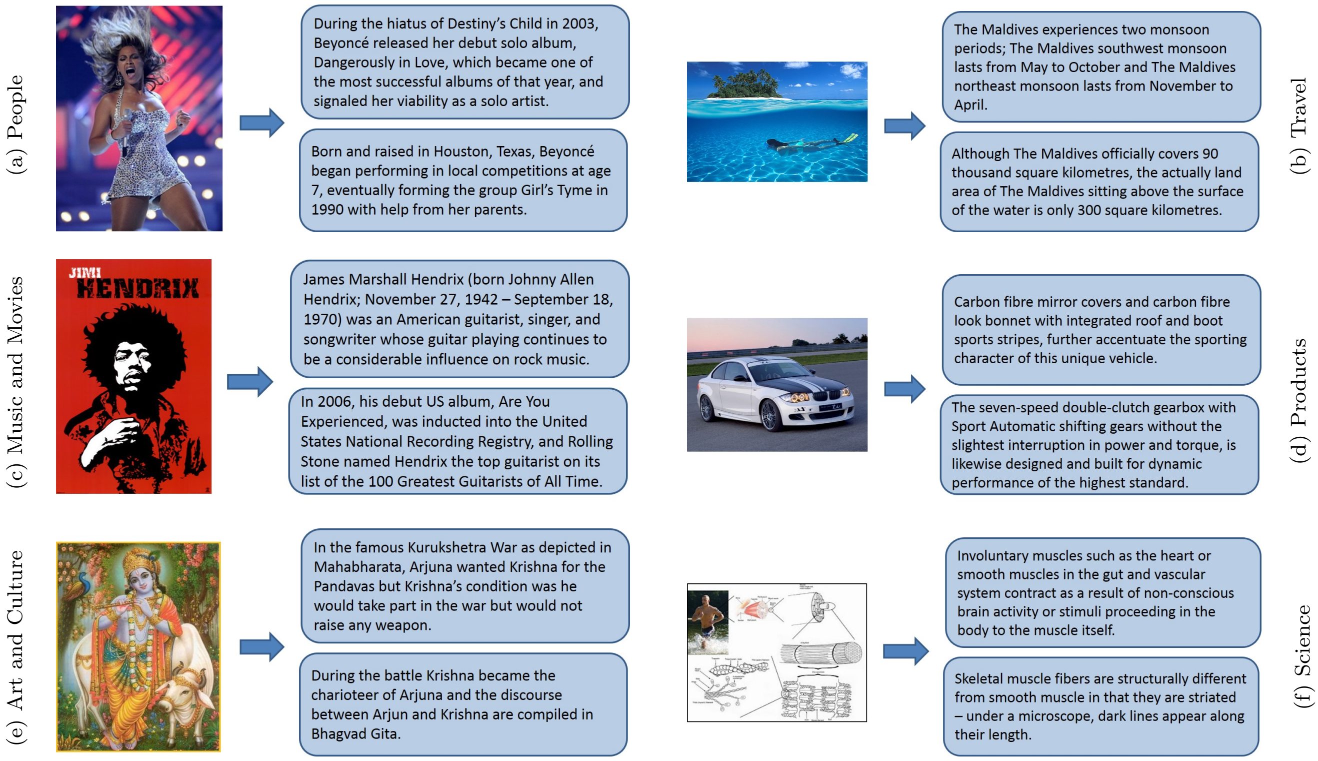
Figure 1 shows a sample of results for a few images. The images are selected to cover many of the important types of images on the web; people, travel, music, etc. For each image, we show two of the many snippets identified for it. The snippets are related to the image in the most general sense: memories or events behind the image, an analogy to establish context, or even descriptions of a historical event in the context of the image. Note, however, that the text generally goes far beyond a simple description of the visual contents of the image. For example, a visual description of Figure 1(b) might be “someone snorkeling in the Maldives.” Instead, our mined snippets include interesting information about the monsoon season and the geography of the islands, information that is not apparent in the image itself.

Our first contribution is to propose a scalable solution to mine the web for multiple relevant, interesting and diverse textual snippets about the image. The scale of our solution has two dimensions: we consider *all* images on the web and for each image we consider *all* web pages that contain the corresponding image. This scale enables reliable identification of multiple high quality snippets for images that are discussed widely on the web. Note that our approach differs considerably from prior work in understanding images [7, 13, 12, 16, 20, 22, 28], where the goal is to synthesize textual descriptions of *the visual content of the image* by recognizing objects, their attributes, and possibly using similar images in image databases to borrow captions.

With the web as a repository of world knowledge, we operationalize our solution based on a key insight: if an image is interesting, multiple people will post it in webpages, including stories related to it. These stories will vary in their

[*] Research conducted while an intern at Microsoft.
[†] Research conducted while an intern at Microsoft.

KDD'14 August 24–27, 2014, New York, NY, USA
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623346 .



(a) People

During the hiatus of Destiny's Child in 2003, Beyoncé released her debut solo album, Dangerously in Love, which became one of the most successful albums of that year, and signaled her viability as a solo artist.

Born and raised in Houston, Texas, Beyoncé began performing in local competitions at age 7, eventually forming the group Girl's Tyme in 1990 with help from her parents.

(b) Travel

The Maldives experiences two monsoon periods; The Maldives southwest monsoon lasts from May to October and The Maldives northeast monsoon lasts from November to April.

Although The Maldives officially covers 90 thousand square kilometres, the actually land area of The Maldives sitting above the surface of the water is only 300 square kilometres.

(c) Music and Movies

James Marshall Hendrix (born Johnny Allen Hendrix; November 27, 1942 – September 18, 1970) was an American guitarist, singer, and songwriter whose guitar playing continues to be a considerable influence on rock music.

In 2006, his debut US album, Are You Experienced, was inducted into the United States National Recording Registry, and Rolling Stone named Hendrix the top guitarist on its list of the 100 Greatest Guitarists of All Time.

(d) Products

Carbon fibre mirror covers and carbon fibre look bonnet with integrated roof and boot sports stripes, further accentuate the sporting character of this unique vehicle.

The seven-speed double-clutch gearbox with Sport Automatic shifting gears without the slightest interruption in power and torque, is likewise designed and built for dynamic performance of the highest standard.

(e) Art and Culture

In the famous Kurukshetra War as depicted in Mahabharata, Arjuna wanted Krishna for the Pandavas but Krishna's condition was he would take part in the war but would not raise any weapon.

During the battle Krishna became the charioteer of Arjuna and the discourse between Arjun and Krishna are compiled in Bhagvad Gita.

(f) Science

Involuntary muscles such as the heart or smooth muscles in the gut and vascular system contract as a result of non-conscious brain activity or stimuli proceeding in the body to the muscle itself.

Skeletal muscle fibers are structurally different from smooth muscle in that they are striated – under a microscope, dark lines appear along their length.

Figure 1: Examples of text snippets found by our algorithm. We just show the top two snippets, although generally many more are found. Note that the snippets are not purely descriptions of the content of the image. Instead, they contain all manner of interesting facts and stories related to the content.

content bringing about diversity. Note that these stories are generally not contained in the image captions (which are most often just descriptive), but the captions can help identify the most interesting stories. Concretely, we mine the web by first clustering all the images in the index of a Web search engine into "near duplicate" groups. For each group, we identify the top K snippets that are the most diverse, interesting and relevant, where the interestingness and relevancy are scored using learned models. The snippet selection problem is posed as the combinatorial optimization problem of subset selection. Since each image set can be processed independently, our algorithm easily parallelizes. We have run our algorithm on the entire web image index of a search engine containing billions of images.

Our second contribution is to demonstrate how the resulting database of images along with their snippets can be used to construct two new applications. Both applications are bulit at web scale(§ 5). The first application (Figure 8 and § 5.1) is a plugin to an internet browser than enables an enhanced reading experience on the web by displaying text snippets on top of images as the user passively browses the web. The second application (Figure 9 and § 5.2) is a real time web image browser that provides a seamless browsing experience through the most interesting images on the web, and their associated stories. Images are inter-connected semantically through the text snippets, unlike approaches that rely on visual similarity [10, 24].

2. RELATED WORK

There are two main bodies of related work. The first set focuses on analysing an image to generate a text description or caption. The second body of work is that of document (i.e. webpage) summarization. We discuss each in turn.

Image caption generation: Possibly the earliest work in this area is [20], which focused on associating word tags with an image by analyzing image regions at multiple granularity. Subsequently, in [12, 16, 28] the semantics of the image pixels are inferred to generate image descriptions. In particular, [12] detects objects, attributes and prepositional relationships in the image. These detections are composed to generate descriptions of the image. In [28], the image is decomposed at multiple granularity to infer the objects and attributes. The outputs of this step are then converted into a semantic ontology based representation that feeds into a text generation engine. A related line of work poses caption generation as a matching problem between the input image and a large database of images with captions [7, 13, 22]. In [7], a learned intermediate semantic space between image features and the tokens in the captions are used to generate sentences for images. In comparison, [22] constructs a data set of 1 million images spanning over 87 classes. Given an input image, this approach transfers captions from the closest matching image in the data set. In [13], phrases extracted from the caption of the matched image are used to generate a descriptive sentence. The goal of all these techniques is to describe the image content while we would like to capture the stories conveyed in the context of the image, and is not restricted by what is contained in the image.

Document Summarization: There is a large body of work in automatic text summarization. See [15, 23] for surveys. In this literature, the goal is to summarize documents, either by identifying key phrases and sentences that are reflective of the focus of the document (extractive approaches) [9, 4, 5] or by paraphrasing the content (abstractive approaches) [3]. In [9], sentences are extracted from one or more articles about the same topic. In [4] key sentences from webpages are extracted for viewing on a mobile device. In another example, [3] summarizes a webpage using a probabilistic model of the gist of the webpage. The underlying goal of all these methods is to identify the central focus of the document to be summarized. In contrast, we would like to identify content related to the image. One can think of using these techniques if we know the region of text that corresponds to image, which is a hard task by itself [2].

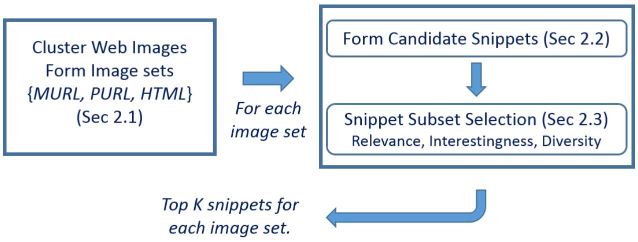

Cluster Web Images
Form Image sets
{MURL, PURL, HTML}
(Sec 2.1)
For each image set
Form Candidate Snippets (Sec 2.2)
Snippet Subset Selection (Sec 2.3)
Relevance, Interestingness, Diversity
Top K snippets for each image set.

Figure 2: Overview of Snippet Mining Algorithm: We first cluster all the images on the web into "near duplicate" groups. For each set of images we extract a set of candidate snippets. We then identify the top K snippets that are relevant, interesting and diverse.

3. SNIPPET MINING ALGORITHM

Our algorithm is based on the insight that if an image is interesting, multiple people will embed it in a webpage and write about it. For each image (along with near-duplicates), one can mine the web for all the webpages containing it in order to identify text snippets that are relevant and interesting and also form a diverse set of text.

With this insight, we have developed a highly scalable solution to mine the web for the interesting text snippets related to an image. Figure 2 provides the overview. The web images are clustered into "near duplicate" groups (§ 3.1), referred to as "duplicate image set" or "image set," for short. While each image set corresponds to a single image, these images reside in multiple webpages, $PURL$ (page URL in which the image resides) with a unique media url ($MURL$). Hence, each image set is represented by a collection of triplets $\{MURL, PURL, HTML\}$ where $HTML$ is the associated HTML source of $PURL$. Using this representation, we identify candidate snippets for each image set (§ 3.2). Then, the top K snippets that are most diverse, interesting and relevant to the image (§ 3.3) are identified by posing the problem as the combinatorial optimization problem of subset selection. We use machine-learned models of relevance and interestingness to score the snippets which are used within the optimization framework (§ 3.3.1).

3.1 Scalable Image Set Identification

We would like to cluster images such that each cluster consists of images that are near duplicate to each other. Images within a near duplicate image set can differ in their sizes, compression ratios, cropping, color transformations and other minor edits. For this purpose, we require a clustering algorithm that covers (a) large variation within a duplicate image cluster while minimizing false positives and (b) is highly scalable for clustering billions of images on the web.

To satisfy both these requirements, we use two-step clustering method [26] that is scalably implemented using hashing techniques within MapReduce framework [6]. The technique combines global and local features, with global descriptors to discover seed clusters with high precision and the local descriptors to grow the seeds to obtain good recall.

For the global descriptors, we first partition each image into 8×8 blocks and compute the mean gray value for each block. The image is then divided further into 2×2 blocks, and a histogram of edge distributions in evenly-divided 12 directions is extracted from each block, plus one dimension of non-edge pixel ratio. The histograms of mean gray values and edge distributions are cascaded into a 116-dimensional raw feature vector. A pre-learnt Principle Component Analysis (PCA) model is then applied and the first 24 dimensions of the PCA-ed vector are retained. The local descriptor of an image is a truncated set of visual words defined on a visual codebook of 1 miilion codewords. The local features proposed in [27] are first extracted and then ranked by their average Difference-of-Gaussian (DoG) responses. The top-ranked 40 features are used as the local descriptors of an image. Two images are near duplicates if they have 6 visual words in common.

Seed clusters are generated in three steps. First, each image is encoded into a 24-bit hash signature by quantizing each dimension of the global descriptor against the PCA mean. These signatures then serve as keys to divide the images in a MapReduce framework (i.e. the Map step). Second, images with identical hash signatures are assigned to the same hash buckets. Each bucket is processed within one vertex of the MapReduce machine cluster (i.e. the Reduce step). Third, pair-wise distance matching is performed within a bucket.

We grow the seed clusters by merging them using local descriptors. For computational tractability, we assume that near duplicate images reside in neighboring buckets in the global feature space, where neighboring buckets are those whose corresponding signatures are different in at most two bits. We randomly select one image from each seed cluster and add to the cluster all the images from neighboring buckets that have 6 or more visual words in common.

3.2 Forming Candidate Snippets

Each duplicate image set is represented by triplets $\{MURL, PURL, HTML\}$ of the image url, page url and the HTML source of the page [1]. Along with this, we parse $HTML$ to obtain a linear ordering of the text and image nodes and this is maintained as W^{PURL}.

Corresponding to each text node[2] in W^{PURL}, we generate a candidate snippet $< s^{n,PURL}, l^{n,PURL} >$, where $s^{n,PURL}$ is the $n, PURL^{th}$ snippet text and $l^{n,PURL}$ is its location in W^{PURL}. The set of candidate snippets generated is

[1] We restrict triplets to be from English-language non-spam websites. We also exclude social networking and photo sharing sites that do not contain enough textual content.

[2] Non-English and/or ill-formed sentences are ignored

$\{< s^{n,PURL}, l^{n,PURL} >\}_{n=1}^{N^{PURL}}$ where N^{PURL} is the number of snippets.

Similarly, we represent an image as follows: For each image node corresponding to the $MURL$, we extract its associated "Alt" text or "Src" text. We denote this text $< m^{PURL}, L^{PURL} >$, where m^{PURL} is the text and L^{PURL} is the location of the image node in W^{PURL}.

3.3 Top K Snippet Selection

For each image, I, let $\mathcal{S}$ be the candidate set of snippets identified from the set of triples $\{MURL,\ PURL,\ HTML\}$ (§ 3.2). From this large candidate set, we would like to select K top snippets, $\mathcal{T} \subset \mathcal{S}$ which are not only *relevant* and *interesting* to I but also exhibit *diversity*. This selection can be guided by the objective function

$$\mathcal{F}(\mathcal{T}|I) = \lambda \sum_{s \in \mathcal{T}} rel(s)int(s) + (1-\lambda)H_0(\mathcal{T}), \qquad (1)$$

that trades-off total gain given by the sum of the product of relevance $rel(s)$ and interestingness $int(s)$ of each snippet with their diversity as measured by the entropy of the set $H_0(\mathcal{T})$ (larger the entropy, larger the diversity).

In our objective (eq. 1), we consider the product $rel(s)int(s)$ instead of sum $rel(s)+int(s)$ because we prefer snippets that are both simultaneously relevant and interesting. We also differentiate between relevance and interestingness and not use a single term since a highly relevant snippet may not be very interesting, and vice versa (See Figure 4).

The optimal subset satisfies:

$$\mathcal{T}^* = \underset{\mathcal{T} \in 2^{|\mathcal{S}|}, |\mathcal{T}|=k}{\arg\max} \quad (\mathcal{F}(\mathcal{T}|I)). \qquad (2)$$

While, one can solve for the optimal subset through exhaustive search over all possible subsets of $\mathcal{S}$, this is clearly prohibitive even for a reasonable size of $\mathcal{S}$.

However, for suitable choices of $rel(s)int(s)$ and $H_0(\mathcal{T})$, Eq. 1 is submodular. That is, it exhibits the property of "diminishing returns" so that the difference in the value of the function that a single element makes when added to an input set decreases as the size of the input set increases. Mathematically, if $X \subset Y \subset \mathcal{S}$, then adding an element $z \in \mathcal{S}$ to both X and Y should satisfy:

$$\mathcal{F}(X \cup z) - \mathcal{F}(X) \geq \mathcal{F}(Y \cup z) - \mathcal{F}(Y) \qquad (3)$$

As long as we choose $rel(s)int(s) \geq 0$ so that $\sum_{s \in \mathcal{T}} rel(s)int(s)$ is a monotonically increasing function, Eq. 1 will be submodular since entropy is submodular [11]. We describe one such choice for $rel(s)int(s)$ in § 3.3.1.

The advantage of submodular functions is that while being computationally difficult, we can design greedy solutions that are at most $(1 - \frac{1}{e})$ times worse than the optimal solution [21]. Using this theoretical guarantee, we design a greedy solution as follows: Initialize with a snippet that has the largest value of $rel(s)int(s)$. Then, iteratively add snippets, one at a time, such that the added snippet (along with previously chosen set of snippets) maximizes Eq. 1.

3.3.1 *Relevance and Interestingness*

The top-K snippet selection formulation (Eq. 1) uses separate functions to score a snippet for relevancy $rel(s)$ and interestingness $int(s)$ to the corresponding image. Here, we describe the instantiation that we use in this paper for these two components. We used separate machine learned models of regularized linear regression.

The regularized linear regressor captures the relationship between features, $\phi(\mathbf{x})$, extracted from the snippet $\mathbf{x}$ and its scores through the functional form $y = \mathbf{w}^T\phi(\mathbf{x})$. The details of the features used are provided momentarily. Given some training data $\mathbf{x}_i \in \mathcal{X}$ that has been annotated with the relevance (or interestingness) scores y_i, the corresponding parameters $\mathbf{w}$ are learned by solving the optimization function:

$$\mathbf{w}^* \;=\; \arg\min_{\mathbf{w}} \lambda ||\mathbf{w}||^2 + \sum_i (\mathbf{w}^T\phi(\mathbf{x}_i) - y_i)^2 \qquad (4)$$

where λ is a regularization parameter. The unique optimum of Eq. 4 can be found via gradient descent [14].

Features used:

First, we describe the vocabulary for representation: Let $\mathcal{M} = \cup_{PURL} \left[m^{PURL} \cup \left[\cup_n s^{n,PURL} \right] \right]$ be the union of the image representation and all the candidate snippets for the image set. Let $\mathcal{V}$ be the unigrams (*sans* stop words) identified in $\mathcal{M}$.

Each snippet s is represented using a $|\mathcal{V}|$ vector, $\mathbf{s}$, such that $\mathbf{s}[k]$ is the number of times that the k^{th} unigram in $\mathcal{V}$ occurs in s . Similarly, for image I, $\mathbf{m}$ is a $|\mathcal{V}|$ vector such that $m[k]$ is the sum of the number of times that the k^{th} unigram in $\mathcal{V}$ occurs in the union of the representations of the images across all $PURL$, $\cup_{PURL} \left[m^{PURL} \right]$.

The feature vector,$\phi(\cdot)$, consists of nine features from the following five groups of features:

Match Score: While $\mathbf{s}$ corresponds to a single snippet, $\mathbf{m}$ is an aggregate over all image representations (captions, Alt Text) across multiple instances of the image. Therefore, we measure similarity between $\mathbf{s}$ and $\mathbf{m}$ as a dot product between $\mathbf{s}$ and normalized $\mathbf{m}$:

$$\text{matchscore}(\text{s}, \text{I}) = \mathbf{s}^T \frac{\mathbf{m}}{||\mathbf{m}||} \qquad (5)$$

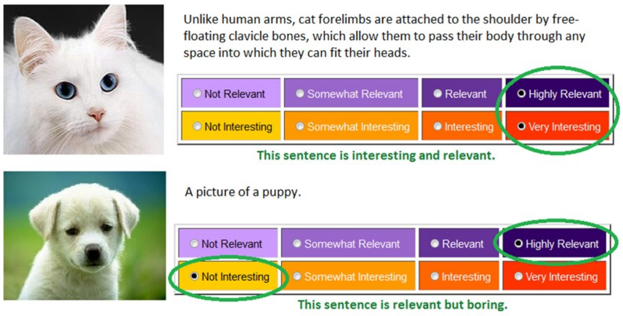
Context Scores: When the text around a snippet is relevant, then this specific snippet is likely to be relevant too. With this intuition, we compute two features which are average of the Match Score on windows around the snippet in question. Let $\rho_{j,n,PURL} = \delta\left[l^{j,PURL} \in [l^{n,PURL} - K, l^{n,PURL} + K] \right]$ be a indicator function which evaluates to 1 if $l^{j,PURL}$ is in the range $[l^{n,PURL} - K, l^{n,PURL} + K]$. The context scores are then given by: $\text{contextScore}(\text{s}^{\text{n},PURL}, \text{I})$

$$= \frac{\sum_{j=1}^{N^{PURL}} \quad \rho_{j,n,PURL} \text{matchscore}(\text{s}^{\text{j},PURL}, \text{I})}{\sum_{j=1}^{N^{PURL}} \quad \rho_{j,n,PURL}} \qquad (6)$$

In our experiments, we used two window sizes, corresponding to $K = 5$ and $K = 30$.

HTML Parse Distance: Good snippets tend to be closer to the image location. Therefore, we also capture distance from the snippet to the image node as:

$$\text{ParseDistance}(\text{s}^{\text{n},PURL}, \text{I}) = l^{n,PURL} - L^{PURL}. \qquad (7)$$



Unlike human arms, cat forelimbs are attached to the shoulder by free-floating clavicle bones, which allow them to pass their body through any space into which they can fit their heads.

This sentence is interesting and relevant.

A picture of a puppy.

This sentence is relevant but boring.

Figure 3: HIT used to collect training data to learn the snippet scoring functions. We showed image-snippet pairs to the judges and asked them to rate the snippet based on its "Relevance" to the image and on how "Interesting" it is. Some sentences are highly relevant, but not very interesting (bottom).

Measure of SPAM: When a snippet contains a lot of repeated words, it is less likely to be relevant or interesting. We use entropy of the words present to capture this idea: spamminess($\mathrm{s}^{n,PURL}$) =

$$-\sum_{k:\mathbf{s}^{n,PURL}[k]>0} \frac{\mathbf{s}^{n,PURL}[k]}{\sum_r \mathbf{s}^{n,PURL}[r]} \log\left[\frac{\mathbf{s}^{n,PURL}[k]}{\sum_r \mathbf{s}^{n,PURL}[r]}\right]. \qquad (8)$$

Linguistic Features: The interestingness of a sentence often depends on its linguistic structure. We use four linguistic features: (1) the length of the sentence with the intuition that longer sentences are more interesting (2) whether the sentence begins demonstrative (such as beginning with "this" or "these" (3) whether the sentence is first person, beginning with "I" or "we" and (4) whether the sentence is definitional, *i.e.,* begins with a pronoun and then includes the word "is" or "are" afterwards.

Data set for learning

To learn these functions, we construct a training set as follows: We randomly chose 250 images. For each image, we assembled all the snippets that are within reasonable distance to the MURL in the HTML parse tree (so that clearly irrelevant snippets are not considered). This resulted in 5443 image-snippet pairs. For each pair, we obtain human judgments independently for relevance and interestingness.

Human judgments: We designed a Human Intelligence Task (HIT) on Amazon Mechanical Turk to label each pair of image and snippet. Figure 3 shows the HIT setup for two example pairs. Given this HIT, the judges were asked to independently score the snippet for relevance and interestingness on a discrete scale between zero and three, where zero means not relevant (not interesting) and three corresponds to very relevant (very interesting). Each pair was judged by approximately 10 judges. The average relevance of the snippets is 1.57 (about half way between "Somewhat Relevant" and "Relevant"). The average interestingness score is 1.17 (a little bit higher than "Somewhat Interesting.").

(a) Somehow, between the first announcement and now The Lorax has a poster.
Relevance 2.0, Interestingness 1.0

(b) The Lorax is a children's book, written by Dr. Seuss and first published in 1971.
Relevance 2.6, Interestingness 1.7

(c) Born on March 2, 1904, Theodor Seuss Geisel, the man who would one day be published under the pen name "Dr. Seuss" began his writing career at Dartmouth College's humor paper.
Relevance 2.0, Interestingness 2.1

Numerical Scores	Not Relevant / Interesting	Somewhat Relevant / Interesting	Relevant / Interesting	Highly Relevant / Very Interesting
Relevance	0.0	1.0	2.0	3.0
Interestingness	0.0	1.0	2.0	3.0

Figure 4: An example image with three snippets and their average human labels. (a) illustrates a snippet that is relevant but not very interesting. (b) is much more interesting. (c) is more interesting than relevant; it is about the author, not his character.

In Figure 4 we include an image to illustrate the human labeling. We include three snippets together with the human scores. Snippet (a) is provided as an example of why it is important to assess the interestingness of the snippets. Snippet (a) is a fairly factual statement about the image. It is scored as relevant, but as not particularly interesting. In comparison, (b) is significantly more interesting. Snippet (c) is notable in that it has a relatively low relevance (compared to the interestingness, which is generally scored more conservatively than relevance.) The sentence is about "Dr. Seuss" whereas the image is one of his characters.

3.3.2 Entropy of Set:

The normalized entropy, $H_0(\mathcal{T})$ in Eq. 1 models the diversity of the chosen sentences. We would like to favor sentences that are diverse in two respects. First, they differ in their vocabulary usage. Second, they are derived from different webpages. Hence, the entropy $H(\mathcal{T})$ of set $\mathcal{T}$ is computed over $N+1$ binary random variables, where N of them constitute the vocabulary used in snippets in $\mathcal{T}$ and the $(N+1)^{th}$ variable models the webpage variability among $\mathcal{T}$. The probability of the n^{th} variable to have a value of 1 is given by the fraction of times the value is observed in the set $\mathcal{T}$. The normalized entropy is given by $H_0(\mathcal{T}) = H(\mathcal{T})/log(N+1)$ which normalizes the entropy across different subsets. As we add snippets, the incremental gain in entropy will diminish due to the fact that the vocabulary is fixed.

3.4 Scaling to the Web

We have run our snippet mining algorithm on the entire web image index of a major search engine with billions of images. We provide details of scaling up the image clustering along with the details of the clustering in (§ 3.1). After the clustering step is performed, the algorithm is embarrassingly parallel and is performed independently for each image set. In fact, on a 100 core cluster, the entire processing took about 45 hours. We can also run the pipeline for delta updates to introduce clusters of new images or additional snippets for the same image set.

4. EVALUATION OF SNIPPET MINING

We now provide an experimental validation of our algorithm using a quantitative comparison with two reasonable baselines (§ 4.1) and quantitative coverage results on popular web queries for image search retrieval (§ 4.2)

4.1 Comparison with Baselines

To the best of our knowledge, there is no prior work on extracting a set of text snippets for an image on the web. There are, however, two reasonable baselines:

- **Query-by-Image and Webpage Summarization** (QbI/WS): Find all occurrences of the image on the web and then use a webpage summarization algorithm to generate a snippet for each page. We compare with a commercial system (http://images.google.com) that follows this approach for a subset of its results. In particular, we use the snippets shown for the top three results from the "Pages that include matching images."

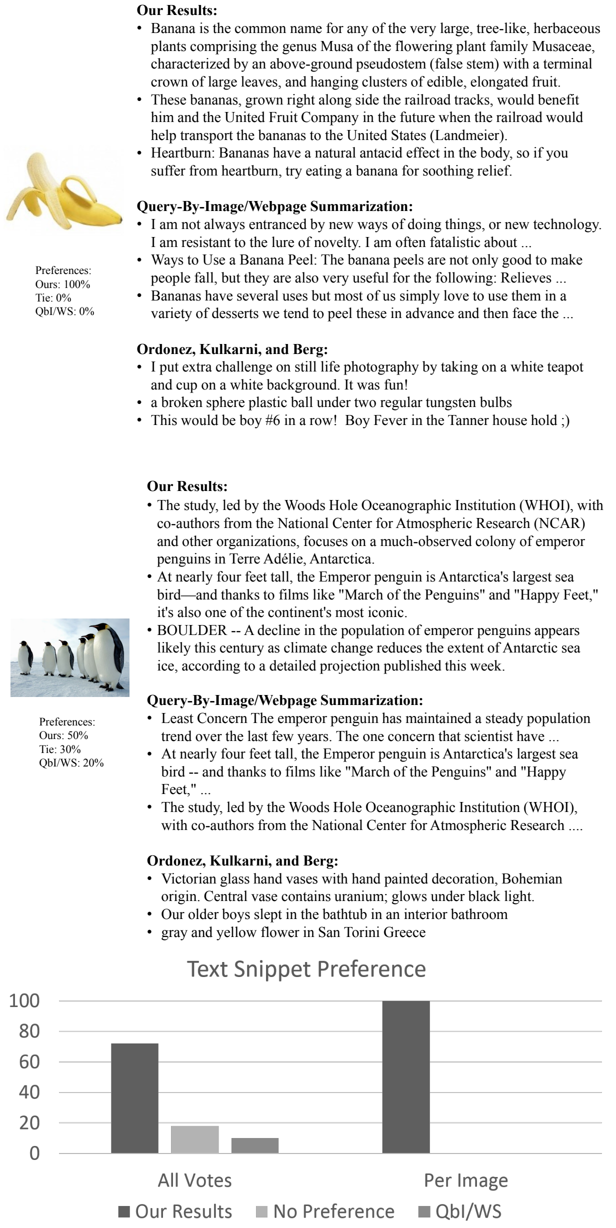
- **Im2Text using Visual Features**: We also compare with the approach of Ordonez, Kulkarni, and Berg [22]. This algorithm matches the image to a database of 1 million Flickr images that already have captions, and transfers the captions from the best matches[3].

Data set: Web image search retrieval algorithm such as QbI/WS are most tuned for head (popular) images and their landing pages. Therefore, images that are most popular on web image search will serve as better baselines. We take the top 10,000 textual queries to a popular search engine and randomly selected 50 images that are among the top ranking results for those queries as the test set.

The results for two images are included in Figure 5. We found that the QbI/WS algorithm always found the correct image and generated plausible results. The results using purely visual features for comparison [22] are far less relevant. This algorithm never really found the correct topic of the image, due to a mismatch between their database (1 million Flickr images on narrow topics of interest) and the web (billions of images). We therefore dropped the Im2Text algorithm [22] from the subsequent quantitative comparison.

Mechanical Turk Setup: In order to perform a quantitative evaluation of our approach with QbI/WS, we conducted pairwise evaluations focused on the overall preference of one ranked list over other. We used Amazon Mechanical Turk with each Human Intelligence Test (HIT) consisting of a pair of ranked lists of snippets, corresponding to our proposed algorithm and QbI/WS. To remove the presentation bias, we considered both orderings of the lists. Each ordering is a separate HIT. We also had each HIT judged by five judges. Each judge was asked to study the two lists and specify which of the two they preferred. They also had the option of choosing "both are comparable". The judges were asked to consider multiple dimensions while making their judgements: relevancy, interestingness, diversity and repetitiveness of content. They were required to spend at least five minutes on the task (we discarded answers from judges that did not conform to this guideline).

Results: Figure 5(bottom) shows the key results. Of all the judgments, 72% exclusively preferred our approach,



Our Results:
- Banana is the common name for any of the very large, tree-like, herbaceous plants comprising the genus Musa of the flowering plant family Musaceae, characterized by an above-ground pseudostem (false stem) with a terminal crown of large leaves, and hanging clusters of edible, elongated fruit.
- These bananas, grown right along side the railroad tracks, would benefit him and the United Fruit Company in the future when the railroad would help transport the bananas to the United States (Landmeier).
- Heartburn: Bananas have a natural antacid effect in the body, so if you suffer from heartburn, try eating a banana for soothing relief.

Preferences:
Ours: 100%
Tie: 0%
QbI/WS: 0%

Query-By-Image/Webpage Summarization:
- I am not always entranced by new ways of doing things, or new technology. I am resistant to the lure of novelty. I am often fatalistic about ...
- Ways to Use a Banana Peel: The banana peels are not only good to make people fall, but they are also very useful for the following: Relieves ...
- Bananas have several uses but most of us simply love to use them in a variety of desserts we tend to peel these in advance and then face the ...

Ordonez, Kulkarni, and Berg:
- I put extra challenge on still life photography by taking on a white teapot and cup on a white background. It was fun!
- a broken sphere plastic ball under two regular tungsten bulbs
- This would be boy #6 in a row! Boy Fever in the Tanner house hold ;)

Our Results:
- The study, led by the Woods Hole Oceanographic Institution (WHOI), with co-authors from the National Center for Atmospheric Research (NCAR) and other organizations, focuses on a much-observed colony of emperor penguins in Terre Adélie, Antarctica.
- At nearly four feet tall, the Emperor penguin is Antarctica's largest sea bird—and thanks to films like "March of the Penguins" and "Happy Feet," it's also one of the continent's most iconic.
- BOULDER -- A decline in the population of emperor penguins appears likely this century as climate change reduces the extent of Antarctic sea ice, according to a detailed projection published this week.

Preferences:
Ours: 50%
Tie: 30%
QbI/WS: 20%

Query-By-Image/Webpage Summarization:
- Least Concern The emperor penguin has maintained a steady population trend over the last few years. The one concern that scientist have ...
- At nearly four feet tall, the Emperor penguin is Antarctica's largest sea bird -- and thanks to films like "March of the Penguins" and "Happy Feet," ...
- The study, led by the Woods Hole Oceanographic Institution (WHOI), with co-authors from the National Center for Atmospheric Research ....

Ordonez, Kulkarni, and Berg:
- Victorian glass hand vases with hand painted decoration, Bohemian origin. Central vase contains uranium; glows under black light.
- Our older boys slept in the bathtub in an interior bathroom
- gray and yellow flower in San Torini Greece

Figure 5: Top and Middle: Qualitative results from a comparison between our algorithm and two baselines. The first baseline is a Query-by-Image/Webpage Summarization algorithm that finds the image on the web and then summarizes each webpage independently. The second baseline is the algorithm of Ordonez, Kulkarni, and Berg [22]. Bottom: A quantitative comparison between our algorithm and the Query-by-Image/Webpage Summarization algorithm. We took the results for 50 images and asked judges which they preferred. No preference was also an option. Bottom Left: 72% of all votes preferred our results, compared to just 10% for the Query-by-Image/Webpage Summarization algorithm. Bottom Right: When we tally up the votes for each image, we found that the judges preferred our results for all 100 images.

[3] We used the software available at http://vision.cs.stonybrook.edu/~vicente/sbucaptions/

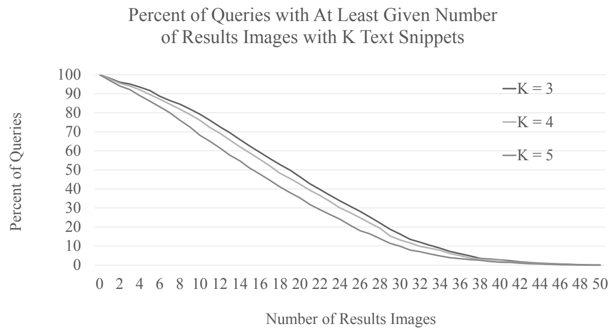


Figure 6: Results on popular queries to an image search engine. We took the top 10,000 queries and considered the top 50 result images for each query. We plot the number of result images with at least K = 3, 4, 5 text snippets (X-axis) against the page-view weighted percentage of queries (Y-axis) that have at least that number of results images with K text snippets. The results show that most queries have 10-30 results images with snippets.

17.8% felt "Both are comparable," while 10.2% preferred the QbI/WS algorithm. We also looked at the individual results. For all images, a (strictly) greater number of judges preferred our approach to the QbI/WS algorithm.

The qualitative results in Figure 5 were chosen to give one example (top) where our results are unanimously preferred, and one example (middle) where a relatively high percentage of judges prefer the Query-by-Image/Webpage Summarization results (20% compared to the average of 10.2%).

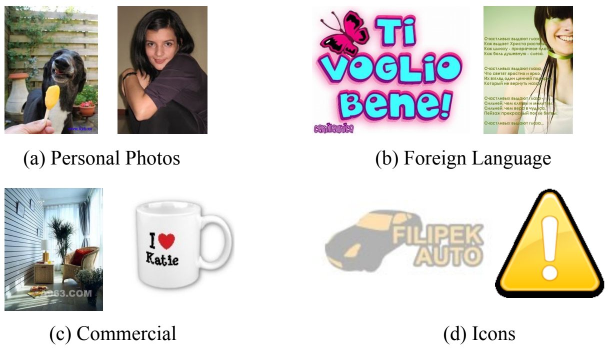
4.2 Coverage for Popular Images

Although we ran our algorithm on the entire web index of a major search engine, we do not obtain results for all images. We impose a threshold on the overall score $rel(s)int(s)$ to remove poor quality results. We now present a set of experimental results to illustrate how likely a snippet will be available for the most popular images; i.e. the ones that are returned as results for the most common queries to Bing image search engine.

Experiment Setup: We consider the top 10,000 most frequent queries issued to Bing image search engine. For each query, we considered the top 50 image results, corresponding to the first page of image search results. We used our approach to identify the top K snippets. Some of these results may not have K snippets associated with them. Next, we compute a metric that captures this property.

Metric: We use the percentage of queries weighted by the frequency with which they were issued. Each query in the top 10,000 was issued by multiple users over the one month sampling interval. When computing the percentage of queries that meet a criterion, we weight by the number of times the query was issued. We refer to this metric (perhaps slightly confusingly) as "page-view weighted."

Results: The results are plot in Figure 6. On the X-axis, we plot the number of result images that have a least $K = 3, 4, 5$ text snippets. On the Y-axis we plot the page-view weighted percentage of queries that have at least that many (X-axis) results images with K text snippets. The results in Figure 6 show that for essentially all[4] queries, there are some results images with text snippets. Most queries have 10–30 results images with text snippets.



(a) Personal Photos

(b) Foreign Language

(c) Commercial

(d) Icons

Figure 7: Examples of common types of images for which our algorithm does not find enough high-quality text snippets. (a) Personal photos frequently do not have high quality text, but are often posted with a short caption. (b) We currently only process English language webpages, so foreign language images do not result in any good snippets. (c) Commercial images are often repeated across multiple webpages with very similar text and are removed as having too much text duplication. (d) Generic icons can rarely be associated with high quality text.

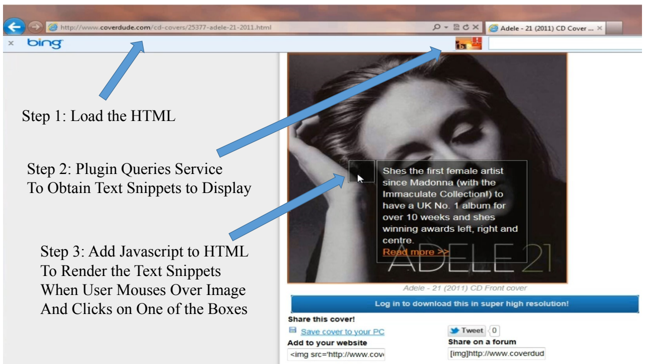
We qualitatively investigated the nature of the images for which no high quality text snippets were found. Example of the sort of images are shown in Figure 7. (a) Personal photos that are posted to blogs and photo-sharing cites rarely have high quality text associated with them, even if they are subsequently reposted multiple times. (b) As we currently only process English webpages, foreign content photos generally do not result in any snippets. (c) Photos from commercial websites often do not result in high quality snippets or are filtered out due to high repetition (often the same image is used on multiple websites with very similar text.) (d) Icons can rarely be associated with high quality text.

5. APPLICATIONS

The database of text snippets mined by our algorithm has a variety of possible applications. The snippets could be shown to users along with image search results. They can be used to improve image search relevance. It may also be possible to use the snippets to discriminate more interesting images from less interesting ones. Images for which there are a lot of interesting snippets (see Figure 1) are generally far more interesting than images for which not much text is found (see Figure 7).

We now propose two other applications. The first consists of augmenting images on the web when a user views the page (§ 5.1). The second is an image browsing application built by connecting images through their snippets (§ 5.2).

[4] We investigated the queries for which there are no results with text snippets. These were mainly cases where the top 50 results images are classified as adult content, but the query was not. We restricted our processing to non-adult images and so obtained no results in these cases.



Step 1: Load the HTML

Step 2: Plugin Queries Service To Obtain Text Snippets to Display

Step 3: Add Javascript to HTML To Render the Text Snippets When User Mouses Over Image And Clicks on One of the Boxes

Shes the first female artist since Madonna (with the Immaculate Collection!) to have a UK No. 1 album for over 10 weeks and shes winning awards left, right and centre.
Read more >>

http://www.coverdude.com/cd-covers/25377-adele-21-2011.html
Adele - 21 (2011) CD Cover ...
bing
ADELE 21
Adele - 21 (2011) CD Front cover
Log in to download this in super high resolution!
Share this cover!
Save cover to your PC
Tweet
0
Add to your website
<img src="http://www.cov
Share on a forum
[img]http://www.coverdud

Figure 8: A screenshot of the web augmentation application (§ 5.1). The plug-in is part of the Bing Bar toolbar at the top. The text results are populated over the image as text pop-ups that show up on mouse-over. Please see http://research.microsoft.com/en-us/um/people/sbaker/i/www/augmentation.mp4 for a video.

These applications both follow a client-server architecture. For both application, we built a Windows Azure [19] service that hosts the database of text snippets extracted using our algorithm. This service is used by the client to query for snippets corresponding to an image and/or for images pertaining to key words that appear in a text snippet.

5.1 Web Image Augmentation

Often when an image is included in a web page, only a brief caption is provided. Enriching these images with additional interesting information can improve the web browsing experience for the user. If the information comes with associated links to its source, the user can easily choose to explore the additional information by clicking on the links. To this end, we developed a 'Bing Bar' plug-in for Internet Explorer [18] that automatically identifies all the images on the current web page, and augments them with any snippets that were found by our algorithm.

Figure 8 shows a screenshot of the application. After the webpage has been rendered, the HTML is passed to the plug-in. The plug-in parses the HTML, extracts the URLs of the images, and queries the Windows Azure service to obtain the snippets for the images. The plug-in then injects Javascript into the HTML that renders the text snippets when the user hovers over the image and then clicks on one of the boxes that are displayed. The plug-in then passes the HTML back to the browser which re-renders the page. Figure 8 shows such an example snippet for an 'Adele' image. For a detailed demo, please see the video http://research.microsoft.com/en-us/um/people/sbaker/i/www/augmentation.mp4.

There are several advantages of this system over other interfaces for viewing query-by-image results. First, the browsing experience is completely passive. The user does not need to initiate a query for every image. The query is initiated automatically, and only when there are interesting results, are they displayed in an un-obtrusive manner. Secondly, this application adds virtual hyperlink on top of the Web, from images to related content on other web pages. This can enrich the browsing experience for the user.

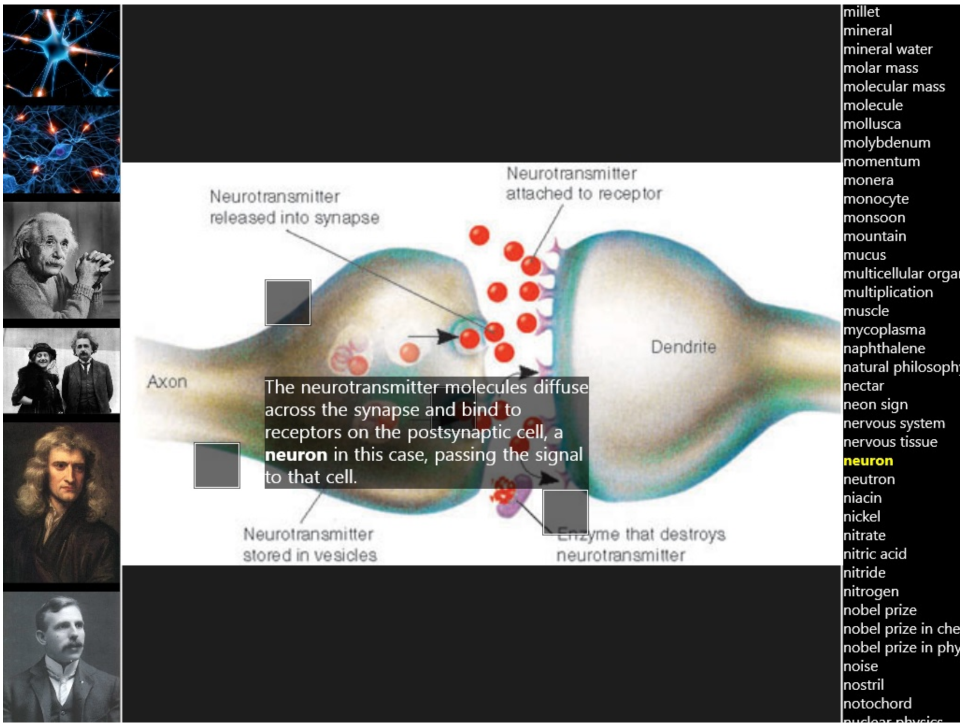
5.2 Semantic Image Browsing

A number of recent works have focused on building effective visualizations of a collection of images that are retrieved in response to a query of interest [10, 24]. In this section, we present a system for a *never-ending* browsing experience that revolves around images and the text snippets associated with them (found using our algorithm). The starting point for the application is either an image or a concept of interest, which can be user-provided or chosen at random. The choice of the next image is based on the text snippets, thereby allowing transitions between images that can be visually unrelated, but semantically related. In particular, we identify concepts in the snippets which are then used as the means for the user to control the choice of the next image in a semantically meaningful manner. See Figure 9.

Browse interface: The interface allows people to browse images on the web. As they browse, the system displays a set of text snippets extracted using our algorithm. These snippets provide interesting information about the current image being viewed. We also detect concepts phrases in the snippet that map to Wikipedia article titles [8, 17, 25], further refined using the techniques proposed in [1]. These concept phrases are used to hyperlink to other images that share the concept phrases in their snippets. Our algorithm for determining the destination of the hyperlink is randomized; if the viewer returns to the same image twice, we want them to explore a different path the second time around. Using an inverted index, we locate K=10 top images at random that have a text snippet that includes the concept being clicked on. There are typically thousands of possible destinations. We then sort these images by the frequency with which the concept appears across all the snippets associated with the image, and choose the image with the highest frequency. We found this algorithm yielded a compelling experience. More sophisticated approaches are left as future work.

Figure 9 contains an screenshot of our application. The current image is a diagram explaining how neural signals are transmitted from an axon to a dendrite in the brain. The displayed text snippet explains the process. The history of recent images is shown on the left side. We began with an image of Ernest Rutherford, then tapped on Isaac Newton, then Albert Einstein twice, once a photo of Albert and his wife, the second time just the famous physicist. We then tapped on brain and neuron before arriving at the current image. The right side of the display includes a list of all the concepts for this subset of science images. The user can tap on any concept to initiate a browsing session at a different starting point. An illustration of the browsing session in Figure 9 is contained in the web video http://research.microsoft.com/en-us/um/people/sbaker/i/www/browsing.mp4.

Evaluation using Mechanical Turk: We evaluated the effectiveness in identifying the next image to browse, based on the semantics conveyed by the snippets. In particular, are two images that are visually unrelated, indeed related through the snippets that describe them? To answer this question, we make use of an evaluation set consisting of 500 pairs, $< x_i, y_i >$, of images (along with the snippets that link x_i to y_i) constructed as follows: for each pair, we choose an image x_i at random. Then, we choose a concept associated with x_i (from the snippets of x_i) at random. We then choose the next image, y_i using the algorithm described above to compute the destination of a link within our application.

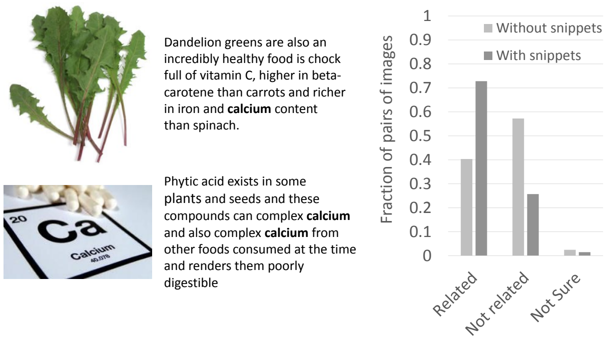


Neurotransmitter released into synapse
Neurotransmitter attached to receptor
Axon
Dendrite
Neurotransmitter stored in vesicles
Enzyme that destroys neurotransmitter

The neurotransmitter molecules diffuse across the synapse and bind to receptors on the postsynaptic cell, a **neuron** in this case, passing the signal to that cell.

millet
mineral
mineral water
molar mass
molecular mass
molecule
mollusca
molybdenum
momentum
monera
monocyte
monsoon
mountain
mucus
multicellular orga
multiplication
muscle
mycoplasma
naphthalene
natural philosoph
nectar
neon sign
nervous system
nervous tissue
neuron
neutron
niacin
nickel
nitrate
nitric acid
nitride
nitrogen
nobel prize
nobel prize in che
nobel prize in phy
noise
nostril
notochord

Figure 9: A screenshot of our semantic image browsing application. The current image is displayed with one of the text snippets. In the snippet, two concepts are detected, neuron and axon. Tapping on either of these concepts takes the user to a related image. The panel on the left shows the browsing history. The panel on the right contains all the concepts for this subset of images, and tapping a concept navigates to an image with that concept. See http://research.microsoft.com/en-us/um/people/sbaker/i/www/browsing.mp4 for a video.



Dandelion greens are also an incredibly healthy food is chock full of vitamin C, higher in beta-carotene than carrots and richer in iron and **calcium** content than spinach.

Phytic acid exists in some plants and seeds and these compounds can complex **calcium** and also complex **calcium** from other foods consumed at the time and renders them poorly digestible

Figure 10: We validated our browsing application using Amazon Mechanical Turk. Left: We showed pairs of images that are connected with a transition in the application. The connection here is the word "calcium." Half the time we showed the corresponding text snippets. The other half we just showed the images. Right: The results show that judges are able to see connections between images more often when they are also shown the text snippets.

We conduct two sets of experiments using Amazon Mechanical Turk. In the first experiment, each Human Intelligence Task (HIT) consists of *only* the pair of images, $< x_i, y_i >$. Each judge was asked to specify if they found the pair of images to be 'related', 'not related' or 'not sure'. In the second experiment (a different HIT), the judges repeated the same experiment, but now, in addition to the pairs of images, they were also shown snippets linking the two images with the connecting concept highlighted.

Results: Figure 10 shows the average number of images that were judged in each of the categories. We can see a huge drop in fraction of 'unrelated' images from 57% to 25% showing the efficacy of our approach in linking semantically related images that are seemingly unrelated. As examples, consider the pair of images shown in Figure 10. In isolation, the connection between the two images is not obvious. With the addition of the text snippet, the viewer immediately finds both images more interesting and sees the 'semantic" relationship between them.

6. CONCLUSION

We have presented a scalable mining algorithm to obtain a set of text snippets for images on the web. The optimal set of snippets is chosen to maximize a combination of relevance, interestingness, and diversity. The relevancy and interestingness are scored using learned models.

There are a number of applications of the resulting text snippets. One possibility is to display the snippets along with image search results. In this paper we proposed two others, a plugin to an internet browser that augments webpages with snippets overlaid on images, and a tablet application that allows users to browse images on the web via hyperlinks embedded in the text snippets. The snippet data can be useful for improving image search relevance.

One suggestion for future work is to analyze the snippets in more detail, for example by clustering, to find groups of related images. The results could be used to broaden the set of snippets and concepts associated with an image, possibly leading to deeper understanding of the content of the images, and more interesting browsing experiences.

7. ADDITIONAL AUTHORS

Rizwan Ansary (Microsoft, rizwan@microsoft.com)
Ashish Kapoor (Microsoft, akapoor@microsoft.com)
Qifa Ke (Microsoft, qke@microsoft.com)
Matt Uyttendaele (Microsoft, mattu@microsoft.com)
Xin-Jing Wang (Microsoft, xywang@microsoft.com)
Lei Zhang (Microsoft, leizhang@microsoft.com)

8. REFERENCES

[1] R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan. Mining videos from the web for electronic textbooks. *International Conference on Formal Concept Analysis*, 2014.

[2] R. Angheluta, R. De Busser, and M.-F. Moens. The use of topic segmentation for automatic summarization. In *Proceedings of the ACL-2002 Workshop on Automatic Summarization*, 2002.

[3] A. L. Berger and V. O. Mittal. Ocelot: a system for summarizing web pages. In *Proceedings of ACM SIGIR*, pages 144–151. ACM, 2000.

[4] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the 10th international conference on World Wide Web*, pages 652–662. ACM, 2001.

[5] W. T. Chuang and J. Yang. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of ACM SIGIR*, pages 152–159. ACM, 2000.

[6] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Sixth Symposium on Operating System Design and Implementation*, pages 137–149, 2004.

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV 2010*. Springer, 2010.

[8] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, 2007.

[9] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, 2000.

[10] Y. Jing, H. A. Rowley, C. Rosenberg, J. Wang, M. Zhao, and M. Covell. Google image swirl, a large-scale content-based image browsing system. In *Multimedia and Expo (ICME), IEEE International Conference on*, pages 267–267. IEEE, 2010.

[11] C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

[12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1601–1608. IEEE, 2011.

[13] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *Proceedings of the Association for Computational Linguistics*, pages 359–368, 2012.

[14] D. C. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[15] I. Mani and M. T. Maybury. *Advances in automatic text summarization*. the MIT Press, 1999.

[16] R. Mason and E. Charniak. Annotation of online shopping images without labeled training examples. *NAACL HLT 2013*, page 1, 2013.

[17] O. Medelyan, D. Milne, C. Legg, and I. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 2009.

[18] Microsoft. Internet Explorer. http://windows.microsoft.com/en-us/internet-explorer/go-explore-ie.

[19] Microsoft. Windows Azure Cloud Services. http://www.windowsazure.com.

[20] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[21] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[22] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

[23] K. Spärck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481, 2007.

[24] G. Strong, E. Hoque, M. Gong, and O. Hoeber. Organizing and browsing image search results based on conceptual and visual similarities. In *Advances in Visual Computing*, pages 481–490. Springer, 2010.

[25] M. Strube and S. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI*, 2006.

[26] X.-J. Wang, L. Zhang, and C. Liu. Duplicate discovery on 2 billion internet images. In *Proceedings of the Big Data Workshop, IEEE CVPR*, pages 429–346. IEEE, 2013.

[27] S. Winder and M. Brown. Learning local image descriptors. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[28] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.