ISSN:0975-9646

Visually Summarizing the Web using Internal Images and Keyphrases

M.V.Gedam, S. A. Takale

Department of computer engineering, PUNE University
Vidya Pratishthan's College of Engg. , India

Abstract— Visual summarization of web pages helps to achieve a friendlier user experience in search and refinding tasks. With the use of this method users quickly get an idea of what the web page is all about and help in recalling the visited web page. The internal images of web page are usually used to describe the content of a webpage. Internal images in web pages are generally worthy for this purpose. Nevertheless, in many web pages dominant internal images are unavailable. So, we move towards a new approach i.e. external image summarization to summarize web pages without any dominant internal images. But, external images are sometimes unreliable because of the way they are retrieved from internet. By taking advantage of internal and external images we propose a scheme for visual summarization. To improve the reliability of external images, relevant external images are retrieved from the Internet by considering textual and visual importance between target webpage and hosting webpage. Relevance, Dominance and typicality of the image with respect to web page are the features of images. We propose an Affinity Propagation Clustering based algorithm to select the best of internal and external images as summarization.

Index Terms— Affinity propagation, Internal images, external image, keyphrases, visual rank, visual summarization

I. Introduction

When The Web has become the largest information repository over the world, thus effectively and efficiently searching the information becomes provides a means for fast and effective Web browsing and information retrieval. The goal of summarization is to produce coherent summaries. This allows for faster and better understanding of the contents of a Website without first browsing its content

A search engine results normally include a list of items with titles, a reference to the full version, and a short description showing where the keywords have matched content within the page. A search engine result page may refer to a single page of links returned, or to the set of all links returned for a search query. Nearly all search engines provide search results using summarization consists of page title, URL and a short textual snippet. On the WWW users frequently revisit information that everyone have previously seen, but "keeping found things found" is difficult when the information has not been visited frequently or recently. Generally webuser look at the bookmarks, history in their browser to find out WebPages that everybody visited before. Research has shown that visual images are much easier and better to remember and understand when compare with words. Almost all applications concerned to summarization of webpage should become more efficient if the web pages are visually summarized. The visual summarization represents a webpage using visual information, such as images and thumbnail. Google integrates the "Sites with images" feature in web search system that summarizes the webpage using the images in the page. This feature provides several images selected from webpage above the textual snippet, from those web users may quickly identify the required webpage.

Most application (Search engines, browser, etc) uses Thumbnail and internal images based approach for visual summarization of webpage. But everyone face problem if WebPages having rich contents and without dominant internal images. These problems are solved by retriving relevant external image from the internet. On other hand, the external images are often unreliable to represent the target webpage because of the imperfection of the way in which they are retrieved. But both the approaches internal and external image based summarization has their respective advantages which may complement each other. So, by taking into consideration both the sources of images, we proposed a clustering based scheme Affinity propagation to select best visual summary for webpage.

Section II briefly describes the existing system related to visual summarization. Section III gives implementation details of proposed approach to generate visual summary. Next section gives the details about dataset used for this system. At last, we conclude this paper and future work

II. Literature Review

People regularly interact with different representations of Web pages. While considering the visual representation of webpage, mostly Thumbnail and Internal image based visual summarization are used in existing products and research communities.

a) Thumbnail based visual summarization

Current commercial web browsers such as Mozilla Firefox and Microsoft Internet Explorer provide a wide range of utilities, such as history lists and bookmarks, which support revisiting previously seen pages on the web. Yet previous research indicates that these utilities are largely unused. WebView; a prototype designed to improve the efficiency and usability of page revisitation. It does this by paying particular attention to how previous pages are integrating many revisitation capabilities into a single display space. But this system did not evaluate implicit bookmarks, dogears, and the hub and spoke view, complex revisitation sequences clearly. Thumbnail is a reduced snap of a webpage depicted in most browsers. For example, Firefox "FastDial" visualizes the bookmarks as thumbnail. Viewzi presents search results using text snippet and thumbnail of webpage for web search. Users found problem in reading text content from small snapshot of webpage which have rich contents. To overcome this problem in search task Dziadosz and Chandrasekar [2] suggest that in the search engine thumbnails of web pages should appear with text snippets .Woodruff *et al.* design textually enhanced thumbnails, which enhance the legibility of the text contents through highlighting some keywords in the scaled-down snapshot. In this way, the problem of reading text information is removed, but the time of loading the snapshot increases dramatically. Erol et al. [3] propose a multimedia thumbnail, one of the kinds of document representation which is a pan-and-zoom movie trailer for the document. In their system, the salient document information (visual and audible) is extracted and synthesized into a playable thumbnail. Multimedia thumbnail address the problem of representing multipage and high resolution documents on small form factor devices. User needs assistance in navigation, but it provides through strict formatting, not via links.

b) Internal image based visual summarization

Internal images are the best representative of its webpage; hence commonly use to visually summarize the webpage. To improve the relevance judgment of web search result, Li. et al. [4] propose an internal image based system which extracts the dominant images from the web page as “image excerpt”. A text snippet is generated by selecting keywords around the matched query terms for each returned page. This system automatically generates image excerpts by considering the dominance of each picture in each web page and the relevance of the picture to the query. Another approach Visual Snippet [5], which is an image generated by composing a dominant image from the web page, salient text (e. g., title), and a watermarked logo from the web page, is an extension of image excerpts. In most industrial products, such as Google’s “Sites With images”, adopt these internal image based summarizations approach. However, the problems in the applicability of such approaches are for a large amount of web pages, dominant internal images are unavailable.

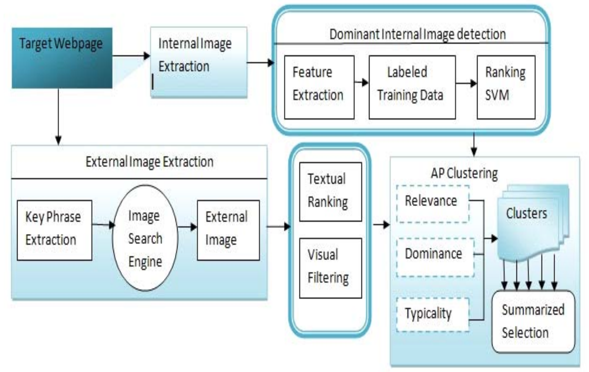


Fig. 1. System Architecture

III . SYSTEM OVERVIEW

Internal and external images are considered as two useful sources for image based summarization. Since they have respective advantages and may complement each other, we jointly took into consideration these two sources of images for summarizing web pages. System architecture for the system as shown in fig. 1 Project is divided into three modules which are as follows

a. Dominant Internal image extraction

The content of the page; such images are called as internal images. These images can be directly used to summarize the webpage. However, a typical web page may have a lot of images, most of which may be decoration, advertisement or logo. Hence we use an algorithm to detect the dominant ones among all of the images in the webpage for visual summarization. Here, we adopt the learning-based algorithm proposed in [6] for this task, which is described as shown in fig.2

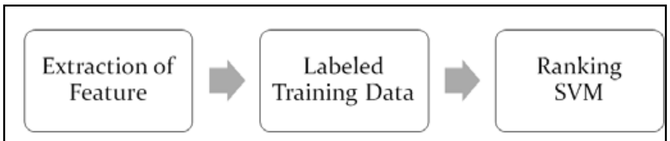


Fig. 2. Dominant Internal image extraction

Features are extracted for each image from three levels, including the image itself, the hosting web page and the hosting web site. These features are summarized in Table 1

TABLE 1
FEATURES FOR DOMINANT INTERNAL IMAGE

SN	Feature Kind	Feature Name
1	Image Level	size, width/height ratio, blurriness,contrast,colorfulness
2	Page Level	relative position,relatuve size, relative width/height ratio
3	Site Level	IsInnerSite

After extracting features from three level of an image, dominance model can be learned from some labeled training samples, which are represented as $(x_{i,k}\ ;\ y_{i,k})$ where $x_{i,k}$ is the extracted feature vector of the image i in the page k and $y_{i,k}$ is its labeled dominance, namely, 0 (non-dominant) and 1 (dominant). Linear Ranking Support vector machine is then adopted to train the ranking model for dominance detection. And finally dominant images from webpage are separate out.

b. External image retrieval

For a large amount of web pages, there are no dominant images. So, to generate a meaningful visual summarization for such web pages we have to retrieve images which can visually represent the web page from the Internet. The idea we design is stated as follows. First, the key phrases, which can represent the salient topics in the web page, are extracted. These key phrases are used as queries to search images relevant to the topics of the web page in an image search engine. Finally the result images are ranked based on the textual similarities between the web pages which host these images and the target web page. Nonrelevant images are further filtered out based on the visual similarities among the result images. The top ranked image is adopted as the most relevant external image for given webpage. Fig. 3 illustrates the work flow of this an approach.

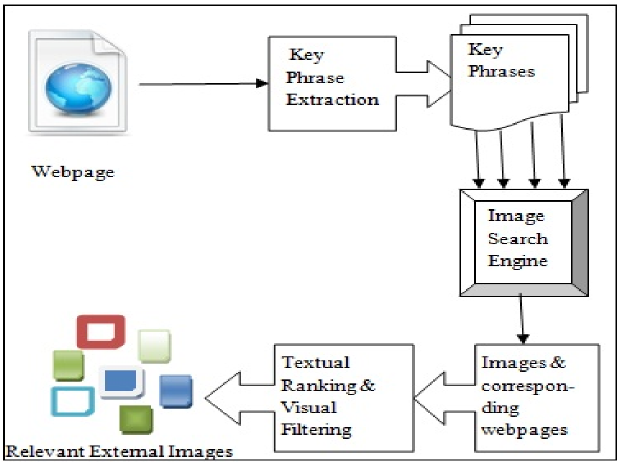


Fig. 3. External image retrieval

1) Keyphrase extraction

Keyphrases provide semantic metadata that characterize and Summarize document. We use these phrases to facilitate Web users grasping the main topic(s) of a Web page. For extraction of Keyphrases from webpage. We use KEX algorithm [7].

2) Ranking and Filtering

After extracting Keyphrases from webpage, these are given to the image search engine. We use an algorithm to re-rank and filter images from the search result. In our algorithm, we first rank result images based on the textual similarities between the target web page and the web pages which host the images. Then we propose an algorithm to filter out visually unimportant images.

Textual similarity: Using cosine similarity based on vector space model (VSM) we calculate the textual similarity between two WebPages. A web page is represented as a vector where each component is a Term Frequency score of a particular term in the page. And Cosine similarity [8] is adopted to measure the textual similarity between each result web page and the target web page. The cosine similarity function can be formulated as:

$$Similarity(T,H)=\frac{\sum_{k=1}^{t} w_{tk}.w_{hk}}{\sqrt{\sum_{k=1}^{t}(w_{tk})^2.\sum_{k=1}^{t}(w_{hk})^2}} \quad (1)$$

Where,

w_{tk} =Weight of term tk in Target webpage

w_{hk} =Weight of term tk in Hosting webpage

Visual Similarity: Visual importance of the result image is calculated using VisualRank [9]. For each image, SIFT features are extracted. The visual similarity between two images is defined as the number of similar SIFT features divided by the average number of SIFT features in the two images. Then a graph is constructed with images as vertices. Each pair of images is connected with an edge while the weight on the edge is defined as the visual similarity.

$$VisualSimilarity(u,v)=\frac{No.ofsimilarSIFTfeature}{Avg.no.ofSIFTfeature} \quad (2)$$

Page Rank is applied on the graph to calculate the importance score for each image. The images whose visual importance score is below a threshold will be filtered out. After ranking and filtering, the top ranked image is adopted as relevant external image used for summarization.

c. Affinity Propagation Clustering

The existing system provides visual summary of webpage are focused on either internal image based summarization or external image based summarization. Each of these approaches has their respective advantages and disadvantages. While focusing on internal image based summarization, though internal images are best to represent its hosting webpage, about 40 percent of webpage doesn't have any dominant internal image. In case of external image based summarization, most images that were extracting from internet are irrelevant because of imperfection in ways they are retrieved. So, we are considering both the kind of images jointly. We apply Affinity propagation clustering [10] on external and internal images that we got from above two modules.

1) Features of good image summarization

An effective image based web page summarization should satisfy the following three properties in order to alleviate user in search and refinding tasks.

- Relevance - This property reflects whether the image is relevant to the topic of input webpage. We are considering both internal and external images for visual summarization, so with consideration of this property we are able to select the summary representing the topics in the webpage. In our project this property is incorporated by computing textual similarity between target webpage and the webpage who host the candidate summary images.
- Dominance- A webpage consists of lots of images representing different topics among them some are dominant one while others are nondominant. We have to select image from the candidate summary images which reflects the dominant topic of webpage. So, by seeing a summary, users quickly get idea about subject of webpage. In the first module we generate model to select the dominant internal image from webpage which describe the dominant topic of target webpage. The studies [7] shows that the dominant images specifically topical dominant images are often observed in attractive areas of webpage with high quality and large size compared with other images in webpage. Here, we use method in first module to

compute dominance of an image.

- Typicality- The above two properties relevance and dominance depict the relationship between a webpage and an image, the typicality describe relationship between image and multiple aspects of a topic in webpage. According to the theory of "graded structure", people often evaluate some images imitate a particular topic than others. Generally typicality can be measured by some central tendency since all images of same topic follow same graded structure. Central tendency normally related to the way in which data incline to cluster around some value. In our project we are measuring typicality using a clustering method.

To integrate all three properties relevance, dominance, typicality into an integrated framework, we adopt affinity propagation clustering (AP). This clustering method used in our project has several benefits. It is very hard to predefine number of clusters in summarization problem, which is not required in case of AP clustering. Also this clustering method easily considers a relevance and dominance property to indicate probability of an image to be an exemplar.

In the clustering, first compute relevance and dominance for each image then we adopt AP to generate exemplar which are served as summarization candidate. Finally, visual summarization selected from exemplars having highest score. The aim of AP is to identify Small no. of images that accurately represent set of images. Given a set of images $I = \{I_i\}_{i=1}^{N}$ and the similarity matrix $S = \{s_{i,k}\}_{i,k=1}^{N}$, Where $s_{i,k}$ is the similarity between image I_i and I_k , the AP algorithm tries to cluster I into P (P < N) groups, each represented by an exemplar. The generated clusters are denoted as $I_e = \{I_{ei}\}_{i=1}^{M}$ and $e(I_i)$ represents the exemplar of image I_i . Initially all images as potential exemplars.

AP clusters images according to two kinds of messages propagated between images:

1. The responsibility $r_{i.k}$ sent from image i to k, reflecting how well-suited k is to serve as the exemplar for i.
2. The availability $a_{i,k}$ sent from image k to i, reflecting how appropriate it would be for image i to choose image k as its exemplar, with considering support from other images that k should be an exemplar.

After messages of all images are updated, image I_k which maximizes $r_{i,k}$ +$a_{i,k}$ is selected as $e(I_i)$.The clustering procedure is terminated after a fixed number of iterations or after I_e stays constant for some number of iterations. The algorithm is shown as follows:

Algorithm: Affinity Propagation

```
Input: I , S
Output: I_e
Initialization: $a_{i,k}^{(0)} = 0, r_{i,k}^{(0)} = 0$
For t =1 to a fixed no. of iteration do
 Update and Damp all responsibility:
   • $r_{i,k}^{(t)} = s_{i,k} - \max_{k \neq k'} \{a_{i,k}^{(t-1)} + s_{i,k'}\}$
   • $r_{i,k}^{(t)} = \lambda r_{i,k}^{(t-1)} + (1-\lambda) r_{i,k}^{(t)}$
 Update and Damp all availabilities :
   • $a_{i,k}^{(t)} = \min\left\{0, r_{k,k}^{(t)} + \sum_{i' \notin \{i,k\}} \max\{0, r_{i',k}^{(t)}\}\right\}$
   • $a_{k,k}^{(t)} = \sum_{i' \neq k} \max\{0, r_{i',k}^{(t)}\}$
   • $a_{i,k}^{(t)} = \lambda a_{i,k}^{(t-1)} + (1-\lambda) a_{i,k}^{(t)}$
 Select exemplars:
 $e(I_i) = I_r$ , where r= $\arg\max_k \{r_{i,k} + a_{i,k}\}$
 Add $I_r$ into exemplar set $I_e$ if $r = i$
   If $I_e$ stay constant for some no. of iteration then
     Return $I_e$
   End if
End for
```

Affinity propagation takes as input a collection of real valued similarities between images, where the similarity s(i,k) indicates how well the images with index k is suited to be the exemplar for image i. Each similarity (3) is computed using a linear model of textual similarity and visual similarity.

$$sim(i,k) = \beta sim^t(i,k) + (1-\beta) sim^v(i,k) \tag{3}$$

The textual similarity between images sim^t is evaluated using the cosine similarity over the TF vectors of their hosting webpage. And Visual Similarity sim^v is evaluated as:

$$sim^v(i,k) = 1 - dist(i,k) / dist_{\max} \tag{4}$$

Where Euclidean distance $dist$ is calculated in feature space as(5)

$$dist(i,k) = \frac{1}{m} \sum_{j=1}^{m} \left(\min_{1 \le f \le n} \sqrt{\sum_{q=1}^{128} (s_{jq} - t_{fq})^2} \right), \tag{5}$$

Where, m is the number of SIFT feature in image I, n is the number of SIFT feature of image k s_{jq} is the qth element of the jth feature vector of image I and t_{fq} is the qth element of fth feature vector of image k.

The responsibilities are evaluated using the rule(6)

$$r_{i,k}^{(t)} = s_{i,k} - \max_{k \neq k'} \{a_{i,k}^{(t-1)} + s_{i,k'}\} \tag{6}$$

In the first iteration, as availabilities are set to zero initially, r(i,k) is set to the input similarity between image i and image k minus the largest of the similarities between image i and other candidate exemplars. In later iterations, when some images are effectively assigned to other exemplars, their availabilities for other images will goes negative as prescribed by the availability update rule(7) :

$$a_{i,k}^{(t)} = \min\left\{0, r_{k,k}^{(t)} + \sum_{i' \notin \{i,k\}} \max\left\{0, r_{i',k}^{(t)}\right\}\right\} \qquad (7)$$

For image i, the value that maximizes a(i,k) + r(i,k) either identifies image i as an exemplar if k = i, or identifies the image that is the exemplar for image i. The message passing procedure may be terminated after a fixed number of iterations, after changes in the messages fall below a threshold, or after the local decisions stay constant for some number of iterations. When updating the messages, it is important that they be damped to avoid numerical oscillations that arise in some circumstances. Each message is set to λ times its value from the previous iteration plus (1 - λ) times its prescribed updated value, where the damping factor λ is between 0 and 1. Here, we used a default damping factor of $\lambda = 0.5$.

In the dominant image detection, we develop a model to detect the dominant image on webpage and dominance score are mapped into [0,1]. In external image extraction , we found that the relevance between an image and a webpage can be evaluated by cosine similarity over term frequency vector of target webpage and hosting webpage.

After performing all iteration of AP algorithm, we have to select visual summary from candidate exemplar. All candidate exemplar represent multiple topics in webpage but we have to choose the one as summary which represent the dominant topic of webpage. The selection of summary is based on two evaluation criteria. First one is number of images in a cluster, if more images are present in a cluster means that more images are similar to candidate exemplar represents its importance. Second is preference, to leverage the relevance and dominance, we also consider preference as an evaluation criterion.

2) Visual Summary Selection:

In our system , the preference for image I_i is calculated as a combination of the relevance score $r(I_i, TW)$ and the dominance score $d(I_i)$:

$$pref(I_i) = [\alpha r(I_i, TW) + (1-\alpha) d(I_i)] sim_{med}$$

Where,

$pref_i$ =preference of exemplar of ith cluster

$prop_i$ =proportion of images in ith cluster

A model is developed based on these evaluation criteria to compute score for each exemplar to select the summary the exemplar with the largest score is selected as the image based summarization

$$score^i = \gamma pref_i + (1-\gamma) prop_i$$

Where,

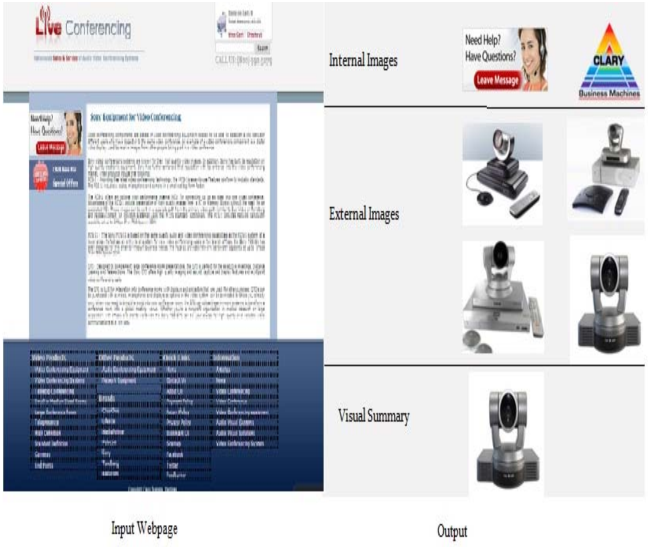
TW= Target Webpage

sim_{med} = median of the similarities in the similarity matrix

α =parameter to trade off the contribution from Relevance and dominance

IV Result And Analysis

If an input given to our system any webpage suppose http://www.live-conferencing.com/articles/sony-equipment-for-video-conferencing.html . The result obtained are shown in above figure. This page is about sales of video conferencing equipment. The output we get from our system is most appropriate to describe the contents of webpage.



The existing product or system regarding visual summarization of webpages consider only either of internal image or external image or sometimes thumbnails . In our system we taking into consideration all types of webpages and all summarization criteria. Thus obtain a better results.

V Conclusion And Future Work

The various kinds of visual summarizations based on thumbnails, internal images, visual snippets and external image are present in existing product. Since their result can complement each other in terms of availability and reliability, we propose a clustering based scheme to jointly select a summary which best presents three properties (relevance, dominance and typicality). The existing approaches considered either internal image or external image. So, the results could complements as contents of webpages changes. Thus, the proposed system considered both the kind of images jointly to select best summary for any kind of webpages.

Acknowledgment

This is to acknowledge and thank all the individuals who played defining role in shaping this journal. I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Prof. S. A. Takale madam for giving her valuable guidance, inspiration and encouragement to embark this paper.

REFERENCES

[1] Binxing Jiao, Linjun Yang, Jizheng Xu, Qi Tian "Visually Summarizing Webpages Through Internal and External Images" , *IEEE Trans. Multimedia.*, vol. 14, no. 6, pp. 1673-1683, Dec. 2012

[2] S. Dziadosz and R. Chandrasekar, "Do thumbnail previews help users make better relevance decisions about web search results?," in *Proc. SIGIR '02*, New York, 2002, pp. 365–366.

[3] B. Erol, K. Berkner, and S. Joshi, "Multimedia thumbnails for documents" in *Proc. 14th Annual ACM Int. Conf. Multimedia*,New York, 2006, pp. 231–240.

[4] Z. Li and L. Zhang, "Improving relevance judgment of web search results with image excerpts," in *Proc. 17th Int. World Wide Web Conf. (WWW2008)*, Apr. 2008, pp. 21–30.

[5] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu, "Visual snippets: Summarizing web pages for search and revisitation," in *Proc. CHI '09*, New York, 2009, pp. 2023–2032.

[6] Q. Yu, S. Shi, Z. Li, J.-R. Wen, and W.-Y. Ma, "Improve ranking by using image information," in *Proc. ECIR 2007*, 2007, pp. 645–652.

[7] M. Chen, J.-T. Sun, H.-J. Zeng, and K.-Y. Lam, "A practical system of keyphrase extraction for web pages," in *Proc. CIKM '05*, New York, 2005, pp. 277–278

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[9] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[10] B. J. Frey and D.Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.