ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Click Prediction and Ranking of Web Images using Sparse Coding and Query Prediction

Meera R

M.Tech Student, Dept. of CSE, Viswajyothi College of Engineering and Technology, Kerala, India

ABSTRACT: The process of arranging images based on their visual similarity or keyword matching is called ranking of images in web. A web image mainly has two important factors associated with it, i) metadata ii) visual features. An efficient method for 'ranking' web images using the visual features and 'query recommendation' using metadata associated with images, is proposed. The technique proposed is a variant of several SEO (Search Engine Optimization) techniques. It also solves the problem of sparse click logs of web images as the system tries to enrich the click logs in the initial stage itself. Thus without ignoring the metadata and visual features of web images an efficient Search Engine Optimization method is suggested, by making use of image processing techniques like selective feature extraction and sparse coding. Traditional search engines make use of the keyword matching techniques to retrieve and rank the images (based on keyword-match count), whereas some other systems make use of content based image retrieval. In both methods either one feature is ignored or simply wasted. The proposed method tries to utilize both associated features of a web image so as to rank the images and to optimize the visualization of search engines.

KEYWORDS: meta-data, click-logs, sparse coding

I.INTRODUCTION

Search engine optimization (SEO) is the process of affecting the visibility of a website or a web page in a search engine's unpaid results - often referred to as "natural," "organic," or "earned" results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. SEO may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines.

As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience. Optimizing a website may involve editing its content, HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of back links, or inbound links, is another SEO tactic.

Every search engines make use of a reference called click logs for ranking web documents and web images. Click logs of a search engine is said to have the query submitted by different users, the URLs clicked by the users, the number of times URLs are clicked and various other details. These logs will reect the importance of different URLs, and how much relevant is a particular web document to a query.

Image search technology at the major search engines does mostly rely upon searches where images are associated with keywords rather than for specific details about the images themselves (such as file size, file type, resolution, etc). Images are indexed by the URL where they appear, and the text associated with the page at that URL. Deciding what images are associated with which queries first depend upon a search engine associating images with keywords that might be used as search queries. A web crawling program travels through the web and aggregates images and text that appears on the same pages as those images. It might take all of the text from those pages and store it in a database, or text that is only a certain distance away from the pictures. It might also look for text that is associated with animage, but is found on different pages (perhaps links to the picture, and possibly text associated with thoselinks). Ranking factors are then used to determine the relevance of a picture to the query, and the order that theseassociated

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

imagesispresented to a searcher. A name detection program might recognize a query as a person's name, and might trigger the use of a face detector program, to show people in response to queries that use people's names.

In the proposed system, metadata [13] alone is not trusted for ranking images, instead metadata is used for query recommendation and visual features associated with the images are used as a ranking factor. Thus both metadata and visual features are not ignored.

II.RELATED WORK

People regularly interact with different representations of Web pages. A person looking for new information may initially find a Web page represented as a short snippet rendered by a search engine. When he wants to return to the same page the next day, the page may instead be represented by a link in his browser history. Previous research has explored how to best represent. Web pages in support of specific task types, but, consistency in representation across tasks is also important.

The related work [2] is all about exploring how different representations are used in a variety of contexts and present a compact representation that supports both the identification of new, relevant Web pages and the refinding of previously viewed pages. The visual snippet generation process involves four steps:

1. Cropping and scaling the salient image. The image is cropped manually along one dimension to an aspect ratio of 4x3 and scaled to 120x90. If no salient image is identified, a snapshot of the page is used instead, appropriately scaled.

2. Scaling the logo. The logo is scaled to fit within a 120x45 rectangle while preserving its original aspect ratio. The logos scale is chosen so that it either falls half of the height or the full width of the visual snippet. If no logo is available, it is omitted.

3.Cropping the title. 30-39 letters to be necessary to provide medium quality.

To provide satisfying summarized search result, they [3] a two-step ranking process. Considering both relevance and diversity in ranking object categories and the object layout was taken into account while selecting the most representative image for each category. The authors also believed that focusing on object queries is a promising direction for further advancing image search reranking and they envision the work in the future as follows: First, they will systematically classify queries into different domains regarding the possibility of image search reranking, and then develop algorithms to solve them respectively. Second, motivated by the object bank image representation they may combine the object vocabulary discovered for the query and the objects from the collection to seek a more comprehensive representation of images and queries. Finally, identify and address the system challenges so as to most efficiently integrate this algorithm into a real-world image search engine.

Web image ranking is a tedious task because of the huge number of images in web and sparse click logs. Click logs [1] are used to know the relevancy of images under a query based on the number of clicks. Click logs of images are said to be sparse as user's usually prefer clicking on web images. Thus the very first point is to enrich the click logs by finding images that has similar features with that of existing images in the click log. Secondly, using sparse coding scores, the images are ranked. Finally from the ranked image's metadata unique keywords are extracted andused for query recommendation. Image reranking is effective for improving the performance of a text-based image search. However, existing reranking algorithms are limited for two main reasons: 1) the textual meta-data associated with images is often mismatched with their actual visual content and 2) the extracted visual features do not accurately describe the semantic similarities between images. Recently, user click information has been used in image reranking, because clicks have been shown to more accurately describe the relevance of retrieved images to search queries. However, a critical problem for click-based methods is the lack of click data, since only a small number of web images have actually been clicked on by users. Therefore, the aim to solve this problem by predicting image clicks.

III.PROPOSED ALGORITHM

The proposed system makes use of both metadata and visual semantics associated with the web images. The metadata part is used for query recommendation system and the visual features are used for ranking of images. Click Prediction

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

is used to enrich the click logs which are said to be sparse. This is carried out by sparse code comparison of the test image with sparse codes of rest of the images in the click logs. Majority matching bits predicts the test image to belong to the click log. Thus enriching the cick logs as shown in figure 1. Once click prediction is carried out and click log is enriched, comes the task of ranking and query prediction. Ranking is purely carried out based on the sparse code comparison, such that images with similar sparse codes are displayed after sorting the sparse hit valued images in descending order. The metadata associated with images are collected on to a file and then stemming and stop word algorithm is applied. Thus the stop words like "the, also, then, thus, you, us, for" etc is removed. Stop words in English language were collected from Wikipedia site. The stop word removal algorithm effectively removes stop words from the metadata and helps to pick unique keywords from the filtered metadata list. Thus synonym queries or keywords are obtained as a result which facilitates users to submit these queries for advanced search. Again the technique comes under a Search Engine Optimization tactic that helps users to accomplish their search tasks quickly.

A. Feature Extraction

The very first step of implementation is to extract features from the image log. Various features extracted are:

- Block Wise Color Moment
- HSV Histogram

Block Wise Color Moment:

Color moments of an image in the Y(Luminance), Cb (Chrominance of blue), Cr (Chrominance of Red) color space are very simple to calculate. Color moments (will be referred as CM from now on) are very effective for color-based image analysis. They are especially important for classification of images based on color, image retrieval, and identification of Image angle (0, 90, 180, or 270 degrees). Generally the first order (mean of color values) and the second order moments (variance of color values) have been used in research. A given image has to be converted first into Y, Cb, and Cr format from the raw R,G,B format.

The function 'rgb2ycbcr' will do this job. After this, An image has to be partitioned into sub-blocks. Deciding optimal number of sub-blocks is a qualitative question and has to be decided as per the type of the application. By scanning each sub-blocks, calculate for each sub-block in Y, Cb, and Cr plane two measures: statistical mean and statistical variance. Each image is sub-divided into 64 patches. Each sub-block size is then 32*32pixels.Color moments for these blocks are nothing but the statistical mean and statistical variance of this 32*32 pixel values. Thus in short each image is divided into 64 patches, and moment extraction is carried out in each of these 64 patches. A triplet of mean, variance, skew is extracted from each block. Thus 64 triplets for a single image is formed.

HSV Histogram:

A color histogram is a representation of the distribution of colors in an image. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space, the set of all possible colors. The color histogram can be built for any kind of color space, although the term is more often used for three-dimensional spaces like RGB or HSV. For monochromatic images, the term intensity histogram may be used instead. For multi-spectral images, where each pixel is represented by an arbitrary number of measurements (for example, beyond the three measurements in RGB), the color histogram is N-dimensional, with N being the number of measurements taken. Each measurement has its own wavelength range of the light spectrum, some of which may be outside the visible spectrum.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

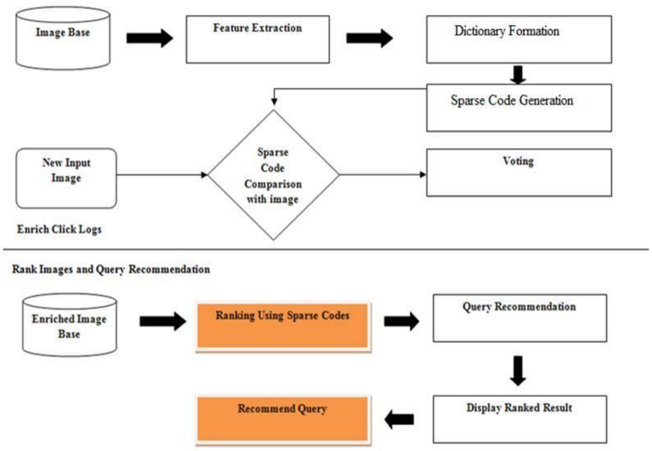

Figure 1: Architecture of the Proposed System

B. Enriching Click Logs

Logs are record of information that trace the user navigation patterns. Log analysis will reveal the usefulness of web documents. The use of data stored in transaction logs of Web search engines, Intranets, and Web sites can provide valuable insight into understanding the information searching process of online searchers. This understanding can enlighten information system design, interface development, and devising the information architecture for content collections. For Web searching, a transaction log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine. AWeb search engine may be a general-purpose search engine, a niche search engine, or a searching application on a single Web site.

The users may be humans or computer programs acting on behalf of humans. Interactions are the communication exchanges that occur between users and the system. Either users or the system may initiate elements of these exchanges. Click logs for images are found to be sparse because user'susually prefer clicking on web documents instead of web images, after submitting the query. Such kind of logs is called sparse logs as it contains insufficient information. Enriching click logs is the very first objective. For those first features of available images in the log is extracted and compared with the new images. These new images are the ones to be predicted whether they belong to the same image log or not. This is done by feature matching.

C. Ranking of Images

There are many ranking techniques, but all of them make use of metadata associated with the images. Another ranking feature can be based on the total number of images on a webpage. For example, an image may be given a higher or lower ranking based on the number of images that are on the same web page as the image. Another ranking feature can be based on the total number of images that are linked to by a particular web page. For example, an image may be given a higher or lower ranking based on the number of images that are linked to by the same web page that the image is located.

Moreover, another ranking feature can be based on the total number of thumbnail images that are located on the same webpage as the ranked image. For example, an image may be given a higher or lower ranking based on the number of thumbnail images that are located on the same page as the image. Furthermore, another ranking feature can be based on the total number of links there are to the URL of the an image. For example, an image can be given a higher ranking if it has a greater number of links to its URL compared to other images. In other embodiments, the image may be given a lower ranking if it has a greater number of links to its URL compared to other images. Another ranking factor can be

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

based on the distance that text within a search query is located on the same web page as an image, such that text that is closer to the image is associated more strongly than text that is further away from the image. The distance that text within a search query is from an image can be based on different distance elements. Distance elements may include the number of intervening words between the text and the image.

D. Sparse Coding

Sparse Coding is a method for finding a small number of basis functions to represent an input signal. In a way it is similar to Principal Component Analysis and Compressed Sensing. The least number of coefficients or basis of the input signal that reduces the complexity is the optimized basis representation of the input signal. Sparse coding is a popular signal processing method and performs well in many applications, e.g. signal reconstruction, signal decomposition, and signal denoising. Although orthogonal bases like Fourier or Wavelets have been widely adopted, the latest trend is to adopt an overcomplete basis, in which the number of basis vectors is greater than the dimensionality of the input vector. A signal can be described by a set of overcomplete bases using a very small number of nonzero elements . This causes high sparsity in the transform domain, but many applications need this compact representation of signals. In computer vision, signals are image features, and sparse coding is adopted as an efficient technique for feature reconstruction. Ithas been widely used in many different applications, such as image classification, face recognition, image annotation , and image restoration .

E. Query Recommendation

A web search query is a query that a user enters into a web search engine to satisfy his or her information needs.A search engine query is a request for information that is made using a search engine. Every time a user puts a string of characters in a search engine and presses "Enter", a search engine query is made. The string of characters (often one or more words) acts as keywords that the search engine uses to algorithmically match results with the query. These results are displayed on the search engine results page (SERP) in order of significance (according to the algorithm). Every search engine query adds to the mass of analytical data on the Internet. The more data search engines collect, the more accurate the search results become and that’s a good thing for Internet users. Initially metadata is fetched into a file. From the metadata of images in the click logs, stop words are removed, stemming is performed and finally unique words are picked up called object queries that are synonyms. These words are suggested to users.

IV.PSEUDO CODE

Let X be the image dataset, X(i) be the i-th image in the dataset. Y(i) be the feature vector matrix of the i-th image.

Step 1: Start
Step 2: Find the sparse codes of database images.
Step 3: Calculate the feature matrix Y_{new} of new input image and sparse code.
Step 4: Compare Sparse(Y_{new}) with Sparse(X_i) , predict the click for new input image based on sparse value hit.
Step 5: Make use of sparse code weights to rank images.
Step 6: From the ranked image set, extract metadata.
Step 7: Perform preprocessing of metadata
Step 8: Extract unique words
Step 9: End.

V.SIMULATION RESULTS

The proposed system is evaluated against a traditional system titled Variant-KNN (VKNN). VKNN is purely a traditional Search Engine like- ranking system, where the entire feature matrix is used to compare and rank. But the size of feature matrix is huge and comparison takes more time as well. Thus increases the elapse time. The proposed system uses an intermediate code representation to make the process of comparison easy and fast.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

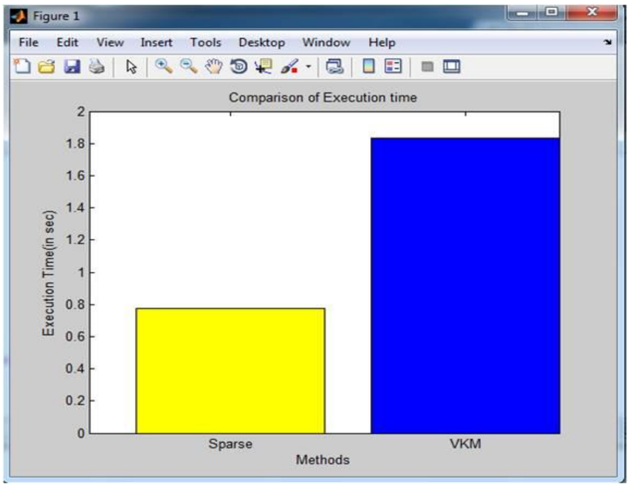

Figure 2: Execution Time Comparison

The VKNN method make use of Euclidean distance measure as the comparison parameter. Same set of features are extracted in both proposed system and VKNN, hence the comparison parameter becomes unique and is the elapse time. K-nearest neighbor algorithm is a classification algorithm but then can be used for ranking purpose as well. Here, a representative image is selected initially and k (say number of images in the log) is retrieved based on the similarity value with that of the representative. The more similar the log image is, high rank is assigned. If same similarity measure arise for two images , then randomly pick one as high ranked one.

The problem with this method is that comparison is carried out directly by making use of the huge feature matrix. Hence elapse time is more as shown in the figure 2.

VI.CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better with the total transmission energy metric than the maximum number of hops metric. The proposed algorithm provides energy efficient path for data transmission and maximizes the lifetime of entire network. As the performance of the proposed algorithm is analyzed between two metrics in future with some modifications in design considerations the performance of the proposed algorithm can be compared with other energy efficient algorithm. We have used very small network of 5 nodes, as number of nodes increases the complexity will increase. We can increase the number of nodes and analyze the performance.

REFERENCES

1. Jun Yu,YongRui and Dacheng Tao, "Click Prediction or web Image Reranking Using Multimodal Sparse Coding",IEEE Transaction,2014.
2. Xiaogang Wang, Shi Qiu, Ke Liu, Xiaoou Tang, "Web Image Re-Ranking by Query Specific Semantic Signatures",IEEE Transaction ,2014.
3. Yang Yang, Linjun Yang, Gangshan Wu, ShipengLi,"Predict Image Relevance Using Query-Context Bag-of-Object Retrieval Model",IEEE Transaction , 2014.
4. Nick Craswell, Martin Szummer," Random Walks Click Graph", Conference on click graphs,SNGIR ACM 2007.
5. http://www.submitexpress.com/search-engine- optimization.html
6. http://www.searchtools.com/info/faceted-metadata.html.

BIOGRAPHY

Meera R, M.Tech Student of Viswajyothi College of Engineering and Technology, 2013-2015 batch. Her area of interests are Search Engine Optimization, Web Extraction and Text Processing. Completed B.Tech in Information Technology in 2009 from KMEA Engineering College.