International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Relevance of Images on Webpage with its Content

Prachi Balla

Department of IT Engineering, AISSMS IOIT, Pune, India.

A B S T R A C T

Web pages contain both related and unrelated images. Classifying these images is an error-prone process due to the many variations in website design. Using multiple web pages provides additional functionality that improves the performance of related image extraction. In this study, we use features extracted from various websites consisting of standard news, gallery, video, and link pages to improve the performance of related image extraction. The easiest way to find relevant images is to extract the largest image on the web page. This approach achieves an F measure score of 0.451 as a baseline. We then use the properties of this dataset to apply different machine learning techniques to find the best machine learning technique. Using the AdaBoost classifier, the optimal f-value is 0.822. Some of these functions have been used in previous web data extraction research. To our knowledge, 15 new features for finding related images were proposed for the first time in this study. Compare the performance of AdaBoost classifiers for different sets of functions.

Keywords: Image classification, image retrieval, feature extraction, web mining.

1. Introduction

Web pages contain irrelevant images along with relevant images. The classification of these images is an error-prone process due to the number of design variations of web pages.

A website consists of many different pages including standard news, gallery pages, link pages. There are a lot of different website layouts that define a web site's structure. Relevant and irrelevant images are included in these layouts.

Relevant images are images related to the content of the page, while irrelevant images are images including advertisements, other links, headers, logos, etc. Images can be of many different sizes and in different parts of a web page.

The internet provides us with a lot of valuable content among HTML tags.

Features are needed for an automatic prediction process which can be obtained by inspecting the web pages. For this task, features like height,width, file extension, file size of an image and other features may contribute to the prediction process.

2. Related studies

1. Determining the most Representative image on the web page

K. Vyas and F. Frasincar

Inf. Sci., vol. 512, pp. 1234–1248,

Feb. 2020.

To decide the maximum consultant photo on a Web web page and try to enhance the overall performance of recognized algorithms with the usage of Support Vector Machines.

The version used is SVM .It no longer executes thoroughly while the records set has extra sound i.e. goal lessons are overlapping.

2. Image retrieval approach based on local texture information derived from predefined patterns and spatial domain information

N. Hor and S. Fekri-Ershad

Int. J. Comput. Sci. Eng., vol. 8, no. 6, pp. 246– 254

Nov./Dec. 2019

For photo retrieval that try and outline photo contents with the aid of using texture, shade or form residences a way is offered primarily based totally on a mixture of neighborhood texture statistics derived from unique texture descriptors.

Considering handiest textural , shade and form functions may be deceptive at times , different residences must additionally be considered.

3. Content-based image retrieval based on combination of texture and color information extracted in spatial and frequency domains

N. T. Bani and S. Fekri-Ershad

Electron. Library, vol. 37, no. 4, pp. 650–666

Aug. 2019

An opportunity technique for green picture retrieval is proposed primarily based totally on a mixture of texture and color data in spatial and remodel domain names jointly.

The color version used for color data is Hue Saturation value color scale which has a few boundaries and may be deceptive at times.

3. Motivation

There are a whole lot of specific internet site layouts that outline an internet site`s structure. Relevant and beside the point pictures are blanketed in those layouts. Relevant pictures are pictures associated with the content material of the page, whilst beside the point pictures are pictures along with advertisements, different links, headers, logos, etc. Images may be of many specific sizes and in specific components of an internet page. For those reasons, it's miles some other undertaking to routinely discover applicable pictures on an internet site.

4. Web traverser and its features

A traverser that downloads pictures one by one and collects charaṢcteristic facts to be used in a system gaining knowledge of techniques. This fact is saved in a tab delimited report with tab delimited properties. In this study, he grouped those functions into 3 categories.

4.1 Text Attributes

Text attributes are taken directly from HTML elements in the web page.

The image element contains img tags and attributes that provide additional information about the image. The most important feature is src which specifies the path to the image. Other attributes such as alt, class, id, height, width, style, sizes, style are optional. These attributes can also have different values on your website.

4.2 Image Attributes

Image features are data extracted from an image.

Traversers use the src characteristic as a key. Additionally, this mechanism improves overall performance and minimizes bandwidth consumption. The houses Width, Height, WidthHeight, ratioWH, FileSize and file_ext are often used inside the literature. The range and altitude features are values taken from the photograph record. Do now no longer confuse those features with attr_width and attr_height. The internet fashion dressmaker has an choice to alternate the attr_width and attr_height values of the internet page. However, you can't alternate the width and top of the photograph record without the use of a photograph processor.

4.3. Last Attributes

After all images have been downloaded, the last feature will be generated.

After all of the pictures of the net web page had been processed, new functions are computed for the ones pictures. Since that is the very last degree of characteristic extraction, it's far known as the very last function. Classifier values are the end result of sorting documents in keeping with diverse criteria. Sorting is in descending order.

5. Experiments

This segment first offers records approximately the dataset and remarks on applicable and unrelated images. The 2nd subsection offers the important overall performance signs used in this study. The 0.33 subsection in short describes device getting to know methods. The very last subsection describes the overall performance consequences of our experiments.

5.1 Dataset

In this study, web traversers download web pages from 200 websites. 100 web pages are downloaded for each website. These websites create a data record of 635,015 images.

22,682 of these images are related images. The dataset is divided into a training dataset and a test dataset to evaluate model performance. To prevent overfitting, 150 sites are used as the training data set and the remaining 50 sites are used as the test data set.

5.2 Evaluation metrics

There are several metrics for evaluating the performance of machine learning techniques. This study focuses specifically on predictive performance for relevant images. All of the metrics below are for the correct evaluation of each image:

- Precision
- Recall
- f − Measure
- Accuracy
- LogLoss

5.3 Machine Learning Methods

To create the prediction model:

- Naive Bayesian (NB) ,
- Support Vector Machines(SVM),
- K-Nearest Neighbours (KNN)], and
- Decision Tree

To improve the performance results of the prediction model:

- Random Forests (RF)
- Adaboost

For feature selection:

- Information gain

5.4. Performance of machine learning methods and simple heuristic techniques

Selecting the largest file size as the relevant image gives the best performance results compared to other heuristic techniques.

AdaBoost and Random Forests give the best results for this task.

5.5. Comparison of feature categories

Features are classified into three categories: text features, image features, and final features. Performance results show that load features contribute more to the predictive model than text features.

5.6. Evaluation of feature selection

Get a lot of information to determine the most influential characteristics of your model.

Using the Last and Image features, the predictive model performance achieves an f-value of 0.822.

6. Conclusion

Image download is not the proper way to get image details. This consumes disk space and makes the download part take more time. Getting the details directly from the img element is more efficient. Image color and texture characteristics can be used to improve model performance. Because these features are directly related to image relevance. Using other boosting techniques such as XGBoost can improve: Overfitting with regularization. Execution speed due to optimization of sorting by parallel operations. Run using the maximum depth of the decision tree.

Finding suitable features, creating large datasets, and evaluating suitable machine learning methods are key problems in web data extraction research.

This study allows web traversers to easily create data sets and extract new features from web pages. Using all features, the predictive model performance result obtained from this classifier has an F-value score of 0.807. You can improve the performance results of your predictive model by removing unnecessary features. Predictive model built without text features has an f-value of 0.822

References

K. Vyas and F. Frasincar, ‘‘Determining the most representative image on a Web page,’’ Inf. Sci., vol. 512, pp. 1234–1248, Feb. 2020.

N. Hor and S. Fekri-Ershad, ‘‘Image retrieval approach based on local texture information derived from predefined patterns and spatial domain information,’’ Int. J. Comput. Sci. Eng., vol. 8, no. 6, pp. 246–254, Nov./Dec. 2019.

N. T. Bani and S. Fekri-Ershad, ‘‘Content-based image retrieval based on combination of texture and color information extracted in spatial and frequency domains,’’ Electron. Library, vol. 37, no. 4, pp. 650–666, Aug. 2019.