

Web content mining

Oral presentation of the project

-----Zheng xianghan,Huwen,Zhangli

1.Short summary

Today,web is an increasingly important resource in many aspects of life: education, employment, government, commerce, health care, recreation, and perhaps will include everything in the future. However,the web page is not perfect designed.For example,using image like mathematical formulas not properly might lead to reduce accessibility of the web pages.Fortunately,more and more people and society are working in this field to develop the accessibility.Look at the WCAG WG(Web Content Accessibility Guidelines Working Group),which is written for Web content developers (page authors and site designers) and developers of authoring tools.

2.Introduction ,describing planned work

The goal of this project is to create a crawler that downloads the images in a web page and tries to classify the content of each image into different

categories, e.g., mathematical formula, logo, buttons, and so on. The focus should be on automatic detection of image usage that reduces the accessibility of a web page.

Planned work:

- Download the Harvestman, and then learn it (some of the Python code in the image crawler)
- Learn Entropy, Linescan, Poisson, Benford
- Programming according to several algorithms above to categorize the picture
- Final report

3. Literature survey

Before we do this project, we look at some related books and some websites, in the following list:

«Digital Image Processing» -----Richard E. Woods

Python Programming language -----WWW.python.org

harvestman.freezope.org

4. Problem statement

- Is the list of the project enough, or needed to be improved, or more advice.
- It is difficult for us to learn some technology in the digital image processing
- The most difficult also the most important thing is to learn the algorithm related to our project
- How to learn the algorithms efficiently
- We want the format style for the final report

5. Planed report contents

In the final report, we expect the report as the five chapters, as the list:

- A short introduction of the project, including the concept of the web content mining, the objectives of the project and the structure of this report etc., will be presented in the first chapter.
- Some algorithms in categorizing the images will be introduced, followed by some basic image processing principle. Several algorithms which are regarded to be better will be emphasized
- Compare each algorithms in the real practice after the program was made, based on the real data, processing speed. The result of compare will output in the form of graph so that it is easy to understand.
- In terms of the above algorithms, we will choose the optimal

algorithm. The final solution is tested at random web pages.

- Chapter five makes up the summary, which is some finishing words about the results and the future of web content mining.

Indicated of roles in the project

Zheng xianghan-----leader of the web content mining project, Cooperate with the other partners in collecting requirements, analyze the project, assign the work to the other partner, and hand in the final report.

Huwen-----Collecting requirements, analyze the project, learn the algorithm in the field of Digital image processing and then teach the other two partner.

Zhangli-----Coding based on some algorithms, and test each of the program. Final to decide which one is the best, base on the response of the speed, run time and occupancy factor of the memory and CPU.

Web page and the contact information

<http://home.hia.no/~xiangz05/>

Zheng xianghan xiangz05@hia.no

Huwen hwcarey@hotmail.com

Zhangli ttbb620@hotmail.com

