

Representative Image Extraction from Web page

Md Imranul Islam

Master's Thesis



UNIVERSITY OF
EASTERN FINLAND

School of Computing

Computer Science

October 2021

UNIVERSITY OF EASTERN FINLAND, Faculty of Science and Forestry, Joensuu
School of Computing
Computer Science

Md Imranul Islam: Representative Image Extraction from Web page.

Master's thesis, 51p.,

Supervisors: Prof. Pasi Fränti, Dr. Radu Marinescu-Istodor, and Nancy Fazal

October 2021

Abstract

We live in the era of technology, where web pages play a vital role. A web application is a blend of different types of information—Its objective is to ensure effective online services. The application combines text, graphs, images, photos, videos, advertisements, and interactive content. It is imperative to extract meaningful information from this information. There are numerous extraction tools worldwide, but most search tools cannot find essential information effectively. Image performs a crucial role to represent a web page. The web page contains various types of images such as logos, representatives, advertisements, and navigation banners, where not all of them are essential for the summarization task. A solution is needed to ignore irrelevant images, and it is required to find the best pictures on web pages that are most relevant to the content. In this thesis, we focus on choosing a representative image that would best describe the content of a web page. We extract all images and classify them based on their features and functionality. Our system identifies representative images Based on Image Features and Functionality. The result analysis accrues 69% accuracy. We have a plan to integrate image classification techniques in machine learning into our system in the future.

Keywords: Representative image, Image extraction, Web page information extraction, Web mining, Mopsi image, Image features.

Acknowledgment

This thesis was written at the School of Computing, University of Eastern Finland, during spring 2021.

I am thankful to the University of Eastern Finland, the IMPIT program, and all the teachers for their support during my master's degree journey. It was an excellent opportunity for me to be part of the IMPIT program, where I studied with students from different countries of the world and learned from the diversity.

I want to express my gratitude and great thank towards my supervisor, Professor -Pasi Fränti, for his guidance and support and for supporting me mentally in a covid-19 pandemic situation. During his teachings, I have to improve my problem-solving and time management skills that will be useful for my future career.

I would like to thank every member of the Machine Learning research group, especially Dr. Radu Mariescu-Istodor and Nancy Fazal, for their support and for making any decision in completing my thesis. I am also thankful to Dr. Oili Kohonen for her enormous support and recommendations on every matter.

Conclusively, I would like to bless my family and friends for giving me mental and financial support during my study life.

List of abbreviations

UEF	University of Eastern Finland
DOM	Document Object Module
WWW	World Wide Web
URL	Universal Resource Locator
CSS	Cascading Style Sheets
HTML	HyperText Markup Language
H1	Heading Size 1
H2	Heading Size 2
H3	Heading Size 3
Tr	Table Row
HTTP	Hypertext Transfer Protocol
href	Hypertext Reference
PHP	Hypertext Preprocessor
JPG	Joint Photographic Expert Group
PNG	Portable Graphics Format
SVG	Scalable Vector Graphics
GIF	Graphics Interchange Format
API	Application Programming Interface
OSM	Open Street Map

Contents

1	Introduction.....	1
1.1	Research Background	2
1.2	Mopsi	4
1.3	Structure of Thesis	4
2	Image Extraction Strategy	6
2.1	Review on Image Extraction.....	6
2.2	Rule-Based Method	7
3	The representative image extraction procedure	9
3.1	Image Extraction.....	10
3.2	Image Features	11
3.3	Categorical Features	12
3.3.1	Representative.....	12
3.3.2	Logos	13
3.3.3	Banners	14
3.3.4	Advertisements	14
3.3.5	Formatting and Icons	15
3.4	Image Scoring	18
3.4.1	Image Size.....	18
3.4.2	Aspect Ratio.....	18
3.4.3	Image Alt and Title	19
3.4.4	Image path and URL.....	19
3.4.5	Image Format.....	19
4	User Interface of WEBIMA 2.00.....	22
5	Experiment.....	25
5.1	Dataset Used	25
5.2	Experimental Setup.....	25
5.3	WebIma 2.0 Performance	32
5.3.1	Accuracy Caparison with Existing Systems	33
5.3.2	Precision, Recall and F-Score	34
5.3.3	Performance Evaluation with Changing Parameters	35
5.4	Discussion.....	36
6	Conclusions.....	38
	References.....	39

List of Tables

Table 1: Rules-based image categorization.	16
Table 2: Rules used for image scoring.....	20
Table 3. Performance comparison between datasets	33
Table 4. WebIma Dataset. Performance Results Comparison.....	34
Table 5. Assessing system parameters performance	36

List of Figures

Figure 1: A sample web page and its relevant content to the user: title and image.....	2
Figure 2: Extracting the representative image	9
Figure 3: Extracted images	10
Figure 4: A sample representative image and its features	11
Figure 5: A sample web page that contains all the five categories of images	12
Figure 6: Sample of representative images.....	13
Figure 7: Sample of logos.....	13
Figure 8: Sample of banners	14
Figure 9: Sample of advertisements.....	14
Figure 10: Sample of formatting and icons images	15
Figure 11: Decision tree for image categories	17
Figure 12: Application homepage.....	22
Figure 13: Extracted images	23
Figure 14: Image categories and features	24
Figure 15: Images score	24
Figure 16: Examples of working web application	26
Figure 17: Example of not working web applications (pop-up window).....	27
Figure 18: Example of not working web applications(pop-up window)	27
Figure 19: Examples of true positive image	28

Figure 20: Examples of false negative and positive images	29
Figure 21: Examples of true positive images.....	30
Figure 22: Examples of false-positive images	30
Figure 23: Examples of true positive images.....	31
Figure 24: Example of false-positive and negative images	31

1 Introduction

In the last few years, the systems of communication and information sharing have been transformed unimaginably by Web applications. All sectors attempt to supply their services through web applications, including traditional methods to meet consumer demand. More than 4.66 billion people use the internet in their everyday lives for personal and organizational purposes, according to Statista¹.

Technology today rules the globe, and knowledge is at our fingertips. There is no doubt that Web pages play a vital function in today's technological world. A web application always contains a wealth of helpful information. The modern dynamic web pages provide their pieces of information in various ways to entice their consumer. This data is available in multiple formats, including text, audio, video, images, and more. In this technological world, every person is busy with their schedule. Sometimes they don't have enough time to read, watch videos, or listen to audio. In this situation, a person can acquire a general impression of the image by looking at it. To deliver the informative elements of a web page to the user, researchers presented strategies such as information retrieval (Yu et al., 2003), central content extraction (Kim et al., 2013), and picture extraction (Kherfi et al., 2004) (see Figure 1).

¹ <https://www.statista.com/statistics/617136/digital-population-worldwide/>

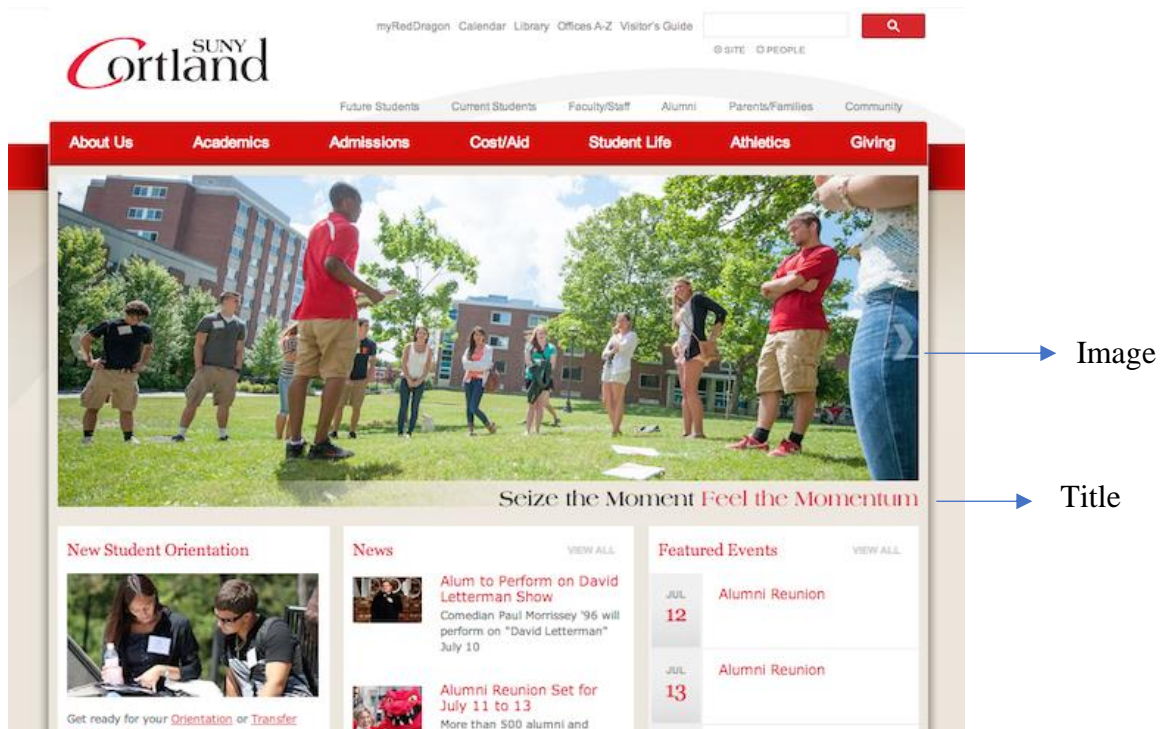


Figure 1: A sample web page and its relevant content to the user: title and image

1.1 Research Background

A web application is always a source of a lot of data. Text, graphs, graphics, photos, videos, advertising, interactive content, and other sorts of information are all combined on a web page. A message is attached to every piece of information that a web application contains. However, several photos are utilized on a web page to draw a user's attention. The picture can quickly obtain the theme message of the online application. However, there are high numbers of images embedded inside web pages. Several are less crucial to the web page's content, such as advertisements, navigational banners, and icons (Azad et al., 2014). A solution is always required to ignore the useless photos and choose a representative image for the web page. For the sake of clarity, we define the best representative embodiment according to (Gali et al. 2015). There are a variety of strategies that emphasize summarization (Money & Agius 2008), interest (Gygli et al., 2013), memorability (Isola et al., 2013), and diversity (Isola et al., 2013). Our research looks for representative photos and focuses on image extraction and feature analysis.

An important consideration for this study is to distinguish between a website and a web page defined by javatpoint². We only look for photographs that are reflective of web pages. In this study, the representative picture of a web page describes the image that most accurately reflects the page's content to the user. Representative photographs use various reasons, including when bandwidth constraints limit the total number of images that can be accessed or when a visual category is needed. A single image must represent an entire category of papers as well as the material that goes with them (Helfman & Hollan, 2000). The main photos for social media platforms like Facebook and Google+ have triggered. It is also essential for location-based service (LBS) such as MOPSI, looking for new restaurants or any services. There are many more familiar examples of its use, e.g., Foursquare, MOPSI, Yelp, Groupon, and Facebook Places

In this thesis, we present a tool that recognizes all photos from websites and scrape all images. The objective of the experiment is to discover the representative images of a web application. The system divides the images into five categories based on their function: representative, logos, banners, advertisements, and formatting icons. We provide a score to the image based on its attributes and categories of importance concerning the web page's content. For Ensuring work in real-time, the approach developed, eliminating the need to save results in a database or query a set of pre-downloaded web pages.

Our method's key benefit is that it does not rely on surrounding content, specific templates, or web page categories—however, the system design functions with a wide range of web pages. As a result, it does not create constrain by the writing style or layout of the web page. The method does not require any training data other than the threshold values to choose—it is built to work in real-time. The system eliminates the need to save results in a database or query a set of pre-downloaded web pages. Our technology is beneficial in various applications, including automatically recognizing

²<https://www.javatpoint.com/webpage-vs-website>

advertisements, saving bandwidth by web crawlers by downloading only the most relevant media items, and automatically transforming online pages for consumption on mobile tiny screen devices.

The proposed technique is implemented in Mopsi (a mobile location-based application) (Fränti et al., 2011), in two places. The first is an interactive tool for adding new services to the database. The database can be used as ground truth for the evaluation of the image extraction. The database contains web pages with their representative images that were selected by users. The second is a tool for showing search results to mobile users.

1.2 Mopsi

Mopsi is a location based application created by the Machine Learning group at the University of Eastern Finland's computing department. Mopsi offers geotagged photographs, location mining and data processing, filtering and retrieval of GPS trajectories, user activity, and moving object exposure from GPS trajectories, among other services (Xie, 2019). Fränti et al. (2010) introduce the concept of location-based search engines and their design. The research also developed a prototype and demonstrated it in Finland with comparison using a typical search. The evaluation found the system's usefulness for practical applications. Waga et al. (2012) provide a system that allows for three different categories of items: services, photographs, and routes. The system's goal is to suggest places to visit in the user's immediate vicinity. The results of the evaluation reveal that the system provides useful information.

1.3 Structure of Thesis

The remainder of this thesis is structured as follows. The following Section 2 presents the related works on image extraction strategy with several methods and examples. The explanation of image extraction with their plans is necessary because we work with image extraction and defining representative images. A discussion on the working

procedure of our system with implementation challenges appeared in Section 3. Section 4 deals with the assessment of the collected data and summarizes the findings from the analysis. The analysis extracts the performance of the developed system WebIma 2.00 and compares it with other existing solutions. Section 5 concludes the investigation with the work to come.

2 Image Extraction Strategy

Studies have long advocated research in the area of image retrieval from the Web. With increasingly complex queries, studies such as (Nie et al. 2012) have explored the usage of visual- and semantic concepts for image search. Furthermore, the challenge of performance prediction for image search is investigated (Nie et al., 2012). However, it is only in the past few years that the importance of determining and retrieving a representative image from specific Web pages recognized, rather than the entire Web (Wynblatt & Benson, 1998). This section discusses the most common image extraction approaches and rule-based methods in literature and practice.

2.1 Review on Image Extraction

Since the very beginning, image ranking has started. According to Datta et al. (2008), there has been a significant amount of effort in indexing photos from the Web since the WebSeek project (Smith & Chang 1997) in 1996. Wang et al. (2012) proposed *ImageKB*, an image knowledge-based for representative images. (Lu et al. 2009) present a block-based image feature representation to reflect the spatial properties for a particular concept. Other existing work has primarily focused on extracting multiple helpful photos from a web page (Fauzi et al. 2009) or collecting online pages (Park et al. 2006) and selecting an image for a specific type of web page a news item. (Joshi & Liu, 2009) focuses on news stories where an appropriate image is embedded in the article block and includes a caption. The approach helps overlook potentially significant web images. For image database categorization, Tsai & Lin (2009) analyze several combinations of image feature representation comprising global, local block-based, and region-based features. However, we find less attention to choosing a picture that symbolizes the complete web page.

There have been some studies on the analysis of web photos, but they have mainly concentrated on two elements. The extraction and recognition of text from web photographs is one example (Antonacopoulos & Karatzas, 2000; Lopresti & Zhou, 2000).

The other is web-based image search and retrieval (Frankel et al.1996; Munson & Tsymbalenko, 2001; Yang et al. 2002). The study in Adam et al. (2010) also finds an image extraction approach that concentrates on online pages published in article style (title and body). This method defines the border of the article and selects an image from this region based on its size and aspect ratio. If the size and aspect ratio is appropriate, it may select advertisement images by accident. Our task is more difficult since we must evaluate all of the pictures on the page and choose one that best represents the entire page.

Lew et al. (2009) give a comprehensive overview of the current state of the art in context-based picture retrieval. Tsymbalenko & Munson (2001) concentrate on locating relevant photographs to a particular query without downloading or analyzing them. It just looks at the words in the source code that surrounds the image element. A study (Hu & Bagga, 2003) discovered the functional categorization of images. However, in our technique, we can choose valuable photos that are not surrounded by text. Image classification is also used, but it is used directly to assist in selecting the best image. Vyas & Frasinicar, (2020) conducted an investigation to discover the most representative image of web page. For increasing the performance, the study uses *Support Vector Machines (SVM)* framework with improved classification. The experiment uses Web-lma dataset to evaluate performance. The framework provides 95% accuracy at the cost of a long processing time, whereas our system provides information within a very short time. However, the system is not working anymore.

2.2 Rule-Based Method

The system must assess the context of the web page to choose the best representative image. Rahman et al. (2001) present a technique that combines structural, contextual and summarization analysis. The objective is to extract useful text from an HTML page. Another technique in (Gupta et al., 2003) navigates a Document Object Model (DOM) tree constructed by recursively analyzing HTML code and retrieving pertinent information, including images. Evaluating the contents of the src and href properties to determine the servers to which the links refer filters away irrelevant data such as

advertisement pictures. The DOM tree removes the link's node if an address fits a list of commonly used advertisement servers. A rule-based classifier is used by Parmar and Gadge (2011) to remove advertisement pictures. The majority of advertisement images on the web page release utilize this strategy. However, we employ a web page's DOM tree with more image properties and defined categories. The study by (Gali et al. 2015) was able to achieve strong, consistent results, and it was the first to officially measure and compare the performance of the few other algorithms in the field. On a self-collected dataset, the algorithm of (Gali et al. 2015) had a 64 percent accuracy, compared to the modest 48 percent and 38 percent achieved by Google+ and Facebook, respectively. This thesis rebuilt (Gali et al. 2015) work because this system is not working correctly nowadays. Tabarcea et al. (2017) investigate ways to gather location-aware data from the internet. The investigation employs gazetteers to locate locations. A simple rule-based method applies for titles and representative images summarizing the search results. These details use for individual search results so that relevant to the user's location. Despite the efforts of various academics in relevant fields, none of the available solutions are directly applicable to our case. To our knowledge, the only ways that exist are the commercial ones deployed in Google+ and Facebook. However, neither of them works well, according to our tests.

In this research, we offer a method for parsing the web page's source code. The proposed method detects all images and the best content representative image from the source based on the parsing. We rely on the functional purpose of the pictures inside the web page and features such as size, aspect ratio, image format, and HTML tag attributes rather than studying the content of the photos or examining the text around them. We divide the pictures into categories in the same way (Hu & Bagga, 2003). Based on picture functionality, we divide the following categories: *representative*, *logos*, *banners*, *ads*, and *formatting*, which includes *icons*. The classes rank in this order based on how important they are to the web page's content. This thesis sort the photos based on the features.

3 The representative image extraction procedure

In this thesis, we develop a system for extracting representative images. We follow the method in (Gali et al., 2015). Image extraction and finding representative images from webpage tools is a real-time application. It selects the best representative image instead of analyzing the content by using the following steps: extract all images, extract their features, categorize all images, Score by image features, and best representative images. We score the image by attributes such as HTML tags, size, aspect ratio, image format, etc. We rank the image categories based on their image functions, such as logos, banners, advertisements, formatting icons, and representatives. Our tools' primary benefit is that they do not depend on enclosing content on specific templates or web page levels. It is, on the other hand, designed to function with a wide range of web pages.

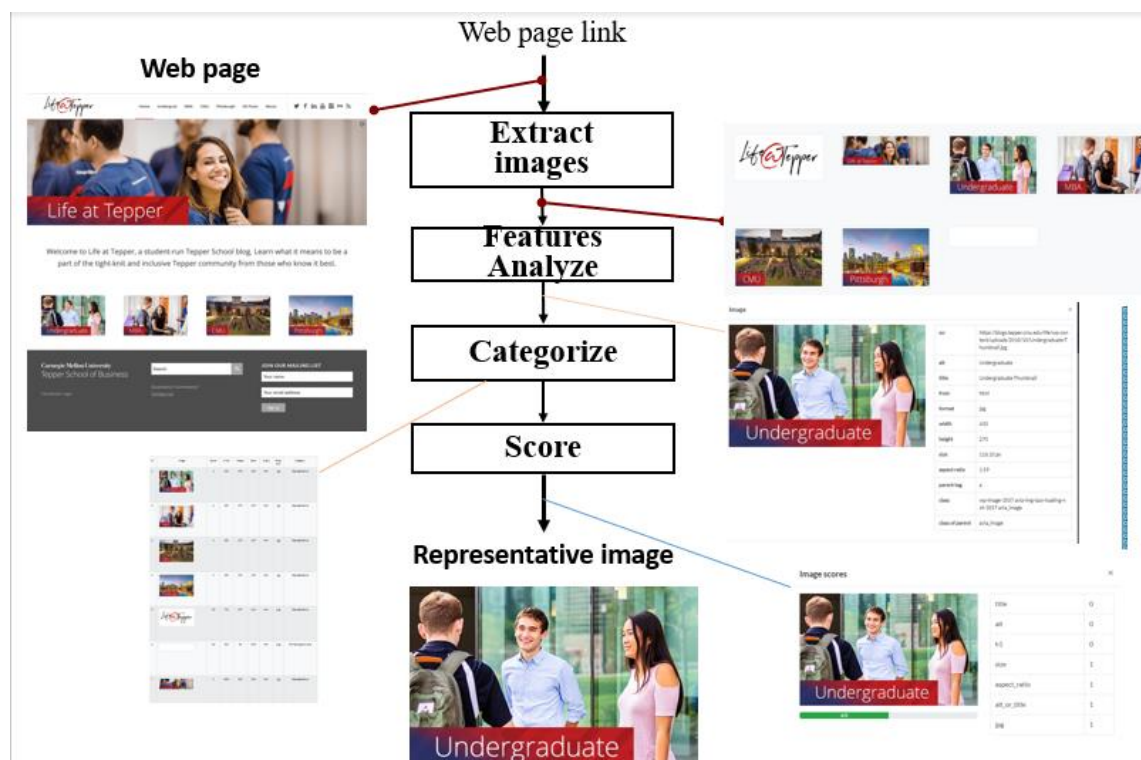


Figure 2: Extracting the representative image

(Gali et al. 2015) has been divided the representative image extraction procedure into four sub-sections described briefly in the below part. As a result, it is not constrained

by the writing style or the web page's layout. The method does not require any training data. In this thesis, we define the threshold values.

3.1 Image Extraction

The algorithm's workflow depicts in Figure 2. First, the URL of the webpage fetches the source code of the webpage via the PHP function and scrapes all the images through the DOM parser³. The Document Object Model (DOM) is a cross-platform and language-independent user interface. It allows programs and scripts to dynamically access and update content and the structure and style of documents⁴.

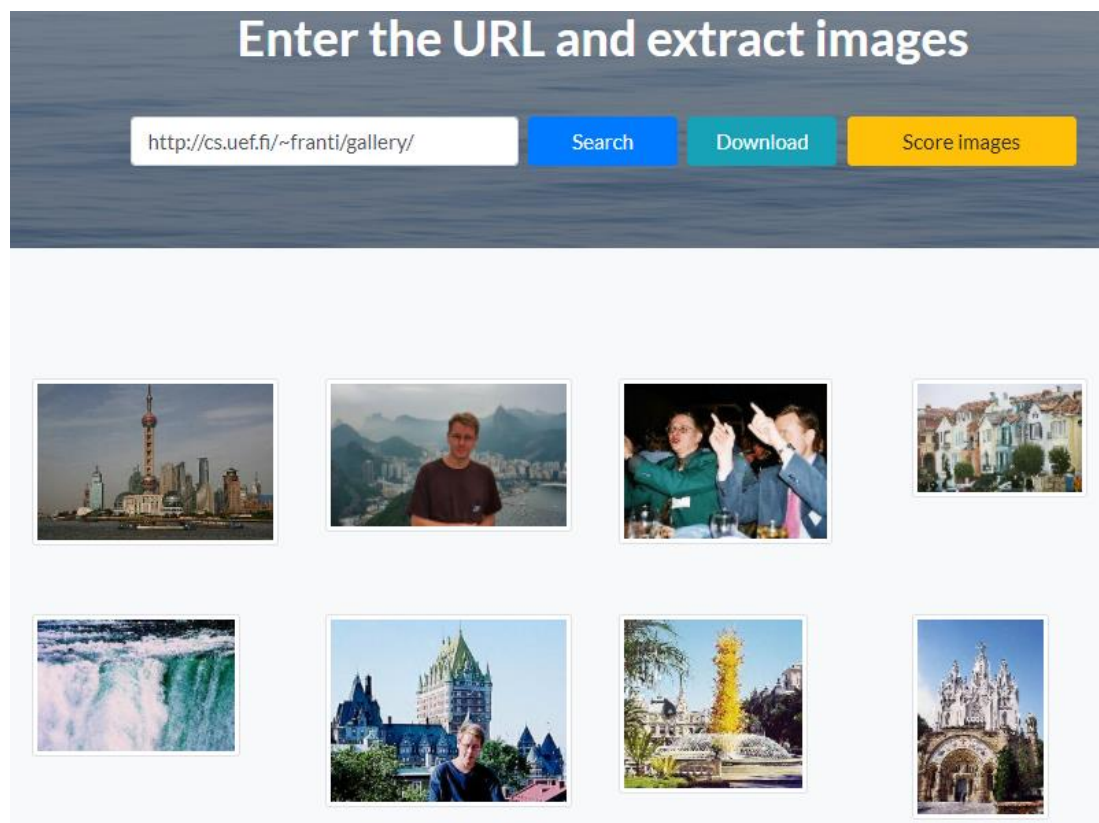


Figure 3: Extracted images

³ <https://simplehtmldom.sourceforge.io/>

⁴ www.w3.org/DOM

We traverse through the DOM tree to find the links on the web page. We look for `` tags for images, `<link>` tags (`type=text/CSS` or `rel=stylesheet`) for Cascading Style Sheets (CSS) files, and `<script>` elements for JavaScript (JS) files. After evaluating the HTML source code, we employ regular expressions to extract pictures from CSS and JS files. Furthermore, if the targeted web page does not contain any images and does not belong to the domain's root page, we must perform the same operation on the root page. The tool extracts the images as shown in Figure 3.

3.2 Image Features

We extract a list of attributes for each image, including `src`, `alt`, `title`, `form`, `format`, `width`, `height`, `size`, and `aspect ratio`, as shown in Figure 4.


	src	http://freshbites.fi/wp-content/uploads/2020/08/h1_pizza.png
	alt	
	title	
	from	html
	format	png
	height	669
	width	669
	size	447.56px
	aspect ratio	1.00
	parent tag	div
	class	attachment_full size_full
	class of parent tag	elementor_widget_container

Figure 4: A sample representative image and its features

We do not download the images because we want to make a real-time application. Therefore we calculate the width and height instead. Use the `` tag attributes to extract the height and width from CSS files or download the image header to calculate the height and width (the first kilobytes of the file that contain the image meta-information).

3.3 Categorical Features

Based on the usages within the web page, we define five image categories presented in Figure 5. After that, we rank them in the following priority order:



Figure 5: A sample web page that contains all the five categories of images

3.3.1 Representative

When extracting images from the source code, types of images directly related to the topic are considered under the representative category. Figure 6 contains the usual variety. This is a representative image for a restaurant.



Figure 6: Sample of representative images.

3.3.2 Logos

It refers to the identity or individuality of an organization or an institution, or a specific thing. In other words, it is a recognizable image that is mainly used to identify the owner of the website of the company or institution. It is used to differentiate the website owner from each other (in Figure 7).



Figure 7: Sample of logos.

3.3.3 Banners

These graphics positions are above, below, or to the sides of the web page's content. The usages of banners are to beautify or inform about a particular topic. This category, which appears in Figure 8. includes headers and footers.

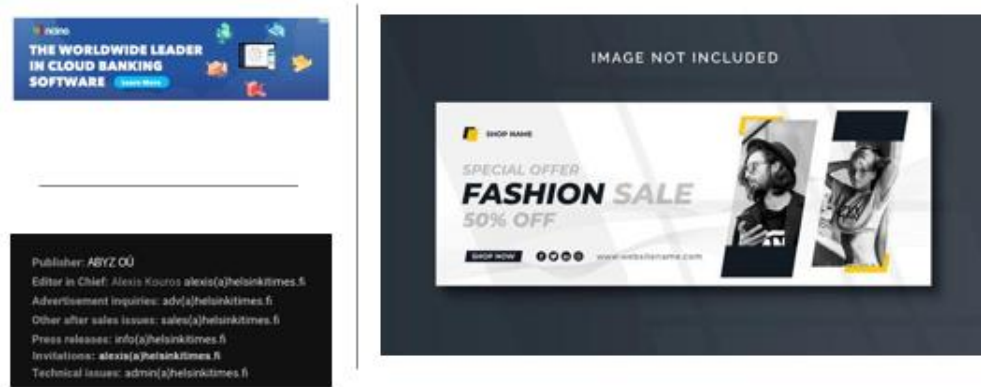


Figure 8: Sample of banners

3.3.4 Advertisements

Commercials are required to promote items or services. The source code of a website has occasionally discovered these types of graphics. The majority of these photographs are unrelated to the theme or substance of the web pages. Figure 9 depicts the category of adverts.



Figure 9: Sample of advertisements

3.3.5 Formatting and Icons

It refers to visuals such as spacers, bullets, borders, backdrops, or photos employed to enhance the visual appeal. These photos are frequently used for decorating purposes. We occasionally come across tiny graphics that aren't designated as logos but serve a functional purpose; these include this category. It could be a link to the home page or a switch to a different language. Figure 10 depicts the icon's variety and formatting.



Figure 10: Sample of formatting and icons images

At first, we sorted all the photographs into categories. Then we sorted them in order of importance. The system selects the image from the highest priority category, including all images within the stated categories.

Table 1 shows the rules for categorizing the extracted photos, with a specified set of keywords for each category. The implemented system classifies the images into a specific category based on a predetermined keyword. Suppose any photos are identified with a particular keyword while extracting images from the source code. In that case, the system keeps the data into a category used to distinguish the type from the others. If the image URL contains the class name of the tag or is identified under the parent element, it goes to that category based on the predefined keyword.

Table 1: Rules-based image categorization.

Category	Keyword	Features
Logo	Logo	
Banner	banner, header, footer, button, Campaign	Ratio>1.8
Advertisement	free, adserver, now, buy, join, click, affiliate, adv, hits, counter, Sell, Discount, offer, ad, Billboard, Campaign, Adverts	
Formatting and Icons	background, bg, spirit, templates, Icon, Format	Width<100 px Height<100 px
Representative		Not in other categories

However, numerous categories can be assigned in a single image because one image can match the requirements of multiple types. The picture is assigned to a single class using a decision tree, as shown in Figure 11. We begin the classification process with logo images because their size and aspect ratio may be appropriate for the Banner and Formatting categories. Following that, we group advertising images based on their ratio or HTML assigned keywords, which may satisfy the formatting or Banner category's requirements. The formatting category then follows because its image ratio may meet the Banner category's criterion.

Furthermore, if the image does not fall into another category, it is considered representative. All HTML, CSS, and JS files have the same priority. With each file, the type gets the same treatment.

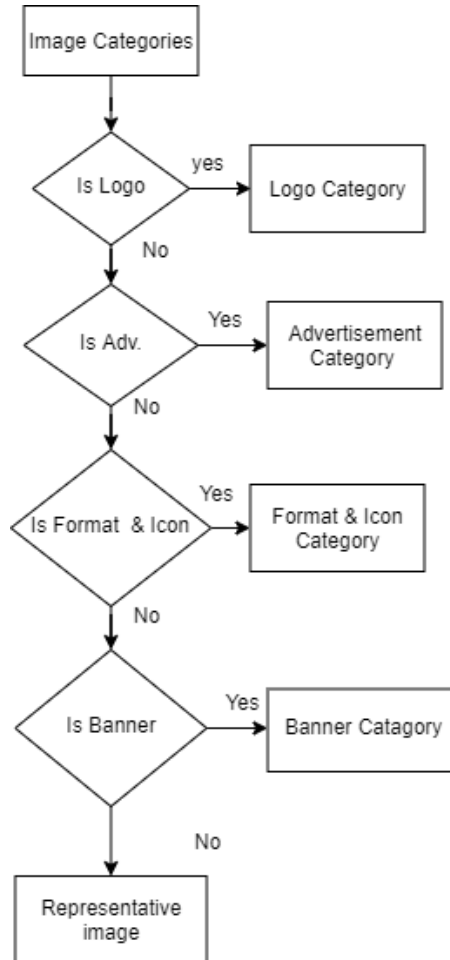


Figure 11: Decision tree for image categories

The predefined keyword "logo" must appear in at least one of the HTML tags attributes (URL, detected classes, element IDs, or parent element IDs), which is a requirement of the logo categories. The presence of any of the following terms in at least one of the HTML tag attributes defines the banners category: "banner," "header," "footer," "Campaign," or "button." The image's aspect ratio must also surpass a threshold of 1.8. In this study, we find the value empirically using a small training sample of 50 web pages. For the Advertisements category, the required prerequisite is that at least one of the HTML tag attributes must have to contain any of the following advertising keywords:

free, now, buy, join, adserver, click, affiliate, adv, hits, counter, Sell, Discount, Offer, ad, billboard, Campaign, adverts.

A separate server or on the same domain might store Advertisement-related images as the web page. Most websites use a different server to store photos on a cloud storage server. Furthermore, many sorts of domains are used to keep the photographs that are necessary for our research. As a result, the image's hosting environment does not follow a constant pattern. To figure out if the image is an advertisement or not, the system could use it.

At least one HTML tag attribute must contain one of the specified keywords: "background," "bg," "sprite," "icon," "format," or "templates," which is a requirement for the Formatting and Icons category. Furthermore, it is a mandatory criterion to fit the type that the height or width is less than a 100-pixel experimentally established threshold.

3.4 Image Scoring

We examine the image's features using a set of rules indicated in Table 2. The system provides a score to the image based on the following criteria:

3.4.1 Image Size

In this study, we use a rule to determine whether or not an image is of sufficient size. If the image fits the following criteria, we consider it to be of appropriate size, where the rule is:

$$Size = width \times height \geq 10.000 px \quad (1)$$

3.4.2 Aspect Ratio

When calculating the image score, we consider which image has an aspect ratio of 1.8 to be more representable. As a result, we must evaluate the characteristics of banners, formatting, or adverts, which are often vast and short or narrow and long. To compute the aspect ratio of the retrieved images, we follow the below formula.

$$Ratio = \frac{\max(width, height)}{\min(width, height)} \leq 1.8 \quad (2)$$

3.4.3 Image Alt and Title

The alt and title elements primarily explain the image's content. When collecting photos from web pages, if any of the pictures include an alt or title attribute, these are prioritized above others. The keywords alt and title are extracts from the image to continue the process. The following step compares them to the keywords on the web page using the title and h1 components. We start by using XPath to extract the content of the web page's related to <title> and <h1> tags. After that, a predefined set of patterns is used, which consists of space and delimiters such as ',', ';', '/', '|', '>', '|', '<', '-', '!', ':', '!', ':!'. These predefined patterns are mainly used to separate the words and the phrases of the web page <title> and <h1> tags. The next step is to remove any special characters from the text, such as '[', '{', '?', '!'. The final step in this procedure is to use string comparison. This technique is mainly applied comparison to match the keywords of image alt and title with the keywords of the web page title and h1s;

3.4.4 Image path and URL

To calculate the Image Path and URL values, you must first parse the image path then extract the keyword. The next step is to match the keywords to the title and h1 tags on the web page. The principal aim of this step is to relate the keywords with the content of the web page. If the term matches any predefined keywords, we consider the image more relevant to the web page's content.

This thesis considers images that are children of <h1> or <h2> in the DOM tree. The suggested image is more suited to its content because the <h1> and <h2> tags describe the web page's major focus. In addition, the DOM tree's sub-tree picture location is significant.

3.4.5 Image Format

The Joint Photographic Experts Group (jpg), Scalable Vector Graphics (SVG), Portable Network Graphics (png), and Graphics Interchange Format (gif) are the four types

of image formats examined for this rule (gif). Because this term is commonly used for images, the jpg format is taken into account when extracting an image from the source code in this situation. However, because it is widely used for compressing photos, the png file format is also appropriate for demonstrating the value of an image. Furthermore, most web pages make use of this style of graphic to depict logos and icons. As a result, the jpg format has a higher priority than the png format, although the SVG and gif formats are also utilized in picture files but are less focused. Graphics and image formatting are the most common uses for these formats. The image scoring of this thesis conducts for 8 points in total.

Table 2: Rules used for image scoring.

Criterion	Points
Image size ≥ 10.000 px	1
Aspect ratio ≤ 1.8	1
Image alt or title set a value	1
Keywords of alt or title also appear in <title> tag	1
Keywords of alt or title also appear in <h1> tag	1
Keywords of image path also in <title> or <h1> tags	1
The image is in the sub-tree of <h1> or <h2> tags	1
Format = jpg, png or gif	1

In this research, all of the rules are regarded as equally important. As a result, with the exception of the previously mentioned image formats, all image formats were assigned the same weight of 1.

In the case of the representative image category, the scores are mainly calculated only for images that belong to the highest priority list. If no images are found under this category, it goes to the next category based on the priority list for calculating the image scores. This procedure is applied for all the categories to calculate the scores.

At the last phase of the research, the scores are summed up. Then the ranking procedure is applied based on their calculated scores. In this case, this technique is applied except the logo category because, in this category, the images are stored based on the size. Furthermore, we consider that the logo which is extracted from the web page has the largest size among the other logos on the page.

4 User Interface of WEBIMA 2.00

This study develops a web data scrapping system, WebIma2.0, for identifying representative images of a web page. The tool is available on the University of Eastern Finland⁵ website. The implemented system supports all the major browsers. However, Google Chrome and Mozilla Firefox are recommended. The system has a user-friendly interface where users can start their search by entering a URL into the text box and hitting the Search button. Figure 12 shows the user interface of WebIma 2.0.

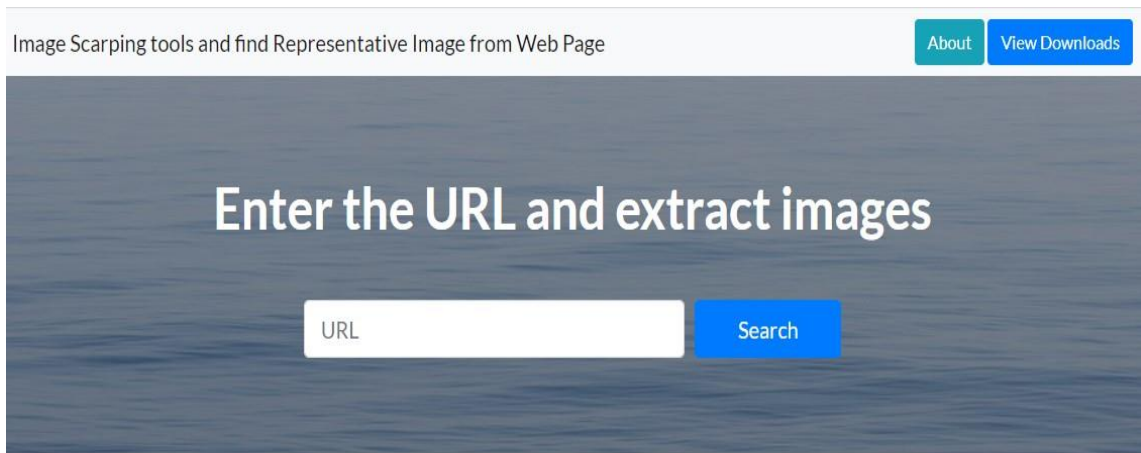


Figure 12: Application homepage

After entering a web link in the search box, the system extracts all possible images using web scrapping techniques. Figure 13 shows an example of WebIma 2.0 image extraction. Figure 13 also offers two other functionalities of WebIma 2.0. The system downloads all extracted images through the download button, whereas the score button helps to score all pictures. WebIma 2.0 finds the image types through the score button and defines the web application's representative images.

⁵ <https://cs.uef.fi/mopsi/WebIma/demo/>

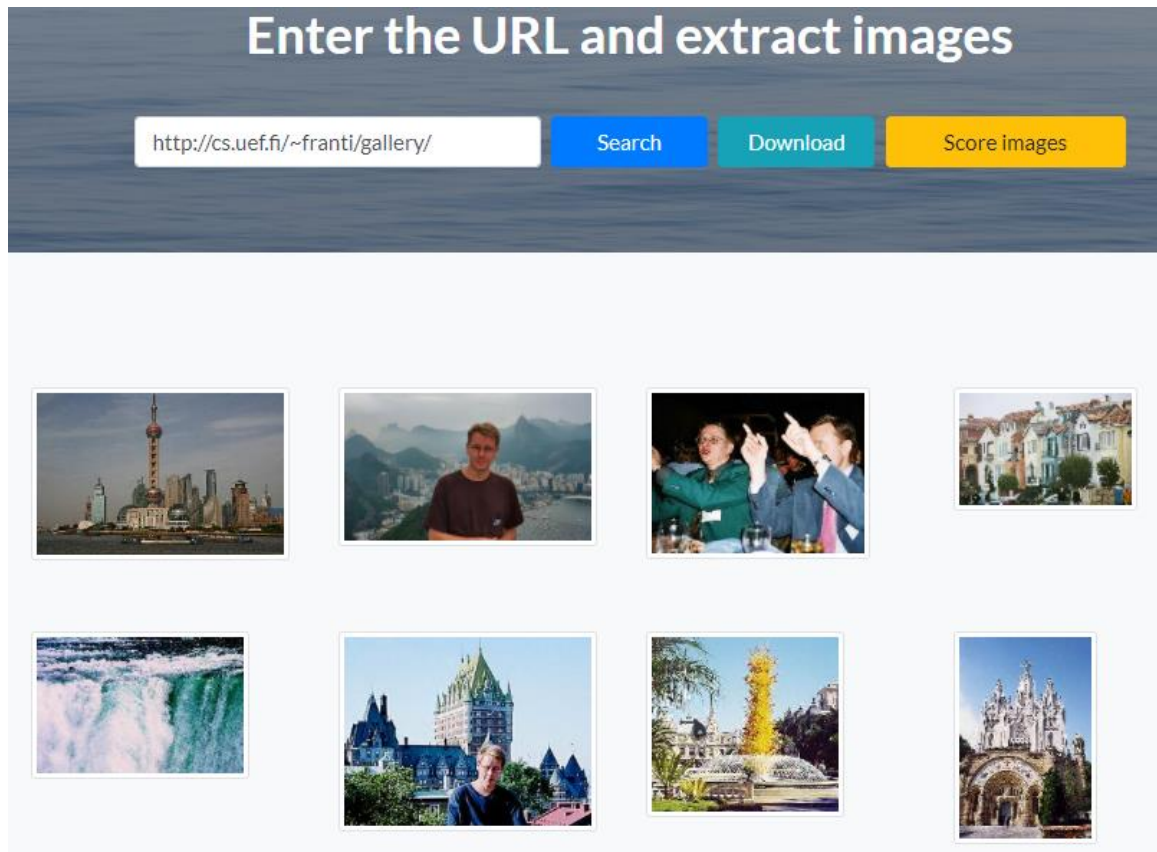


Figure 13: Extracted images

Figure 14 shows the details of the score function. The function scores the images and defines several critical features. Figure 14 shows the essential features are score, height, weight, origin, ratio, and image extension types. The score button also defines image categories like representative, formatting, and icons. Figure 14 shows that the image scores are in descending order.

Image Scarping tools and find Representative Image from Web Page About View Downloads

Enter the URL and extract images

Search Download Score images



ID	Image	Score	Width	Height	Ratio	Origin	Extension	Category
1		6	242	161	1.50	html	jpg	Representative

Figure 14: Image categories and features

This study examines the image's features using a set of rules based on image size, aspect ratio, image alt and title, image path and URL, image format, image tag. Figure 15 shows image scoring.



title	1
alt	0
h1	1
size	1
aspect_ratio	1
alt_or_title	1
jpg	1

4/8

Figure 15: Images score

5 Experiment

To collect the web application links database, we follow two previous datasets OSM and WebIma dataset. In this study, the implemented tool extracts the pictures. The implemented tool selects the first three images for every web application. The top three images have been selected based on their score. The system automatically selects the top scorer's three representative images. The whole working procedure is described in detail in the below sections.

5.1 Dataset Used

In this thesis, we selected the previous dataset named OSM and Weblma. The ground truth of WebIma 2.0 information stores on a website⁶. The WebIma dataset has 810 web application links, whereas the OSM dataset has 255 web links. The investigation is conducted on 1065 web applications where WebIma 2.0 selects 1576 images as top representative images. For each web application, WebIma2.0 selects only the top three scorer images as representative images.

5.2 Experimental Setup

In this thesis, we develop a web scraping tool. Web scraping is a technique of extracting organized web data. Pricing monitoring, price intelligence, news monitoring, lead creation, and market research are just a few of the many web scraping applications. People and enterprises use web data extractors for making better decisions in the context of a large amount of publicly available web data. In this thesis, we use scraping for extracting representative web images. If the public website you want to acquire data from doesn't have an API, or if it has, but gives you restricted access to the data,

⁶ <https://cs.uef.fi/mopsi/WebIma/demo/doc/>

our developed web scraping tool is a good option. Figure 16 shows an example of a working web application.

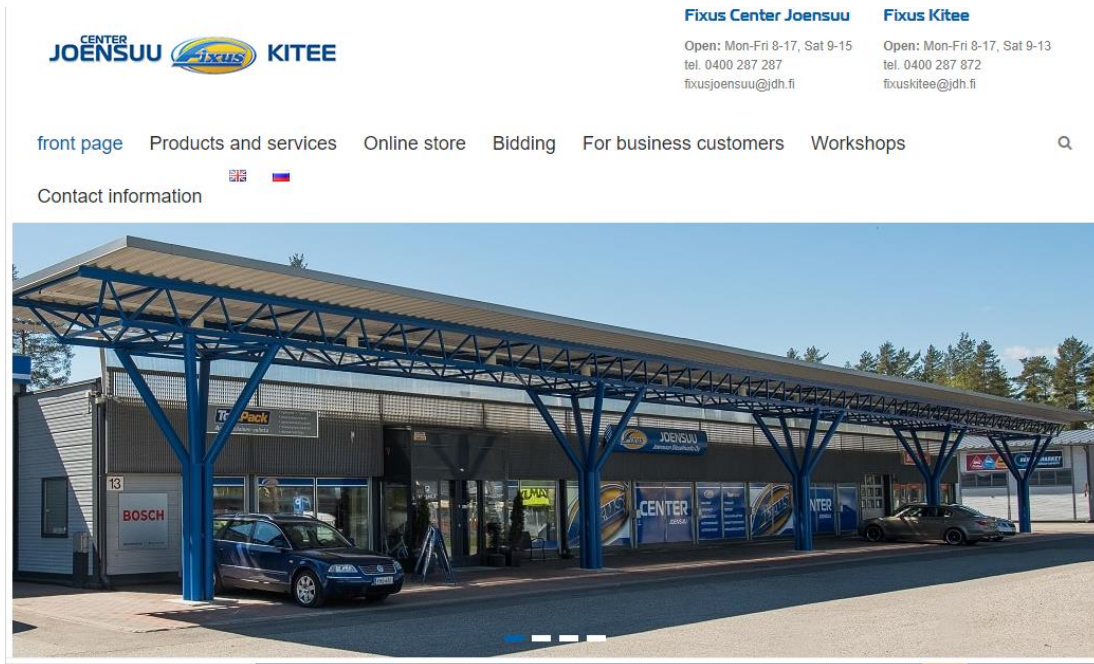


Figure 16: Examples of working web application

The tool has some limitations for data extractions. The system does not work for web applications that show pop-up windows⁷ have no image tag⁸. Those websites are now updated like the image comes from backend technology, unlike the previous web page in Figure 16. On those pages, I didn't find any image tag. That's why tools are not working for those web pages. In Figure 17, we show an example of a pop-up window web application. We also see other links are not working for the pop-up window. Figure 18 shows the image of the pop-up window web application. In general, when we enter those types of links, it asks for some permission and opens a pop-up window. When we confirm, we can browse those links. So those two types of web pages my tools are not working.

⁷ <https://www.pippurimylly.fi/>

⁸ <https://www.helsinki.fi/fi>

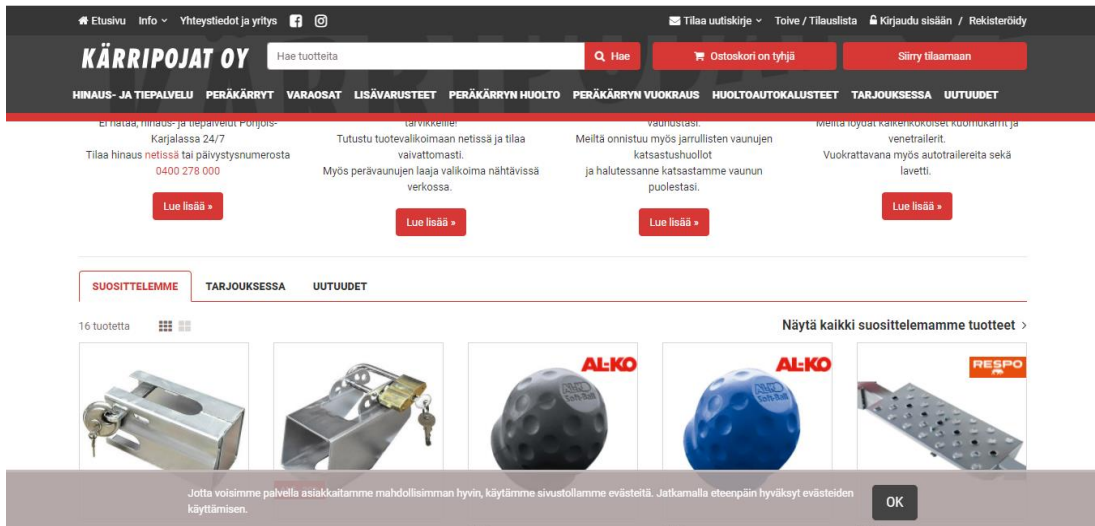


Figure 17: Example of not working web applications (pop-up window)

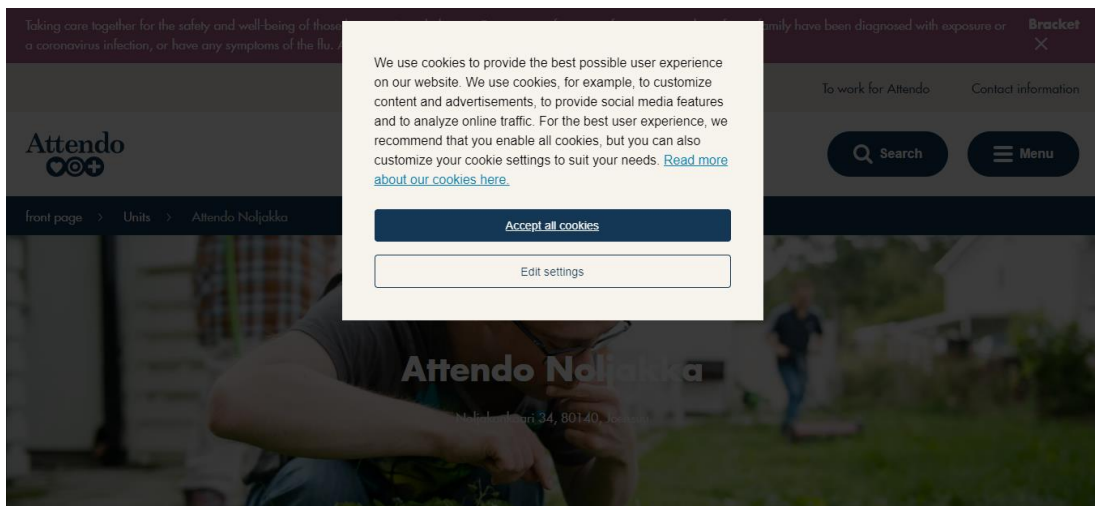


Figure 18: Example of not working web applications (pop-up window)

This thesis uses several website images as examples of true positive, false positive, and false negative. This study defines those examples of scenarios shown below descriptions.

In scenario 1, Figure 19 shows an example of accurate positive images. Figure 19 images selected from supermarket website⁹. A supermarket is a self-service store divided into sections and offers a broad selection of food, beverages, and home items. In our day-to-day life, we are using supermarkets that is the reason behind selecting those images. So, we can easily understand the scenario. We prefer Figure 19 images as true positive because supermarket sells day to day life products. So, for a supermarket, those images represent the web application.



Figure 19: Examples of true positive image

In investigating true positive, false positive, and false negative, we use the same web application for better comparison. In Figure 20, the higher portion titled false-positive, shows the false-positive image example. The images are defined as representative through WEBIMA 2.0. However, if we see those images, we can easily understand that those images have no connection in the supermarket context. we, therefore, define those images as false positive. Figure 20 shows lower portion provides examples of false negatives as we say that we select all images from a supermarket web application. The photos represent the web application because those images are promoting products. Supermarkets have the common feature to promote and sell day to day life valuable things. However, in this thesis, WEBIMA 2.0 selects formatting and icons images which are false negative information.

⁹ <https://www.lidl.fi/>

False Positive Images



False Negative Images



Figure 20: Examples of false negative and positive images

In scenario two, this thesis chooses examples of a hotel¹⁰. Figure 21 shows examples of true positive images. WebIma 2.0 chooses Figure 21 images as representative images which represent the hotel buildings. As those images represent the hotel and the implemented system defines those images as a representative, this analysis defines them as true positive images.

¹⁰ <https://www.greenstar.fi/>



Figure 21: Examples of true positive images

This study also defines the false-positive images of the same website¹¹. This web application provides hotel services, and the information may always be about hotels and different hotel-related services. However, Figure 21 shows the hotel web application's false-positive images because those images do not represent the hotel services.



Figure 22: Examples of false-positive images

In scenario 3, this thesis chooses the images of an e-commerce web application¹²—the aims of the applications to promote a variety of clothing products. So, the image of clothing product promotions and sales represents the web application. Figure 23 shows the true positive images, which are defined by WebIma 2.0.

¹¹ <https://www.greenstar.fi/>

¹² <https://www.imagewear.fi/>



Figure 23: Examples of true positive images

False Positive Images



False Negative Images



Figure 24: Example of false-positive and negative images

Figure 24 shows examples of false-positive and negative images. Figure 24 the higher portion titled false-positive images provides the information of false-positive images because those images do not show e-commerce related information. Figure 24 the

lower part titled false-negative offers the information of false-negative images because the images show that models promote products. So, those images represent the web application. However, WebIma 2.0 defines those images as formatting and icons.

5.3 WebIma 2.0 Performance

The selected images provide an essential ground truth of our purpose. In general, user choices may be subjective, making it impossible for any practical algorithm to get 100% accuracy even if the algorithm has been trained for that specific web application and identified the user who created the image choice. However, the ground truth collection method is often valuable for assessing our system's performance. This thesis finds 69% accuracy of WebIma 2.0 with 97% image extraction percentages. The performance comparison for the two datasets is displayed in Table 3. Equation 1 has been used to calculate accuracy, whereas 2 and 3 have been used for false positive and negative.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

In this thesis, we apply our implemented tool on several domain web applications to check the performance of the developed device as we say that this thesis used two datasets. So there is always a question that comes to mind: Is there any performance difference between those two datasets. Table 3 shows the performance differences. The performance difference is shallow. The implemented tool accuracy with the combination of that two datasets is 69%. The false-positive and negative ratios are 16% and 15%, respectively.

True Positive = Identifies representative images correctly.

False Positive = Identify logos, banners, advertisements, and formatting and icons as representative images.

False Negative = Do not recognize the representative image.

The two dataset results have a lower percentage of accuracy difference. Weblma dataset accuracy is 68%, where the OSM dataset has 71% accuracy. The false-positive and negative 17% and 15% for the Weblma dataset. On the other hand, OSM has 16% and 14% false positive and negative respectively. Here essential points reveal false positives and negatives. In all those cases, the false positive is always higher than the false negative. This analysis also compares precision and recall, where Table 3 shows that OSM has 82% precision and 84% recall. Weblma dataset gets 80% and 82% precision and recall, respectively.

Table 3. Performance comparison between datasets

	Accuracy	False Positive	False Negative	Precision	Recall
Weblma	68%	17%	15%	80%	82%
OSM	71%	16%	14%	82%	84%
Weblma+OSM	69%	16%	15%	81%	82%

5.3.1 Accuracy Comparison with Existing Systems

This investigation compares Weblma 2.00 performance with other existing tools. The performance of other existing systems (Weblma, Google+, and Facebook) have been selected from (Gali et al. 2015). This analysis finds some web links are not in service. This thesis evaluates the accuracy of Weblma 2.00 without the evaluation of those web links. Table 3 provides information on the accuracy and image extraction ratio. This evaluation helps us to compare the effectiveness of Weblma 2.00 Table 4 summarizes the findings. It shows that our approach correctly identified 1273 images from 1850 images. The accuracy of our system is 69% whereas Weblma is 64%, Google+ 48%, and Facebook 39%. Weblma, Google+, and Facebook performed for 1002 images, according to Gali et al. 2015). The results also revealed that our parser could retrieve photos from 97 percent of the websites. The image extraction Performance is a bit lower than Weblma.

Table 4. WebIma Dataset. Performance Results Comparison

	Accuracy	Extracted Images
WebIma 2.00	69%	97%
WebIma	64%	99%
Google+	48%	92%
Facebook	39%	90%

5.3.2 Precision, Recall and F-Score

In this section, we calculate the precision, recall, and f-score of our implemented tool. The f-score, often known as the f-measure, is a statistical measure of a test's quality. The F-score calculation is conducted through precision and recall. The number of correctly identified positive results is divided by the number of positive outcomes, including those incorrectly specified, called precision measurement. On the other hand, recall means the number of correctly identified positive results divided by the number of samples identified as positive. A higher precision algorithm produces more relevant results than irrelevant ones, while a high recall algorithm delivers the most relevant results (whether or not irrelevant ones can also return). The highest possible value of the F-score is 1, indicating perfect precision and recall, and the lowest potential value is 0 if either the precision or the recall is zero.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

In our investigation, we find 81% precision, and the recall is 82%. After the evaluation of precision and recall, we evaluate the F-score. The assessment gets an 82% F-score.

5.3.3 Performance Evaluation with Changing Parameters

In this study, we implement a web scraping tool to extract images from web applications. When we get pictures from web applications, we find representative photos. The objective behind searching representative images is to get an overview of a web application very quickly. There are several parameters that we define to select the representative. In this section, this thesis changes the parameter values or the entire parameter to check the effect on the performance of the developed WebIma 2.00. Table 5 shows the comparison of performance based on changes in parameters. The changes of parameters affect accuracy, precision, and recall. This experiment uses 200 randomly selected web applications.

Table 5 shows that some parameter changes do not affect accuracy much. The shift in SVG, png, or gif 0.5 to 1 point generates 68% accuracy with 74% precision and recall 89%. Other parameters like aspect ratio, jpg, and image size change do not significantly differ with system observed accuracy. Aspect ratio 1.8 to 1.00, jpg 1 to 0.5, and no points in image size get 67% accuracy. This investigation observes that with the same accuracy have different precision and recall. For aspect ratio, jpg and not points on image size get the same precision 74%, whereas recall is 89%, 88%, and 87%, respectively.

The limit keyword and image height and width change get 65% and 63% accuracy, respectively. The evolution of some parameters affects performance in a big ratio. The limit of advertisement keywords and removing formatting and icons keywords get a lower accuracy rate. They get 56% and 54% accuracy, respectively. For limited advertisement, precision and recall are 61% and 88%, respectively, whereas formatting and icons get 58% and 88% precision and recall. Table 4. confirms that the developed tool with no change of parameters extracts more accuracy.

Table 5. Assessing system parameters performance

Parameters Description	Accuracy	Precision	Recall
Aspect Ratio 1.8 → 1.00	67%	74%	89%
jpg 1 → 0.5 point	67%	74%	88%
svg, png or gif .5 → 1 point	68%	74%	89%
Height < 100 px → <300 Width < 300 px → <300	63%	68%	88%
Limit Advertisement keywords: free, ad-server, now, buy, join	56%	61%	88%
Limit keywords: banner, header	65%	71%	89%
Removed Formatting and Icons keywords	54%	58%	88%
Set no points for image size	67%	74%	87%

5.4 Discussion

We are living in a world of technology where web applications contain massive information. The web applications also carry a vast number of images, videos, and other interactive content. With the change of time, that information made a colossal dataset. Different techniques are introduced to deliver the informative parts of the web applications where web scraping is one of them. In this thesis, we developed a web scraping tool.

In the investigation of the performance of our implemented system, the first investigation conducts for the difference of performance among the combination of OSM and Weblma, only OSM and only Weblma. Through the analysis, we get confirmation that our advanced tool works better for the OSM dataset. The accuracy is 71%, whereas Weblma 2.0 has 69% accuracy.

Table 4 summarizes the findings. It shows that Weblma 2.00 correctly identified 1273 images from 1850 images. We compare Weblma 2.00 with Weblma, Google+, and Facebook. Through the analysis, we got the information that our tool has better accuracy than others. The accuracy of our system is 69%, whereas Weblma is 64%, Google+ 48%, and Facebook 39%. Table 3 summarizes the findings. It shows that Weblma 2.00 correctly identified 1273 images from 1850 images. We compare Weblma 2.00 with Weblma, Google+, and Facebook. Through the analysis, we got the information that our tool has better accuracy than others. The accuracy of our system is 69%, whereas Weblma is 64%, Google+ 48%, and Facebook 39%.

In this thesis, we conduct the analysis of precision, recall, and F-score. When the precision and recall are high, that indicates that the algorithm is doing very well. Our investigation found 81% precision, and the recall is 82%, which means our system works well. We also do the F-score to decide the test quality. The assessment gets an 82% F-score

In Table 5 we can see that some parameter changes do not affect accuracy more. The main differences show in false positive and negative. In the case of our developed tool, the difference between false positive and negative is shallow. However, when we change the parameter values, then we see that false-positive have a higher ratio than the false-negative result. Another point was that the limit of advertisement keywords and removing if formatting and icons keywords get a lower accuracy rate. They get 56% and 54% accuracy, respectively.

6 Conclusions

In this thesis, we introduced a method to extract representative images from a web page. Our method is implemented in Mopsi . We evaluate the performance through an experiment where the system runs 894 website links. The results show that WebImage achieved 69% accuracy whereas the previous system, Google+, and Facebook provide (64%), (48%) and (39%) accuracy, respectively. The experiment also reveals that the method works better on the OSM dataset. Through the analysis, we found out that two parameters have a significant impact on the performance: removed formatting and icons keywords and limit advertisement keywords. With the change of these two parameters, the accuracy dropped to 60%. Therefore, they must remain unchanged to maintain the good performance of WebIma 2.0. In addition, we discover that our method does not work on particular websites, such as pop-up window/no image tag. Future research should focus on these challenges. The existing method can also be improved by training the parameters for better outcomes. However, a vast amount of training data, which are currently not available is needed. Nonetheless, the data we used included 894 websites, resulting in this performance.

References

Adam, G., Bouras, C., & Pouloupoulos, V., (2010). Image Extraction from Online Text Streams: A Straightforward Template Independent Approach without Training. *In Advanced Information Networking and Applications Workshops (WAINA)*, 24th International Conference, pp. 609-614. IEEE.

Antonacopoulos, A., & Karatzas, D. (2000). An anthropocentric approach to text extraction from WWW images.

Azad, H. K., Raj, R., Kumar, R., Ranjan, H., Abhishek, K., & Singh, M. P. (2014). Removal of Noisy Information in Web Pages. *In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*. ACM.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2), 1-60.

Fränti, P., Chen, J., & Tabarcea, A. (2011). Four Aspects of Relevance in Sharing Location-based Media: Content, Time, Location and Network. *In WEBIST*, pp. 413-417.

Fränti, P., Kuittinen, J., Tabarcea, A., & Sakala, L. (2010). MOPSI location-based search engine: concept, architecture and prototype. *In Proceedings of the 2010 ACM symposium on applied computing* (pp. 872-873).

Fauzi, F., Hong, J. L., & Belkhatir, M. (2009). Webpage segmentation for extracting images and their surrounding contextual information. *In Proceedings of the 17th ACM international conference on Multimedia*, pp. 649-652. ACM.

Frankel, C., Swain, M. J., & Athitsos, V. (1996). *Webseer: An image search engine for the world wide web*. Technical Report 96-14, University of Chicago, Computer Science Department.

- Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. *In Proceedings of the IEEE International Conference on Computer Vision* (pp. 1633-1640).
- Gali, N., Tabarcea, A., & Fränti, P. (2015). Extracting Representative Image from Web Page. *In WEBIST* (pp. 411-419).
- Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P., (2003). DOM-based content extraction of HTML documents. *In Proceedings of the 12th international conference on World Wide Web*, pp. 207-214. ACM.
- Hu, J., & Bagga, A., (2003). Functionality-Based Web Image Categorization. WWW (Posters), 2003.
- Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable?. *IEEE transactions on pattern analysis and machine intelligence*, 36(7), 1469-1482.
- Joshi, P. M., & Liu, S., (2009). Web document text and images extraction using DOM analysis and natural language processing. *In Proceedings of the 9th ACM symposium on Document engineering*, pp. 218-221. ACM.
- Kherfi, M. L., Ziou, D., & Bernardi, A., (2004). Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Computing Surveys (CSUR)*, 36(1), pp. 35-67.
- Kim, M., Kim, Y., Song, W., & Khil, A. (2013, August). Main content extraction from web documents using text block context. *In International Conference on Database and Expert Systems Applications* (pp. 81-93). Springer, Berlin, Heidelberg.
- Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1), 1-19.

- Lu, Z., Ip, H. H., & He, Q. (2009). Context-based multi-label image annotation. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (pp. 1-7).
- Lopresti, D., & Zhou, J. (2000). Locating and recognizing text in WWW images. *Information Retrieval*, 2(2), 177-206.
- Helfman, J. I., & Hollan, J. D., (2000). Image representations for accessing and organizing Web information. In *Photonics West 2001-Electronic Imaging*, pp. 91-101. International Society for Optics and Photonics.
- Money, A. G., & Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, 19(2), 121-143.
- Munson, E. V., & Tsymbalenko, Y. (2001). To search for images on the Web, look at the text, then look at the images. In *Proceedings of the First International Workshop on Web Document Analysis* (pp. 39-42).
- Nie, L., Wang, M., Zha, Z. J., & Chua, T. S. (2012). Oracle in image search: a content-based approach to performance prediction. *ACM Transactions on Information Systems (TOIS)*, 30(2), 1-23.
- Nie, L., Yan, S., Wang, M., Hong, R., & Chua, T. S. (2012). Harvesting visual concepts for image search with complex queries. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 59-68).
- Park, G., Baek, Y., & Lee, H. K. (2006). Web image retrieval using majority-based ranking approach. *Multimedia Tools and Applications*, 31(2), pp.195-219.
- Parmar, H. R., & Gadge, J., (2011). Removal of Image Advertisement from Web Page. *International Journal of Computer Applications*, 27(7).
- Rahman, A. F. R., Alam, H., & Hartono, R. (2001). Content extraction from html documents. In *1st Int. Workshop on Web Document Analysis (WDA2001)* (pp. 1-4).

Smith, J. R., & Chang, S. F. (1997). Image and video search engine for the world wide web. In *Storage and Retrieval for Image and Video Databases V* (Vol. 3022, pp. 84-95). International Society for Optics and Photonics.

Tsai, C. F., & Lin, W. C. (2009). A comparative study of global and local feature representations in image database categorization. In *2009 Fifth International Joint Conference on INC, IMS and IDC* (pp. 1563-1566). IEEE.

Tabarcea, A., Gali, N., & Fränti, P. (2017). Framework for location-aware search engine. *Journal of location Based services*, 11(1), 50-74.

Tsymbalenko, Y., & Munson, E. V., (2001). Using HTML metadata to find relevant images on the world wide web. *Proceedings of internet computing*, 2, pp.842-848.

Vyas, K., & Frasincar, F. (2020). Determining the most representative image on a Web page. *Information Sciences*, 512, 1234-1248.

Waga, K., Tabarcea, A., & Fränti, P. (2012). Recommendation of points of interest from user generated data collection. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)* (pp. 550-555). IEEE.

Wang, X. J., Xu, Z., Zhang, L., Liu, C., & Rui, Y. (2012). Towards indexing representative images on the Web. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 1229-1238).

Wynblatt, M., & Benson, D. (1998). Web page caricatures: Multimedia summaries for WWW documents. In *Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No. 98TB100241)* (pp. 194-199). IEEE.

Xie, M. (2019). Trajectories medoid and clustering. University of Eastern Finland , 1-56.

Yang, J., Li, Q., & Zhuang, Y. (2002). Octopus: aggressive search of multi-modality data using multifaceted knowledge base. *In Proceedings of the 11th international conference on World Wide Web* (pp. 54-64).

Yu, S., Cai, D., Wen, J. R., & Ma, W. Y., (2003). Improving pseudo-relevance feedback in web information retrieval using web page segmentation. *In Proceedings of the 12th international conference on World Wide Web*, pp. 11-18. ACM.