Web Scraping Data Collection from Leather Craft E-Commerce

Catur Nugroho1*, Avip Kurniawan2, Bei Harira Irawan3, Deddy Prihadi4, April Lia Hananto5, Tukino6

1 Informatics Department, Universitas Siber Asia, 12550, Jakarta
2 Informatics Department, Krisnadwipayana University, 13077, Bekasi
3,4 Digital Business Department, Pancasakti Tegal University, 52121, Tegal
5,6 Information System Department, Buana Perjuangan Karawang University, 41361, Karawang

1*caturnugroho@lecturer.unsia.ac.id

This article contributes to:

Highlights:
• information on these leather materials we get from skin types: Leather, synthetic (polyurethane polymer) which is generally used for belts, cases, wallets, bracelets, jackets, key chains, sandals, gloves, shoes, sofas, sheet leather stickers, watch straps, bags, hats

Abstract
In the field of information research leather crafts are an important part. Especially about leather crafts, about the most dominant materials used and to assist process making of other products. With the emerge of e-commerce and digital computing, gathering information from digital sources and structuring it in a convenient way can provide meaningful information. Data sources are obtained by using web scraping in extracting information from websites based on text by presenting data statistics, the results of this study show the dominant use of synthetic leather materials, which is Polyurethane polymers can be applied to all types of leather handicraft products.

Keywords: market study; web scraping; python; fashion leather; e-commerce

1. Introduction

The Internet now is the new access for importance business. for knowing how to use the Internet and for use different opportunities that can bring us can be one of the keys to success in area e-commerce and e-marketing [1]. In the modern digital world, E-Commerce websites have become very popular [2], The new trend of technology put us to change the way we’re doing our business [3].

Web scraping It is the definition automated extraction of information online [4]. Web scraping the process of retrieving a semi-structured document from the internet, generally in the form of web pages in a markup language such as HTML or XHTML. The term describes the automated process of accessing websites and downloading certain information, such as: price, of individual products [5], Web scraping is faster and it is more flexible in terms of obtaining data [6], extracting information from open sources with tools available online users is a very complex task, most of this is done with the manual process required to select the information the user needs from the web page [7].
A promising method for collecting data in real time, overcoming some of the limitations of price data sources used today. With today’s applications mostly focusing on consumer (online) pricing [8], consumer online shopping behavior to meet consumer needs and further success in this key market [9]. The scope of applications for web scraping is expanding as more and more pricing data is published online. The research findings aimed to reveal the current situation and shed light on the trends in [10] Leather Craft.

At present, new approach for architecture recovery using ensemble clustering [11]. The diverse use conditions of leather and increasing consumer demands are becoming challenges for the leather industry, Leather inherits all-natural leather bio-mimetic properties such as flexibility, sanitization, cold resistance, biocompatibility, biodegradability, and other cross-domain functions, unremitting attention in multi-functional bio-based materials [12]. The findings describe the scope of leather-based functional materials that are commonly used in order to help manufacturers and marketers become better at understanding the function of leather in the industry, there are four criteria for searching for leather bag recommendations, namely price, genuine or level of leather authenticity, bag dimensions in inches, gender or gender. Various leather shoe products ranging from casual shoe models, formal to sneakers models [13]. A system for fashion retail stores based on a multi-clustering approach of items and user profiles in online and physical stores [14].

2. Methods

The collection of this research data from the Tokopedia marketplace, which is one of the top MSME’s marketplace in Indonesia according to good stats shown in Figure 1 [15], offering variety of products according to needs and also handicrafts leather products [16].

![Mainstay Marketplace in Indonesia 2023](Image)

Reliable marketplace data in Indonesia originating from surveys on internet usage by MSMEs and corporations conducted by the Association of Indonesian Internet Service Providers (APJII) 2023 is presented in Figure 1.

Web scraping can be done using three most common ways: using non-free software, using specialized web scraping services for paid, or using opensource software such as Python, JavaScript, or R. Web scraping algorithms are designed effectively by combining several libraries already provided by some software (examples shown in Table 1).

Technically, it is done with opensource tools with Anaconda Navigator 2.3.2 (analysis and data science platform). The fundamental difference between static content and dynamic content lies in the key: static content can be inserted into HTML text, if you want to do an access such as downloading a file from the internet and extract that information [17]. Dynamic content requires the action of rendering on the browser environment (search engine) in order for content to appear and possibly more sophisticated interaction through the website [18]. When users use a browser to access the website manually [19]. The differences described are likely almost the same, when a scraper is used to access a website, some elements may or may not be accessible based on static/dynamic rendering and access methods. First, we analyze the target website pages for their structure and gather some table elements that will be modeled into the data [20].
The explanation in the table 1 is stated as follows: Beautiful Soup can be run using several programming languages, one of which is Python which will be used in developing this web scraping application and is very easy to implement with good documentation [21]. Selenium Websites that produce dynamic content using JavaScript require browser support. This system design uses the web scraping method to retrieve scientific article data, which uses the Python programming language using the selenium library [22]. Kimurai is a modern web scraping framework written in Ruby Dynamic scraping content to scrape and interact with JavaScript rendered websites. Drop is PHP’s scraping features. RSelenium provides R bindings for the Selenium WebDriver API, Available on R programming. Request is Statically collect raw HTML data of a web page. Scrapy is Data collection via proxy and VPN, for complex highly efficient use. The data collected at the end of the job is extracted into CSV format for further analysis to display other statistical packages. As for information search by relying on the search feature using keywords by typing “Fashion Leather”, technically to start collecting and extracting data in this way can be done easily.

How the order of information collection leather craft data using web scraping techniques by flowchart of web scraping process is presented in Figure 2.

2.1. Data Collection

In doing data collection in we call the web scraping technique, the process is done semi-structured from a web. The data retrieval process is done automatically, by writing coding scripts with Python programming language, using Anaconda Navigator and Beautiful Soap library and data miner tools by relying on Chrome Driver to be able to visit marketplace websites and data search processes. The process of finding this data is done by typing a keyword to go the web page that matches the keyword, then the web scraping process is carried out with the available libraries and copying into an excel or CSV file [23].
2.2. Data Classification

Before the web scraping process is carried out, we make modifications that need to be done to be able to present data in the form of statistical calculations, examples of the need to modify this data are, the type of skin collected, so that it looks how synthetic skin, original leather, cowhide, canvas leather, looks dominant from the use of other skins from animal materials. All the lists on color, where we found a variety of colors, and the most dominant (basic color) were brown and black. We did this gender differentiation process because we wanted to see how this leather handicraft product wanted to be targeted so that we got male results as the target sales of this handicraft product.

2.3. Web Scraping on The Website

Web scraping is the process of collecting information from URLs, Meta tag information and web page titles, etc [19]. Availability of all information in a set of web pages, navigation and parsing on the entire web page of the website for one command [17]. The work in the web scraping process is generally by first identifying HTML tags on the site on the web page, as well as designing the model structure and all components relevant to the data by retrieving data on the target website detail page so as to produce a data source [24].

Pseudocode genetic web scraping:

**Input:** set of URLs to scrape, \( u \)

**Output:** set of scraped data, \( d \)

1: import the scraping library
2: initialize the scraper
3: for each URL, \( u_i \) in \( u \) do
4: for each pattern, \( p_i \) in \( p \) do
5: use scraper to download the website content from \( u \)
6: extract the pattern \( p_i \) from that content
7: add extracted data to \( d \)
8: end for
9: end for

The retrieval of this data is product name, price and link (for modification), web scraping in the package library is formed in a python file .ipynb extension, Python library used for collection applications that have sophisticated packages in data collection [25]. In the final stage, extract the data into CSV file format for further analysis with other statistical packages. The web scraping process is shown as shown in Figure 3.
The result of web scraping still contains parts of the HTML document that are not needed, such as the \, \, and \ signs. So, it is done with text processing for this case text processing that has generated a model and saved as a CSV file extension, data filtering that has been collected as many as 3552 text-based products and managed with Microsoft Excel version 16.0 (2016). In the use of filtering data is then processed by conducting analysis. In order to get a descriptive description and depiction of analysis that is easy to understand and understand [26].

3. Results and Discussion

This study presents an overview of statistical data and graphs that are relevant to the results on the collection of leather craft data. From collecting data obtained as many as 3,552 products, we analyzed our work with the acquisition of statistics presented by observing the characteristics of belts (16%), cases (1%), wallets (75%), bracelets (2%), jackets (1%), key hangers (0.2%), sandals (0.2%), gloves (0.2%), shoes (1%), sofas (0.2%), sheet leather stickers (0.2%), watch straps (2%), bags (2%), caps (0.2%) from the acquisition of collection data that has been analyzed, and presentation with table graphs are shown in Figure 4 below.

![Figure 4. Number of Types of Leather Handicrafts in The Marketplace](image)

The presentation of statistics for the most commonly used leather ingredients is seen in Figure 5, the products in the data above are divided by various kinds of leather materials used to obtain products, information on these leather materials we get from skin types: Leather, synthetic (polyurethane polymer) which is generally used for belts, cases, wallets, bracelets, jackets, key chains, sandals, gloves, shoes, sofas, sheet leather stickers, watch straps, bags, hats and is an imported product, Leather canvas is widely used in wallets because the material is easy to shape, genuine leather because it is durable, cowhide because the material is durable and produced in an area, crocodile skin, bison skin, lizard skin, leopard skin, stingray skin, sheep skin, woven leather.

<table>
<thead>
<tr>
<th>BASIC MATERIAL LEATHER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Python snake Leather</td>
<td>1</td>
</tr>
<tr>
<td>Genuine Crocodile Leather</td>
<td>4</td>
</tr>
<tr>
<td>Bison Leather</td>
<td>18</td>
</tr>
<tr>
<td>Lizard Leather</td>
<td>3</td>
</tr>
<tr>
<td>Sheep Leather</td>
<td>5</td>
</tr>
</tbody>
</table>
| Leather Original Cow-hide | 1053
| Leopard Leather        | 2   |
| Stingray Leather       | 1   |
| woven leather          | 3   |
| Synthetic leather      | 1732|
| Canvas Leather         | 710  |
The type of use is usually based on the gender of the wearer. The proportion in wallets, belts, gloves this category is widely targeted and dominant in male users, while in female users is dominant in the production of bags, sandals, dresses, and in both men and women are dominant for case products, key chains, watch straps, bracelets, sofas, sheet stickers, making statistics are explained in the following figure.

User type domination data based on gender is presented in Figure 6.

To estimate color frequency are done in various ways in describing a leather craft product. The color categories correspond to the main colors that are widely used brown and black dominants and can be seen as shown in the following Figure 7:

4. Conclusion

From the web scraping process can illustrate the data collection there is a text-based field that describes 3553 types of leather crafts on large marketplace websites, based on data sets using the keyword "leather fashion" which shows the procedural constraints of obtaining data sources concluded in this study. In the results we do the dominance of the use of many materials using synthetic leather, canvas leather, real cow leather in leather handicraft products, on the dominant color frequency in brown and black because it is a natural color that can be applied to be made into products according to the analysis we did, and in users of leather handicraft products this is dominated for the male sex, although women are also many products and between the two, and the type of craft is dominant for purse products, belts, and from the statistics shown the use of materials that are predominantly synthetic materials can be applied to all types of products: Leather, synthetic (polyurethane polymer) which are generally used for belts, cases, wallets, bracelets, jackets, key hangers, sandals, gloves, shoes, sofa, sheet leather sticker, clock strap, bag and hat.
5. Authors' Declaration

Authors’ contributions and responsibilities - The authors made substantial contributions to the conception and design of the study. The authors took responsibility for data analysis, interpretation, and discussion of results. The authors read and approved the final manuscript.

Funding - No funding information from the authors.

Availability of data and materials - All data are available from the authors.

Competing interests - The authors declare no competing interests.

Additional information - No additional information from the authors.

6. References


