E-Weaver: Sustainable Clothing Aggregation and Recommendation System

Matthew Merenich[1], Colin Murphy[2], Dayun Piao[3], Zifeng Wang[4]

mgm344@drexel.edu[1], cjm486@drexel.edu[2], dp636@drexel.edu[3], zw438@drexel.edu[4]

Drexel University: DSCI-591 Capstone I

Abstract

The fashion industry has been grappling with the demands of consumers to shift towards a more "sustainable" manufacturing process and offer consumers better options to meet their moral beliefs. Although many tools are used by industry to measure clothing sustainability, few brands consider the entire "cradle-to-grave" life-cycle of textiles to better their product and practices; more established household brands experience difficulty replacing timely, cost-effective methods for "eco-friendly" ones. Thus we propose a sustainable clothing recommendation system called *E-Weaver*, which will incorporate state-of-the-art image processing methods for feature extraction and a novel sustainability metric to assist consumers with purchasing more environmentally-conscious products. This paper will focus on the data collection framework for scraping articles of clothing from several different brands to be used in the recommendation system. Additionally, the development of an item-brand level sustainability metric is described. The overall *E-Weaver* framework is outlined and the subsequent applications of the work in this paper are proposed.

1 Introduction

A recommendation system, or a recommender system, is a subclass of information filtering system that seeks to recommend the most suitable items (products or services) to certain users by predicting the users' interest or preference on such type of items (Bobadilla et al. 2013). Over the past two decades, more and more recommendation systems have been designed and implemented for real-world applications, which might be divided into eight main domains: e-government, e-business, e-commerce/e-shopping, e-library, e-learning, e-tourism, e-resource services and e-group activities (J. et al. 2015) .

As one of the most vigorously developing sections, e-commerce, particularly fashion recommendation systems (FRS) are designed to provide personalized recommendations and respond quickly to the consumer's choices, and have therefore contributed significantly to the expansion of e-commerce sales (Chakraborty, Hoque, and Surid 2020). For FRS, useful techniques should include item image parsing, clothing and body shape identification and fashion attribute recognition (Chakraborty et al. 2021). In particular, fashion retailers have shown growing interest in adopting the image-based FRS, which utilize machine learning algorithms and models to effectively detect or recognize scenarios and clothing features (Ullah, Zhang, and Khan 2020).

Recommendations systems rarely consider a user's preference for ethical consumerism, such as sustainability. The concept of "sustainability" can be addressed in several different contexts and scales - from the totality of Earth's carrying capacity over centuries (maximum ability to sustain and support life) to the individual lives/lifestyles of the average human within their local environment. The overarching theme being that the strategic use of current resources to support present needs without affecting the needs of future generations. The immense size and complexity of our world's ecosystem requires that solutions be sub-divided into specialized domains to support the overall effort, of which can be found in every part of society.

This paradigm can be extended to the domain of retail clothing (i.e. fashion, design and textiles). The fashion industry itself is estimated to be responsible for 8-10% of the global CO_2 emissions (Quantis 2018). Furthermore, the recent phenomenon of "fast-fashion" has lead many consumer to become much more conscious of their purchasing habits - opting for items that meet their moral and or ethical beliefs. In response, many brands have adopted various types of "sustainability" campaigns to address consumer concerns; however, the validity of such claims can be difficult to verify as

there are no regulatory bodies to do so.

To address this gap in retailer-consumer relationship/trust, we propose the sustainability-focused recommendation system: *E-Weaver*. *E-Weaver* will help consumers find more sustainable clothing that also meets the aesthetics of their original preferred item. We have developed data collection pipeline to aggregate clothing-article data (e.g. image, color, cost, etc.) from various brands and an index to compare items according to different sustainability criteria. As an initial proof-of-concept for clothing items in general, our system will be limited to recommending T-Shirts.

2 Related Work

2.1 Recommendation Systems

Recommendation system emerged as an independent research area in the mid-1990s when researchers started to focus on the recommendation problems that explicitly rely on the rating structure (Adomavicius and Tuzhilin 2005). During the two decades of developments, the recommendation techniques kept on evolving from commonly used collaborative filtering (Schafer et al. 2007), content-based(Pazzani and Billsus 2007) and knowledge-based techniques (Burke 2000), to recently proposed advanced approaches such as social network-based recommendation systems (He and Chu 2010), fuzzy recommendation systems (Zhang et al. 2013), context awareness-based recommendation systems (Adomavicius and Tuzhilin 2011) and group recommendation systems (Masthoff 2011).

2.2 Measuring Garment Sustainability

Over the past decade, growing criticism of the fashion industry's lack of consideration for environmental and social impacts due to their practices has led researchers and industry leaders to quantitatively evaluate the apparel life-cycle. Many efforts have focused on the development of empirical and generalizable metrics, with the most widely used metric being the Higg Material Sustainability Index (MSI) developed by the Sustainable Apparel Coalition (SAC 2021). The Higg MSI is a "cradle-to-gate" scoring assessment, which considers the impacts from extraction or production of raw material, yarn formation methods, textile formation methods and colouration for each fabric type (Radhakrishnan 2015).

As a result of its simplicity to compared materials using a single distilled score has led to Higg MSI becoming the industry standard tool; however, many researchers have criticized the Higg MSI as an insufficient life-cycle analysis (LCA) tool and a misleading measurement of sustainability (Laitala, Klepp, and Henry 2018). This primarily attributed to the fact that the Higg MSI does not take into account the entire "cradle-to-grave" life-cycle, and thus, the lifespan use and disposal of the garment are not considered. The inclusion of lifespan use or "duration of service" analysis – that is, amount of use and energy consumed by washing/drying – has been shown to alter the overall impact of a garment (Watson and Wiedemann 2019). The Waste and Resources Action Programme (WRAP) found that even a three month extension of use per item would lead to a 5–10% reduction in each of the carbon, water and waste footprints (WRAP 2012).

Measuring the lifespan use of a garment can be difficult due to many impact factors beyond simply the length of time an item is kept (*for example, an item worn once a year over 5 years would have have far less of a functional unit than an item worn everyday for a year*). Additionally, the frequency of washing the item and the associated temperature(s), time(s) plays an import role when measuring the overall environmental impact of a garment. To generalize the habits of use, several studies have been conducted concerning lifespan for a garment based on material and category. By combining various different demographic surveys, user habits across various cultures is also considered (Nielsen 2012; Laitala, Klepp, and Henry 2017). This evidence also serves as additional motivation for our proposed system – not only by introducing more sustainable alternatives for clothing, but also recommending items that are aesthetically similar to the intended item, of which may increase the likelihood of the customer retaining it for longer.

3 Methodology

3.1 Data Source

Our dataset is a combination of several sub-sets scraped from various different clothing brand websites. From these sites we obtained meta-data for individual T-shirt products and associated images. Due to the differences in website design and the companies privacy policy, the common meta-data features of the final dataset were limited to the follow: color, gender, price, material, brand, and image. All auxiliary meta-data was still collected for possible use later in the development of the project.

The web scraping code was developed using the BeautifulSoup and Selenium libraries. The code requests information from each E-commerce website and scrapes data in multiple nested loops. After data is collected from each website, they are cleaned and converted to pandas dataframe, then concatenated to form the final dataset. Code can be found in the project GitHub Repo.

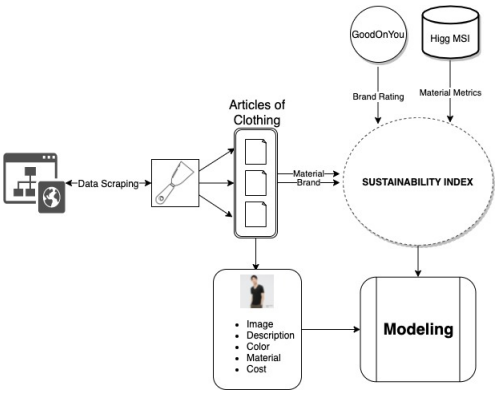

Figure 1: Schematic diagram of the data collection pipeline.

Data was obtained from both large brand name companies that are often associated with fast-fashion, and smaller companies that have a greater sustainability focus. These will serve as the training data for the eventual model, therefore, we chose several different data sources to offer a better representation of varying sustainability levels. The list of brands collected and approximate number of items are listed in Table 1.

Additionally, data from two other sites was obtained in order to build our sustainability metric for ranking items – the aggregate brand rating site GoodOnYou.eco and the industry standard MSI from higg.org. The data from these two sites will be fused together to create a metric that takes into account both the item specific metric (higg.org, Higg MSI) and an entire brand metric (GoodOnYou.eco).

Table 1: Dataset sources

Brand Name	**# of Items Collected**
Uniqlo	406
Tentree	93
The Good Tee	15
H&M	377
Zara	536
Dedicated	377
Outland Denim	22
The Standard Stitch	21
Fair Indigo	87
Zerum	100
Living Craft	73
Total	2107

3.2 Sustainability Metrics

As previously mentioned, the primary motivation of the *E-Weaver* system is to create a sustainability-conscious recommendation system. To achieve this we have developed a comprehensive framework to score articles by several criteria, weighted by relative importance. The framework brings together established metrics in both academic and industry settings as mentioned in review – combining sources to create a pseudo-cradle-to-grave LCA. This is the metric that the recommendation system will be trying to maximize along with aesthetic similarity.

The framework is divided into two main focuses: Item and Brand.

- **Item** - a fusion of the Higg MSI and academic surveys of consumer habits and uses, providing a score based on material blend and garment type. This is intended to be a more granular metric of sustainability.

  The scaled and weighted sustainability indicators of the per-item scores and their relative material ratios are summed up to form the overall score (Eq.1).

$$\text{Item Sustainability} = \sum_i \sum_j w_i \frac{CR_{ij}}{NR_i}(n_j) \quad (1)$$

  Where the characterized result (CR) of indicator, i, for material, j, is divided by the normalization reference (NR) of the corresponding indicator, i. The Higg MSI normalizes according to the industry's annual impact. The division is later multiplied by the weight factor (w) of indicator i – for our current system, this will simply be set to "1" for all factors. And to account for items consisting of several different materials, the material score is scaled according to the relative content (n) of the material, j, in the item.

- **Brand** - GoodOnYou.eco aggregated rating of the overall brand. This is a more generalized metric of the brand as a whole – e.g. given two garments of the same material can be separated by brand rating.

The methodology we have developed should not be viewed as a perfect ranking for any sustainability purpose; however, the flaws and shortcomings of this are a result of the typical trade offs often faced in comparative analyses – assumptions must be made in order to aggregate disparate variables from across the entire life-cycle. Overall, there is no single material or brand option that performs best in all sustainability categories and so any ranking choice results in a compromise.

4 Exploratory Data Analysis, Data Cleaning

4.1 Color

For the color attribute, the top five most popular colors are: White, Black, Blue, Green and Gray. There are 362 unique colors in total. For data simplicity, similar colors are grouped together (e.g. meteorite gray and gray are all considered as gray).

Note: The color scrapped from the website is used for baseline description of dataset, not final modeling input. Instead, the color will ultimately be derived from the image.

4.2 Gender

The brands in our dataset manufacturer both men's and women's T-shirts, with many individual items being specifically labeled for the respective gender. However, it was also discovered that some items were not labeled at all or were labeled as "kids". In Figure 2, the items counts of each label are represented and the relative share of the labels can be clearly seen.

Due to the sparse instances of "kids" items and the focus of our E-Weaver system, these items will not be included in the final dataset. Additionally, the limited instances of "unisex" are to be relabeled as both "men" and "women" categories.

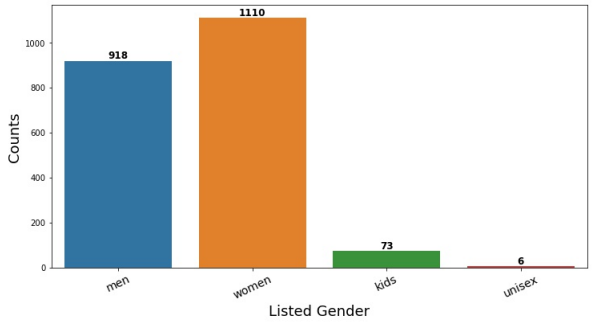

Figure 2: Bar plot showing counts of genders listed for individual items.

4.3 Price

For the price attribute, a KDE plot was made to show the distribution by brand (Figure 3). T-shirts priced at $35 have the most counts, and mostly between $19 to $40 range. Our KDE plots shows an array of price ranges for the shirts collected in the acquisition phase, which should translate to robust recommendations from our model next semester. The brand distribution is as expected: specialty brands such as The Good Tee have the most expensive options, while fast fashion brands (Uniqlo, H&M, Zara) are towards the left half of the price distribution.

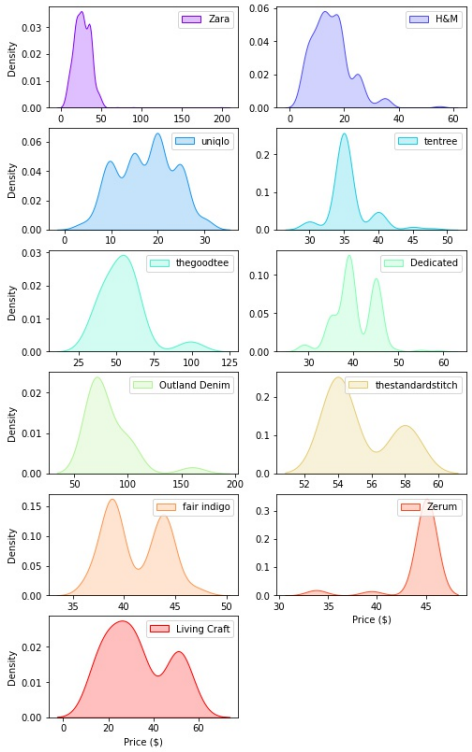

Figure 3: KDE plots of brand price.

Prices for all brands were also plotted (Figure 4). The most common t-shirt prices were between $33 and $38. Our distribution scales left, signifying most options are within a range we expect a customer to be comfortable making a purchase. Most prices fall within $8 and $47 with certain shirts offered by luxury brands we scraped being as expensive as $199. As with our brand analysis, we can conclude our dataset contains a wide breadth of price options up for recommendation.

4.4 Material

For the material attributes, each material is extracted from the item web page and represented as a percentage of the overall composition. A pie chart was made to show the distribution of each material among all T-shirts (Figure 5).

As expected, cotton makes up the majority of the materials in all the T-shirts around 78%. The second most common material is polyester at about 11%, followed by viscose, polyamide, and arylic each accounting for less than 2%. The remaining materials, listed as 'others' in Figure 5 represent the the materials that make up less than 1% of the

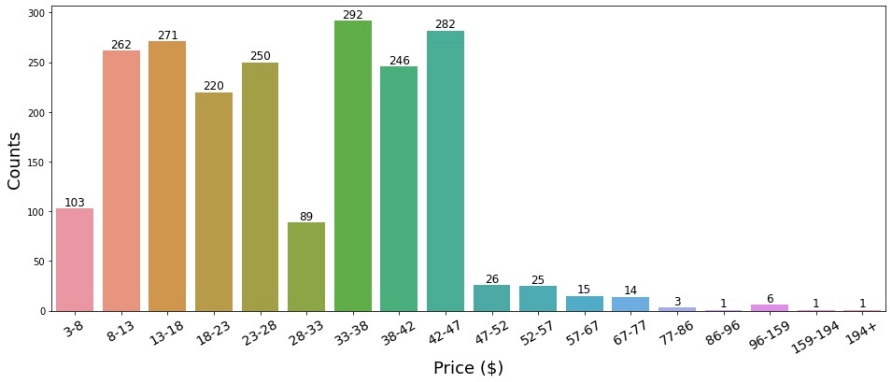

Figure 4: Histogram plot of price bins.

dataset.

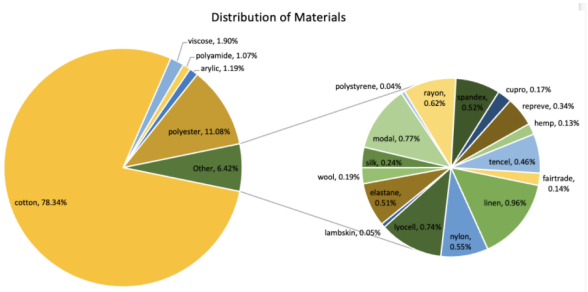

Figure 5: Pie chart of materials distribution.

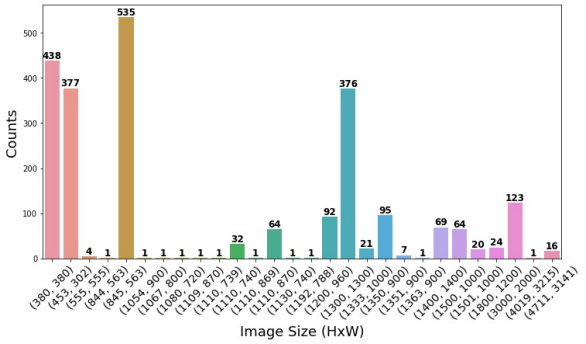

Figure 6: Bar plot showing counts of unique image sizes.

The distribution materials was not found to be uniform among brands – large companies like Zara consisted almost entirely of cotton while other smaller brands tended to include several uncommon materials and blends. Thus, the distribution materials that are less than 1% of the dataset reflect the nuance of smaller brands that create a more diverse dataset for material composition. Among these minor materials, the top three materials are: Linen(0.96%), Modal(0.77%), Lyocell(0.74%).

4.5 Images

As expected, the image format and dimensions differed substantially by brand. The quantity of each unique image size are shown in Figure 6.

It was discovered that some new items listed may be listed for sale by a brand before an actual photo of the item can be posted. As a result a grayscale placeholder image is posted instead, of which the scrapper had still collected. These instances could be identified as they were formatted differently from all of the other images in the brand, and typically did not contain a third dimension for RGB channels. These item instances were discarded from the dataset due to loss of information that would make it impossible for image feature extraction.

In order to have consistent performance in the modeling stage, it is desired to have less variance in the image size (and decrease computational complexity), thus a 300x300x3 reformatting of all images was performed.

5 Pre-Processing

The raw data scraped from different websites/brands contained different data structures or data types, and also varied in their completeness and consistency. Therefore, careful quality checks and pre-processing are essential for further analysis. Current pre-processing has focused on cleaning inconsistent signs, generalization of features, parsing key information and dealing with missing values.

- **Price:** Clean the '$' unit or "EURO", or sometimes 'USD' by applying specific cleaning function on price column and store values as floating point number.

- **Materials:** Material meta-data is collected along with other text. Used Regular Expression to find patterns within the text and extract the precise material information.
- **Gender:** Many sites do not designate the item's gender within the scraped data, and required indirect collection of this data. For example, The Good Tee and TenTree contained gender information within the url of the item (each gender had its own landing page customers could search through). URL and description splitting and extraction was needed to correctly label all items. Even more, The Good Tee had "unisex" items that could be worn by all genders - the same process was used to apply this attribute.
- **Items:** Most brands we chose have a designated landing page for t-shirts making for simple scraping. However, niche companies with a smaller clothing assortment may have just gender landing pages with all products available for the customer to purchase (a more robust website is not yet justified). In these instances, there are more than just t-shirts (The Good Tee contained dresses, accessories, and other general items) which required item name extraction from either the description or the item's name.
- **Missing values:** Drop sparse columns where a majority of scrapped items did not have values for, as well as instances missing the critical attributes for our system: material, cost, gender, brand, and image.

6 E-Weaver System Framework

The E-weaver recommendation system's goal is to provide a ranking of similar articles submitted by the user (both aesthetically and monetarily), yet also more sustainable articles.

The system will consist of several different "stages" and will utilize both the clothing item dataset and the sustainability metric dataset. A diagram of the entire system framework is shown in Figure 7. The system input will request the item image, material, price, gender, and brand. With the exception of the image, all inputs will feed directly into the clustering stage.

6.1 Image Feature Extraction

In many classical item clustering methods, unsupervised learning is applied to the images themselves to infer their similarities. However, this can be computationally intensive and it can be difficult to infer what precise design details make the items similar – providing only a *course* object-level classification.

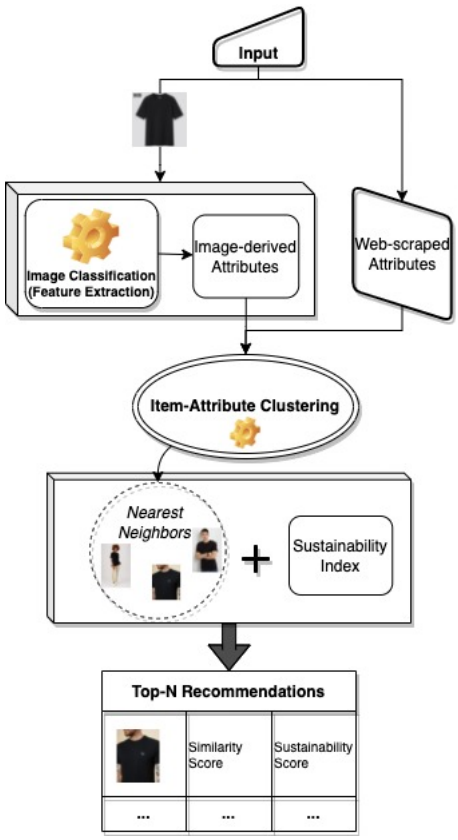

Figure 7: Schematic diagram of entire system framework – image classification and clustering.

Our E-Weaver system is intended to be a customer-facing interface, thus providing details of the model's clustering choices is important. Therefore, we propose applying an intermediate feature-extraction step prior to clustering for images – providing *fine-grain* attributes to describe the design and aesthetics of the item. Due to the relatively small size of our dataset, the use of transfer learning methods will be used to accomplish this feature extraction task. Several potential fashion domain-specific methods have been proposed to accomplish this task (Guo et al. 2019; Jia et al. 2020). A few examples of the primary attributes to be extracted from the image will be color, neckline, pattern, sleeve, and style.

6.2 Item-Attribute Clustering

With the image-derived attributes, along with the material, price, gender, and brand user inputs, the similarity of items can be calculated by some metric. From these, we can create *neighborhoods* of the nearest N neighbors (where N is some integer, e.g. 3, 5, or 10) and return them as our recommenda-

tions to the user.

6.3 Sustainability Ranking

Given the recommendations, we now have a typical clothing recommendation system; however, the context of sustainability must still be incorporated to our system. Here we can now use our brand and material-level sustainability metric developed in Section 3.2. The nearest N neighbors can then be ranked by the sustainability metric and the similarity score (or similar clothing attributes) can also be listed to assist the user.

7 Conclusions

The focus of this work has been project scoping and data acquisition with the following accomplishments:

1. The scraping of data from several online clothing brand sources and combining them into a uniform dataset.
2. Performing EDA on collected attributes including several data cleanling operations to remove web-scraping artifacts.
3. The development of a specialized sustainability metric for ranking to incorporate both item and brand level details for a nuanced comparison of items of the same material composition.
4. The standardization / pre-processing of each dataset for further use in modeling as a single data-pipeline.

7.1 Further Work

Throughout this process, the conceptualization of the E-Weaver as a clothing recommendation platform has also continued – the development of which will be the focus of our work moving forward, divided into the following areas:

- **Continuation of data collection**: new brand sources; integrating data quality checks.
- **Fashion image feature extraction**: transfer Learning - use pre-trained models "as-is" or train on dataset with new layer.
- **Clustering**: try several different clustering methods for image-derived dataset to group similar items.
- **Sustainability Ranking**: investigate the proposed sustainability ranking scheme and consider bias for our limited set.
- **User Interaction/Feedback**: perform limited A/B testing or user survey tests to gauge the effectiveness of recommendations; make system alterations accordingly.

References

[Adomavicius and Tuzhilin 2005] Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17:734–749.

[Adomavicius and Tuzhilin 2011] Adomavicius, G., and Tuzhilin, A. 2011. Context-aware recommender systems. In *Recommender Systems Handbook*. Springer, US. 217–253.

[Bobadilla et al. 2013] Bobadilla, J.; Ortega, F.; Hernando, A.; and Gutiérrez, A. 2013. Recommender systems survey. *Knowledge-Based Systems* 46:109–132.

[Burke 2000] Burke, R. 2000. Knowledge-based recommender systems. *Encyclopedia of Library and Information Systems* 69:175–186.

[Chakraborty et al. 2021] Chakraborty, S.; Hoque, M.; Rahman Jeem, N.; Biswas, M.; Bardhan, D.; and Lobaton, E. F. 2021. Recommendation systems, models and methods: A review. *Informatics* informatics8030049.

[Chakraborty, Hoque, and Surid 2020] Chakraborty, S.; Hoque, M.; and Surid, S. 2020. A comprehensive review on image-based style prediction and online fashion recommendation. *Journal of Modern Technology and Engineering* 5:212–233.

[Guo et al. 2019] Guo, S.; Huang, W.; Zhang, X.; Srikhanta, P.; Cui, Y.; Li, Y.; Adam, H.; Scott, M. R.; and Belongie, S. 2019. The imaterialist fashion attribute dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

[He and Chu 2010] He, J., and Chu, W. 2010. A social network-based recommender system (snrs). In *Data Mining for Social Network Data*. Springer, US. 47–74.

[J. et al. 2015] J., L.; D., W.; M., M.; Wang, W.; and Zhang, G. 2015. Recommender system application developments: a survey. *Decision Support Systems* 74:12–32.

[Jia et al. 2020] Jia, M.; Shi, M.; Sirotenko, M.; Cui, Y.; Cardie, C.; Hariharan, B.; Adam, H.; and Belongie, S. 2020. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European conference on computer vision*, 316–332. Springer.

[Laitala, Klepp, and Henry 2017] Laitala, K.; Klepp, I. G.; and Henry, B. 2017. Use phase of wool apparel: a literature review for improving lca. *Product Lifetimes And The Environment*.

[Laitala, Klepp, and Henry 2018] Laitala, K.; Klepp, I. G.; and Henry, B. 2018. Does use matter? comparison of environmental impacts of clothing based on fiber type. *Sustainability* 10(7):2524.

[Masthoff 2011] Masthoff, J. 2011. Group recommender systems: combining individual models. In *Recommender Systems Handbook*. Springer, US. 677–702.

[Nielsen 2012] Nielsen. 2012. Global wardrobe audit—all countries.

[Pazzani and Billsus 2007] Pazzani, M., and Billsus, D. 2007. Content-based recommendation systems. In *The Adaptive Web*. Springer, Berlin Heidelberg. 325–341.

[Quantis 2018] Quantis. 2018. Measuring fashion: Insights from the environmental impact of the global apparel and footwear industries.

[Radhakrishnan 2015] Radhakrishnan, S. 2015. The sustainable apparel coalition and the higg index. In *Roadmap to sustainable textiles and clothing*. Springer. 23–57.

[SAC 2021] 2021. Higg materials sustainability index.

[Schafer et al. 2007] Schafer, J. B.; Frankowski, D.; Herlocker, J.; and Sen, S. 2007. Collaborative filtering recommender systems. In *The Adaptive Web*. Springer, Berlin Heidelberg. 291–324.

[Ullah, Zhang, and Khan 2020] Ullah, F.; Zhang, B.; and Khan, R. U. 2020. Image-based service recommendation system: A jpeg-coefficient rfs approach. *IEEE Access* 8:3308–3318.

[Watson and Wiedemann 2019] Watson, K., and Wiedemann, S. 2019. Review of methodological choices in lca-based textile and apparel rating tools: key issues and recommendations relating to assessment of fabrics made from natural fibre types. *Sustainability* 11(14):3846.

[WRAP 2012] WRAP. 2012. Valuing our clothes: the true cost of how we design, use and dispose of clothing in the uk.

[Zhang et al. 2013] Zhang, Z.; Lin, H.; Liu, K.; Wu, D.; Zhang, D.; and Lu, J. 2013. A hybrid fuzzy-based personalized recommender system for telecom products/services. *Information Sciences* 235:117–129.