Stylatrix: an interactive model-based system for fashion exploration and outfit discovery

Will J. Sun
Harvard University
Cambridge, MA, USA
willsun@post.harvard.edu

Krzysztof Z. Gajos
Harvard University
Cambridge, MA, USA
kgajos@eecs.harvard.edu

ABSTRACT

Like most aesthetic domains, fashion can be characterized by intrinsic and extrinsic characteristics, such as outfit structure, color, pattern, and material as well as sociocultural connotations. However, most fashion recommendation systems incorporate limited domain-specific knowledge, instead relying upon standard item-based collaborative filtering approaches. Though such systems might help someone looking for a specific item of clothing, they do not help individuals envision outfits or discover and conceptualize their style preferences. To enable outfit-centric decision support with fashion, we started by conducting a formative study to investigate how people judge outfit similarity. The results of this study showed that humans consider fashion through a holistic lens, taking into account both intrinsic and extrinsic features. Based on this understanding, we used machine learning to model users' subjective impressions of outfit similarity. Experimental results validated the robustness of our constructed similarity metric, which serves as the foundation of Stylatrix, an interactive, model-based system that enables fashion exploration, style makeovers, and outfit querying. User evaluation demonstrated that Stylatrix effectively captures key components of how humans perceive style, supporting meaningful interactions with fashion concepts.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

fashion, style, decision support, exploration, discovery

1. INTRODUCTION

Fashion plays a major role in defining an individual's identity. For better or for worse, individual style reflects the milieu in which one lives and the preferences that one holds. For some, fashion is a world of high art, while for others, it means perusing photos of the dresses showcased on the Oscar red carpet. Regardless, the experience of deciding what to wear every morning is near-universal.

Surprisingly, the technology available to consumers to assist and enable fashion-related choices remains at a basic level. Online communities and fashion websites — such as Pinterest, Stylitics, Polyvore, Chictopia, Chicisimo, and Clothia — help individuals examine new trends, participate in fashion-related dialogue, and create curated collections of appealing clothes and outfits. These websites are popular amongst fashion experts and enthusiasts, but they rely upon laborious, non-automated, human-powered means to deliver relevant content. On the other hand, most apparel brands and clothing retailers use item-based collaborative filtering to deliver relevant recommendations to a general audience of online shoppers. While this automated approach is useful for context-specific shopping decisions (i.e. a consumer looking for a black lace top), it does not directly help consumers determine the items necessary for a desired outfit. Clothing retailers assist with style inspiration to the extent that they manually curate a handful of outfit ideas on the front page of their website.

We consider *outfits* as the fundamental unit of fashion. Outfits convey a scope of aesthetic meaning, just as a song does for music, or a canvas for painting. A reductionist, item-based approach therefore cannot adequately model style. Moreover, individuals currently lack a means to systematically explore their style preferences and discover interesting outfit ideas. Though outfit-centric collaborative filtering systems may reveal useful latent structure, this structure is not directly accessible to users and fails to enable exploration and discovery interactions important to conceptualizing fashion goals and needs. We sought to develop a system that would complement existing fashion recommenders by enabling meaningful and efficient exploration of diverse yet relevant outfits.

To enable this, we needed a computational model of outfit similarity. Outfits possess intrinsic characteristics, like outfit structure (i.e., articles of clothing), color, pattern, and material, as well as extrinsic ones, like the socio-cultural connotations or occasion in which it is worn. As Davis [6] describes, fashion suffers from "low semanticity." Neither the set of descriptive characteristics, whether intrinsic or extrinsic, nor the set of descriptive styles is well-defined. Therefore, we consider two approaches to constructing a model of style: the prescriptive and the emergent [1]. Whereas a prescriptive strategy involves systematically identifying relevant characteristics and categories of a domain, the emergent extracts higher-order descriptors from free-form human input. In order to determine whether a generalizable model of style was even possible, we conducted a formative study to explore how people examine outfits and make style judgments. Results from this study indicated that individuals share a reasonably universal notion of style similarity and use a holistic view of outfits, taking into account both intrinsic and extrinsic features, to make similarity judgments.

For our analysis, we gathered a set of 526 female outfits from Chictopia.com. We then represented these outfits with both emergent and prescriptive representations, using topic modeling and a structured feature vector. To construct a style similarity metric, we turned to comparative queries to elicit human similarity intuitions. Drawing from previous work on distance metric learning from relative comparison queries [13], we created a metric that is robust at predicting similarity relationships between outfits.

With this similarity metric, we constructed Stylatrix, an outfit-centric decision support system that is sensitive to style similarities. Our system enables two primary tasks:

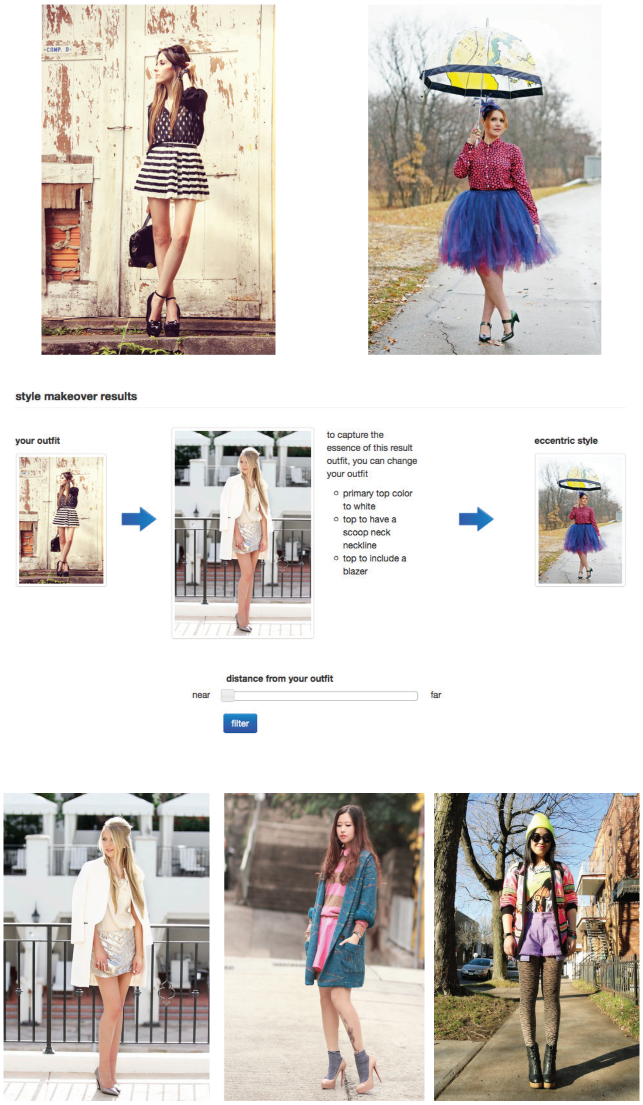
1. Outfit discovery: given an outfit, Stylatrix returns outfits that are stylistically similar.
2. Style makeover: given an outfit as a starting point (perhaps one similar to an individual's current style) and a desired style goal, Stylatrix provides outfit ideas intermediate to both points (Figure 1). Intermediate outfits serve as points of inspiration for individuals striving to envision a new look for themselves.

In this paper, we demonstrate that human similarity intuitions are generalizable and that perceptions of style do indeed involve both intrinsic and extrinsic characteristics of outfits. We present a style-based outfit similarity metric that uses both emergent and prescriptive outfit representations, and show that this metric is effective at modeling style similarity. Using this metric, we constructed Stylatrix, a system that enables novel, style-based interactions with outfits. User evaluation results indicate that the system supports the interactions desired.

2. RELATED WORK

The literature on computation in the fashion domain is limited, but among relevant papers, low-level visual features are frequently used to describe outfits. Iwata et al. [10] perform region detection on top and bottom pieces of an outfit and extract visual features like color, texture, and local descriptors (e.g. SIFT). They then construct a probabilistic topic model that examines these features, which provides outfit completion recommendations based on the top or bottom with the closest topic proportions. Yamaguchi et al. [16] use computer vision for pose detection and outfit parsing, with their algorithm recognizing different pieces of an outfit. They employ their algorithm to build a retrieval system for visually similar outfits. Yuan [17] also uses visual similarity to determine whether articles of clothing possess similar colors or patterns, a task especially useful for visually-impaired individuals when deciding what to wear.

While computer vision approaches can construct intrinsic features in an automated manner, these features are cur-



to capture the essence of this result outfit, you can change your outfit

- primary top color to white
- top to have a scoop neck neckline
- top to include a blazer

Figure 1: Style makeover. A query outfit and outfit representative of the "eccentric style" are shown above. The results interface and three sample results are displayed below. The interface allows users to navigate through the space of outfits between their starting outfit and their goal style. Additionally, the interface presents salient features that capture what makes the result outfits different from the query outfit; individuals can use this information when considering what clothing items are necessary to begin their makeover. Notice that from left to right, the result outfits become noticeably more eccentric and less similar to the query outfit. The first two results possess distinctive structural similarities to the query outfit (all three include skirts and heels). On the other hand, the first result outfit incorporates shiny silver items, while the second and third are even more eccentric in their use of color and pattern.

rently fairly limited and do not provide the necessary specificity for learning style-based models. Fashion's inherent subjectivity limits the generalizability of computer-generated models, while its lack of structure poses issues for human-generated models. Yet machine learning approaches can help discover underlying patterns and structure not immediately perceptible to humans, and humans can help identify subjective distinctions. Human-machine models attempt to combine the benefits of both human and computational understanding while mitigating their shortcomings; our work builds off of this intuition.

Perhaps the most fully-developed fashion recommendation system is "What Am I Gonna Wear", a scenario-oriented system produced by Shen et al. [14]. Articles of clothing are annotated with a six-tuple to encode various dimensions of style (i.e. luxurious, formal, funky, elegant, trendy, and sporty), while a database of English sentences is used to broaden the system's semantic understanding of occasions. Users can then specify their personal preferences and the occasion they seek an outfit recommendation for. Though the system offers limited interactions, it nonetheless introduces manually-encoded style vectors and text descriptors as potential representations of clothing items and extrinsic concepts in the fashion domain.

3. SCOPE

Our work focused on Western female fashion, which is remarkable in its dynamism and variety. Compared to male fashion, females have more structurally unique tops, bottoms, and accessories to array themselves with. We also chose to consider the outfit as the fundamental unit of analysis. Other researchers have explored individual articles of clothing; for example, Iwata et al. built their outfit completion system around the top-bottom split [10]. However, reducing an analysis down to individual articles of clothing, or even portions of an outfit, removes from consideration the interplay between items in an outfit. Finally, we decided to work exclusively with street style outfits. Such outfits are fashionable, everyday outfits captured by photographers. Though several websites feature large databases of user-uploaded street style outfits, we chose Chictopia.com as the source for female outfits. This choice mirrors that of Yamaguchi et al. [16], who also used Chictopia to collect images for their computer vision analysis. We scraped over 1,000 outfits and associated metadata from the Chictopia website. All scraped outfits were posted to the website between September to November 2012, thereby focusing the subset to primarily fall-season outfits.

4. OUTFIT SIMILARITY INTUITIONS

To construct an accurate computational model of outfit similarity, we began by conducting a formative user study to investigate how people assess outfit similarities. We wished to confirm whether a notion of outfit similarity was generalizable; if not, then it would be impossible to generate a useful computational model of style. In addition, we also explored how people examine outfits and make similarity judgments. Understanding how individuals consider outfit similarity, whether through an intrinsic, extrinsic, or holistic view, would inform how we construct our model of style similarity.

Participants, Procedures, and Tasks. We recruited a total of 49 individuals (29 female, 20 male) on Amazon Mechanical Turk. Each participant rated the degree of similarity between pairs of outfits, on a scale of 0-100, with low scores indicating dissimilarity. 12 pairs of outfits were randomly selected, and each participant rated the same set of 12 outfits, in the same order. We randomly assigned participants to four experimental groups: control, holistic, intrinsic, or extrinsic. The task design was identical for all groups with the exception of the instructions:

- **Control**: "Rate and describe the similarity of the outfit pair."
- **Holistic**: control text + "Please consider similarity from a holistic standpoint, taking into account both *intrinsic* (e.g. outfit structure) and *extrinsic* (e.g. social and cultural connotations) factors."
- **Intrinsic**: control text + "Please consider similarity by exclusively examining *intrinsic* characteristics of the outfit (e.g. outfit structure, color, material)."
- **Extrinsic**: control text + "Please consider similarity by exclusively examining *extrinsic* characteristics of the outfit (e.g. social, cultural, emotional associations)."

By comparing the judgments provided in the control condition and the remaining three conditions, we can glean an insight into what criteria people naturally take into account when judging similarity between outfits. This study also allows us to evaluate the universality of outfit similarity judgments.

Results. We measured inter-rater agreement within each condition using Krippendorff's alpha [8], a metric that depends on the reproducibility of results. Krippendorff's alpha considers all pairs of ratings between raters for a given object (e.g., an outfit pair), with $\alpha = 1$ signifying perfect agreement and $\alpha = 0$ signifying agreement no greater than random chance. We observed moderate agreement between individuals in all experimental conditions, suggesting evidence of a universal concept of style (Table 1). Furthermore, we noted that there was substantially higher levels of inter-rater agreement in the control and holistic conditions than in either extrinsic or intrinsic conditions.

We also performed a correlation analysis comparing mean ratings of each pair in the control condition to each of the three remaining conditions. Participants' natural similarity intuitions were substantially more closely correlated with the ratings provided in the holistic condition ($r^2 = .92$) than with either intrinsic ($r^2 = .86$) or extrinsic ($r^2 = .87$) conditions. These results are summarized in Table 1. Additionally, we computed the correlation between ratings from the intrinsic and the extrinsic condition: $r^2 = .91$.

Discussion. Our results demonstrate that substantial inter-rater agreement exists on outfit similarity judgments. In the control condition, where individuals made outfit similarity judgments without any special instructions, we observed a Krippendorff's alpha value of .53. This value is considered to reflect moderate agreement, indicating that there is a substantial universal component to how people reason about outfit similarities. However, this value also suggests that individual differences in similarity judgments exist.

	N	Mean	SD	SE	alpha	r^2
Control	12	31.2	25.9	7.49	0.53	n/a
Holistic	14	33.8	27.6	7.37	0.52	0.92
Intrinsic	11	31.5	28.8	8.67	0.46	0.86
Extrinsic	12	40.4	30.4	8.77	0.45	0.87

Table 1: Average similarity ratings, correlation to average control group ratings, and Krippendorff's alpha for each experimental condition

We also show that individuals tend toward a holistic view of outfits when considering outfit similarity. That is, people consider both the similarities in basic appearance as well as the socio-cultural connotations of the stylistic choices expressed in each outfit. Although there is a strong correlation between similarity judgments based on intrinsic features alone and judgments based solely on extrinsic features, participants who were instructed to consider both sets of features produced judgments most similar to those made by participants who received no special instructions.

Based on the inter-rater agreement results, the goal of modeling human judgments of outfit similarity appears worthwhile. Our results also suggest that a holistic model, which captures both intrinsic and extrinsic properties of the outfits, has the best chance of accurately capturing how people naturally evaluate outfit similarities.

5. REPRESENTING OUTFITS

Results from our formative study confirmed the idea that individuals tend to view outfits in a holistic manner. We therefore considered two different outfit representations that would capture the holistic characteristics of outfits: the emergent and the prescriptive. Through the use of topic modeling, we constructed an emergent representation of outfits by identifying high-level features from text descriptions of outfits. We also constructed a prescriptive representation, where relevant aspects of an outfit's appearance were described in a systematic manner. While an emergent representation is based on how people naturally discuss outfits, the prescriptive representation provides a structured, semantic account of an outfit. These approaches offer different valuable viewpoints of representing outfits.

5.1 Emergent representation

We constructed an emergent outfit representation by using topic modeling, which extracted meaning from two forms of text description: outfit tags and item-based description. We hypothesized that tags would help capture extrinsic characteristics of outfits (e.g. occasion, culture), while longer textual description would focus on salient components of the outfit. With the help of 299 Turkers (216 females, 83 males), we collected tags and descriptions on 526 outfits. Each outfit was only tagged or described once. Given limited resources, breadth was chosen over depth; since the vocabulary that people use to describe fashion is large, we assumed that more outfits would allow more words to enter the vocabulary of analysis and increase word co-occurrences. Including the basic descriptions associated with each outfit on Chictopia, with stopwords removed and tags and descriptions combined, the average document length for each outfit was around 25 words.

LSA (k = 12): 0.370*jeans + -0.273*tights + 0.221*denim + -0.212*scarf + 0.202*skinny + -0.200skirt + -0.181*gray + 0.180*silver + -0.180*sweater + 0.164*gold
LDA (k = 12): 0.053*high + 0.049*long + 0.041*leggings + 0.038*skirt + 0.032*white + 0.030*pumps + 0.029*heels + 0.027*black + 0.023*trendy + 0.020*purse

Table 2: Sample topics constructed through topic modeling

For topic modeling, we considered both latent semantic analysis (LSA) [7] and latent Dirichlet allocation (LDA) [2], which model text descriptions as derived from k pertinent topics. Outfit documents were represented as bags-of-words and transformed using term frequency-inverse document frequency to normalize word counts. We used the Gensim toolkit [12] to perform both LSA and LDA topic modeling. In terms of parameters, LSA requires a selection of k, while LDA requires a selection of k as well as appropriate hyperparameters α and β for the Dirichlet distributions used. We chose to tie our k for LSA to the k determined for LDA for easier comparison.

In order to determine the optimal k, we performed both quantitative and qualitative evaluation of the topics produced. We first measured the lower perplexity bound of a held-out set of outfits [9]. For our hold-out analysis, we examined 100 outfits and trained an LDA model on the remaining 426 outfits using different fixed, symmetric hyperparameters. Based on our hold-out analysis, $\alpha = \frac{50}{k}, \beta = 0.1$ appeared to minimize perplexity. Including an examination of the actual topics generated at different k, we determined $k = 12$ as the number of topics that would both minimize perplexity and maximize coherence of topics. Evident from the keywords that described some of the topics (Table 2), topic modeling captured meaningful simple and higher-level concepts. For example, the LSA topic relates jeans and denim together, while the LDA topic provides a vivid depiction of a trendy outfit: skirt with leggings, high heels, and purse.

5.2 Prescriptive representation

The emergent representation has the benefit of capturing what people consider salient when reasoning about outfits. We complemented that representation with a prescriptive one, which involved a hand-crafted representation of outfits that uses a systematic set of relevant structural attributes. Whereas Shen et al. [14] used custom style-based vectors to construct their recommendation system, we extend this concept to focus on the intrinsic features that describe outfits.

Our prescriptive representation decomposes outfits into a top, a bottom, and accessories. Derived from expert discussion on websites like Chictopia, as well as the text descriptors described in the previous section, we decided to highlight outfit structure, color, pattern, fit, and material. Each of the three sections is characterized by a primary and secondary color, distinctive materials or patterns, and other relevant intrinsic features. The semantics of this representation make data collection efforts easier; rather than the free-form descriptions required for our emergent representation, we can ask an individual to describe an outfit by answering a small set of questions.

Under the hood, we chose to construct this representation

Outfit Bottom	
Primary color	16 colors
Secondary color	16 colors
Material	lace, leather, denim, knit, sheer
Size	fitted, loose
Dress length	ankle, floor
Skirt cut	pleated, flared, straight
Dress type	sundress, sweater dress, empire waist
Structure	short shorts, shorts, miniskirt, skirt, dress, minidress, pants, jeans, leggings

Table 3: Prescriptive representation, bottom of an outfit

as a 199-dimension binary feature vector (Table 3). This vector takes advantage of distributed encodings, which split each value of a categorical variable into individual binary features. Distributed encodings are ideal for categorical variables where the distance between values is uncertain, as is the case for many of our features.

Both authors initially participated in encoding, but after demonstrating a high level of agreement between independent encodings on a subset of outfits, the remainder was completed by the first author.

6. SIMILARITY METRIC

Using our constructed outfit representations, we proceeded to construct a predictive model of style similarity judgments. Specifically, given a representation of the differences between two outfits, we sought to model how humans would judge the similarity of two outfits. A natural approach for training such a model would be to collect human labels in the form of similarity ratings: given two outfits, participants would be asked to rate how similar or different the outfits are on some scale. This data would then be used in standard regression models to learn a model of outfit similarity.

However, there exists compelling evidence that rating queries produce unreliable results for complex judgment tasks and our results from Section 4 corroborate that. Instead, comparison queries, where people are asked to compare two outcomes, have been shown to be more robust in several domains [4]. This is the approach we chose. To elicit human judgments of outfit similarity, we presented individuals with triplet queries. With triplet queries, individuals are shown three outfits, i, j, and k, and are asked to choose whether i is more similar to j or k. Though this leads to a less standard learning problem, Schultz and Joachims [13] have developed a support vector machine that can learn a distance metric from relative similarity triplets. We used this algorithm to construct our style similarity metric.

6.1 Learning from similarity triplets

We collected triplets in a manner similar to that described by Schultz and Joachims [13] and Tamuz et al. [15], where individuals are presented with three outfits and asked the question: "is i more similar to j or k." However, we modified the Schultz-Joachims-Tamuz triplet query to allow for individuals to select *any* similar pair from within the three outfits. In order to reconstruct the comparative structure of triplet similarity queries, we randomly chose one outfit out of the similar pair to serve as the "base," such that base outfit i and similar outfit j are more similar than base outfit i and dissimilar outfit k. Overall, 4058 triplets were collected, and 203 Turkers contributed (121 females, 82 males).

Schultz and Joachims [13] use similarity triplets to collect relative, qualitative feedback and learn a parameterized Mahalanobis distance metric with a SVM. Using a custom kernel function, the Schultz-Joachims SVM finds the maximum-margin separator between similar and dissimilar pairs of outfits, such that the distance between outfits in the similar pair is less than the distance between outfits in the dissimilar pair. Therefore, during training, the SVM learns from pairs of pairs (i.e. a similar pair of outfits and a dissimilar pair of outfits, with a shared base outfit). The learned weights serve as the basis for our similarity metric

$$D(\vec{x}, \vec{y}) = \vec{w}^T \|\vec{x} - \vec{y}\|^2$$

where $\vec{w}$ is the extracted weight vector and $\vec{x}$ and $\vec{y}$ correspond to representations of two outfits. Thus, the metric takes a difference between two outfits and uses a linear model to compute their similarity, with lower D corresponding to lower distance or increased similarity.

We constructed outfit feature vectors using our prescriptive representation and also incorporated our emergent topic modeling representations. We calculated proportions for each of the $k = 12$ topics for each outfit and appended them onto the prescriptive feature vector. To make it consistent with Schultz and Joachims' algorithm, for training, we converted each triplet i, j, k into vector $\vec{x}_{ijk}$ by concatenating the i, j vector and the i, k vector, which were composed of the differences between vector i and j or k.

Evaluation. Because we did not have the actual human ratings of distances between outfits at our disposal, we relied on prediction of similarity triplets as our evaluation metric. Measuring the proportion of correctly classified triplets serves as a reasonable proxy for our metric's performance. Using cross-validation on a training set of 7200 triplets (3600 triplets, duplicated to form two classes), we considered three ways of incorporating emergent features (LSA, LDA, LSA / LDA) and found that the LSA emergent representation yielded the best results.

Using $C = 1.0$, we evaluated the similarity metric on both a prescriptive outfit representation and a prescriptive + emergent representation. We trained our SVM with a training set of 7200 triplets and tested it on an unseen test set of 916 triplets. Results are shown in Table 4.

Because our training data primarily consisted of triplet queries that had been answered by a single person, and because our formative study indicated that individuals exhibit moderate, but not complete, agreement on similarity judgments, we decided to also test our metric on sets of consensus triplets. These triplets were answered by 20 different Turkers. We determined the consensus answer to be the pair most frequently selected as similar and defined the degree of agreement of a consensus triplet to be the proportion of individuals who selected this similar pair (out of three possible similar pairs, ij, ik, jk). Our results in Table 4 suggest that the metric performs very well on non-noisy triplets. Moreover, the results indicate that including the emergent representation in the similarity metric improves performance.

Test Set		Performance of Outfit Representations		
Degree of Agreement	Size of Test Set	Prescriptive	Prescriptive + Emergent	Prescriptive + Predicted
Individual triplets	916 triplets	0.60	0.61	0.60
Consensus (0.5 agreement)	84 triplets	0.65	0.70	0.67
Consensus (0.6 agreement)	58 triplets	0.69	0.74	0.72
Consensus (0.7 agreement)	48 triplets	0.67	0.73	0.73
Consensus (0.8 agreement)	30 triplets	0.77	0.83	0.83
Consensus (0.9 agreement)	16 triplets	0.81	0.88	0.88

Table 4: Performance of similarity metric on individual and consensus similarity triplets, measured by proportion of correctly classified triplets

6.2 Predicting emergent features

Given that a prescriptive + emergent dual representation is ideal, we also considered the feasibility of applying our metric to new outfits. Because obtaining text description is a costly task, to require all outfits in our data set to be properly annotated is a stringent requirement. Thus, we considered using the prescriptive representation to predict the higher-order features captured in our emergent topic-modeling representation. After all, people use intrinsic features to assess extrinsic ones; could an automated approach do the same?

Using ridge regression with optimal parameters chosen through cross-validation, we predicted sets of 12 higher-order features and used these to construct pseudo-prescriptive + emergent outfit representations. We tested this representation on the same individual and consensus test sets (Table 4). The results demonstrated that a similarity metric using predicted emergent features performed nearly as well as a metric using explicitly-provided emergent features and was more successful than the metric using the prescriptive representation alone.

7. STYLATRIX

As discussed in the introduction, the goal of our work was to design a system that enabled fashion exploration and outfit discovery. Having built a robust style similarity metric, we proceeded to use the metric as the foundation of Stylatrix. Inspired by previous work on use of examples in support of creative activities [11] and on example critiquing for exploring complex trade offs [3, 5], we decided to focus on interactions that allowed for dynamic exploration of diverse but relevant examples. Since the "low semanticity" of fashion limits the capabilities of individuals to specify fashion-related queries, designing interactions around outfit examples seemed to be an apt decision.

Stylatrix enables two primary interactions: *outfit discovery* through style-based outfit browsing and *style makeovers*. Both interactions are supported by style-based example critiquing, filtering, and ranking. Since Stylatrix can describe the extent to which an outfit is representative of a style based on its similarity to "prototype" outfits (i.e., how hipster or formal an outfit is), users can use these tools to further organize and explore outfit collections.

Outfit discovery. Stylatrix supports outfit browsing, which can be used to identify interesting outfits based on style similarity. Consider the query in Figure 2. The top three results returned for this outfit include two outfits which exhibit strong structural similarity (i.e. black top, black pants, black heels), as well as one which differs quite a bit in structure but nonetheless follows a similar urban-chic look. Users can filter results based on intrinsic features, perform style-based example critiquing (Figure 3) to discover similar outfits that are closer to some style (e.g. more hipster, more goth), and re-rank results based on how close they are to a style (e.g. rank results based on hipsterness).

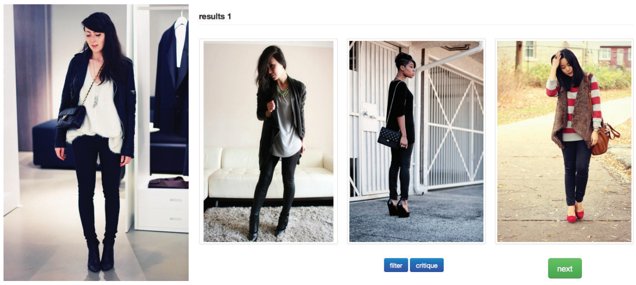


Figure 2: Outfit browsing. A query outfit is shown to the left, while the top-3 most similar outfits are displayed in the interface.

Style makeover. As illustrated in Figure 1, Stylatrix also supports style makeovers. Though individuals seeking a style makeover may find inspiration throughout daily life, it is less clear how an individual can use elements of her existing wardrobe to construct outfits that are closer to the desired style. The style makeover interaction enables users to specify an starting outfit (perhaps one they might own already) and a desired style to move towards. Results are then conceived as *intermediate* outfits, which are stylistically similar to both the query outfit and style. These outfits can inspire users and demonstrate how they can begin to change their wardrobe (e.g. to appear just a little more formal or a little more bohemian). The interface displays salient features that the user can focus on in their makeover purchases, and the slider allows the user to move within the space of intermediates.

Technical implementation. Using the similarity metric defined in Section 6, we can determine which outfits are most stylistically similar to a particular query outfit: given an outfit $\vec{x}$, we compute $D(\vec{x}, \vec{y})$ for all outfits $\vec{y}$ in the dataset and return a list of outfits with increasing D. We use this procedure to deliver relevant results for outfit browsing.

For all style-based interactions, we defined ten common styles for use in Stylatrix: vintage, eccentric, business, night out, grunge, schoolgirl, goth, punk, hipster, and bohemian. These

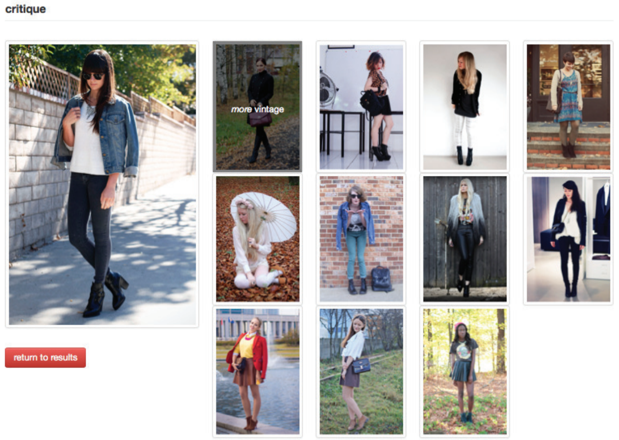


Figure 3: Style-based example critiquing. A result outfit is shown on the left, and users can explore outfits that are more similar to a style by clicking on the images to the right.

styles were based on the most common style names used on Chictopia. Each style is specified by several "prototype" outfits, which we have hand-picked from the dataset as concrete representatives of those styles. Similarity to a given style therefore reduces to similarity to the prototype outfits. We modify D to include two endpoints: the query outfit $\vec{x}$ and style prototype $\vec{s}$:

$$D(\vec{i}, \vec{x}, \vec{s}) = \alpha * D(\vec{i}, \vec{x}) + \beta * D(\vec{i}, \vec{s})$$

Given an outfit $\vec{x}$ and a style $\mathcal{S}$ represented by prototypes $\vec{s} \in \mathcal{S}$, we can compute D for all outfits $\vec{i}$ and identify suitable intermediates. We set an upper bound on dissimilarity, thereby preventing outfits that are highly dissimilar to either query outfit $\vec{x}$ or style prototype $\vec{s}$ from being considered as potential intermediates. α and β correspond to constants that can be adjusted in order to focus search on outfits closer to outfit $\vec{x}$ or prototype $\vec{s}$.

For both style-based critiquing and style makeover interactions, we identify the closest prototype outfit to the result outfit, compute D with differing α, β values for closest prototype $\vec{s}$, and select the top results from each computation to construct a series of intermediate outfits between outfit $\vec{x}$ and closest prototype $\vec{s}$. We append onto this list the intermediate outfits computed for all other prototypes for style $\mathcal{S}$, in order of decreasing similarity of prototype to query outfit. For style rankings, we rank how much an outfit represents a style (e.g. how hipster an outfit is) by computing the average similarity to each of the prototypes. Finally, the salient features presented in the style makeover interface are identified based on the largest-magnitude features from our computation of D.

7.1 User evaluation

In order to evaluate how well our constructed similarity metric supports the interactions described above, we conducted an informal user evaluation of Stylatrix. We recruited 8 undergraduates (6 females, 2 males, all aged between 20 and 21) to participate in a study on outfit browsing. Participants received monetary compensation for their work. Before beginning the study, we interviewed participants about their basic habits related to fashion and level of interest in style. Though most participants described minimal engagement with fashion, two individuals were eager consumers of fashion-related media content (e.g. Fashion Week, street style blogs).

We requested that each participant evaluate the outfit discovery interaction by making three separate queries. Participants were then shown both randomly-generated results and results generated by the similarity metric. They were asked to rate the relevance of each set of results using a 7-point Likert scale, where 1 corresponded to strong disagreement and 7 corresponded to strong agreement with the statement, "these results were relevant to my query." Random results received an average Likert score of 3.0, while the similarity metric results received an average Likert score of 4.8. These results suggest that the similarity metric can deliver novel and relevant outfit ideas.

Participants also explored the style makeover interaction in an unconstrained manner. All participants mentioned that the intermediates presented were both interesting and relevant, though were occasionally disappointed to see repeat outfits upon further querying. This latter point can be addressed by increasing the size of our outfit set, so as to provide novel results with each query. Figure 4 displays two queries and sample results from intermediates. In the first query, a user is attempting to move from a casual denim outfit to a more business-like style; whereas the first intermediate similarly incorporates a dark skirt and black leggings, the second intermediate is even more business-like with a dress shirt, sweater, and gray dress pants. The second query involves a movement from a casual, hipster outfit toward the goth style: the first intermediate uses more black, while the second incorporates a certain edginess lacking in both the query outfit and first intermediate. These examples epitomize the types of scenarios that the style makeover interaction can enable.

After completing the study, we briefly discussed with participants their general impressions. Some participants noted that the results they were provided would, on first glance, seem unrelated, but upon closer examination, they would realize a reasonable rationale style similarity. For example, one individual described how a query outfit with pants led to results that included skirts. Though the structure of those outfits were immediately different, she could nonetheless see how they were stylistically similar. Stylatrix aims to create experiences similar to the one described by the participant, which is to allow individuals to reflect upon how they view style and conceptualize their style preferences.

Overall, participants expressed their interest in the interactions enabled by Stylatrix, and described how useful a fully-featured system could be. A couple participants discussed how they would have liked to see characteristics like price and occasion built into the system. Participants generally enjoyed the results they were provided, even as non-experts with a pragmatic approach to fashion. Combined with our quantitative evaluation results, we conclude that Stylatrix effectively enables meaningful interactions that assist users with discovering and conceptualizing their fashion preferences.

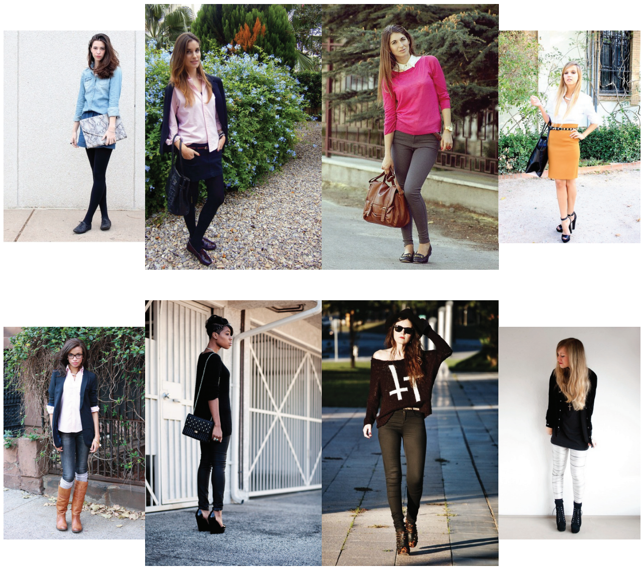


Figure 4: Intermediates for style makeover. Query outfit is shown on the left and desired style (business and goth) on the right. Two intermediates are shown per query, which possess distinctive structural and style similarities to both query outfit and desired style.

8. CONCLUSION

The work presented here describes an interactive, model-based system that offers novel fashion exploration and outfit discovery interactions. Results from a formative study served as the foundation for our work, suggesting that style was a holistic concept and that there was a moderate level of agreement among individual assessments of outfit similarities. Using a combined prescriptive and emergent representations of outfits, derived from human labeling and topic modeling of text descriptors, as well as similarity intuitions elicited through triplet similarity queries, we constructed a robust style similarity metric. We used this similarity metric to develop Stylatrix, an interactive system that enables outfit browsing and style makeover interactions. Qualitative and quantitative user evaluation results suggest that Stylatrix delivers results relevant to user goals.

Adding further outfits to our dataset currently requires annotating the outfit with a small number of well-defined descriptive labels. This task is easily amenable to crowdsourcing or could be completed by online volunteers on sites like Chictopia, as individuals naturally share their thoughts about outfits. By demonstrating that emergent, higher-order features can be successfully inferred from a prescriptive outfit representation, we obviate the need for the more demanding task of providing textual descriptions for the outfits.

As suggested by our user study participants, we intend on augmenting Stylatrix with more contextual elements; for example, occasion, mood, and weather. Additionally, Stylatrix could also support outfit completion interactions: given some attributes of an outfit, Stylatrix could return clothing items that result in an outfit that best meets certain style criteria. In the current style makeover interaction, Stylatrix highlights salient features or items that can be changed; outfit completion follows naturally from this.

Unlike traditional collaborative filtering approaches, Stylatrix allows users to navigate through the outfit space and construct their style preferences. Through an outfit-centric approach, Stylatrix presents users with outfit ideas and inspiration that are diverse, yet relevant. Stylatrix serves as a complement to item-based recommendation systems. Given the fairly sparse literature in this domain, this paper provides a solid foundation for future research into novel, model-based fashion interactions.

9. REFERENCES

[1] J.-J. J. Aucouturier and F. Pachet. Representing Musical Genre: A State of the Art. *Journal of New Music Research*, 32(1):83–93, Mar. 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] R. D. Burke, K. J. Hammond, and B. C. Young. The FindMe Approach to Assisted Browsing. *IEEE Expert: Intelligent Systems and Their Applications*, 12(4):32–40, July 1997.

[4] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *Proc. ECIR'08*, pages 16–27, Berlin, Heidelberg, 2008. Springer-Verlag.

[5] L. Chen and P. Pu. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1-2):125–150, 2012.

[6] F. Davis. *Do Clothes Speak? What Makes Them Fashion?* Routledge Student Readers Series. Taylor & Francis Group, 2007.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *JASIS*, 41(6):391–407, 1990.

[8] A. F. Hayes and K. Krippendorff. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[9] M. D. Hoffman, D. M. Blei, and F. R. Bach. Online Learning for Latent Dirichlet Allocation. In *Proc. NIPS '10*, pages 856–864. Curran Associates, Inc., 2010.

[10] T. Iwata, S. Wanatabe, and H. Sawada. Fashion Coordinates Recommender System Using Photographs from Fashion Magazines. In *Proc. IJCAI*, pages 2262–2267, 2011.

[11] B. Lee, S. Srivastava, R. Kumar, R. Brafman, and S. R. Klemmer. Designing with interactive example galleries. In *Proc. CHI '10*. ACM Request Permissions, Apr. 2010.

[12] R. Rehurek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[13] M. Schultz and T. Joachims. Learning a Distance Metric from Relative Comparisons. In *Proc. NIPS '03*. MIT Press, 2003.

[14] E. Shen, H. Lieberman, and F. Lam. What am I gonna wear?: scenario-oriented recommendation. In *Proc. IUI '07*, pages 365–368, New York, NY, USA, 2007. ACM.

[15] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively Learning the Crowd Kernel. *CoRR*, abs/1105.1, 2011.

[16] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *Proc. CVPR*, pages 3570–3577, 2012.

[17] S. Yuan. A system of clothes matching for visually impaired persons. In *Proc. ASSETS '10*, pages 303–304, New York, NY, USA, 2010. ACM.