Learning to Rank Similar Apparel Styles with Economically-Efficient Rule-Based Active Learning

Conference Paper · April 2014

DOI: 10.1145/2578726.2578773

CITATIONS
7

READS
444

3 authors:

Mariane Moreira
Federal University of Alfenas
14 PUBLICATIONS **82** CITATIONS
SEE PROFILE

Jefersson A. dos Santos
Federal University of Minas Gerais
141 PUBLICATIONS **4,192** CITATIONS
SEE PROFILE

Adriano Veloso
Federal University of Minas Gerais
220 PUBLICATIONS **4,528** CITATIONS
SEE PROFILE

Learning to Rank Similar Apparel Styles with Economically-Efficient Rule-Based Active Learning

Mariane Moreira Jefersson A. dos Santos Adriano Veloso
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{mariane, jefersson, adrianov}@dcc.ufmg.br

ABSTRACT

Increasingly, people define and express themselves in online social networks, such as Facebook and Instagram, by uploading photos showing the clothes they wear. As a result, such online social networks are becoming major sources of inspiration, with users looking for others with similar clothing style. In this paper, we propose a novel learning to rank (L2R) algorithm for finding similar apparel style given a query image. L2R algorithms use a labeled training set to generate a ranking model that can later be used to rank new query results. These training sets, however, are costly and laborious to produce, requiring human annotators to assess the relevance of candidate images in relation to a query. Active learning algorithms are able to reduce the labeling effort by selectively sampling an unlabeled set of images and choosing the subset that maximizes a learning function's effectiveness. Specifically, our proposed L2R algorithm employs an association rule active sampling algorithm to select very small but effective training sets. Further, our algorithm operates on visual (e.g., image descriptors) and textual (e.g., comments associated with the image) elements, in a way that makes it able (i) to expand the query image (for which only visual elements are available) with textual elements, and (ii) to combine multiple elements, being visual or textual, using basic economic efficiency concepts. We conducted a systematic evaluation of the proposed algorithm using every-day photos collected from Instagram, and we show that our L2R algorithm reduces by two orders of magnitude the need for labeled images, and still improves upon the state-of-the-art models by 4-8% in terms of mean average precision.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory, Experimentation

Keywords

Apparel style, Active Learning, Online Social Networks

ICMR'14 April 01–04 2014, Glasgow, United Kingdom

1. INTRODUCTION

Networking sites, such as Facebook and Instagram, allow their users to express themselves in many different ways by creating and sharing content. A particular way of expression being increasingly adopted by members of these sites is to post photos that show their latest looks and clothes. Typically, comments appear shortly after the image is posted, and thus online social networks are becoming major sources of clothing inspiration, with users looking for others with similar apparel style and fashion sense.

The problem, in this case, is that a substantial time is spent searching for specific looks, often related to a style. A user may navigate for hours, and there is no guarantee of finding the desired content. In this paper we tackle this problem by using a retrieval based approach − given a query image, we find images with similar clothing styles in a large database of images posted in online social networks. In this process, accurately ranking the returned images is of paramount importance, since users consider mostly the top-most results. Therefore, we consider learning to rank (L2R) algorithms [21], which deliever superior performance when compared with more traditional solutions [9]. L2R algorithms rely on labeled training sets to build ranking models that are used to rank results at query time [22]. To create these training sets, human annotators must evaluate query-result pairs, and provide a relevance judgment (i.e., relevant, not relevant). Clearly, it is costly and laborious to label large training sets for L2R algorithms.

Our Approach to Reduce Labeling Effort. Active learning algorithms have been proposed to help dealing with the labeling effort problem in L2R [27, 8, 26]. The motivation behind active learning is that it may be possible to learn effective ranking models by labeling only data instances (i.e., images) that are "informative" to the learning algorithm. Images that are similar to each other may offer redundant information to the learning algorithm, and thus one should avoid labeling similar images. In a typical active learning scenario [7], images are selected from an unlabeled set one at a time and labeled by a human expert. Every time a new image is labeled and included into the training set, a new ranking model is produced and the active learning algorithm again chooses another image from the unlabeled set. This process is repeated until the ranking model converges, and the training set is finally built. In particular, our proposed algorithm represents images using visual (e.g., local image descriptors) and textual (e.g., comments) elements, and adopts a novel "one-to-many" similarity measure based on association rules [1, 2], in order to assess the amount of information associated with a candidate image with respect to the images already selected and labeled (i.e., the current training set). Intuitively, a candidate image is informative to the learning algorithm if there is no similar image in the current training set, and thus it is worth

to label and include this candidate image into the training set. At the end of the process, the resulting training set is composed of the most diverse (or "dissimilar") training images.

Our Approach to Improve Ranking Performance. Once the training set is built, the learning algorithm becomes able to produce a ranking model for any query image. The produced model comprises features derived from visual and textual elements of images in the training set. Such features include those based on off-the-shelf image descriptors and style-related terms within posts accompanying the image. The full potential of the model, however, is prevented as only visual elements are available in the query image. Thus, in order to fully exploit the ranking model, our proposed algorithm expands the query image with textual features. More specifically, query expansion is performed by uncovering any association that may exist between visual and textual features.

The last highlight of the proposed L2R algorithm is a novel approach for combining multiple elements of the image in order to further improve retrieval performance. Our L2R algorithm interprets each image element as a dimension in an $n-$dimensional scattergram, and consequently each returned image can be viewed as a point in the scattergram. This enables us to exploit a central concept of Economics − Pareto Efficiency − in order to find a proper (or economically efficient) balance between all image elements. The Pareto Efficiency criterion informally states that "when some action could be done to make someone better off without hurting anyone else, then it should be done." This action is called Pareto improvement, and society is said to be Pareto-Efficient if no such improvement is possible. We exploit the concept of Pareto Efficiency by separating images that are not dominated by any other image in the scattergram. These images compose the Pareto frontier [24], and correspond to cases for which no Pareto improvement is possible, and are therefore more likely to be relevant with respect to the query as they excel in at least one image element.

Contributions and Findings. In practice, we claim the following benefits and contributions over existing solutions:

- A practical and effective active learning approach that can be used to produce training sets for L2R algorithms devised to rank similar apparel styles.
- A fast and effective approach to expand the query image with textual features, enabling the full potential of the produced ranking models.
- An effective approach to combine multiple elements of the image. The proposed approach exploits the notion of Pareto Efficiency in order to separate images that excel in at least one dimension of comparison (shape, color, texture, or text), or images that offer a proper balance between different dimensions.
- A systematic set of experiments, using a collection of everyday photos crawled from Instagram, reveals that our L2R algorithm reduces by two orders of magnitude the need for labeled images, and still improves upon the state-of-the-art models based on combining multiple image elements [9], by 4-8% in terms of mean average precision.

The paper is structured as follows. Related work is presented in Section 2. We introduce our L2R algorithm, its components, algorithms and concepts in Section 3. Experimental evaluation, as well as the effectiveness of the proposed algorithm, are discussed in Section 4. Finally, in Section 5 we conclude the paper.

2. RELATED WORK

In recent years, there has been an increasingly interest on the intersection of online social networks and fashion related issues. Tu and Dong [30], for instance, present a system that helps customers finding the most suitable clothing choices using information from social networking sites. Iwata et al. [15] proposed a system that recommends clothes using full-body photos collected from fashion magazines. Specifically, given a photograph of a fashion item (e.g., tops) as a query, the system must recommend photographs of other fashion items (e.g., bottoms) that are appropriate with regard to the query. Tokumaru et al. [29] proposed a system, named "Virtual Stylist", which aims to help users finding out outfits that might fit them well. Also, Hidayati et al. [14] present approaches to automatically recognize clothing genre, with an initial focus on upperwear clothes. The work of Vogiatzis et al. [32] describes the recommendation of clothes based on the similarity between users and models appearing in fashion magazines.

There also works that exploit the combination of visual and textual elements. The work of Shen et al. [25], for instance, introduces the recommendation of outfits for specific occasions based on a textual input that defines the occasion and how the user wants to look like. Also, the work of Liu and Cheng [5] defines an approach to retrieve images of clothes in a virtual closet, also according to textual input. Yagamuchi et al. [33] studied the clothing parsing problem using a retrieval based approach. For a query image, they find similar styles from a large database of tagged fashion images and use these examples to parse the query. Finally, work of Dao et al. [6] proposes a watershed-based method with support from external data sources and visual information to detect social events in web multimedia.

Liu et al. [20] addressed the problem of cross-scenario clothing retrieval. Given a photo captured in a general environment (e.g., on street), the problem is to find similar clothing in online shops. Fu et al. [12] address the problem of large scale cross-scenario clothing retrieval with semantic-preserving visual phrases. Another approach related to cross-scenario retrieval is proposed by Yannis et al. [17], which present a scalable approach to automatically suggest relevant clothing products. In this case, the query is an image, while products from online shopping catalogs are presented. Li-Jia Li et al. [19] proposed a segmentation approach to remove background pixels. Their approach follows a hierarchical generative model that, given an image, classifies the overall scene, recognizes and segments each object component, as well as annotates the image with a list of tags. In this paper we followed a similar approach to remove background pixels, in order to ensure more quality to the visual feature extraction process.

In this paper, we employ different concepts, not exploited in the aforementioned works. First, instead of using simple distance metrics or creating specific descriptors for the sake of retrieving similar images (or images showing similar clothing style), we combine many such distances/descriptors using an efficient learning to rank approach. Many L2R algorithms exist in the literature [22, 16, 13, 9], and some of them are used as baseline in this work. In order to reduce labeling effort while producing training sets, we introduce an active learning approach which drastically reduces label requirements without compromising the effectiveness of the final ranking model. Further, we introduce a query expansion approach, so that the query image is reformulated in order to include textual elements. Finally, we propose economically-efficient approaches for aggregating rankings produced by different (visual or textual) elements.

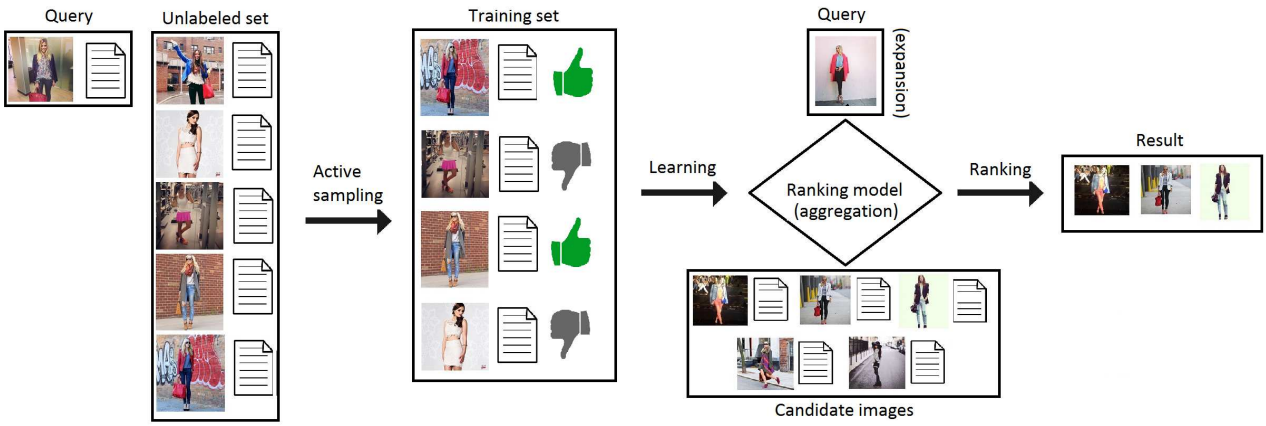

Figure 1: Overview of the proposed L2R algorithm.

3. RANKING SIMILAR APPAREL STYLES

In this section we present our L2R algorithm for ranking similar apparel styles with respect to a query image. Our algorithm, named as AL2R+RL2R, comprises approaches for reducing labeling effort as well as approaches for improving ranking performance. Figure 1 shows an illustrative example involving the main components and steps that will be discussed in this section. We start discussing visual and textual elements used to represent images.

3.1 Visual and Textual Elements

Images posted in online social networks (in particular those related to clothing) may contain both visual and textual elements, and each modality may be analyzed in a variety of ways. For instance, visual elements can be analyzed based on color, texture, shape, and so on. Textual elements, in turn, include style-related terms in the comments. These elements may provide predictive signals to L2R algorithms. Specifically, we observed that: (i) images showing similar apparel styles are likely to share similar visual signals (e.g., color, texture and shape), and (ii) people tend to post similar annotations for images showing similar aparel styles.

Image Descriptors. Visual elements are strongly based upon the concept of image descriptors. A descriptor expresses perceptual qualities of an image, and is composed by (i) a feature-vector that represents image properties, such as color, texture and shape, and (ii) a similarity measure (or distance function) that computes the similarity between two images as a function of the distances between their corresponding feature-vectors. Both the feature-vector and the distance function affect how the descriptor encodes the perceptual qualities of the images.

There is a multitude of descriptors available in the literature [28, 31, 23, 4, 34, 18], that we used to represent visual elements. Clearly, different descriptors produce different rankings. Also, the best descriptor to apply is data-dependent, and unlikely to be known before query time. Further, it is intuitive that different descriptors may provide complementary information about images, so that the combination of multiple descriptors is likely to improve ranking performance. However, the optimal combination of descriptors is, again, data-dependent and unlikely to obtain in advance.

Textual Information. Textual content includes tags and terms appearing in comments associated with the image. An apparel expert helped us to define a vocabulary containing terms related to different clothing styles, such as sporty (e.g., sweats, t-shirts, tennis) and casual (e.g., shorts, pants, flip-flops). After filtering out all terms not in the style vocabulary, the remaining textual content is described with TF-IDF vectors. The TF-IDF transformation weights each term according to its discriminative capacity. Textual similarity between two images is assessed using the standard cosine and BM25 similarity measures [3].

3.2 Rule-Based Active Learning for L2R

We provide a set of query images as input to our L2R algorithm. Associated with each query image, we also provide a set of sample images which are represented by the corresponding (visual and textual) similarities to the query image. The relevance of each sample image with respect to the query image is also informed as input. This information is used as training so that our L2R algorithm produces a ranking model that maps similarity values to the relevance of images with respect to the query image. When a new query image is given, the relevance of the returned images is estimated according to the learned model. Next we discuss the detailed steps of our L2R algorithm. Then, in order to reduce labeling effort, we introduce our active learning algorithm.

The RL2R Algorithm. The rule-based learning to rank (RL2R) algorithm uses association rules [1] to rank images showing similar apparel styles given a query. The RL2R algorithm takes as input a labeled training set $\mathcal{D}$ composed of records of the form $< q, d, r >$, where q is a query image, d is a returned image, and r is the relevance of d to q. Images are represented as a list of m similarity values $\{f_1, f_2, \ldots, f_m\}$, where each f_i gives the similarity between d and q according to an image descriptor or the textual content (as discussed in Section 3.1). Similarity values are discretized [10] and then assigned to similarity intervals,[1] in order to allow for the enumeration of association rules. Relevance r is drawn from a set of discrete and ordered possibilities $\{r_0, r_1, \ldots, r_k\}$ (e.g., 0: not relevant, 1: somewhat relevant, 2: relevant, 3: very relevant).

The test set $\mathcal{T}$ consists of records $< q, d, ? >$ for which only the query q and the returned image d are known, whereas the relevance of d to q is unknown. From the training set $\mathcal{D}$, the algorithm extracts a rule-set $\mathcal{R}$ composed of relevance rules which are used to estimate the relevance of the images.

[1]Hereafter we refer each f_i as the corresponding interval.

Definition 1. A relevance rule has the form:

$$\overbrace{\{\ f_j \wedge \ldots \wedge f_l\ \}}^{\text{Similarity intervals}} \quad \xrightarrow{\theta} \quad \overbrace{\{\ 0, 1, \ldots, k\ \}}^{r_i}$$

where $j \geq 1$ and $l \leq m$.

These rules can contain any mixture of similarity intervals in the antecedent and a relevance level in the consequent. The strength of the association between antecedent and consequent is measured by a statistic θ, which is known as confidence [1] and is simply the conditional probability of the consequent given the antecedent. Basically, each relevance rule $\{X \rightarrow r_i\} \in \mathcal{R}$ is a vote given for relevance r_i. Given an image $d \in \mathcal{T}$, a relevance rule is a valid vote if it is applicable to d.

Definition 2. A relevance rule $\{X \rightarrow r_i\}$ is said to be applicable to image $d \in \mathcal{T}$ if all intervals in X are in d, that is, $X \subseteq d$.

We denote as $\mathcal{R}_d$ the set of relevance rules in $\mathcal{R}$ that are applicable to image d. Thus, only rules in $\mathcal{R}_d$ are considered as valid votes when estimating the relevance of d with respect to q. Further, we denoted as $\mathcal{R}_d^{r_i}$ the subset of $\mathcal{R}_d$ containing only rules predicting relevance r_i. Votes in $\mathcal{R}_d^{r_i}$ have different weights, depending on the confidence of the corresponding rules. Given an image d, the weighted votes for relevance level r_i are averaged, resulting in the score for r_i, as shown in Equation 1:

$$s(d, r_i) = \frac{\sum \theta(X \rightarrow r_i)}{\mathcal{R}_d^{r_i}}\text{, where } X \subseteq d. \tag{1}$$

The likelihood of $d \in \mathcal{T}$ having relevance level r_i is obtained by normalizing the scores, as expressed by $\hat{p}(r_i|d)$, shown in Eq. 2:

$$\hat{p}(r_i|d) = \frac{s(d, r_i)}{\sum_{j=0}^{k} s(d, r_j)}. \tag{2}$$

Finally, the relevance of image d with respect to q is estimated by a linear combination of the likelihoods associated with each relevance level, as expressed by the ranking function $rel(d)$, which is shown in Equation 3:

$$rel(d) = \sum_{i=0}^{k} (r_i \times \hat{p}(r_i|d)). \tag{3}$$

The value of $rel(d)$ is an estimate of the relevance of image d using $\hat{p}(r_i|d)$. This estimate ranges from r_0 to r_k, where r_0 is the lowest relevance and r_k is the highest one.

The AL2R Algorithm. The training set $\mathcal{D}$ may contain similar images, that is, images sharing most of their similarity intervals. For the RL2R algorithm, similar images offer redundant, useless, and possibly detrimental information during training. In the following we discuss the AL2R algorithm, a selective sampling algorithm which builds a training set by iteratively minimizing the amount of redundancy in it. The basic idea behind AL2R is that $|\mathcal{R}_d|$ is an indication of how much information image d shares with the current training set. Specifically, if $\mathcal{R}_d$ contains few rules, then d is highly informative since it is likely to be different from images in the current training set. Otherwise, if $\mathcal{R}_d$ contains many rules, then many rules in $\mathcal{R}$ are applicable to d, indicating that there may be various images in the current training set that are similar to d, and thus d is not worth the effort of labeling it.

The AL2R algorithm works as follows. Given a large set of unlabeled images $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$, the algorithm selects highly informative images to compose the training set $\mathcal{D}$ such that $|\mathcal{D}| \ll |\mathcal{U}|$. Initially, $\mathcal{D}$ is empty and thus the algorithm cannot extract any rule from it, so it randomly selects an image from $\mathcal{U}$. The selected image is labeled and included into $\mathcal{D}$ (but it also remains in $\mathcal{U}$). Then, at each round, the algorithm applies the sampling function $\gamma(\mathcal{U})$ which returns an image $u_i \in \mathcal{U}$ according to Equation 4:

$$\gamma(\mathcal{U}) = \{u_i \text{ such that } \forall u_j : |\mathcal{R}_{u_i}| < |\mathcal{R}_{u_j}|\} \tag{4}$$

Once $u_i \in \mathcal{U}$ is inserted into $\mathcal{D}$, the number of rules for images in $\mathcal{U}$ that share similarity intervals with u_i will increase in the next round. But the number of rules for those images in $\mathcal{U}$ that do not share any similarity interval with u_i will remain unchanged. Therefore, the number of rules extracted for each image in $\mathcal{U}$ can be used as an approximation of the amount of redundant information between images already in $\mathcal{D}$ and images in $\mathcal{U}$. The sampling process iterates so that, at round i, the selected image is denoted as $\gamma_i(\mathcal{U})$, and it is likely to be as dissimilar as possible from the images already in $\mathcal{D} = \{\gamma_{i-1}(\mathcal{U}), \gamma_{i-2}(\mathcal{U}), \ldots, \gamma_1(\mathcal{U})\}$. The result is a very small training set based on a *diversity* criterion: the more diverse images we have in the training set, the more we cover the similarity space with the smallest possible amount of labeled images.

The algorithm stops when all available images in $\mathcal{U}$ are less informative than any image already inserted into $\mathcal{D}$. This occurs when AL2R selects an image which is already in $\mathcal{D}$. When this condition is reached, AL2R will keep selecting the same image over and over again, and there is no gain with the inclusion of these images. Thus the algorithm stops, and the final training set $\mathcal{D}$ becomes available to the RL2R algorithm.

3.3 Query Expansion

Only visual elements are available in a query image. As a result, rules produced by the RL2R algorithm do not consider textual elements, that is, features in the antecedent of the rule come exclusively from similarity intervals associated with image descriptors. In order to improve ranking performance, we propose a novel approach to reformulate the original query image by adding similarity intervals associated with textual content. Specifically, we exploit the association between visual similarity intervals and textual similarity intervals in the training set $\mathcal{D}$ using expansion rules.

Definition 3. An expansion rule has the form:

$$\overbrace{\{\ f_j \wedge \ldots \wedge f_l\ \}}^{\text{Visual similarity intervals}} \quad \xrightarrow{\theta} \quad \underbrace{\{\ f_{\text{Cos}} \wedge f_{\text{BM25}}\ \}}_{\text{Textual similarity intervals}}$$

where $j \geq 1$ and $l \leq m$.

These rule can contain any mixture of visual similarity intervals in the antecedent and textual similarity intervals in the consequent.

Basically, we pick the expansion rule $\{X \xrightarrow{\theta} Y\}$ with highest θ value, and expand the original query image with the corresponding cosine and BM25 intervals in Y. Now, the reformulated query also comprises textual elements and the ranking model can be fully exploited.

3.4 Economically-Efficient Rank Aggregation

As discussed, the relevance of an image is judged based on the similarity to the query in terms of their visual features (and possibly the text surrounding both images). The usual approach to combine these similarities in order to produce the ranking model is to jointly consider all similarity values and simply let the L2R algorithm find a proper combination of them.

However, as a natural split exists in the sense that similarities can be grouped into different types (i.e., shape, color, texture and text), we may assume that similarity values associated with different types are conditionally independent. We may also assume that similarity values within each type are sufficient, in the sense that accurate ranking models can be produced using similarity values within each type alone [9] (i.e., only color, shape, texture, or text). Under these circunstances, similarity values within each type provide different, complementary information about the relevance of the image, and we may take advantage from this in order to improve ranking performance. Specifically, instead of combining all available similarity values into a single ranking model $rel(d)$, we may produce a separate ranking model for each type of similarity, and then aggregate these different models into a final one.

Definition 4. We define the relevance space as a 3-dimensional scattergram, in which each dimension corresponds to the relevance estimate with respect to a query, whether in terms of shape ($rel_s(d)$), color ($rel_c(d)$), or texture ($rel_t(d)$). Images are placed in the relevance space according to the correspoding relevance estimates, as shown in Figure 2. Thus, each image a is a point in such space, and is given as $< rel_s(a), rel_c(a), rel_t(a) >$. Textual similarity is used as an additional dimension of the relevance space if query expansion is performed.

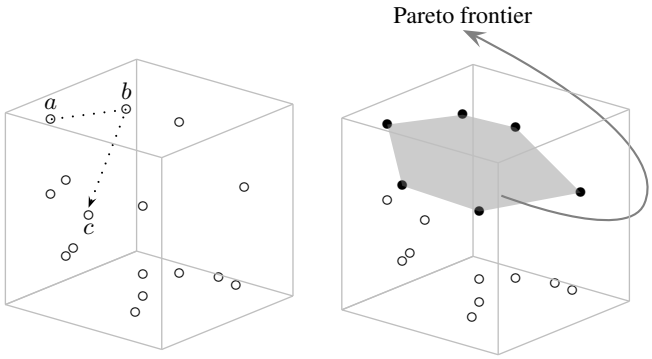

Figure 2: Illustrative example. Points are images, and are represented by the corresponding relevance to the query in terms of shape ($rel_s(d)$), color ($rel_c(d)$) and texture ($rel_t(d)$). Textual similarity may also be considered as an additional dimension in case of performing query expansion. (Left) The dominance operator: neither a or b dominates each other, but b dominates c. (Right) Points lying in the Pareto frontier.

The relevance space is used to aggregate multiple relevance estimates for an image into a stronger one. Our aggregation approach is based on a basic concept from Economics called Pareto Efficiency, which states that when the society is economically efficient, any changes made to assist one person would harm another. In our case, Pareto Efficiency is related to the notion of dominance in the relevance space.

Definition 5. Image a is said to dominate image b iff both of the following conditions are hold:

- $rel_s(a) \geq rel_s(b), rel_c(a) \geq rel_c(b)$ and $rel_t(a) \geq rel_t(b)$
- $rel_s(a) > rel_s(b)$ or $rel_c(a) > rel_c(b)$ or $rel_t(a) > rel_t(b)$

Therefore, the dominance operator relates two images so that the result of the operation has two possibilities as shown in Figure 2 (Left): (i) one image dominates another or (ii) the two images do not dominate each other.

Definition 6. The Pareto frontier is composed of all images that are not dominated by any other image in the relevance space. Formally, the Pareto frontier is a list of p images $\{d1, d2, \ldots, d_p\}$ such that there is no pair (d_i, d_j) for which d_i dominates d_j.

The Pareto frontier contains either images that excel in at least one dimension of the relevance space, or images with a proper balance of the considered relevance estimates with respect to an arbitrary query. The final step involves sorting these images according to the final relevance estimate.

Definition 7. Let $dom(d)$ returns the number of images in the relevance space that are dominated by image d. The final ranking given an arbitrary query q is an ordered list $\{d1, d2, \ldots, d_p\}$ such that there is no pair (d_i, d_j) for which $dom(d_i) > dom(d_j)$, given that $i > j$. That is, most dominant images appear first in the ranking.

In the next section we evaluate the ranking performance of the proposed AL2R+RL2R algorithm.

4. EXPERIMENTS

In this section we present the experimental results for the evaluation of the proposed AL2R+RL2R algorithm in terms of ranking performance and labeling effort. We consider two different approaches for learning rankings with AL2R+RL2R. The first one, which is the usual one, is to consider all similarity values together, as a single vector of similarity values. We denote this approach as "Feature Combination", as the model produced by AL2R+RL2R is a combination of all available similarity values. The second approach is based on our proposed rank aggregation strategy, discussed in Section 3.4. Results reported in Sections 4.4 and 4.5 were obtained following the "Feature Combination" approach. In summary, we show that:

- AL2R+RL2R achieves state-of-the-art ranking performance using a fraction of the original training set, when searching similar apparel styles (Sec. 4.4)
- Query expansion enables the use of textual elements, and improves ranking performance with no increase in terms of labeling effort (Sec. 4.5)
- Rank aggregating using the Pareto Efficiency concept leads to further improvements in ranking performance (Sec. 4.6)

4.1 Dataset

Utilizing the Instagram API that gives access to public profiles, we have crawled a large amount of images and associated comments. In total we crawled 1,627,482 images with their corresponding comments. After filtering those images with meaningfull human pose [11] and with sufficient number of comments (≥ 10),

we reach a total of 58,222 images. Background pixels were then removed [19], and a set of descriptors based on color [28], shape [23] and texture [31] were employed to encode the images. Finally, (i) we randomly selected 150 out of the 58,222 images to be used as query-images, and (ii) for each query-image an expert manually selected 10 relevant and 30 not relevant images. The expert decided the relevance of an image based on its similarity to the query image and also on their apparel styles.

4.2 Baselines

We considered the following algorithms in order to provide strong baseline comparison:

- L2R algorithms from the LETOR benchmark.[2] In particular, we consider two of the best available algorithms, namely RANK-SVM [16, 13] and ListNet [35].
- L2R algorithm proposed in [8], which will be referred to as ELR (Estimated Loss Reduction). The ELR algorithm applies active learning in order to reduce labeling effort during training.
- L2R algorithms proposed in [9]. In particular, we consider the best overall performer, namely CBIR-SVM.

4.3 Evaluation Procedure

To evaluate the ranking performance of the algorithms, we have used the standard MAP (Mean Average Precision) measure, which gives a summarized measure of the precision x recall curve. MAP is sensitive to the entire ranked list of images. For detailed discussion about MAP, the reader is referred to [3]. To evaluate the labelling effort of the algorithms we simply used the number of training examples required to produce the ranking model.

To evaluate the ranking performance of AL2R+RL2R, we conducted five-fold cross validation. Thus, the dataset was arranged into five folds, including training and test. At each run, four folds are used as training set, and the remaining fold as test set. The results reported are the average of the five runs. Parameters used are those that lead to the best results for each algorithm.

4.4 Ranking Performance and Labeling Effort

Our first experiment shows the trade-off between labeling effort and ranking performance. Labeling effort is given as the fraction of the unlabeled set $\mathcal{U}$ that is labeled, while ranking performance is given in terms of MAP. Figure 3 shows, for each algorithm, the corresponding MAP number and labeling effort. Algorithms such as RL2R, CBIR-SVM, RANK-SVM and ListNet, demand the maximum labeling effort. In contrast, AL2R+RL2R and ELR employ different active learning approaches, and as a result, these algorithms need much less effort. AL2R+RL2R shows superior ranking performance with less labeling effort, in comparison with ELR. Further, MAP numbers obtained by AL2R+RL2R are competitive to the numbers obtained by the best performers.

Next, we investigate the ranking performance of each algorithm as the number of labeled images increases. Specifically, we randomly selected a fraction of $\mathcal{U}$ to be labeled and used for training the corresponding ranking models. Figure 4 shows the results obtained. As can be seem, RL2R, CBIR-SVM, RANK-SVM, and ListNet, perform poorly when only few labeled images are made available for training. In such scenario, ELR and AL2R+RL2R provide much better MAP values. As expected, MAP values increase with the number of labeled images increases. The ranking performance of AL2R+RL2R is only surpassed by RL2R when around 50% of the images in $\mathcal{U}$ are labeled. Similarly, ListNet requires labeling around 70% of the images in $\mathcal{U}$ in order to surpass the ranking performance of AL2R+RL2R.

[2] http://research.microsoft.com/~letor

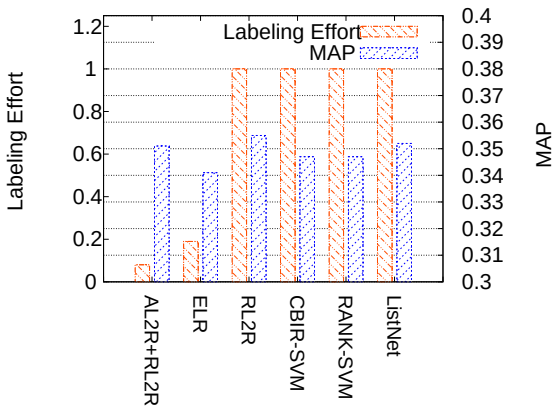

Figure 3: (Color online) Reference ranking performance associated with different L2R algorithms. These numbers are used to assess the gains provided by AL2R+RL2R.

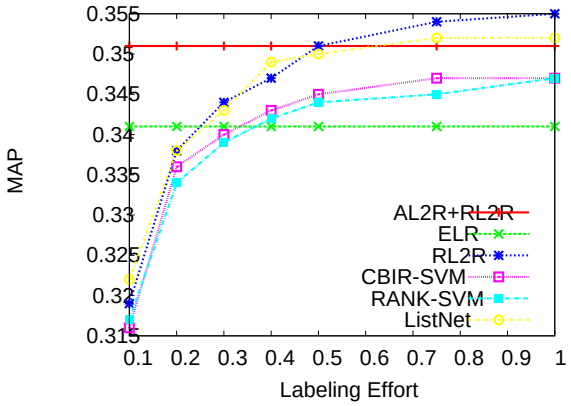

Figure 4: (Color online) – MAP numbers for different L2R algorithms, as a function of labeling effort.

Figure 5 shows the gains provided by AL2R+RL2R, in terms of MAP, as a function of a labeling reduction factor. Clearly, for higher reduction factors, such as 92%, the gains are almost negligible, and sometimes negative. On the other hand, significant gains are provided by AL2R+RL2R when the labeling reduction factor ranges from 20% to 60%. The correlation between labeling reduction and MAP gains is almost linear, and thus, large gains are obtained even with important reduction in terms of labeling effort.

4.5 Query Expansion

Figure 6 shows MAP values and labeling effort associated with different apparel styles, and obtained using AL2R+RL2R. Clearly, the need for labeled images varies greatly depending on the style. This is intuitive, since some styles allow more complex and diverse combinations. AL2R+RL2R requires a few labeled images for learning "beach-wear" and "sporty" styles, but many labeled images for learning "preppy" and "grunge styles."

MAP values also vary accross different styles. Query expansion provides an average increase of 3% in MAP values. Styles such as

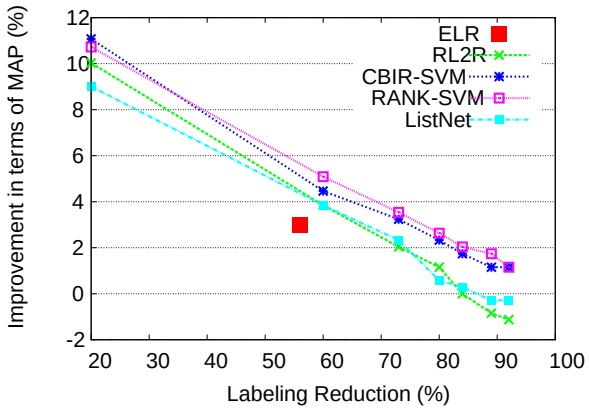

Figure 5: (Color online) Gains in terms of MAP for increasing labeling reduction factors.

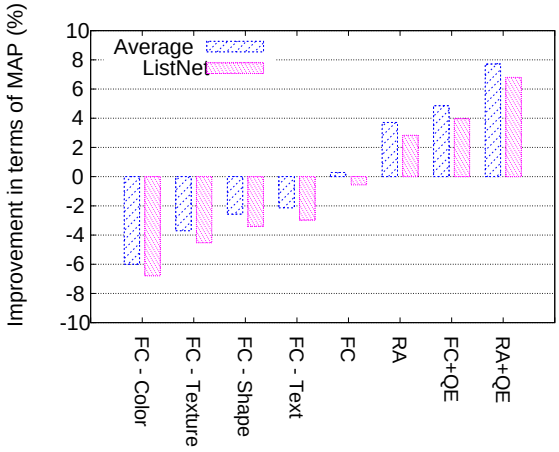

Figure 7: (Color online) MAP improvements provided by AL2R+RL2R relative to the average and best baseline performance. FC − Feature Combination. RA − Rank Aggregation. QE − Query Expansion.

"vintage" and "sexy" are the most benefited ones. We also computed an upper bound for MAP values, which is given as the MAP value obtained if we always reformulate the query with the correct textual similarity intervals. On average, upper bound values are 5% higher than the values obtained using our query expansion approach.

4.6 Economically-Efficient Rank Aggregation

In the last set of experiments we evaluate the effectiveness of our proposed rank aggregation approach. Figure 7 summarizes the improvements provided by different configurations of AL2R+RL2R. It is worth noticing that AL2R+RL2R required only 8% of the images in $\mathcal{U}$ to be labeled, while the baseline required all images in $\mathcal{U}$ to be labeled.

We first evaluate the use of different image elements (i.e., color, texture, shape, text) in isolation. That is, for each of these elements, AL2R+RL2R produced a specific ranking model which is the combination of the similarity intervals associated with the corresponding element. For instance, in the figure, "FC − Color" represents the ranking model produced by AL2R+RL2R by combining the available color-based similarity intervals. Clearly, combining all image elements leads to much better results than the ones obtained by using image elements in isolation. Further, our proposed rank aggregation approach leads to even better results, providing gains of 4% when compared with the average baseline performance. Additional gains are obtained by performing query expansion. Specifically, gains of 8% are obtained by performing query expansion and using our rank aggregation approach.

5. CONCLUSIONS AND FUTURE WORK

This paper focused on approaches for efficiently finding images showing similar apparel styles. Apparel and fashion are popular topics in networking sites, such as Instagram, and thus users of these sites are typically looking for specific content related to clothing. Our proposed algorithm follows a learning to rank strategy for image retrieval, that is, a ranking model is learned from a training set composed of labeled images. In particular, we studied two issues: (i) labeling effort, and (ii) ranking performance. We propose an active learning approach (AL2R) for reducing labeling effort based on a diversity criterion: the more diverse (or dissimilar) images we have in the training set, the more we cover the similarity space with the smallest possible amount of labeled images. Also, we propose approaches for improving ranking performance with query expansion and rank aggregation. Query expansion is performed by exploiting association between visual and textual elements, while rank aggregation is based on a basic concept from Economics, namely Pareto Efficiency. Comparison against the state-of-the-art L2R algorithms reveals the superiority of our proposed algorithm, which provides a 92% reduction in labeling effort while providing improvements ranging from 4% to 8%, in terms of MAP. Future works include the development and validation of more visual descriptors.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[3] R. Baeza-Yates and B. R-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, 1999.

[4] A. Çarkacioglu and F. Yarman-Vural. SASI: a generic texture descriptor for image retrieval. *Pattern Recognition*, 36(11):2615–2633, 2003.

[5] C. Cheng and D. Liu. An intelligent clothes search system based on fashion styles. In *ICMLC*, pages 1592–1597, 2008.

[6] M. Dao, G. Boato, F. D. Natale, and T. Nguyen. Jointly exploiting visual and non-visual information for event-related social media retrieval. In *ICMR*, pages 159–166, 2013.

[7] J. de Freitas, G. Pappa, A. Soares, M. Gonçalves, E. de Moura, A. Veloso, A. Laender, and M. de Carvalho. Active learning genetic programming for record deduplication. In *CEC*, pages 1–8, 2010.

[8] P. Donmez and J. Carbonell. Optimizing estimated loss reduction for active sampling in rank learning. In *ICML*, pages 248–255, 2008.

[9] F. Faria, A. Veloso, H. de Almeida, E. Valle, R. Torres, M. Gonçalves, and W. Meira Jr. Learning to rank for content-based image retrieval. In *Multimedia Information Retrieval*, pages 285–294, 2010.

[10] U. Fayyad and K. Irani. Multi interval discretization of

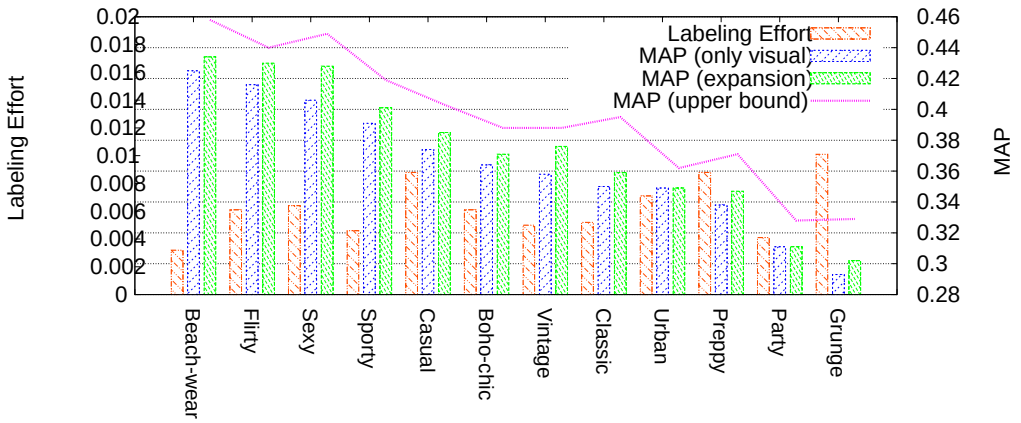

Figure 6: (Color online) Ranking performance and labeling effort associated with different apparel styles.

continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1027, 1993.

[11] V. Ferrari, M. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[12] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*, pages 420–431, 2012.

[13] R. Herbrich, T. Graepel, and K. Obermayer. *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA, 2000.

[14] S. Hidayati, W. Cheng, and K. Hua. Clothing genre classification by exploiting the style elements. In *MM*, pages 1137–1140, 2012.

[15] T. Iwata, S. Wanatabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, pages 2262–2267, 2011.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[17] Y. Kalantidis, L. Kennedy, and L. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*, pages 105–112, 2013.

[18] D. Lee and H. Kim. A fast content-based indexing and retrieval technique by the shape information in large image database. *J. of Systems and Software*, 56(2):165–182, 2001.

[19] L. Li, R. Socher, and F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009.

[20] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337, 2012.

[21] T. Liu. Learning to rank for information retrieval. In *SIGIR*, page 904, 2010.

[22] Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *Learning to Rank Workshop in conjuntion with SIGIR*, 2007.

[23] F. Mahmoudi, J. Shanbehzadeh, A. Eftekhari-Moghadam, and H. Soltanian-Zadeh. Image retrieval based on shape similarity by edge orientation autocorrelogram. *Pattern Recognition*, 36(8):1725–1736, 2003.

[24] F. Palda. *Pareto's Republic and the new Science of Peace*. Cooper-Wolfling, 2011.

[25] E. Shen, H. Lieberman, and F. Lam. What am i gonna wear?: scenario-oriented recommendation. In *IUI*, pages 365–368, 2007.

[26] I. Silva, J. Gomide, A. Veloso, W. M. Jr., and R. Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *SIGIR*, pages 475–484, 2011.

[27] R. Silva, M. Gonçalves, and A. Veloso. Rule-based active sampling for learning to rank. In *ECML/PKDD (3)*, pages 240–255, 2011.

[28] R. Stehling, M. Nascimento, and A. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *CIKM*, pages 102–109, 2002.

[29] M. Tokumaru, M. Muranaka, and S. Imanish. Examination of adapting clothing search system to users subjectiity with interactive genetic algorithms. In *CEC*, pages 1036–1043, 2003.

[30] Q. Tu and L. Dong. An intelligent personalized fashion recommendation system. In *ICCCS*, pages 479–485, 2010.

[31] M. Unser. Sum and difference histograms for texture classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(1):118–125, 1986.

[32] D. Vogiatzis, D. Pierrakos, G. Paliouras, S. Jenkyn-Jones, and B. Possen. Expert and community based style advice. *Expert Syst. Appl.*, 39(12):10647–10655, 2012.

[33] K. Yamaguchi, M. Kiapour, and T. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, page to appear, 2013.

[34] J. Zegarra, N. Leite, and R. Torres. Wavelet-based feature extraction for fingerprint image retrieval. *Journal of Computational and Applied Mathematics*, 2008.

[35] C. Zhe, T. Qin, T. Liu, M. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007.