Multimodal Sparse Linear Integration for Content-based Item Recommendation

Qiusha Zhu[1], Zhao Li[2], Haohong Wang[2], Yimin Yang[3] and Mei-Ling Shyu[1]

[1]Department of Electrical and Computer Engineering, University of Miami
[2]2700 Augustine Drive, TCL Research America
[3]School of Computing and Information Sciences, Florida International University

***Abstract*—Most content-based recommender systems focus on analyzing the textual information of items. For items with images, these images can be treated as another information modality. In this paper, an effective method MSLIM is proposed to integrate multimodal information for content-based item recommendation. It formalizes the probelm into a regularized optimization problem in the least-squares sense and coordinate gradient descent is applied to solve the problem. Aggregation coefficients of items are learned in an unsupervised manner during this process, based on which k-nearest neighbor (k-NN) is used to generate the top-N recommendations of each item by finding its k nearest neighbors. A framework of using MSLIM for item recommendation is proposed accordingly. Experiments on self-collected handbag dataset show that MSLIM outperforms the comparison methods and how the parameters of the model affect the final recommendation results.**

***Index Terms*—recommendation; sparse linear; multimodal integration;**

I. Introduction

Content analysis for multimedia retrieval has been a hot research topic for more than two decades. It includes areas such as semantic concept detection [1], image and video classification [2], recommendation [3] etc. Differing from tradition information retrieval, recommendation goes beyond providing accurate results since recommender systems are centered around users. User profiles or behavior data as an important information are usually preferred comparing to information of item contents since user profiles can better capture their interests and behaviors.

Recommendation approaches can be categorized into three categories: content-based recommendation which focuses on analyzing the content of items; collaborative filtering which utilizes user profiles such as ratings or clicks to recommend items for like-minded users; and hybrid recommendation which incorporate both approaches. Latent factor model (LFM) is the one of the most popular technique in recommendation and has been adopted in many state-of-the-art recommendation models [4][5]. They involve analyzing user profiles, typically in the form of the user-item matrix. Some recently proposed frameworks bring contents of items into consideration, such as [6][7] extend LFM to incorporate item features, and [8] considers item features as side information and integrates them into users' implicit feedbacks. However, in many situations user profiles are not available or very sparse. which is referred as the cold-start problem in recommendation. The pure collaborative filtering methods would fail in these scenarios, and the enriched or extended approaches can only handle the cold-start problem to some extent for they rely on factorizing the user-item matrix or use it to optimize the models. As a result, we have to rely on content-based recommendation before enough user profiles can be collected.

The pure content-based recommendation methods don't need user profiles or behavior data and the recommendation is based on the content of items. Typically features are extracted to represent each item as a feature vector, then item relations are discovered in the extracted feature space or a projected feature space. Currently the content in most recommender systems is limited to descriptions or metadata associated with items, such as text descriptions and tags. While for items with multimedia information, such as images and videos, their visual contents are not often utilized. This is because items are usually organized with correct descriptions available. Meanwhile, there is a "semantic gap" between low-level visual features (e.g., color, edge, texture etc.) and high-level human perception of inferred semantic concepts. However, nowadays with the exponential growth of multimedia data, a large proportion of the data are unorganized which means their textual information could be incomplete or non-existent or even incorrect. For these data, either a lot of efforts need to be spent in manually annotating them or automatical tagging methods have to be applied [9][10] in order to describe them in text. Another option is to explore the visual content directly and utilize the visual information together with the textual information and information of other modalities if available to improve the content-based recommendation.

In this paper, a multimodal sparse linear integration method (MSLIM) is proposed to facilitate content-based item recommendation. Aggregation coefficients of items are learned in an unsupervised manner from multiple modalities and recommendation results are generated using k-nearest neighbor (k-NN) based on the coefficients between items. The contribution of this study can be summarized into two folds:

- An effective yet generic method MSLIM is proposed to integrate multimodalities to mine pair-wise correlations between items.
- A framework adopted MSLIM is proposed accordingly which learns item correlations based on textual informa-

tion from item description and visual information from images, and then applies k-NN for item recommendation.

The rest of paper is organized as follows: Section II discusses related work of addressing this issue. The detailed problem formalization and solution are presented in Section III followed by the experimental results in SectionIV. Conclusion is drawn in Section V.

II. Related Work

Sparse LInear Methods (SLIM) for top-N recommendation are first introduced in [11] which generates recommendation results by aggregating from user purchase or rating profiles. A sparse aggregation coefficient matrix is learned by solving a ℓ_1-norm and ℓ_2-norm regularized optimization problem. The final recommendation is the linear combination of the original user profiles weighted by the learned sparse aggregation coefficients. Later the authors extend SLIM to incorporate item content information [8], but the basic model is the same. The extended method is called SSLIM, which is short for SLIM with item Side information. Experiments on various datasets demonstrate high quality recommendations, and the sparsity of the coefficient matrix allows SLIM to generate recommendations very fast. Compared to SLIM and SSLIM, our proposed method is more generic and can be used in general information retrieval task. The focus in this paper is to utilize multimodalities of item content to handle recommendation scenarios when user profiles are not available. Therefore, rather than using the learned coefficients to linearly combine user profiles as in SLIM or user profiles with item side information as in SSLIM, we directly use the learned coefficient matrix of items to generate the recommendations. Because each entry in the coefficient matrix is the coefficients between items and indicates their similarity.

In the field of multimedia retrieval [12] [13], information of multimodalities have been utilized to complete each other and have shown promising results in tasks such as semantic concept detection, speech recognition, multi-sensor fusion and etc. Its core issue is multimodal fusion. Current methods typically fall into one of the three categories: (1) Early fusion which is feature-level fusion. It involves concatenating features from different modalities which results a simple model but can also easily reach to very high dimensions in the feature space. (2) Late fusion which is decision-level fusion. This category is further divided into the rule-based methods, the estimation-based methods, and the classification-based methods. Compared to early fusion, late fusion offers scalability and allows to choose suitable learning method for each modality. However, it can not utilize the feature-level correlations from different modalities and require to make local decisions first. (3) Hybrid fusion which involves both early fusion and late fusion. More detailed discussion of multimodal fusion can be found in [14]. A comparison between early fusion and late fusion is done in [15], and experiments on broadcast videos for video semantic concept detection show that late fusion tends to slightly outperform early fusion for most concepts, but for those concepts where early fusion performs better, the gain is more significant.

Our proposed MSLIM belongs to the early fusion category, but rather than directly concatenating the features, it learns sparse linear aggregation coefficients between items based on features extracted from multiple modalities. A comparison with rule-based late fusion method is conducted to evaluate MSLIM. Details are presented in Section IV.

III. The Proposed Method

Based on the work in [11] [8], an effective and generic method MSLIM is proposed to integrate multimodal information for content-based top-N recommendation. It aims to learn a sparse coefficient matrix from multimodalities in an unsupervised manner. The problem is formalized into a regularized optimization problem in the least-squares sense and a framework of integrating textual and image visual information for item recommendation is proposed accordingly.

A. *Problem Formalization*

Assuming there are P modalities, and each modality is represented by a feature-item matrix $\boldsymbol{S}^p$ and $p \in [1, P]$, where each row is a feature or an attribute and each column is an item. If there are totally M items and A^p features/ attributes for the p-th modality, then the dimension of $\boldsymbol{S}^p$ is $A^p * M$. Let A denote the dimension of matrix including all the modalities which is equal to $\sum_{p=1}^{p=P} A^p$. An entry in $\boldsymbol{S}^p$ is denoted as s_{ij}^p which could be a nominal or a numeric value for the i-th feature of the j-th item from the p-th modality. The j-th column of $\boldsymbol{S}^p$ is denoted as $\boldsymbol{s}_j^p$ while the i-th row is denoted as $(\boldsymbol{s}_i^p)^T$. In this paper, all the vectors are denoted using bold lower-case letters (e.g., $\boldsymbol{s}_i^p$ and $\boldsymbol{s}_j^p$), and matrices are denoted using bold upper-case letters (e.g. $\boldsymbol{S}$). Upper-case letters are used for dimension of vectors and matrices (e.g. K and $A^p * M$), and lower-case letters are used for indices and entries of vectors or matrices (e.g. s_{ij}^p is an entry and p, i and j are indices).

Each feature in $\boldsymbol{S}^p$ represented by a row vector $\boldsymbol{s}_i^p$, and each entry $\boldsymbol{s}_{ij}^p$ is updated as a sparse aggregation of $\boldsymbol{s}_i^p$. As shown in Equation(1), $\boldsymbol{x}_j$ is a sparse aggregation coefficient column vector of length M. It is learned by integrating information from other modalities $\boldsymbol{S}^l$ where $l \neq p$. The matrix form is shown in Equation(2), and $\boldsymbol{X}$ is the corresponding sparse aggregation coefficient matrix with $\boldsymbol{x}_j$ as its column. $\boldsymbol{X}$ contains $M \times M$ coefficients of items, which are obtained by updating each $\boldsymbol{S}^p$ using other modalities $\boldsymbol{S}^l$ where $l \neq p$. This is achieved by minimizing the difference between $\boldsymbol{S}^p$ and the updated $\boldsymbol{S}^p$, as expressed as Equation(2).

$$(\boldsymbol{s}_{ij}^p)^T \leftarrow (\boldsymbol{s}_i^p)^T \boldsymbol{x}_j \quad (1)$$

$$\boldsymbol{S}^p \leftarrow \boldsymbol{S}^p \boldsymbol{X} \quad (2)$$

The problem can be formalized into an optimization problem presented in Equation(3), where $\| \cdot \|_1$ and $\| \cdot \|_F^2$ are the matrix ℓ_1-norm and Frobenius norm respectively. The term

$\|\boldsymbol{S}^p - \boldsymbol{S}^p\boldsymbol{X}\|_F^2$ measures how well the update fits. The term $\|X\|_F^2$ and $\|X\|_1$ are the ℓ_F-norm and ℓ_1-norm regularization terms, respectively, and β and λ are their regularization parameters. A larger regularization parameter imposes a severe regularization. ℓ_1-norm is introduced to get a sparse solution of $\boldsymbol{X}$ [16], which can make the updating process of Equation(2) very fast, especially when dealing with big data. ℓ_F-norm can prevent model from overfitting. The two regularization terms together lead the optimization problem to an elastic net [17], which balances between the lasso using ℓ_1-norm and ridge regression using ℓ_F-norm. The first constraint $\boldsymbol{X} \geq 0$ ensure all the coefficients of $\boldsymbol{X}$ are non-negative, so that the learned $\boldsymbol{X}$ represents positive relations between items. The second constraint $diag(\boldsymbol{X}) = 0$ is applied to avoid trivial solutions [8], that is the optimal $\boldsymbol{X}$ is an identical matrix such that an item is always best related to itself and not related to any other item.

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \textstyle\sum_{p=1}^{p=P} \frac{\alpha^p}{2}\|\boldsymbol{S}^p - \boldsymbol{S}^p\boldsymbol{X}\|_F^2 \\ & + \frac{\beta}{2}\|\boldsymbol{X}\|_F^2 + \lambda\|\boldsymbol{X}\|_1 \\ s.t. \quad & \boldsymbol{X} \geq 0, \\ & diag(\boldsymbol{X}) = 0 \end{aligned} \tag{3}$$

According to [8], Equation(3) can be decoupled into a set of the same optimization problems since each column of $\boldsymbol{X}$ is independent from each other. Therefore, solving Equation(3) is equivalent to solve Equation(4), where $\boldsymbol{S} = [\sqrt{\alpha^1}\boldsymbol{S}^1, \cdots, \sqrt{\alpha^P}\boldsymbol{S}^P]^T$, and it can be parallelized using each processor to handle a column of $\boldsymbol{X}$. Let Y denote the cost function in Equation(4). Using coordinate descent [18], the partial derivative of $\boldsymbol{Y}$ with respect to the i-th entry of $\boldsymbol{x}_j$ is derived as Equation(5). Therefore, x_{ij} can be calculated accordingly if let the partial derivative equal to 0. The update of x_{ij} is shown in Equation(6), where Υ is the soft-thresholding operator. The aggregation coefficients of items calculated by integrating multiple modalities are represented by the $M \times M$ matrix $\boldsymbol{X}$. For each item, its neighbors are defined as items having large coefficients with this item, and thus k-NN can be adopted as the recommendation algorithm. In other words, content-based recommendation is achieved by obtaining similar items from multiple modalities.

$$\begin{aligned} \min_{\boldsymbol{x}_j} \quad & \tfrac{1}{2}\|\boldsymbol{s}_j - \boldsymbol{S}\boldsymbol{x}_j\|_2^2 \\ & + \tfrac{\beta}{2}\|\boldsymbol{x}_j\|_2^2 + \lambda\|\boldsymbol{x}_j\|_1 \\ s.t. \quad & \boldsymbol{x}_j \geq 0, \\ & x_{jj} = 0 \end{aligned} \tag{4}$$

$$\begin{aligned} & \frac{\partial Y}{\partial x_{ij}} \\ & = -\sum_{h=1}^{h=A} s_{hi}(s_{hj} - \sum_{g=1}^{g=M} s_{hg}x_{gj}) + \beta x_{ij} + \lambda \\ & = -\sum_{h=1}^{h=A} s_{hi}(s_{hj} - \sum_{g\neq i} s_{hg}x_{gj} - s_{hi}x_{ij}) + \beta x_{ij} + \lambda \\ & = -\sum_{h=1}^{h=A} s_{hi}(s_{hj} - \sum_{g\neq i} s_{hg}x_{gj}) + \sum_{h=1}^{h=A} s_{hi}^2 x_{ij} + \beta x_{ij} + \lambda \\ & = -\sum_{h=1}^{h=A} s_{hi}(s_{hj} - \sum_{g\neq i} s_{hg}x_{gj}) + (\sum_{h=1}^{h=A} s_{hi}^2 + \beta)x_{ij} + \lambda \end{aligned} \tag{5}$$

$$\begin{aligned} & x_{ij} \leftarrow \frac{\Upsilon(\sum_{h=1}^{h=A} s_{hi}(s_{hj} - \sum_{g\neq i} s_{hg}x_{gj}), \lambda)}{\sum_{h=1}^{h=A} s_{hi}^2 + \beta}, \\ & where \quad \Upsilon(z,\lambda) = \begin{cases} z - \lambda & \text{if } z > 0 \text{ and } |z| > \lambda \\ z + \lambda & \text{if } z < 0 \text{ and } |z| > \lambda \\ 0 & \text{if } |z| \leq \lambda \end{cases} \end{aligned} \tag{6}$$

The proposed MSLIM can incorporate user profiles if available, which is equivalent to the method in [8] if considering all the content information of items as side information. If using $\boldsymbol{U}$ to denote users' implicit or explicit feedbacks, then there is an addition term in Equation 3 which is $\mu\|\boldsymbol{U} - \boldsymbol{U}\boldsymbol{X}\|_F^2$. The solution is the same as depicted in Equation 4 with $\boldsymbol{S}$ replaced by $[\sqrt{\mu}\boldsymbol{U}, \boldsymbol{S}]^T$.

To evaluate the efficiency of MSLIM is to analyze the computational complexity of Equation(2). The updating of $\boldsymbol{S}$ or the computing of $\boldsymbol{X}$ is $O(n_{r_s} \times A \times n_{c_x})$, where n_{r_s} is the average number of non-zero values in the rows of $\boldsymbol{S}$, A is the number of rows of $\boldsymbol{S}$, and n_{c_x} is the number of columns of $\boldsymbol{X}$. Therefore the computational complexity of MSLIM depends on the feature sparsity of the initial $\boldsymbol{S}$ and the number of features from different modalities as well as the number of items. Considering the independent property of the columns of $\boldsymbol{X}$ as presented in Equation(4), the computation of $\boldsymbol{X}$ can be parallelized, which can reduce the computational complexity can be reduced to $O(n_{r_s} \times A)$ if using n_{c_x} processes.

B. A System Framework

A framework utilizing MSLIM for item recommendation is presented. One modality is the textual information of items, and the other modality is the visual information of items. In this framework, they are item descriptions and images respectively. Therefore, in Equation(6), $\boldsymbol{S} = [\sqrt{\alpha^1}\boldsymbol{S}^1, \sqrt{\alpha^2}\boldsymbol{S}^2]^T$, where $\boldsymbol{S}^1$ is the textual feature-item matrix, and $\boldsymbol{S}^2$ is the visual feature-item matrix.

1) Feature Extraction: Visual features of images includes color, texture, shape etc, such as HSV color histogram, histogram of oriented gradients (HOG) and popular scale-invariant feature transform (SIFT) [19]. The aim of this paper is not to compare various features, but to validate the effectiveness of MSLIM which can improve content-based recommendation by integrating features from difference modalities. Hence, we only extract one visual feature which is CEDD [20] feature due to its good balance between accuracy and complexity. It is a compact composite descriptor that incorporates both color and texture information in one histogram. The CEDD extraction system is composed of two units, texture unit and color unit, and three fuzzy systems. First, the image is separated into a preset number of blocks (usually 1600 for compromising between the image detail and the computational complexity). Then each of the blocks passes through all the units as follows. In the texture unit, each image block is classified into one of the 6 texture categories by applying with 5 digital filters. In the color unit, the image block is converted into the HSV color space and fed into two fuzzy systems with a set of rules, obtaining a 24-bins histogram (with each bin representing one color). Finally, the overall histogram contains $6 \times 24 = 144$ regions.

For textual features, we extract keywords/ terms from descriptions of items, and use binary value to represent the presence of a feature. Take feature "leather" for example, 1 means "leather" exists in the item's descriptions while 0 means the opposite. For descriptions in English, standard procedures such as stop word removal and stemming are usually applied to preprocess the terms. WordNet [21] can also be used to validate English words due to ubiquitous typos, especially in user-contributed social media data such as image tags from Flickr. The descriptions we collected for our bag dataset are in Chinese, and more details are given in Section IV-A. There are totally 509 binary features extracted which cover materials, brands, colors, styles, structure and etc.

Normalization is performed on extracted features to convert their scales and to ensure that they are suitable for general data analysis. For visual features, min-max normalization is adopted to scale the feature values between 0 to 1. For textual features, we do not apply any normalization since the extracted features are binary.

2) A Practical Framework: Features extracted from each modality are fed into the sparse linear integration module and generate the aggregation coefficients of items. Then k-nearest neighbor (k-NN) is used to find the neighbors of each item based on the aggregation coefficients. Recommendation results are generated by sorting the similarity scores of neighbors in descending order. A framework summarized these procedures is presented in Figure 1. Textual features and visual features are extracted from descriptions and images of items respectively. If there are other modalities, such as descriptions from other websites or images from a different view point, then features can be extracted accordingly and fed into the sparse linear integration module.

IV. Experiments

To evaluate MSLIM for content-based recommendation, images and descriptions of handbags are collected to perform handbag recommendation, and user rating data are collected as ground truth for judgement. We first investigate how parameters affect the recommendation results, and then use the best parameter settings to conduct further comparison. The comparison includes methods using a single modality, as well as the method which linearly combines the results from each single modality. To ensure fair comparison, k-NN is used as the recommendation algorithm for all methods.

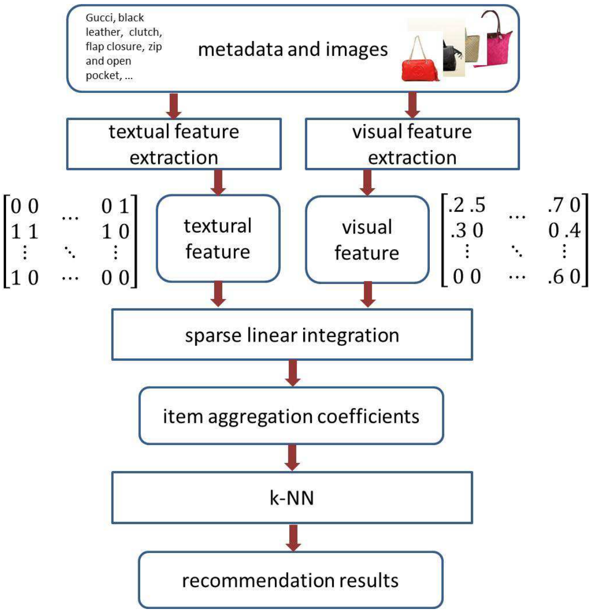
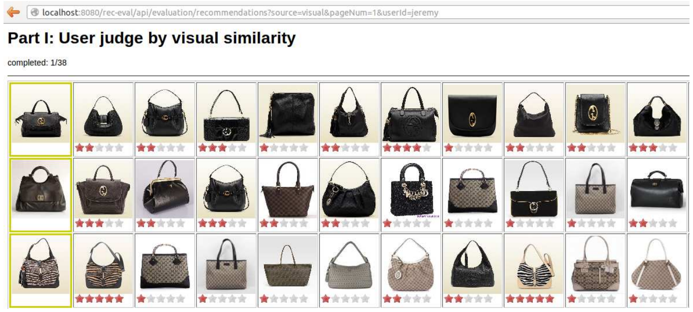

Gucci, black leather, clutch, flap closure, zip and open pocket, ...
metadata and images
textual feature extraction
visual feature extraction
0 0 … 0 1
1 1 1 0
⋮ ⋱ ⋮
1 0 … 0 0
textural feature
visual feature
.2 .5 … .7 0
.3 0 0 .4
⋮ ⋱ ⋮
0 0 … .6 0
sparse linear integration
item aggregation coefficients
k-NN
recommendation results

Fig. 1. The MSLIM framework for content-based multimedia recommendation

A. Data Collection

The dataset of handbags are collected based on the bags appeared in a fashion show video which is created by gluing parts from several videos. Specifically, the first sequence (0-27 sec) is the Gucci fall-winter 2010 women's wear show; the second sequence (28-58 sec) is from some advertising videos downloaded from the Gucci web site; the rest of the video (59-end) is the Prada fall-winter 2012 women's wear show. Google image search based on keywords such as "prada fall winter 2012 women's wear show handbag" is used to find the exact bags appeared in the video. Next Google image search using the images from the keyword search results is utilized to find visually similar bags which form the recommendation dataset. There are totally 440 bags, and both the images and their descriptions are collected as two modalities. Since the descriptions of bags are relatively structured, that is they are described in similar way such as materials, brands, color, styles etc., thus textual features are extracted based on this structure.

To collect the ground truth, we design a web-based interface for users to provide ratings. Each bag is used as a target item which means the bag is the one the user is interested in or wants to purchase. 20 other bags are recommended to the user for him or her to rate from 0 to 5. 0 means the user is not interested in the recommended bag while 5 means the recommended bag is very similar to the target bag. There are actually two parts in this user judgement process. The first part is using visual information alone and presenting the images of the top 20 recommended bags to the users, as showed in Figure 2. The second part is adding textual information and both the important descriptions and the images of the top 20 recommended bags are presented to the users, as showed in Figure 3. In both web interfaces, the target bag is the first one in each row which is highlighted in yellow box. Only the first top 10 recommended bags can be seen from the figures due to the size of the window, but there are actually 20 bags in each row. The reason we design it in two parts is to avoid bias when judging using visual or textual information alone. 11 users participate this judging task, and ratings from both parts are collected. For each target bag, its recommended bags with an average rating equal to or above 3.0 is considered as a relevant recommendation which indicates the interest from users.

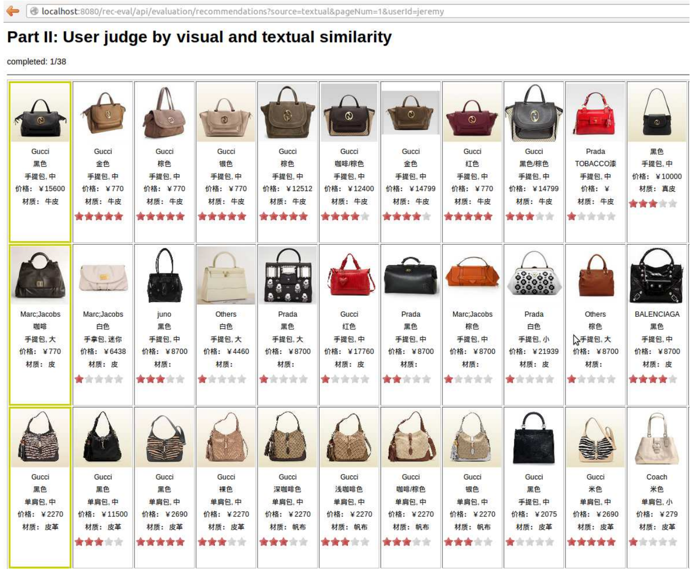


localhost:8080/rec-eval/api/evaluation/recommendations?source=visual&pageNum=1&userId=jeremy

Part I: User judge by visual similarity

completed: 1/38

Fig. 2. The interface of collecting ratings for bags using visual information



localhost:8080/rec-eval/api/evaluation/recommendations?source=textual&pageNum=1&userId=jeremy

Part II: User judge by visual and textual similarity

completed: 1/38

Gucci 黑色 手提包, 中 价格： ¥15600 材质： 牛皮	Gucci 金色 手提包, 中 价格： ¥770 材质： 牛皮 ★★★★★	Gucci 棕色 手提包, 中 价格： ¥770 材质： 牛皮 ★★★★★	Gucci 银色 手提包, 中 价格： ¥770 材质： 牛皮 ★★★★★	Gucci 棕色 手提包, 中 价格： ¥12512 材质： 牛皮 ★★★★★	Gucci 咖啡/棕色 手提包, 中 价格： ¥12400 材质： 牛皮 ★★★★☆	Gucci 金色 手提包, 中 价格： ¥14799 材质： 牛皮 ★★★★☆	Gucci 红色 手提包, 中 价格： ¥770 材质： 牛皮 ★★★★★	Gucci 黑色/棕色 手提包, 中 价格： ¥14799 材质： 牛皮 ★★★☆☆	Prada TOBACCO漆 手提包, 中 价格： ¥ 材质： 牛皮 ★☆☆☆☆	黑色 手提包, 中 价格： ¥10000 材质： 真皮 ★★★☆☆
Marc;Jacobs 咖啡 手提包, 大 价格： ¥770 材质： 皮	Marc;Jacobs 白色 手拿包, 迷你 价格： ¥6438 材质： 皮 ★☆☆☆☆	juno 黑色 手提包, 中 价格： ¥8700 材质： ★★★☆☆	Others 白色 手提包, 大 价格： ¥4460 材质： ★☆☆☆☆	Prada 黑色 手提包, 大 价格： ¥8700 材质： ★☆☆☆☆	Gucci 红色 手提包, 中 价格： ¥17760 材质： 皮 ★☆☆☆☆	Prada 黑色 手提包, 中 价格： ¥8700 材质： ★★☆☆☆	Marc;Jacobs 棕色 手提包, 中 价格： ¥8700 材质： ★☆☆☆☆	Prada 白色 手提包, 小 价格： ¥21939 材质： 皮 ★☆☆☆☆	Others 棕色 手提包, 大 价格： ¥8700 材质： ★☆☆☆☆	BALENCIAGA 黑色 手提包, 中 价格： ¥8700 材质： 皮 ★★★★☆
Gucci 黑色 单肩包, 中 价格： ¥2270 材质： 皮革	Gucci 黑色 单肩包, 中 价格： ¥11500 材质： 皮革 ★★★☆☆	Gucci 黑色 单肩包, 中 价格： ¥2690 材质： 皮革 ★★★★★	Gucci 裸色 单肩包, 中 价格： ¥2270 材质： 皮革 ★★★☆☆	Gucci 深咖啡色 单肩包, 中 价格： ¥2270 材质： 帆布 ★★★☆☆	Gucci 浅咖啡色 单肩包, 中 价格： ¥2270 材质： 帆布 ★★★☆☆	Gucci 咖啡/棕色 单肩包, 中 价格： ¥2270 材质： 帆布 ★★★☆☆	Gucci 银色 单肩包, 中 价格： ¥2270 材质： 帆布 ★★★☆☆	Gucci 黑色 手提包, 中 价格： ¥2075 材质： 皮革 ★☆☆☆☆	Gucci 米色 单肩包, 中 价格： ¥2690 材质： 皮革 ★★★★★	Coach 米色 单肩包, 小 价格： ¥279 材质： 皮革 ★☆☆☆☆

Fig. 3. The interface of collecting ratings for bags using visual and textual information

B. Evaluation Metrics

For evaluation metrics, precision is the most commonly used metric in the top-N recommendation. It measures the percentage of correctly predicted items, denoted as prec@n, and can be calculated based on Equation (7). TP is the number of relevant items or true positive (TP) in the n recommended items, and n is usually set to 5 or 10.

$$prec@n = \frac{TP}{n} \tag{7}$$

Another widely adopted metric is the area under the ROC curve (AUC), which is a more general ranking measure. It is equal to the probability that a model will rank a randomly chosen positive item higher than a randomly chosen negative one. The best possible value of AUC is 1, and any non-random ranking that makes sense would have an AUC value greater than 0.5. Other common metrics for ranking include mean average precision (MAP) and normalized discounted cumulative gain (NDCG), which are also used in this paper. MAP is the mean of the average precision scores for each query, while average precision (AP) is computed as a function of recall, as shown in Equation (8). Q is the total number of queries, FN is the number of false negative, and $\Delta(t)$ is an indicator function equaling 1 if the item at rank t is a relevant one, zero otherwise. NDCG measures the gain of an item based on its position in the result list according to Equation (9). rel_t is the graded relevance of an item at position t, and IDCG is the ideal DCG which is the maximum possible value of DCG.

$$\begin{aligned} MAP &= \tfrac{\sum_{q=1}^{q=Q} AP(q)}{Q} \\ where \quad AP(q) &= \tfrac{\sum_{t=1}^{t=n} prec@t \times \Delta(t)}{TP+FN} \end{aligned} \tag{8}$$

$$\begin{aligned} NDCG &= \tfrac{DCG}{IDCG} \\ where \quad DCG &= rel_1 + \sum_{t=1}^{t=n} \tfrac{rel_t}{\log_2(t)} \end{aligned} \tag{9}$$

C. Compared Methods

The proposed MSLIM is first compared with methods using the same k-NN recommendation algorithm but a single modality in order to prove that the integration of multiple modalities helps the final recommendation. The method using textual information only is denoted as textual method (TM) while visual method (VM) denotes the method using visual information only. However, to further evaluate the improvement, a rule-based late fusion method is adopted as a comparison method. It uses Equation (10) to linearly weight and combine the recommendation scores from each modality, where w^p is the weight of modality S^p and $f(\cdot)$ is a recommendation algorithm which takes feature-item matrix as input and outputs the recommendation scores. As mentioned before, $P = 2$ since there are two modalities in our dataset. This method is denoted as LWM, which stands for linear weighted method.

$$\begin{aligned} &\textstyle\sum_{p=1}^{p=P} \quad w^p \times f(\boldsymbol{S}^p) \\ &where \quad \textstyle\sum_{p=1}^{p=P} w^p = 1 \end{aligned} \tag{10}$$

D. Experimental 1: Parameter Tuning

The parameters involved in MSLIM are α^1, α^2, β and λ a shown in Equation (3) where $P = 2$. There are no paramete in TM and VM, and for LWM, the parameters are the weigh of textual modality w^1 and the weights of visual modality w^2

Let's start from LWM first. We decrease the value of w from 1 to 0 with step of 0.1, and the value of w^2 is increase from 0 to 1 with step of 0.1 accordingly. Figure 4 presents th performance of LWM using the aforementioned 7 metrics wit the weight of visual modality w^2 increasing from 0 to 1. A can be seen, AUC reaches the highest when $w^2 = 0.2$, and it value at 0.1 is relatively high. While for the rest metrics, thei values slightly increase or stay the same when w^2 increas from 0 to 0.1, and drop dramatically when w^2 continue increasing from 0.1 to 0.2. Therefore, we choose w^2 to b 0.1 by considering the performance on all the metrics. Th value of w^1 is set to 0.9 to ensure their summation is equal to 1. These parameters indicate that the information from textual modality is more reliable or accurate since w^1 is much larger than w^2.

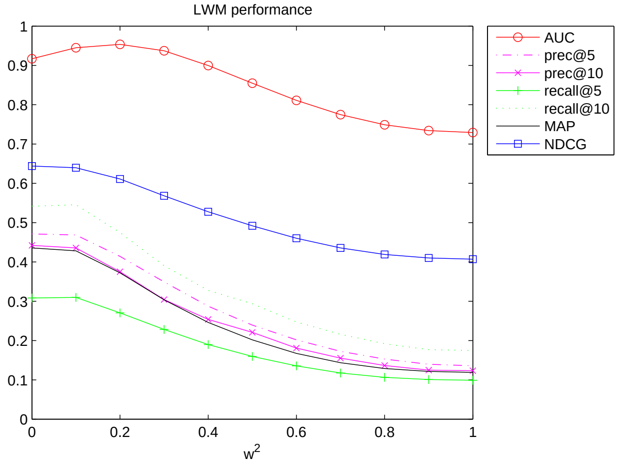


Fig. 4. The performance of LWM varied by w^2

For MSLIM, we first tune the parameter of α^1 and α^2 with fixed β and λ since α^1 and α^2 play a local role in integrating textual and visual modality. Therefore, we set $\beta = 1.0$ and $\lambda = 0.01$ empirically first, and fix α^1 to 1.0 and only vary α^2. Figure 5 shows the performance of MSLIM with α^2 set to $\{0.1, 0.5, 1.0, 2.0, 5.0\}$ and the other parameters are set to the fixed values. As shown in the figure, on average, $\alpha^2 = 0.5$ gives the best overall performance when considering all the metrics.

The next step is fixing α^1 and α^2 to the optimal value we find which are 1.0 and 0.5 respectively, and then using grid search to find the optimal value of β and λ. The search range for β is from 0.001 to 10.0 with points at $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$, and the search points for λ are at $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. Figure 6 and Figure 7 display the performance using AUC and prec@5 from different views with β and λ as two variables. The performances varied by β and λ depict similar pattern on the rest metrics. As can be seen from both figures, there is a big decrease when λ keeps increasing after 0.01, and the performances drop to 0 after λ reaches beyond 0.05. It's because the ℓ_1-norm parameter λ controls the sparsity of the coefficient matrix. If λ is too large which means high sparsity, then there is no item can be recommended since the coefficients with the target item are all 0. For β, the performances increase from a relatively low value to the maximum when β increases from 0.001 to 1.0, and stays almost stable when β keeps increasing from 1.0 to 10. This indicates a small ℓ_2-norm regularization improves model performances but after a certain point, in this case when $\beta = 1.0$, it doesn't affect the performances anymore. From both figures, we can see the maximum performance forms a flat area bounded by $\lambda \in (0, 0.01]$ and $\beta \in [1.0, 10]$. Hence, we fix λ to its upper bound 0.01, and β to its lower bound 1.0 as the empirical values we decide when tuning α^1 and α^2. In fact, any value of λ and β within the aforementioned boundaries could assure the maximum model performances.

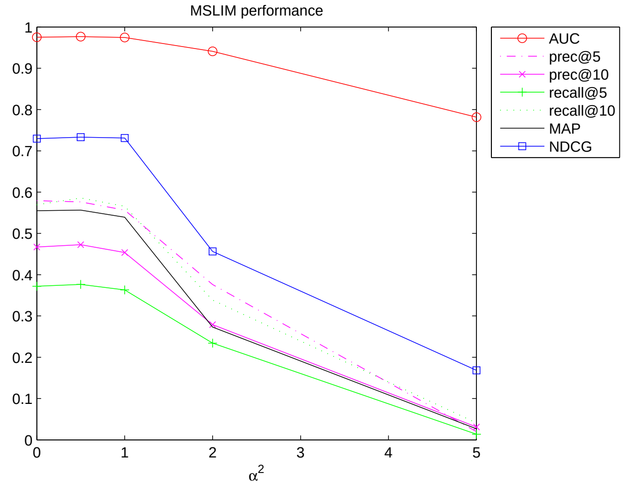


Fig. 5. The performance of MLIMS varied by α^2

E. Experimental 2: Comparison Results

The comparison results of MSLIM against TM, VM, and LWM are shown in Figure 8. VM using visual information results inferior performance compared to the other three methods. One reason is that we only use one visual feature which is CEDD. It achieves relatively good performance compared to other visual features, but one single visual feature is very limited. If introducing more types of visual features, the performance of VM would be better. The other reason is that the semantic gap between low-level visual features and high-level semantic concepts. Take the brand of a bag for example, it's not easy to capture the pattern of a brand using visual features, but from the textual point of view, the exact words of a brand are probably already contained in the item descriptions. MSLIM achieves the best results on all the metrics, followed by LWM and then TM. Its absolute

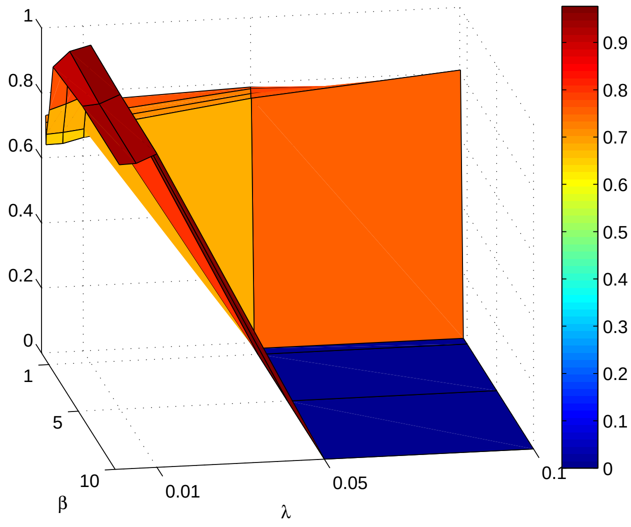


Fig. 6. AUC of MSLIM varied by β and λ

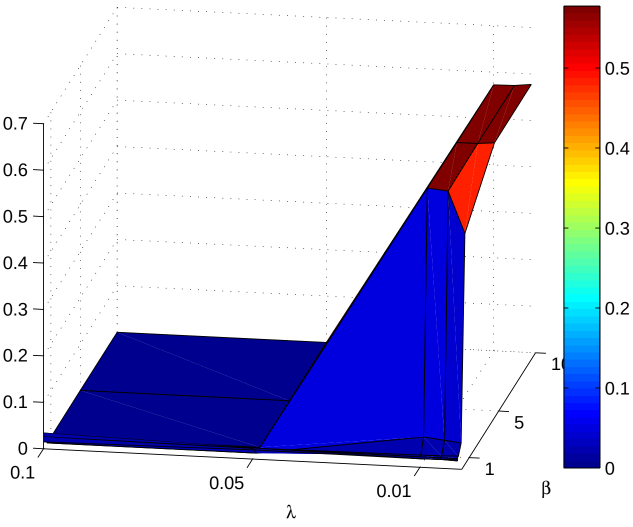


Fig. 7. prec@5 of MSLIM varied by β and λ

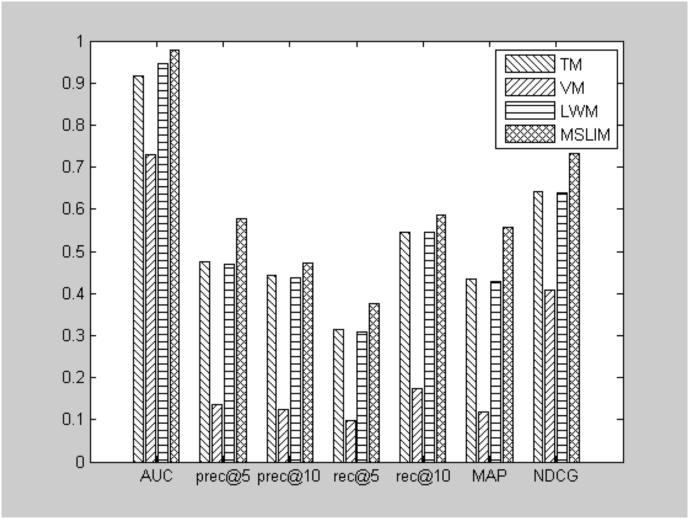


Fig. 8. The comparison performance

TABLE I
IMPROVEMENTS BY MSLIM COMPARED TO TM, VM AND LWM

Metric	TM	VM	LWM
AUC	0.0600	0.2475	0.0316
prec@5	0.0998	0.4403	0.1076
prec@10	0.0288	0.3494	0.0371
rec@5	0.0630	0.2777	0.0669
rec@10	0.0404	0.4112	0.0401
MAP	0.1229	0.4374	0.1283
NDCG	0.0911	0.3260	0.0932
avg	0.0723	0.3556	0.0721

improvements compared to TM, VM, LWM are summarized in Table I. The last row shows the average increase over all the seven metrics. MSLIM outperforms VM by a large margin, which is 0.3556, so does TM and LWM, but with a slightly smaller margin. The results from TM and LWM are close, and are outperformed by MSLIM by about 0.072 on average.

V. Conclusion

In this paper, a multimodal sparse linear integration method MSLIM is proposed for content-based item recommendation. It formalizes the integration problem into a regularized optimization problem in the least-squares sense. Coordinate gradient descent is applied to solve the problem and parallel computing can be used to speed up the process. Aggregation coefficients of items are learned in an unsupervised manner during this process, based on which k-nearest neighbor (k-NN) is used to calculate the neighbors and generate the top-N recommendation results. Evaluation compares MSLIM with other three methods and proves its effectiveness on a handbag dataset.

One limitation of MSLIM is that the number of features from different modalities should be in the similar scale, otherwise the aggregation coefficients learned would lean toward the modality with more features and thus contain more information from it. To solve this issue, one option is to apply feature selection technique to reduce feature dimensions and make sure the features from different modalities are in the same scale. Another limitation of our current work is that we learn the full $M \times M$ item coefficient matrix, which is tedious for top-N recommendation. Instead, we can only take the necessary neighbors into consideration.

References

[1] T. Meng and M.-L. Shyu, “Leveraging concept association network for multimedia rare concept mining and retrieval,” in *IEEE International Conference on Multimedia and Expo*, July 2012, pp. 860–865.

[2] Q. Zhu, L. Lin, M.-L. Shyu, and D. Liu, “Utilizing context information to enhance content-based image classification,” *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 2, no. 3, pp. 34–51, 2011.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[4] S. Rendle and L. Schmidt-Thieme, "Online-updating regularized kernel matrix factorization models for large-scale recommender systems," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 251–258.

[5] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 880–887.

[6] D. Agarwal and B.-C. Chen, "Regression-based latent factor models," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 19–28.

[7] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, and L. Schmidt-Thieme, "Learning attribute-to-feature mappings for cold-start recommendations," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, 2010, pp. 176–185.

[8] X. Ning and G. Karypis, "Sparse linear methods with side information for top-n recommendations," in *Proceedings of the sixth ACM conference on Recommender systems*, 2012, pp. 155–162.

[9] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones, "Automatic tagging and geotagging in video collections and communities," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011, pp. 51:1–51:8.

[10] S. Siersdorfer, J. San Pedro, and M. Sanderson, "Automatic video tagging using content redundancy," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 395–402.

[11] X. Ning and G. Karypis, "Slim: Sparse linear methods for top-n recommender systems," in *2011 IEEE 11th International Conference on Data Mining (ICDM)*, 2011, pp. 497–506.

[12] D. Liu and M.-L. Shyu, "Effective moving object detection and retrieval via integrating spatial-temporal multimedia information," in *Proceedings of the 2012 IEEE International Symposium on Multimedia*, 2012, pp. 364–371.

[13] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology*, 2013, in press.

[14] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, 2010.

[15] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 58, pp. 267–288, 1996.

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, 1999, pp. 1150–1157.

[20] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Proceedings of the 6th international conference on Computer vision systems*, 2008, pp. 312–322.

[21] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.