Augmenting a Content-based Recommender System with Tags for Cultural Heritage Personalization

Pierpaolo Basile, Fabio Calefato, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, Massimo Bux, Cataldo Musto, and Fedelucio Narducci

Universitá degli Studi di Bari, Dipartimento di Informatica
via E. Orabona, 4 - 70126 - Italy
{basilepp,calefato,degemmis,lops,semeraro,bux,musto,narducci}@di.uniba.it

Abstract. Cultural heritage personalization and Web 2.0 joint research efforts have recently emerged in the attempt to build social and collaborative approaches to solve the problem of filtering content in the context of art museums. One way to tackle the problem of recommending artifacts to visitors is to take into account not only the official textual descriptions, but also the user-generated content, namely the tags, which visitors could use to freely annotate relevant works. The main contribution of the paper is a strategy that enable a content-based recommender system to infer user interests by using machine learning techniques both on static content and tags. Experiments were carried out by involving real users who annotated paintings from the Vatican picture-gallery. The main outcome is an improvement in the predictive accuracy of the tag-augmented recommender system compared to a pure content-based approach.

1 Introduction

The importance of providing digital access to cultural heritage collections has been already acknowledged by museums for almost four decades [1]. More recently, museums have also recognized the importance of providing visitors with personalized access to artifacts [2]. Cultural heritage personalization refers to supporting visitors in the selection and filtering of preferred artifacts and their corresponding descriptions, and in the creation of personalized tours. For example, the PEACH project (Personal Experience with Active Cultural Heritage) [3] is a joint Italian-Israeli research collaboration for intelligent information presentation in museums. The goal of PEACH is to build an active, multimedia visitor guide, with strong personalization of all the information provided, so as to ensure that visitors, by expressing their affective attitude, are allowed to accommodate the museum tour according to their own interests and pace.

Because recommender systems have proved to be useful in helping users access to desired information (especially in domains where they are not expert or familiar with), they have found their way also in the context of museums, to support visitors in fulfilling a personalized experience and tour when visiting artworks collections. For instance, the CHIP project (Cultural Heritage Information Personalization) [4] is a research effort for enhancing personalized access to the collections of the Rijksmuseum in Amsterdam. CHIP combines Semantic Web technologies and content-based algorithms for deducing visitors' preference from a set of scored artifacts and then, recommending other artworks and related content topics. In particular, the recommendations of artworks are based on three properties, namely author, genre, and period.

When providing recommendations in cultural heritage context, information about collections must be taken into account because it can be as important as the artifacts themselves. Furthermore, the recent Web 2.0 (r)evolution has radically changed the role of people from passive consumers of information to that of active contributors who create and share new content. One of the forms of user-generated content (UGC) that has drawn more attention from the research community is *tagging*, which is the act of annotating resources of interests with free keywords, called *tags*, thus building a socially-constructed classification schema, called a *folksonomy* (folks + taxonomy). The Steve.museum consortium [5] has begun to explore the use of social tagging and folksonomy in cultural heritage personalization scenario, to increase audiences engagement with museums' collections. Supporting social tagging of artifacts and providing access based on the resulting folksonomy open museum collections to new interpretations, which reflect visitors' perspectives rather than curators' ones, and helps to bridge the gap between the professional language of the curator and the popular language of the museum visitor. Preliminary explorations conducted at the Metropolitan Museum of Art of New York have shown that professional perspectives differ significantly from those of naïve visitors. Hence, if tags are associated to artworks, the resulting folksonomy can be used as a different and valuable source of information to be carefully taken into account when providing recommendations to museum visitors. In this paper we have begun to investigate how to effectively combine existing content-based filtering algorithms with UGC, in the context of cultural heritage personalization. The goal of the paper can be formulated in form of a research question as follows:
In the context of cultural heritage personalization, does the integration of UGC (i.e., tags) cause an increase of the prediction accuracy in the process of recommending artifacts to users?

Content-based recommender systems analyze a set of documents, previously rated by an individual user, and learn a model or profile of user interests based on the *features* of the documents rated by that user[6]. The profile is exploited to recommend new relevant items. This paper presents an approach in which the process of learning user profiles is performed both on static content and UGC. This research was conducted within the CHAT project (Cultural Heritage fruition & e-learning applications of new Advanced multimodal Technologies), that aims at developing new systems and services for multimodal fruition of cultural heritage content. We gathered data from the collections of the Vatican picture-gallery, for which both images and detailed textual information of paintings were available, and letting users involved in the study both rate and annotate them with tags.

The paper is structured as follows. Section 2 provides a description of our recommender system and how it handles users' tagging activity when building user profiles. Section 3 provides the description of the experimental session carried out to evaluate the proposed idea, and a discussion of the main findings. Finally, Section 4 draws conclusions and provides directions for future work.

2 A Content-based Recommender System handling User Tags

ITem Recommender (ITR) [7] is a content-based recommender system, developed at the University of Bari. The inceptive idea behind this paper is to include folksonomies in ITR by integrating *static* content describing the artworks of the collection with *dynamic* user-generated content. Tags are collected during the training step, by letting users: 1) express their preferences for items by entering a numerical rating and 2) annotate rated items with free tags.

Figure 1 shows the general architecture of ITR. The recommendation process is performed in three steps, each of which is handled by a separate component. First, given a collection of documents, a preprocessing step is performed by the *Content Analyzer*, which uses the WordNet lexical database to perform *Word Sense Disambiguation* (WSD) on both static and dynamic content to identify correct senses, corresponding to concepts identified from words in the text. Then, a learning step is performed by the *Profile Learner* on the training set of documents, to generate a probabilistic model of the user interests. This model is the personal profile including those concepts that turn out to be most indicative of the user's preferences. Finally, the *Recommender* component implements a naïve Bayes text categorization algorithm, which is able to classify new documents as interesting or not for a specific user by exploiting the probabilistic model learned from training examples.

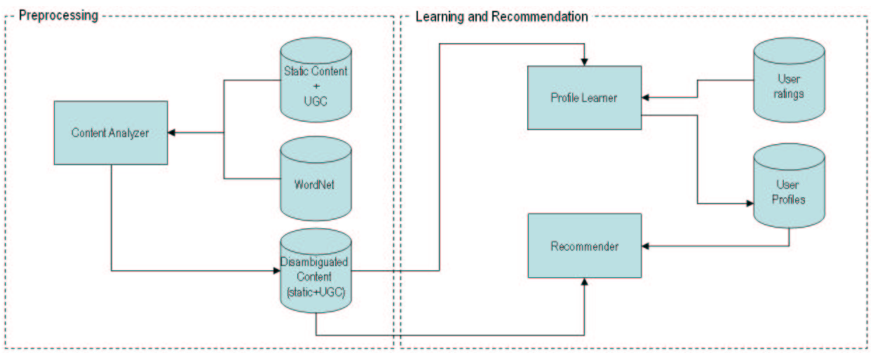

Fig. 1. ITR architecture

2.1 Content Analyzer: Semantic Indexing of Static and Dynamic Content

We propose a document representation that can be exploited as a starting point to build *semantic* user profiles based on the senses (meanings) of words found in the training documents.

There are two crucial issues to address: First, a repository for word senses has to be identified; second, any implementation of sense-based document representation must solve the problem that, although words occur in a document, meanings do not, since they are often hidden in the context. Therefore, a procedure is needed for assigning senses to words: The task of WSD consists in determining which sense of an ambiguous word is invoked in a particular use of the word [8]. As for the sense repository, we adopted WordNet version 2.0. The basic building block for WordNet is the synset (SYNonym SET), a structure containing sets of words with synonymous meanings, which represents a specific meaning of a word. Our WSD algorithm, called JIGSAW, takes as input a document $d = [w_1, w_2, \ldots, w_h]$ encoded as a list of words in order of their appearance, and returns a list of WordNet synsets $X = [s_1, s_2, \ldots, s_k]$ $(k \leq h)$, in which each element s_j is obtained by disambiguating the *target word* w_i based on the *semantic similarity* of w_i with the words in its context. Notice that $k \leq h$ because some words, such as proper names, might not be found in WordNet, or because of bigram recognition.

Semantic similarity computes the relatedness of two words. We adopted the Leacock-Chodorow measure [9], which is based on the length of the path between concepts in a IS-A hierarchy. Since WSD in not the focus of the paper, we do not provide here the complete description of the strategy adopted. More details are reported in [10]. What we would like to point out here is that the WSD procedure allows to obtain a synset-based vector space representation, called bag-of-synsets (BOS), that is an extension of the classical bag-of-words (BOW) model. In the BOS model a synset vector, rather than a word vector, corresponds to a document.

The ITR system is capable of providing recommendations for items in any domain (e.g., films, music, books), as long as item properties can be represented in form of *textual slots*. Hence, in the context of cultural heritage personalization, an artwork can be generally represented by at least three slots, namely *artist*, *title*, and *description*. Besides, provided that museum visitors have a digital support to annotate artifacts, tags can be easily stored in a fourth slot, say *tags*, which is not static as the other three slots because tags evolve over time.

In systems supporting social tagging, the number of tags used to annotate a given resource tend to grow initially, and then to decrease because users tend to reuse existing tags, especially the most common ones. This phenomenon is known as tag convergence [11]. However, being free annotations, tags also tend to suffer from syntactic problems, like polysemy and synonymy, which hinder tag convergence. One way to cope with such a problem is to apply WSD to tags as well. This process allows the document representation model to evolve from using tags as mere keywords or strings, to using *semantic tags* and, consequently, *semantic folksonomies* of concepts.

The text in each slot is represented by the BOS model by counting separately the occurrences of a synset in the slots in which it appears. More formally, assume that we have a collection of N documents. Let m be the index of the slot, for $n = 1, 2, ..., N$, the n-th document is reduced to four bag of synsets, one for each slot:

$$d_n^m = \langle t_{n1}^m, t_{n2}^m, \ldots, t_{nD_{nm}}^m \rangle$$

where t_{nk}^m is the k-th synset in slot s_m of document d_n and D_{nm} is the total number of synsets appearing in the m-th slot of document d_n. For all n, k and m, $t_{nk}^m \in V_m$, which is the vocabulary for the slot s_m (the set of all different synsets found in slot s_m). Document d_n is finally represented in the vector space by four synset-frequency vectors:

$$f_n^m = \langle w_{n1}^m, w_{n2}^m, \ldots, w_{nD_{nm}}^m \rangle$$

where w_{nk}^m is the weight of the synset t_k in the slot s_m of document d_n and can be computed in different ways: it can be simply the number of times synset t_k appears in slot s_m or a more complex TF-IDF score. All the text operations performed on documents are provided by a NLP tool developed at University of Bari, called META [12]. Our idea is that BOS-indexed documents can be used in a content-based information filtering scenario for learning accurate, *sense-based* user profiles, as discussed in the following section.

2.2 Profile Learner: Learning User Profiles from Static Content and UGC

We consider the problem of learning user profiles as a binary Text Categorization task [13] since each document has to be classified as interesting or not with respect to the user preferences. Therefore, the set of categories is restricted to c_+, that represents the positive class (user-likes), and c_- the negative one (user-dislikes). The induced probabilistic model is used to estimate the *a posteriori* probability, $P(c_j|d_i)$, of document d_i belonging to class c_j.

The algorithm adopted for inferring user profiles is a Naïve Bayes text learning approach, widely used in content-based recommenders [6], which is not described here due to space limitations. What we would like to point out here is that the final outcome of the learning process is a probabilistic model used to classify a new document in the class c_+ or c_-. Given a new document d_j, the model computes the a-posteriori classification scores $P(c_+|d_j)$ and $P(c_-|d_j)$ by using probabilities of synsets contained in the user profile and estimated in the training step. An example of user profiles is depicted in Figure 2.

The profile contains the user identifier and the *a-priori* probabilities of liking or disliking an item, apart from its content. Moreover, the profile is structured in two main parts: *profile_like* contains features describing the concepts able to deem items relevant, while features in *profile_dislike* should help in filtering out not relevant items. Each part of the profile is structured in four slots, resembling

```xml
- <profile>
    <user>1</user>
    <Probability like="0.685714" dislike="0.314285"/>
  - <profile_like>
    - <title>
        <feature value="12946864" frequency="0.25"/>
        <feature value="9157420" frequency="0.25"/>
        <feature value="1537400" frequency="0.25"/>
        ...
      </title>
    - <author>
        <feature value="pinturicchio" frequency="0.25"/>
        <feature value="raffaello" frequency="0.25"/>
      </author>
    - <description>
        <feature value="517984" frequency="0.2168465218030137"/>
        <feature value="5626840" frequency="0.22035475747290048"/>
        <feature value="9407931" frequency="0.2168465218030137"/>
        <feature value="3947822" frequency="0.20554267981301802"/>
        ...
        <feature value="2372283" frequency="0.20554267981301802"/>
      </description>
    - <tag>
        <feature value="01150130" frequency="0.25"/>
        <feature value="raffaello" frequency="0.25"/>
      </tag>
    </profile_like>
    <profile_dislike> ... </profile_dislike>
  </profile>
```


Fig. 2. A fragment of user profile

the same representation strategy adopted for artworks. Each slot reports the features (WordNet identifiers) occurring in the training examples, with corresponding frequencies computed in the training step. Frequencies are used by the Bayesian learning algorithm to induce the classification model (i.e. the user profile) exploited to suggest relevant artworks in the recommendation phase.

3 Experimental Evaluation

The goal of the experimental evaluation was to compare the predictive accuracy of our recommender system when 1) user profiles are learned from static content only; 2) both static content and UGC are used in the learning process.

In addition, to properly investigate the effects of including social tagging in the recommendation process, a distinction has to be made between considering, for an artifact rated as interesting by a user, either the whole folksonomy (i.e., the community tags used by all visitors to annotate that artifact), or only the tags entered by that user for that artifact (i.e., the user's contribution to the whole artifact folksonomy). For this purpose, we designed several experiments, described in the following.

3.1 Users and Dataset

The dataset considered for the experiments is represented by 45 paintings chosen from the collection of the Vatican picture-gallery. The dataset was collected using screenscraping bots, which captured the required information from the official website[1] of the Vatican picture-gallery. In particular, for each element in the dataset an image of the artifact was collected, along with three textual properties, namely its title, artist, and description.

We involved 30 users who volunteered took part in the experiments. The average age of the users was in the middle of twenties. None of the users was an art critic or expert.

Users were requested to interact with a web application (Figure 3), in order to express their preferences for all the 45 paintings in the collection. The preference was expressed as a numerical vote on a 5-point scale (1=strongly dislike, 5=strongly like). Moreover, users were left free to annotate the paintings with as many tags as wished. For the overall 45 paintings in the dataset, 4300 tags were used.

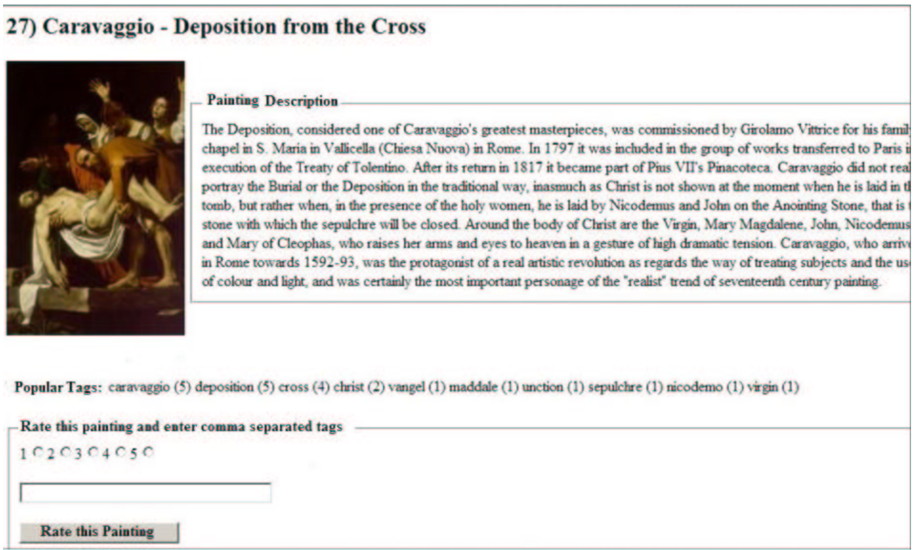

Fig. 3. Gathering user ratings and tags

3.2 Design of the Experiment and Evaluation Metrics

Since ITR is conceived as a text classifier, its effectiveness can be evaluated by classification accuracy measures, namely *Precision* and *Recall* [14].

[1] `http://mv.vatican.va/3_EN/pages/PIN/PIN_Main.html`

Precision (Pr) is defined as the number of relevant selected items divided by the number of selected items. Recall (Re) is defined as the number of relevant selected items divided by the total number of relevant items available. F1 measure, a combination of precision and recall, is also used to have an overall measure of predictive accuracy:

$$F1 = \frac{2 \times Re \times Pr}{Pr + Re}$$

We adopted these specific measures because we are interested in measuring how *relevant* a set of recommendations is for a user. In the experiment, a painting is considered as *relevant* by a user, if the rating is greater than or equal to 4, while ITR considers a painting as relevant if the a-posteriori probability of class likes is greater than 0.5. We designed 5 different experiments, depending on the type of content used for training the system:

– Exp #1: Static Content - only title, artist and description of the painting, as collected from the official website of the Vatican picture-gallery
– Exp #2: Personal tags - only tags provided by a specific user on a specific painting
– Exp #3: Social tags - all the tags provided by all the users on a specific painting
– Exp #4: Static Content + Personal tags
– Exp #5: Static Content + Social tags

All experiments were carried out using the same methodology, consisting in performing one run for each user, scheduled as follows:

1. select the appropriate content depending on the experiment being executed;
2. split the selected data into a training set *Tr* and a test set *Ts*;
3. use *Tr* for learning the corresponding user profile;
4. evaluate the predictive accuracy of the induced profile on *Ts*.

The methodology adopted for obtaining *Tr* and *Ts* was the K-fold cross validation [15], with $K = 5$. Given the size of the dataset (45), applying a 5-fold cross validation technique means that the dataset is divided into 5 disjoint partitions, each containing 9 paintings. The learning of profiles and the test of predictions were performed in 5 steps. At each step, 4 (K-1) partitions were used as the training set *Tr*, whereas the remaining partition was used as the test set *Ts*. The steps were repeated until each of the 5 disjoint partitions was used as the *Ts*. Results were averaged over the 5 runs.

3.3 Discussion

Results of the 5 experiments are reported in Table 1, averaged over the 30 users.

The main finding is that the integration of UGC (whether social or personal tags) causes an increase of precision in the process of recommending artifacts

Table 1. Results of the K-fold Cross Validation

Type of content	**Precision**	**Recall**	**F1**
Exp #1: Static Content	75.86	94.27	84.07
Exp #2: Personal Tags	75.96	92.65	83.48
Exp #3: Social Tags	75.59	90.50	82.37
Exp #4: Static Content + Personal Tags	78.04	93.60	85.11
Exp #5: Static Content + Social Tags	78.01	93.19	84.93

to users. More specifically, precision of profiles learned from both static content and tags (hereafter, augmented profiles) outperformed the precision of profiles learned from either static content (hereafter, content-based profiles) or just tags (hereafter, tag-based profiles). The improvement ranges between 2% and 2.40%. Another interesting finding is that precision of content-based profiles is comparable with that of tag-based profiles. Although this result may suggest that just tags are sufficient for providing accurate recommendations, a decrease of recall (-1.62% with personal tags, -3.77% with social tags) actually shows that static content cannot be neglected even if tags are available. The higher decrease of recall registered with social tags leads to conclude that community tags introduce some noise in the recommendation process (relevant paintings are filtered out due to wrong advice by other users). The general conclusion of the comparison between content-based profiles and augmented profiles is that a significant increase of precision corresponds to a slight and physiological loss of recall. The overall accuracy of augmented profiles (F1 about 85%) is considered satisfactory.

4 Conclusions and Future Work

In this paper we have investigated how to effectively combine existing content-based filtering algorithms with UGC, in the context of cultural heritage personalization. The main contribution of the paper is an approach in which machine learning techniques are adopted to infer user profiles both from static content, as in classical content-based recommender, and UGC, namely tags provided by users to freely annotate artworks. The main outcome of the experiments performed to evaluate the proposed approach is that the integration of UGC causes an increase of precision in the process of recommending artifacts to users.

By definition, social tags used for annotating a painting include personal tags. However, the findings from the experiments with social tags ran counter our expectation because, as compared to the use of personal tags only, a decrease of precision and recall was observed. To gain more insights on the effects of community-generated content, we need to 1) perform an analysis of what tags are used to build the folksonomies and how they affect the user profile generation; 2) replicate the experiments with a more heterogeneous community, involving experts in the art domain so as to identify differences with the tagging activity of naïve users.

Acknowledgements

This research was funded by MIUR (Ministero Universitá e Ricerca) under the contract Fondo per le Agevolazioni alla Ricerca, D.L: 297/99 "CHAT - Cultural Heritage fruition & e-learning applications of new Advanced (multimodal) Technologies" (2006-2008).

References

1. Ellin, E.: Museums and the computer: An appraisal of new potentials. Language Resources and Evaluation **4**(1) (1969) 25–30
2. Aroyo, L., Kuflik, T., Stock, O., Zancanaro, M., eds.: Proc. of Workshop on Personalized Access to Cultural Heritage (PATCH '07). (2007)
3. Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T., Not, E., Rocchi, C.: Adaptive, intelligent presentation of information for the museum visitor in peach. UMUAI **17**(3) (2007) 257–304
4. Wang, Y., Aroyo, L., Stash, N., Rutledge, L.: Interactive user modeling for personalized access to museum collections: The rijksmuseum case study. In Conati, C., McCoy, K.F., Paliouras, G., eds.: User Modeling 2007, Proceedings of the 11th International Conference. Volume 4511 of LNCS., Springer (2007) 385–389
5. Trant, J., Wyman, B.: Investigating social tagging and folksonomy in art museums with steve.museum. In: Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland. (May 2006)
6. Mladenic, D.: Text-learning and related intelligent agents: a survey. IEEE Intelligent Systems **14**(4) (1999) 44–54
7. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. User Model. User-Adapt. Interact. **17**(3) (2007) 217–255
8. Manning, C., Schütze, H.: 7: Word Sense Disambiguation. In: Foundations of Statistical Natural Language Processing. The MIT Press (1999) 229–264
9. Leacock, C., Chodorow, M., Miller, G.: Using corpus statistics and wordnet relations for sense identification. computational linguistics. Computational Linguistics **24**(1) (1998) 147–165
10. Semeraro, G., Degemmis, M., Lops, P., Basile, P.: Combining learning and word sense disambiguation for intelligent user profiling. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07. (2007) 2856–2861 M. Kaufmann, San Francisco, California. ISBN: 978-I-57735-298-3.
11. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: Proceedings of the 20th conference on Computer supported cooperative work, New York, NY, USA, ACM (2006) 181–190
12. Basile, P., de Gemmis, M., Gentile, A., Iaquinta, L., Lops, P., , Semeraro, G.: META - MultilanguagE Text Analyzer. In: Proc. of the Language and Speech Technnology Conference - LangTech 2008, Rome, Italy. (2008) 137–140
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
14. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1) (2004) 5–53
15. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. (1995) 1137–1145