Inf Technol Tourism (2016) 16:5–21
DOI 10.1007/s40558-015-0035-y

ORIGINAL RESEARCH

Analyzing user reviews in tourism with topic models

Marco Rossetti[1] · **Fabio Stella**[1] · **Markus Zanker**[2]

Received: 7 March 2015 / Revised: 24 September 2015 / Accepted: 5 November 2015 /
Published online: 17 November 2015

Abstract User generated content in general and textual reviews in particular constitute a vast source of information for the decision making of tourists and management and are therefore a key component for e-tourism. This paper provides a description of the topic model method with a particular application focus on the tourism domain. It therefore contributes different application scenarios where the topic model method processes textual reviews in order to provide decision support and recommendations to online tourists as well as to build a basis for further analytics. In the latter case the delivery of additional semantics helps digging into the enormous amounts of data that are continuously collected in present time. The contribution therefore consists of new models based on the topic model method and results from experimenting with user generated review data on restaurants and hotels.

Keywords Business intelligence · User reviews · Topic models · Recommender systems

This article is an extended version of the following paper published at ENTER'15 (Rossetti et al. 2015).

✉ Markus Zanker
mzanker@acm.org

Marco Rossetti
rossetti@disco.unimib.it

Fabio Stella
stella@disco.unimib.it

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

[2] Department of Applied Informatics, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

1 Introduction

Web 2.0 applications transformed the Internet from an information source to an opinion source (Dippelreiter et al. 2008; Schmallegger and Carson 2008). Every piece of information, whether it is a product offered in an online store or a post in a social network, can be commented or rated in some way (Litvin et al. 2008; Xiang and Gretzel 2010). In an economy heavily based on customer experience, such as tourism, individual decisions are strongly influenced by the written evidences of the experiences others already made—i.e. online customer reviews (Pang and Lee 2008; Zehrer et al. 2011; Ye et al. 2011).

Yoo and Gretzel 2009) researched the motivation of travelers to share their experiences with others via reviews. Users are mainly motivated by enjoyment of sharing their experiences and reviving them when writing. Furthermore also helping and supporting the travel service provider in case of positive experiences or saving others from bad services are important motives, while venting negative feelings or exercising collective power does not play an important role. These reviews help mainly in the pre-trip phase to generate ideas and to support decision making by narrowing down the set of choices (Gretzel and Yoo 2008). Users trust in reviews heavily depends on the credibility of the source (Yoo et al. 2009). Consequently this article focuses on user reviews that have been published on the widely-used consumer platforms tripadvisor.com and yelp.com.

From an IT perspective the automated exploitation of these opinions in order to provide advice and decision support led to tremendous research efforts in fields such as Machine Learning (ML) and Semantic Web (SemWeb). ML focuses on the construction and study of models that learn regularities and patterns from known data in order to best possibly predict unknown information without a principal need to understand the semantics of the data. In contrast SemWeb focuses on capturing and modeling the semantics of the data on the web and tries to derive "unknown data" by principles of reasoning and logics. In this paper we propose the application of the topic model method (Blei et al. 2003) to the task of analyzing user reviews. The topic model method is an approach that is clearly rooted in statistical ML and automatically extracts sets of terms with a coherent meaning (called topics) from document corpora such as reviews. Thus, although the method is agnostic of the semantics of the terms occurring in the documents themselves it automatically groups those terms where most presumably semantic ties exist between them. Therefore, it has the potential to, at least partially, bridge the gap between the two aforementioned principal research directions towards processing the data harvested from the Web.

The aim of this paper is an extension of the topic model method to derive interpretable user and item models that can be analyzed and exploited for deriving predictions and recommendations in the tourism domain. Topic model methods can be exploited to capture different dimensions from users' reviews that refer to what they like and what they do not like, and, at the same time, what are items strengths and weaknesses. The information extracted can be combined across users and items to predict preferences and it can be analyzed to get insights with an high business

value. In Sect. 2 we illustrate our contributions with a motivating example and continue with a more technical description. In Sect. 3 we present empirical evidence for the practical utility of our propositions by giving results from experimental evaluations, while in Sect. 4 the results obtained are discussed and future developments are planned.

2 The topic-criteria and sentiment models

2.1 Motivating example

The work presented in this paper is motivated by the idea that reviews express different viewpoints or dimensions of the experience that a user made with an item. Therefore extracting and interpreting these dimensions can be exploited to increase the accuracy of systems that automatically process these reviews in order to improve users' experience on such platforms or to extract some form of business value from this review data. In the following we illustrate the propositions of this paper with a fictitious example on reviews on accommodation services. Note that superscripts indicate the relation of a term to a topic. Let's assume Alice is a young woman who likes to travel around on a budget. She wrote the following two reviews on hotels she stayed at:

Hotel 1: The hotel was right in the ***center of the city*** ^{L}*, at* ***walking distance*** ^{L} *from the* ***city center*** ^{L}*! Huge* ***breakfast*** ^{F} *with nice* ***food*** ^{F}*! Rating: 5*

Hotel 2: I stayed in this hotel with my friends, the room was cheap, but the ***shower*** ^{R} *was broken and the* ***mattress*** ^{R} *was very hard! Rating 2*

From these reviews we can get an idea of Alice's taste and the "topics" she cares about, when staying in a hotel. In the literature a topic model (TM) (Blei 2012) is a statistical machine learning approach that tries to extract thematic information from large corpora of natural language documents. Topics are defined as sorted lists of words with a coherent semantic meaning that can be extracted from documents. In Table 1 we provide examples for such lists of terms.

Now Alice's reviews can be mapped on these (pre-extracted) topics based on what she mentioned in the reviews. Note, that we are building user profiles solely based on the content of the user's reviews that indicate what are the topics the user likes to talk about and ignore the specific rating values. In this small example Alice

Table 1 An example of potential topics extracted from hotel reviews

Topic "Location"	Topic "Food"	Topic "Rooms"	Topic "Business"
Walking_distance	Breakfast	Shower	Executive_lounge
Station	Service	Bathroom	Floor
City_center	Restaurant	Tub	Executive_floor
Metro	Bar	Mattress	Hilton
Close	Food	Flat_tv	Conrad

seems to care about the topics Location, Food and Rooms because the conditional probability of occurrence of the terms related to these topics are rather high in her reviews.

Another hotel (Hotel 3) received the following reviews from different users:

User 1: The staff in the ***executive lounge*** ^{B} *is very professional and the* ***location*** ^{L} *is very* ***close*** ^{L} *to the* ***metro station*** ^{L}*. Rating: 5*

User 2: The room was nice, with a ***flat tv*** ^{R}*, but the* ***breakfast*** ^{F} *was so poor! I didn't have enough* ***food*** ^{F}*. Rating 3*

Now, given these reviews and ratings, we can compute scores for each topic and map items and users onto the same "topic" space. Based on these two reviews the Hotel 3 might achieve a high rating w.r.t. the topics "Location" and "Business", but only a low one for "Food" and "Rooms".

How can the tourism domain benefit from applying this approach? First, when Alice is looking for a hotel recommendation, the item profile of Hotel 3 can be matched against Alice's profile in order to check if this item would be a plausible proposition. As Alice is, amongst others, interested in the topics "Food" and "Rooms" on which Hotel 3 is not scoring high, this hotel might not be a formidable recommendation.

Second, the automated extraction of topics and the building of item profiles with scores on each topic is an opportunity to assess the strengths and weaknesses of each item as they are perceived by users. This way item profiles based on collected reviews allow tourists to compare different service providers as well as provide a source for business analytics for management. Moreover, the automatic topic discovery process can also bring new interesting insights, as it can find connections between words that are typical of the tourism domain. For instance, if we look at the words that compose the topic labeled "Location", we can find useful information to understand what users usually consider to be important for this aspect.

Third, based on analysis of what the user is writing we can estimate the rating the user would probably assign to the item. Such a scenario could either help to make rating values more consistent with reviews or enable a business analytics application to derive numeric scores from text, where no rating value is given (e.g. in posts on social networks or email feedback).

It's important to specify that topics are not automatically labeled, as they are extracted with an unsupervised technique that is not aware of the meaning of the topics. Several works have tried to label topics (Magatti et al. 2009; Lau et al. 2011; Aletras and Stevenson 2014). However, the match between user and item profiles can still be exploited, even if the meaning of the topics is not known.

2.2 Topic model

As already mentioned in the previous section the topic model method summarizes natural language documents based on thematic information denoted as *topics*. Historically, the first technique that tried to extract thematic information from text documents was the latent semantic indexing technique (LSI) (Deerwester et al.

1990), which consists of the factorization of the term frequency-inverse document frequency (TF-IDF) matrix (Manning et al. 2008). This model was extended by the probabilistic LSI (pLSI), which provided a statistical foundation based on the likelihood principle and defines a generative model of the text data termed *aspect model*. It is a latent variable model that associates an unobserved class variable which each observation, i.e. the occurrence of a word in a document, as described in (Hofmann 1999) in detail. The main contribution to the topic model approach happened with the introduction of the Latent Dirichlet Allocation technique (LDA) (Blei et al. 2003), which describes a full generative model for topics and text. Every topic corresponds to a probability distribution over the corpus dictionary (i.e. a controlled vocabulary of terms) and every document is associated with a probability distribution over topics. The generative process of a generic document d consists of the following steps:

- a topic distribution θ_d is randomly generated;
- for each word position in d:
  - a topic k is extracted from θ_d;
  - a word w is selected with a given probability from topic k;

The aim of the LDA model is to invert this generative process: the occurrences of words in the documents are the observed variables, while the topic structure is hidden. By exploiting techniques of statistical inference and sampling, these probability distributions are inferred by observing the frequency of words within documents. Since then the topic model method was extended in several ways, i.e. for dealing with topics evolution over time (Blei and Lafferty 2006b), topics correlation (Blei and Lafferty 2006a) and networks of correlated documents (Chang and Blei 2009).

Wang et al. (2010) proposed a probabilistic generative model similar to LDA applied to textual reviews on hotels to estimate opinion ratings on topical aspects such as *cleanliness* or *sleep quality*, a problem defined as *Latent Aspect Rating Analysis* (*LARA*). Each review is split into sentences, and each sentence is supposed to be about a specific aspect. The proposed generative model assumes that for each sentence a user decides which aspect he/she wants to write and chooses the words to write, carefully based on the decision made. To assign one or more aspects to each sentence a bootstrap procedure is defined: an initial seed of aspect keywords is provided, and based on these initial corpora of keywords sentences are assigned to different aspects. The empirical experiments show that the proposed method is able to estimate aspect ratings and to discover interesting cases where the overall ratings are the same, but aspect ratings are different. Furthermore, review analysis opens a range of possible applications, such as opinion summarization on topical aspects, ranking of entities based on these aspect ratings and the analysis of the rating behavior of reviewers.

Agarwal and Chen (2010) introduced a matrix factorization method for recommender systems where items have a natural bag-of-word representation termed *fLDA*. Words from these item descriptions are associated with a discrete latent factor termed *topic*. Topics extracted from item descriptions and user metadata are used as priors to regularize item and user latent factors. The posterior distribution of item and user factors depends on both the prior and user ratings on items, since the LDA model is exploited to regularize item latent factors, and the Gaussian linear regression regularizes user latent factors. The proposed model is accurate and able to deal with warm-start and cold-start scenarios, as textual data related to new users and new items can be used to compute recommendations on them. Furthermore, it provides interpretable latent factors that can explain user-item interactions.

Wang and Blei (2011) defined an extension of LDA for recommending scientific articles called *collaborative topic regression* (*CTR*). Topic model and matrix factorization are merged into a single method, where item latent factors are obtained from adding an offset latent variable to the item topic distribution. The latent variable is optimized with an expectation-maximization (EM) algorithm that tries to identify the maximum likelihood estimates of these model constituents. This method is capable of providing in-matrix and out-of-matrix predictions, where the former case constitutes predictions of items already rated by some users, while the latter signifies computing ratings for novel and unrated items. Even in this case the method is able to provide interpretable latent factors that can be used to profile users and items.

Recently, McAuley et al. (2013) merged matrix factorization and topic models in order to estimate the ratings from textual reviews on different datasets. The *Hidden Factors as Topics* (*HFT*) approach consists of two steps: first, latent factors for rating prediction are fitted, and second, topic assignments are updated binding item topic distributions and item latent factors. In this work all the reviews associated with an item are merged into a single document. The proposed approach not only leads to more accurate predictions on recommendations, but can also solve side problems. First, it deals with the cold-start problem, exploiting item topics for items with only a few ratings. Second, it is able to discover and automatically categorize items in different categories based on the topics discussed in the reviews. Third, it can identify representative reviews, which can be shown to users as a summary of item characteristics. The proposed approach was tested on a set of huge datasets scraped from the web: 35 millions review from *Amazon*,[1] 6 millions review from *ratebeer*[2] and 220 thousand reviews from *Yelp*.[3]

Another extension to LDA is the Joint Sentiment-Topic model (JST) (Lin et al. 2012). In contrast to the majority of sentiment analysis models which are based on classification models, this model is able to extract sentiment and topics simultaneously from text in an unsupervised way. The main difference with respect to the LDA model is that JST adds an additional sentiment layer between the document

[1] http://www.amazon.com.

[2] http://www.ratebeer.com.

[3] http://www.yelp.com.

and the topic layer. In this way a four level hierarchy is defined where documents have distributions on sentiment labels, sentiment labels have distributions on topics and topics have distributions on words.

The following subsections shows how to exploit the LDA and JST models to attach semantics to user reviews in the tourism domain. After that, three application scenarios for these proposed models are presented.

2.3 The topic-criteria model

The first model we propose in this paper is the topic-criteria (TC) model, which exploits the topic model method to extract latent features from textual reviews and discuss its application for several application scenarios in tourism. The difference between this approach and other approaches which use or extend topic model methods for Recommender Systems is that in this case the classic LDA is applied to process reviews and the extracted topics are exploited to define user and item profiles. This particular step makes the method very intuitive in its formulation as well as in the meaning of the computed information. Let us define R to be the set of ratings, and D to signify the set of textual reviews. r_{ij} is the rating given by user i to item j, while d_{ij} is its associated review. For simplicity, let R_i be the set of ratings given by user i, while let R^j be the set of ratings given to item j. The analogous notation is defined for reviews, i.e. D_i denotes the set of reviews given by user i and D^j reviews about item j. The probability of the topic z given the document d_{ij} is indicated with $P(z|d_{ij})$. Finally, we reserve the letter i to indicate users, j to indicate items and Z to denote the number of topics. The user profile is constructed by aggregating the topic distributions of all the reviews written by the user, without considering the associated ratings. The main idea is that to profile a user we are only interested about what aspects of an, for instance, accommodation the user writes. Therefore, the user model consists of those topics that the user seems to care about in her/his reviews. The rating values are not needed for this purpose. The user profile is computed by aggregating the topic distributions of the user's reviews, as shown in Fig. 1. The user profile (UP) for user i is a numeric degree on each topic z (from 1 to Z) that defines the relevance of topic z for user i (see Eq. 1).

$$UP(i,z) = \frac{\sum_{d_{ij} \in D_i} P(z|d_{ij})}{|D_i|} \tag{1}$$

Item profiles are built by using both: topic distributions and numeric ratings, because the topics signify the aspects the user cared about in her/his review and the rating values indicate how satisfied the user was with respect to these aspects. Thus, the main idea is that if an item has reviews that frequently mention a specific topic, we have to consider the ratings to understand if this topic is a strong or a weak point of this item. The item profile can be built as a numeric score from 1 to 5 for each extracted topic aggregating the topic distributions and the related ratings, as illustrated by Fig. 2. As in the previous case, the item profile (IP) for item j can be computed as defined in Eq. (2)

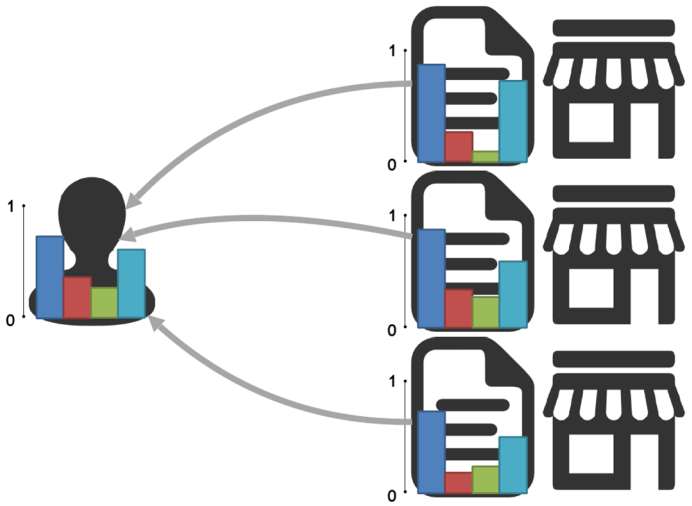

Fig. 1 User profile creation from topic distributions

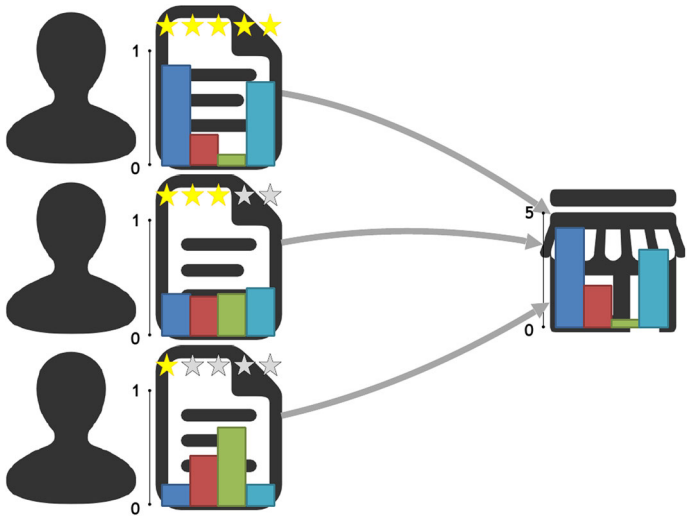

Fig. 2 Item profile creation from topic distributions and ratings

$$IP(j,z) = \frac{\sum_{d_{ij} \in D_i} P(z|d_{ij}) * r_{ij}}{\sum_{d_{ij} \in D_i} P(z|d_{ij})} \tag{2}$$

Since user profiles define the interest of the users in different topics and item profiles indicate how well an item does with respect to each topic, the combination of both profiles should allow us to estimate a user rating for an unseen item. The match between a user and an item profile is computed by the sum of the products for

each topic, as defined in Eq. (3). In order to improve the prediction accuracy of the approach, a topic weight is added to assign more value to those topics that are more influential for the estimation of rating values.

$$\hat{r_{ij}} = \sum_{z=1}^{Z} UP(i,z)IP(j,z)w_z \tag{3}$$

These weights are optimized by minimizing the loss function with the gradient descent approach, as shown in Eq. (4). Note that λ is a regularization parameter that penalizes more complex models in order to avoid data overfitting.

$$\min_{w} \sum_{r_{ij} \in R} \left(r_{ij} - \sum_{z=1}^{Z} UP(i,z)IP(j,z)w_z \right)^2 + \lambda \|w\|_F^2. \tag{4}$$

The model was implemented exploiting the MALLET[4] implementation of the LDA model. The preprocessing step was addressed using MALLET capabilities: textual content was tokenized considering all non-alphabetic characters as separators, characters were converted to lowercase and common English stopwords have been removed.

2.4 The topic-sentiment criteria model

The second model we propose in this paper is the Topic-Sentiment Criteria (TSC) model, which exploits the Joint Sentiment-Topic model (Lin et al. 2012) to extract sentiment and polarized latent features from textual reviews. By extracting sentiment from user reviews we are able to identify useful predictors for the overall rating, as well as to discover representative reviews and textual features for items and users.

We partly repeat the formal definitions from the previous subsection for ease of understanding. As already defined R denotes the set of ratings, and D signifies the set of textual reviews. r_{ij} is the rating given by user i to item j, while d_{ij} is its associated review. Furthermore R_i denotes the set of ratings given by user i, while R^j signifies the set of ratings given to item j. The analogous notation is defined for reviews, i.e. D_i denotes the set of reviews given by user i and D^j represents the set of reviews about item j. For each review we have a probability distribution on the sentiment space S that is composed by the sentiments *neutral*, *positive* and *negative*. We indicate the probability of sentiment s in the document d_{ij} as $P(s|d_{ij})$, and $\sum_{s \in S} P(s|d_{ij}) = 1$. For each sentiment s a topic space is defined and the probability of the topic z_s given the document d_{ij} is indicated with $P(z_s|d_{ij})$. Note that z_s indicates the topic z in the sentiment space s, and $\sum_{z_s=1}^{Z_s} P(z_s|d_{ij}) = 1$, where Z_s is the number of topics in the sentiment space s.

If we knew the real topic and the real sentiment distribution for each document, we would be able to estimate the final rating using topic and sentiment probability values as predictors, as shown in Eq. (5):

[4] http://mallet.cs.umass.edu/.

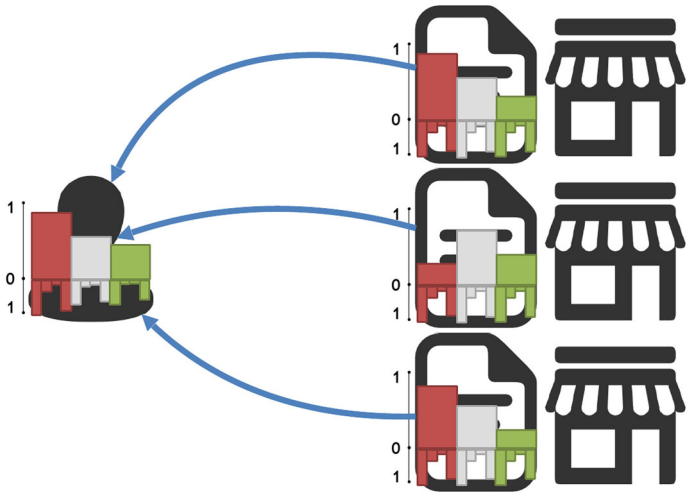

Fig. 3 User profile creation from topic distributions

$$\hat{r}_{ij} = \sum_{s \in S} \sum_{z_s=1}^{Z_s} P(z_s|d_{ij}) P(s|d_{ij}) b_{z_s} \tag{5}$$

where b_{z_s} is the coefficient for topic z_s. However, usually the real topic distribution for an unseen user-item couple is not known and must be estimated by combining the knowledge collected about the user and the item. To address this task, we compute and combine user and item profiles.

In analogy to the Topic-Criteria model the user profile is computed by aggregating the topic distributions of the user's reviews, as shown in Fig. 3. The user profile (UP) for user i is a numeric degree on each topic z_s (from 1 to Z_s) that defines the relevance of topic z_s for user i (see Eq. 6). The probability of topic z_s is multiplied by the probability of the sentiment s. In this way the full user profile sums to 1, i.e. $\sum_{s \in S} \sum_{z_s=1}^{Z_s} UP(i, z_s) = 1$.

$$UP(i, z_s) = \frac{\sum_{d_{ij} \in D_i} P(s|d_{ij}) P(z_s|d_{ij})}{|D_i|} \tag{6}$$

Item profiles are built in the same way: topic distributions for each sentiment are aggregated taking even the sentiment probability into account. As in the previous case, the item profile (IP) for item j can be computed as defined in Eq. (7):

$$IP(j, z_s) = \frac{\sum_{d_{ij} \in D^j} P(s|d_{ij}) P(z_s|d_{ij})}{|D^j|} \tag{7}$$

Similar to the Topic-Criteria model the combination of both profiles serves as an estimate for the user rating for an unseen item (Fig. 4). The match between a user and an

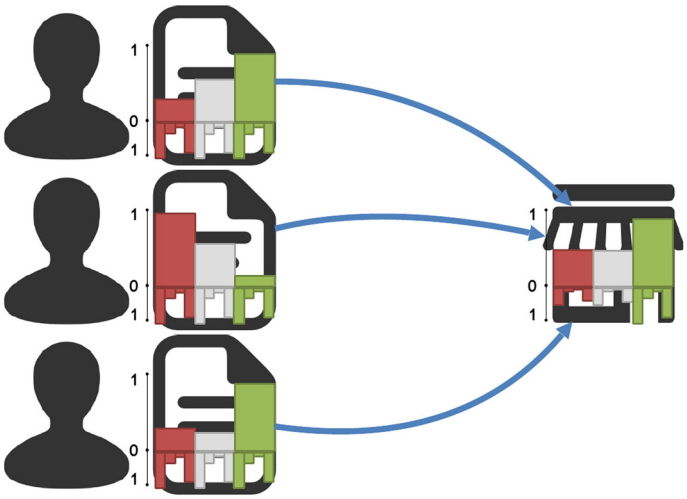

Fig. 4 Item profile creation from topic distributions and ratings

item profile is computed in three different ways: product, minimum and maximum. The product aggregation, as defined in Eq. (8), consists of the multiplication of each topic component of the user profile by the same component of the item profile. The minimum aggregation, as defined in Eq. (9), takes the minimum of the user profile and item profile components with respect to each topic, while the maximum aggregation, as defined in Eq. (10), takes the maximum value. The aggregated weights are then normalized such that the sum of weights for each topic and for each sentiment sum up to one.

$$w_{ij}^{Prod}(z_s) = \frac{UP(i,z_s)IP(j,z_s)}{\sum_{s\in S}\sum_{z_s=1}^{Z_s} w_{ij}^{Prod}} \tag{8}$$

$$w_{ij}^{Min}(z_s) = \frac{min(UP(i,z_s),IP(j,z_s))}{\sum_{s\in S}\sum_{z_s=1}^{Z_s} w_{ij}^{Min}} \tag{9}$$

$$w_{ij}^{Max}(z_s) = \frac{max(UP(i,z_s),IP(j,z_s))}{\sum_{s\in S}\sum_{z_s=1}^{Z_s} w_{ij}^{Max}} \tag{10}$$

The weight $w_{ij}(z_s)$ estimates the real value of $P(z_s|d_{ij})P(s|d_{ij})$ and can be used as predictors to estimate the rating that user i would give to item j, as defined in Eq. (11). The coefficients b_{z_s} are learned considering the real topic distributions and applying linear regression in order to estimate the real rating, as shown in Eq. (5).

$$\hat{r}_{ij} = \sum_{s\in S}\sum_{z_s=1}^{Z_s} w_{ij}(z_s)b_{z_s} \tag{11}$$

The model was implemented exploiting the C++ implementation of the JST model (Lin et al. 2012) provided by the authors.[5]

The preprocessing step was addressed with the *NLTK* Python library.[6] Textual content was tokenized considering all non-alphabetic characters as separators, converted to lowercase and stemmed using the *NLTK* implementation of the Porter stemmer. Common English stopwords have been removed.

2.5 Application scenarios

In this paper we propose three different application scenarios for our proposed approaches and provide empirical evidence based on available data:

1. Rating prediction and recommendation: user profiles represent the degree of interest of users in extracted topics. Item profiles express an item's scoring on each topic. Thus, the match between user and item profiles indicates how appropriate an item might be for a user.
2. Analytics and interpretation: the topic model method provides a natural characterization and interpretation of user and item profiles. When interpreting (selected) topics as item features or characteristics a system can transparently display to a user the model that is internally used for personalizing content. Furthermore, items can be compared to each other from several perspectives as if multi-criteria ratings from users would be known, where each item is assessed according to different dimensions such as quality of service, value for money, rooms, cleanliness or location (Jannach et al. 2014).
3. Suggest ratings for review: the proposed approach can also be exploited to suggest a rating given a textual review and a user profile. For instance, the system can propose a rating based on what the user currently writes as a review, or it can assess the coherence of the review content and the rating value assigned to a particular item.

3 Empirical evaluation

For assessing the proposed approach in these three scenarios two datasets were used: the *Yelp*[7] dataset and the *TripAdvisor*[8] dataset. The Yelp dataset is provided by Yelp for the *Yelp Dataset Challenge*[9] and it contains reviews and ratings given by users of the Yelp website to business activities, mainly restaurants. The TripAdvisor dataset (Jannach et al. 2014) was crawled from the popular website and it contains reviews about hotels in different cities. The TripAdvisor dataset contains also more

[5] https://github.com/linron84/JST.

[6] http://www.nltk.org.

[7] http://www.yelp.com.

[8] http://www.tripadvisor.com.

[9] http://www.yelp.com/dataset_challenge.

Table 2 Dataset summary

	YELP-5-5	YELP-10-10	TA-3-3	TA-5-5
#Users	9382	3802	13048	1850
#Items	3733	2413	12342	1774
#Ratings	145735	101416	83395	14656
Sparsity	0.0042	0.0111	0.0005	0.0045

fine-granular user feedback that not only encompasses an overall rating value but also ratings on more specific dimensions such as value for money, cleanliness or rooms. In order to experiment with different levels of data sparseness (i.e. the share of unknown entries in the full user-item rating matrix) we identified data subsets that have at least n known ratings for each user and each item. This processing leads to the datasets described in Table 2.

For each user we randomly selected 80 % of the ratings for training and the remaining ones were used for testing. With the exception the of the *TA-3-3* dataset, there we selected 66 % for training and used the remaining ratings for testing.

3.1 Scenario 1: rating prediction and recommendation

The rating prediction accuracy was evaluated with the classic error measure for machine learning, the Root Mean Squared Error (RMSE). The proposed TC model was tested without considering topic weights (TC) and alternatively optimizing topic weights (TC-O). In the TC-O method we also considered the average rating of each user by subtracting it from the original rating Eq. (2) and adding it back to the estimated rating Eq. (3). For the Topic-Sentiment Criteria model results for all three aggregation operators (product, min, max) are given. We tested different values of the number of topics with respect to the dataset considered, and we found that on the Yelp dataset we had better results with a higher number of topics (around 50–100), while on TripAdvisor a smaller number was better (around 10–30). The topic model variants were evaluated against three classic Collaborative Filtering (CF) algorithms: the K-Nearest Neighbor User Based (KNN-UB), the K-Nearest Neighbor Item Based (KNN-IB) and the Probabilistic Matrix Factorization (PMF) acting as baselines.

Neighborhood models, also known as memory-based models, are the most common approach to CF (Herlocker et al. 1999). In the user-based case the idea is to suggest items which are liked by users with similar tastes, while in the item-based one the system recommends items similar to the items liked by the user (Sarwar et al. 2001). Based on a parameter selection step with Pearson correlation as a similarity measure in the user-based approach and the cosine similarity in the item-based approach we set the number of neighbors to 10 in both cases. Probabilistic Matrix Factorization (Mnih and Salakhutdinov 2007) is a model-based approach which tries to factorize the user-item matrix with a probabilistic perspective. Although several extensions of this model have been developed, the classic PMF is still a good baseline for the CF. Due to a parameter selection phase we set the number of latent factors to 10.

Table 3 shows RMSE values for the baselines and the TC and TSC models. The proposed approaches achieve RMSE rates that are at least comparable to the classic

Table 3 RMSE values for the different methods on the four datasets

Algorithm	YELP-5-5	YELP-10-10	TA-3-3	TA-5-5
KNN-IB	1.0709	1.0249	1.0531	0.9601
KNN-UB	1.1088	1.0424	1.0715	0.9447
PMF	1.0956	1.0389	1.0373	0.9946
TC	1.0706	1.0247	1.0625	0.9719
TC-O	**1.0599**	**0.9955**	1.0916	0.9776
TSC-Prod	1.0797	1.0303	1.0716	0.9989
TSC-Min	1.0846	1.0373	1.0108	0.9527
TSC-Max	1.0832	1.0336	**0.9977**	**0.9443**

Lowest RMSE values are in bold

CF approaches: on the YELP datasets the TC models achieve lower RMSE values than all other approaches, while on the Tripadvisor datasets the additional model layer that considers sentiment values reaches highest average accuracy in terms of RMSE. However, the advantage of the proposed approaches does not solely lie in being as accurate as or slightly better than other CF approaches, but in employing richer user models. Only review content is exploited to model users, therefore novel ways to explain users why a specific item is recommended become possible. The system could explain to a fictitious user Alice that it assumes that she puts a lot of emphasis on topics B and D when looking for an accommodation and that a specific item is particularly high appraised w.r.t. these two topics in reviews of other users.

3.2 Scenario 2: analytics and interpretation

Based on user and item models that are built from textual content even more application scenarios become thinkable. For instance, it cannot only be analyzed which topics are important to a particular user or segments of users, but also the relative strengths and weaknesses of items can be compared to each other. Based on the Tripadvisor dataset we identified exemplary topics that can be related to the dimensional rating values. For instance, in case of a low rating value for the "cleanliness" dimension, the topics associated with that dimension can provide hints about the reasons. On the other hand, in case of high ratings we can explore which topics the users particularly appreciated. In order to identify which topics are important for a particular rating dimension we performed a non-parametric test to compare the overall rating distribution and the rating distribution of the top-k reviews strongly associated with a topic. A test rejecting the null hypothesis means that the presence of the topic has a positive (or negative) impact on the rating. We applied a two-sample Kolmogorov-Smirnov test with significance level equal to 5 %.

Table 4 shows two illustrative examples of topics strongly correlated with either the rating dimension cleanliness or business. For this analysis we split reviews based on a specific destination for the purpose of reducing the fragmentation of topics. Other subsamples can be extracted dividing the reviews by item types such as specific hotels or by user segment such as "senior couples" or "families on a budget" in analogy to Jannach et al. (2014).

Table 4 Illustrative examples for selected Topics related to multi-criteria dimensions

Topic related to..	
Cleanliness in reviews on Orlando hotels	Business in reviews on New York hotels
Dirty mold bugs smelled smell filthy	Internet_access wireless_internet
Carpet musty stained disgusting	Business_center computers
Bed_bugs black mildew moldy stains	Free_wireless business boarding
Bites dust musty_smell refund	Gym center print free_internet_access

Table 5 RMSE values for the Scenario 3

	YELP-5-5	YELP-10-10	TA-3-3	TA-5-5
TC	1.0718	1.0258	1.0663	0.9783
TC-O	1.0600	0.9976	1.0932	0.9826
TSC	**0.8137**	**0.8424**	**0.7401**	**0.7372**

3.3 Scenario 3: suggest ratings for review

The third scenario refers to the ability of our approach to predict a rating given the textual review, for instance, interactively when the user just entered the review text. Such an approach can be used to propose a rating right after the user finished writing his/her review. To estimate the rating of a review, topics are extracted from the text and the topic distribution is multiplied with the item profile in order to estimate the rating. Since the CF methods considered in Scenario 1 cannot predict ratings based on textual input, we only compute accuracy results using RMSE for our proposed methods (see Table 5). It is interesting to notice that the RMSE values for the TC model are only slightly higher than the ones already obtained for Scenario 1. The small difference can be explained by the fact that a user profile is more informative (aggregates several reviews) than the topic distribution of the single review and therefore a profile better represents user's interests. However the TSC method, that considers the expressed sentiment in reviews, clearly outperforms the TC model (highlighted in boldface). Also note, that we do not need to differentiate between different aggregation parameters for the TSC method, because the model is applied on a single review without the need to aggregate over the user model.

4 Discussion and conclusions

This paper explores the application of the topic model method in the tourism domain. The paper's contribution is twofold; first, a novel Topic-Criteria model and an extended Topic-Sentiment Criteria (TSC) model are proposed that extract topics and in case of the TSC model also sentiment to build rich user models. This way user models can be interpreted in terms of the topics users care about when writing reviews and platform providers can match this information with the language they

are using to describe the tourism domain. In addition, also items are modeled by a rating for each topic that indicates how well they are performing with respect to each topic in the eyes of their customers. This approach is, for instance, in analogy to the work of Xiang et al. (2009) that compared queries of travelers with website content. However, in contrast to the work of Xiang et al. the application of topic model methods follows a fully automated workflow.

We proposed a set of concrete application scenarios such as rating prediction and recommendation exploiting the proposed topic model approaches. They show not only the potential to increase the accuracy of different prediction mechanisms due to the exploitation of the content from textual reviews, but they also promise an improved user experience by potentially providing additional transparency and explanations. Content can be exploited to give reason to users why a specific item is proposed. Second, we also contribute empirical evidence for the practical relevance of the proposed technical approach by describing the three usage scenarios: Rating Prediction and Recommendation, Analytics and Interpretation and Suggest Ratings for Review and by exploiting available datasets to assess the predictive accuracy of the approach. It remains to note, that the presented results constitute only a first step of our work agenda that will include hybridizing the method with other well-known techniques and developing the application scenarios further. Another remark with respect to the Analytics and Interpretation scenario is the risk of cherry-picking as a general threat to the validity of explorative findings from large datasets. A possible extension of this work can be the application of the supervised LDA machine learning technique (Mcauliffe and Blei 2008) selecting reviews as learning input based on user features or rating values. In this way the identification of topics will be guided by predefined criteria such as rating dimensions and they can therefore be even better semantically interpreted.

Acknowledgments Authors acknowledge the financial support from the European Union (EU), the European Regional Development Fund (ERDF), the Austrian Federal Government and the State of Carinthia in the Interreg IV Italien-Österreich programme (project acronym O-STAR).

References

Agarwal D, Chen BC (2010). flda: matrix factorization through latent dirichlet allocation. In: Proceedings of the third ACM international conference on web search and data mining, pp 91–100

Aletras N, Stevenson M (2014) Labelling topics using unsupervised graph-based methods. In: Proceedings of the association for computational linguistics, pp 631–636

Blei D, Lafferty J (2006a) Correlated topic models. Adv Neural Info Process Syst 18:147

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84

Blei DM, Lafferty JD (2006b). Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, pp 113–120

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Chang J, Blei DM (2009) Relational topic models for document networks. In: International conference on artificial intelligence and statistics, pp 81–88

Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. JASIS 41(6):391–407

Dippelreiter B, Grün C, Pöttler M, Seidel I, Berger H, Dittenbach M, Pesenhofer A (2008) Online tourism communities on the path to web 2.0: an evaluation. Info Technol Tour 10(4):329–353

Gretzel U, Yoo KH (2008) Use and impact of online travel reviews. In: Information and communication technologies in tourism 2008 (ENTER). Springer, Heidelberg, pp 35–46

Herlocker JL, Konstan JA, Borchers A, Riedl J (1999) An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 230–237

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 50–57

Jannach D, Zanker M, Fuchs M (2014) Leveraging multi-criteria customer feedback for satisfaction analysis and improved recommendations. Inf Technol Tour 14(2):119–149

Lau JH, Grieser K, Newman D, Baldwin T (2011). Automatic labelling of topic models. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1, pp 1536–1545

Lin C, He Y, Everson R, Rüger S (2012) Weakly supervised joint sentiment-topic detection from text. IEEE Trans Knowl Data Eng 24:1134–1145

Litvin SW, Goldsmith RE, Pan B (2008) Electronic word-of-mouth in hospitality and tourism management. Tour Manag 29(3):458–468

Magatti D, Calegari S, Ciucci D, Stella F (2009) Automatic labeling of topics. In: Intelligent systems design and applications, 2009. ISDA'09. Ninth international conference on, pp 1227–1232

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, New York

McAuley J, Leskovec J (2013) Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on recommender systems, pp 165–172

Mcauliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, pp 121–128

Mnih A, Salakhutdinov R (2007) Probabilistic matrix factorization. In: Advances in neural information processing systems, pp 1257–1264

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Rossetti M, Stella F, Cao L, Zanker M (2015) Analyzing user reviews in tourism with topic models. In: Information and communication technologies in tourism 2015 (ENTER). Springer, Heidelberg, pp 47–58

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, pp 285–295

Schmallegger D, Carson D (2008) Blogs in tourism: changing approaches to information exchange. J Vacat Mark 14(2):99–110

Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 448–456

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp 783–792

Xiang Z, Gretzel U (2010) Role of social media in online travel information search. Tour Manag 31(2):179–188

Xiang Z, Gretzel U, Fesenmaier DR (2009) Semantic representation of tourism on the internet. J Travel Res 47(4):440–453

Ye Q, Law R, Gu B, Chen W (2011) The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput Human Behav 27(2):634–639

Yoo K-H, Gretzel U (2009) What motivates consumers to write online travel reviews? Inform Technol Tour 10(4):283–295

Yoo KH, Lee Y, Gretzel U, Fesenmaier DR (2009) Trust in travel-related consumer generated media. In: Information and communication technologies in tourism 2009 (ENTER). Springer, New York, pp 49–59

Zehrer A, Crotts JC, Magnini VP (2011) The perceived usefulness of blog postings: an extension of the expectancy-disconfirmation paradigm. Tour Manag 32(1):106–113