Beyond Relevance: Explicitly Promoting Novelty and Diversity in Tag Recommendation

FABIANO M. BELÉM, CAROLINA S. BATISTA, RODRYGO L. T. SANTOS, JUSSARA M. ALMEIDA, and MARCOS A. GONÇALVES, Universidade Federal de Minas Gerais

The design and evaluation of tag recommendation methods has historically focused on maximizing the relevance of the suggested tags for a given object, such as a movie or a song. However, relevance by itself may not be enough to guarantee recommendation usefulness. Promoting novelty and diversity in tag recommendation not only increases the chances that the user will select "some" of the recommended tags but also promotes complementary information (i.e., tags), which helps to cover multiple aspects or topics related to the target object. Previous work has addressed the tag recommendation problem by exploiting at most two of the following aspects: (1) relevance, (2) explicit topic diversity, and (3) novelty. In contrast, here we tackle these three aspects conjointly, by introducing two new tag recommendation methods that cover all three aspects of the problem at different levels. Our first method, called *Random Forest with topic-related attributes*, or RF_t, extends a relevance-driven tag recommender based on the Random Forest (*RF*) learning-to-rank method by including new tag attributes to capture the extent to which a candidate tag is related to the topics of the target object. This solution captures topic diversity as well as novelty at the attribute level while aiming at maximizing relevance in its objective function. Our second method, called *Explicit Tag Recommendation Diversifier with Novelty Promotion*, or *xTReND*, reranks the recommendations provided by any tag recommender to jointly promote relevance, novelty, and topic diversity. We use RF_t as a basic recommender applied before the reranking, thus building a solution that addresses the problem at both attribute and objective levels. Furthermore, to enable the use of our solutions on applications in which category information is unavailable, we investigate the suitability of using latent Dirichlet allocation (LDA) to automatically generate topics for objects. We evaluate all tag recommendation approaches using real data from five popular Web 2.0 applications. Our results show that RF_t greatly outperforms the relevance-driven *RF* baseline in diversity while producing gains in relevance as well. We also find that our new *xTReND* reranker obtains considerable gains in both novelty and relevance when compared to that same baseline while keeping the same relevance levels. Furthermore, compared to our previous reranker method, *xTReD*, which does not consider novelty, *xTReND* is also quite effective, improving the novelty of the recommended tags while keeping similar relevance and diversity levels in most datasets and scenarios. Comparing our two new proposals, we find that *xTReND* considerably outperforms RF_t in terms of novelty and diversity with only small losses (under 4%) in relevance. Overall, considering the trade-off among relevance, novelty, and diversity, our results demonstrate the superiority of *xTReND* over the baselines and the proposed alternative, RF_t. Finally, the use of automatically generated latent topics as an alternative to manually labeled categories also provides significant improvements, which greatly enhances the applicability of our solutions to applications where the latter is not available.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services

---

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web grant 573871/2008-6), and by the authors' individual grants from CNPq, CAPES, and FAPEMIG.

Authors' address: F. M. Belém, C. S. Batista, R. L. T. Santos, J. M. Almeida, and M. A. Gonçalves, Department of Computer Science, Federal University of Minas Gerais, Av. Antônio Carlos 6627, CEP 31270-010, Belo Horizonte—MG, Brazil; emails: {fmuniz, cbatista, rodrygo, jussara, mgoncalv}@dcc.ufmg.br.

DOI: http://dx.doi.org/10.1145/2801130

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Tag recommendation, relevance, novelty, topic diversity

ACM Reference Format:
Fabiano M. Belém, Carolina S. Batista, Rodrygo L. T. Santos, Jussara M. Almeida, and Marcos A. Gonçalves. 2016. Beyond relevance: Explicitly promoting novelty and diversity in tag recommendation. ACM Trans. Intell. Syst. Technol. 7, 3, Article 26 (February 2016), 34 pages.
DOI: http://dx.doi.org/10.1145/2801130

1. INTRODUCTION

Web 2.0 applications have become a rich source of user-generated content. Each page on a Web 2.0 application often comprises a main *object* (e.g., a video, image, audio, or text) and several sources of data associated with it, referred to as its features. The *textual features* of an object are well-defined blocks of text, such as title, tags, description, and user comments, used to describe the object's content [Belém et al. 2014]. Among all textual features, tags have gained special importance, as they offer an effective data source for information retrieval (IR) services, such as search [Li et al. 2008] and classification [Figueiredo et al. 2012], and may capture user interests reasonably well [Li et al. 2008]. In this context, there is a large interest in developing strategies to recommend tags to users, providing relevant and useful tag suggestions for a target object and indirectly improving the quality of tag-based IR services.

Tag recommendation methods have historically focused on maximizing the *relevance* of the recommended tags [Belém et al. 2014; Lipczak and Milios 2011; Wu et al. 2009; Yin et al. 2013; Song et al. 2011; Chen and Shin 2013; Lops et al. 2013]. When recommending tags for a target object, relevance refers to how well the recommended tags describe the contents of the target object. However, relevance by itself may not be enough to guarantee recommendation usefulness [Vargas and Castells 2011]. Indeed, the utility of a recommended item (or tag, specifically) depends on the other items in the list of recommendations [Clarke et al. 2011; Vargas and Castells 2011], due to the possible redundancy among them. Recommending tags that bring novel and diverse information with regard to previously ranked tags may promote more complementary information, helping to cover multiple aspects or topics related to the target object and, indirectly, improving the results of tag-based IR services. In particular, *diversity* is an important aspect because multimedia objects on the Web 2.0 may be multifaceted—that is, they may be related to various aspects and topics [Belém et al. 2013b]. *Novelty*, on the other hand, can increase serendipity, coverage, and recall of services that use the more "specific" (yet relevant) recommended tags.

To further illustrate the benefits of novelty and diversity in tag recommendation, Table I shows examples of recommendations produced for two objects (i.e., movies) extracted from MovieLens by three recommenders: one focused on relevance only; a second one that directly incorporates diversity; and a third one that considers relevance, diversity, and novelty aspects.[1]

Let us focus first on the tags recommended for the movie *X-Men: The Last Stand*. The relevance-driven recommender suggested the relevant tags "comics," "superhero," and (although more vague) "based," referring to the fact that the movie is based on Marvel Comics' X-Men. But it also suggested the general tag "dvd." Notice also that despite being driven by relevance, the recommender suggested an apparently irrelevant tag (as far as we can tell), "ummarti2006." The second recommender, in turn, which incorporates diversity aspects, also suggested the tags "dvd" and "comics," but together with "genetics" and "biology," which may be seen as other important subjects of the movie

[1]These are real recommendations produced by some of our proposed methods and baselines, which will be presented in Section 4.

Table I. Examples of Tag Recommendations for Two MovieLens Objects

Title	Relevance Only Recommender	Relevance + Diversity Recommender	Relevance + Diversity + Novelty Recommender
X-Men: The Last Stand	dvd, comics, ummarti2006 super-hero, based	dvd, genetics, biology, comics, mckellen	genetics, dvd, biology, mckellen, marvel
Harry Potter and the Sorcerer's Stone	award, shaw, scary, action, based	award, based, shaw, potter, action	award, shaw, potter, wizard, based

plot (a fiction related to evolution). Those two tags help to cover other topics related to the movie, increasing the diversity of the recommendations. The tag "mckellen," also suggested by the second recommender, can also be considered relevant, as it refers to one of the main actors of the film, Ian McKellen. We also note that all recommended tags are, to some extent, relevant to the movie, which illustrates a "good side effect" of promoting diversity: ensuring that at least one relevant tag for each topic related to an object will be suggested may help to demote too general or noisy tags, improving the relevance of the recommendations. In fact, our experimental results corroborate this hypothesis, as we shall see. Finally, the third recommender, which fully exploits all three aspects, brought one more novel and specific tag, "marvel," which represents well the creators of the movie's universe, not to mention that it helps to cover the "comic" topic in a more specific way.

Considering the movie *Harry Potter and the Sorcerer's Stone*, the relevance-driven recommender suggested the tags "award," possibly referring to the several awards the movie won; "shaw," a possible reference to actress Fiona Shaw, who played the character Petunia Dursley; and "based," referring to the fact that the movie is based on a book. For this same example, the second recommender additionally recommended "potter," which is a very specific tag that helps to discriminate this object from the others. Finally, the third recommender suggested an even better set of tags: besides tags already suggested by the other methods, this recommender also suggested the tag "wizard," which is one of the main subjects of the movie.

Although these examples illustrate that diversity and novelty are important aspects for tag recommendation, we are aware of only two previous attempts to address these aspects in the specific context of tag recommendation. In Belém et al. [2013a], we defined the diversity of a list of recommended tags implicitly, as the average semantic distance [Figueiredo et al. 2012; Markines et al. 2009] between each pair of tags in the list, such that a set of synonyms or semantically similar words has low diversity. Novelty, in turn, was defined as the inverse of the popularity of the tag, since too general and overused words tend not to be complementary and discriminative for an object. In contrast, in Belém et al. [2013b], we defined diversity explicitly by exploiting a set of topics represented by categories commonly available in Web 2.0 applications. Diversity was then defined in terms of the capacity of the recommended tags to cover different topics of the target object. Specifically, we proposed the *Explicit Tag Recommendation Diversifier* (*xTReD*), which reranks tag recommendations to promote tags that better contribute to covering the topics of the target object. Both implicit and explicit approaches were extensively evaluated, and the explicit approach was found to be more effective to increase tag recommendation diversity [Belém et al. 2013b].

In common, these two approaches exploit at most two of the following aspects: (1) relevance, (2) explicit topic diversity, and (3) novelty. We here build upon these previous efforts by proposing to tackle all three aspects conjointly. This is a challenging task, as some of these goals may be contradictory. For instance, focusing too much on recommending novel tags may ultimately harm relevance.

Specifically, we propose two new and complementary tag recommendation methods that tackle the problem at different levels. Our first proposal, called *Random Forest with topic -related attributes*, or RF_t, extends a relevance-driven approach based on the Random Forest (*RF*) learning-to-rank (L2R) method. Although the *RF* recommender [Belém et al. 2014] already includes some novelty aspects as tag attributes, RF_t also includes new attributes that capture the extent to which a candidate tag is related to the topics (e.g., categories) of the target object. This solution captures topic diversity and novelty at the *attribute level* while trying to maximize relevance in its objective function. Our second proposal, called *Explicit Tag Recommendation Diversifier with Novelty Promotion* (*xTReND*), is a new reranking strategy that reorders the results produced by any tag recommender to jointly and directly promote relevance, novelty, and topic diversity. Here we use RF_t as the basic recommender used by *xTReND*. Thus, our second solution addresses the target problem at both attribute and objective levels.

Moreover, considering that category information may be unavailable in some Web 2.0 applications, we alternatively adopt latent Dirichlet allocation (LDA) [Blei 2012] to automatically generate topics for objects, thus enhancing the applicability of our solutions. LDA was used because it considers that a document (the textual features of an object, in our case) can be represented as a mixture of multiple topics.

We evaluate our strategies using real data from five popular Web 2.0 applications: Bibsonomy, LastFM, MovieLens, YahooVideo, and YouTube. In our evaluation, we address the following five questions:

Q1: Do our new topic-related attributes contribute to produce better tag recommenders?
Q2: Is our new explicit diversifier and novelty promoter (xTReND) effective if compared to the state-of-the-art tag diversifier (xTReD)?
Q3: How does xTReND perform compared to RF_t, which also captures relevance, diversity, and novelty?
Q4: Is the use of latent topics a viable alternative to our solutions when the target application does not possess an explicit category system to organize content?
Q5: To which extent can we effectively promote novelty and explicit diversity without harming relevance in tag recommendation?

Contributions. The main contributions of this article over prior work, including our own previous studies [Belém et al. 2013a, 2013b, 2014] , are threefold:

(1) We propose two new tag recommendation solutions that jointly address relevance, explicit diversity, and novelty. As a step to design such solutions, we also propose two new tag diversity ranking attributes: topic coverage and topic similarity. Such proposals are unique in the literature.
(2) We use latent topics as an alternative to exploiting category information, which greatly extends the applicability of our solutions.
(3) We perform a thorough evaluation of our new solutions, comparing them with state-of-the-art approaches. Our experimental study investigates the trade-offs among the three main objectives—relevance, diversity, and novelty—assessing to which extent one can improve diversity and novelty without harming relevance. Considering such trade-offs, our results demonstrate the superiority of our new methods over *RF* and *xTReD* as well as the improvements of *xTReND* over the proposed alternative RF_t, exploiting either explicit categories or latent topics as source of topic information.

The rest of this article is organized as follows. Section 2 states our target problem. Section 3 presents the metrics exploited as attributes by the tag recommenders, which in turn are introduced in Section 4. Our experimental methodology and results are

discussed in Sections 5 and 6, respectively. Section 7 discusses related work, whereas Section 8 offers conclusions and directions for future work.

2. PROBLEM DEFINITION

2.1. Novelty and Diversity in Tag Recommendation

In a recommendation task, the novelty of an item (e.g., a tag, a movie, or any type of element being recommended) captures how different this item is from all other items observed in a given context [Vargas and Castells 2011]. For instance, this context can be the items that have been observed by a single user or by a group of users, or even all items in the application. The novelty of a recommended item has been evaluated under two main perspectives in the literature. One captures how rare the given item is compared to the others. From this perspective, the novelty of an item is estimated by the inverse of its popularity in the given context [Vargas and Castells 2011]. The other perspective captures how different (or dissimilar) the item is from the others. Under this perspective, novelty can be estimated by the average distance of the item to the other items in the given context [Vargas and Castells 2011]. In the particular context of tag recommendation, the dissimilarity-based novelty of a candidate tag can also be evaluated at the topic level, as to whether the tag brings a new topic that is associated with the target object to the list of recommendations (see further discussion later). Novelty is an important factor because, in general, the purpose of a recommendation system is to expose the user to a relevant experience (i.e., item) that she would not find easily on her own.

As in Belém et al. [2013a], we here define the novelty of a tag from the perspective of its popularity in the application. In other words, we estimate the novelty of a tag by the inverse of the frequency at which the tag is used in the collection. A term used as tag a large number of times tends to be a more "obvious" recommendation (if relevant at all), thus being of little use (if any) to improve the description of the target object provided by its tag set. We note that according to this definition, noisy terms such as typos may be considered highly novel. However, our methods will jointly exploit novelty, relevance, and diversity, thus minimizing the chance of recommending noise.

The diversity of a list of recommended items can be defined implicitly or explicitly. The former refers to how different the recommended items are from one another [Vargas and Castells 2011; Nehring and Puppe 2002]. Based on this idea, we have previously defined diversity in the specific context of tag recommendation as the average pairwise semantic distance between the top recommended tags such that a list of synonyms or semantically related words present low diversity [Belém et al. 2013a]. An explicit definition of diversity usually exploits a taxonomy, such as a set of categories or topics. In that perspective, a list of recommended items is diverse if it presents items that cover different topics. This approach has been applied in search result diversification [Santos et al. 2010, 2012], being useful to increase the chance that at least one document will satisfy the information need of the target user. This perspective was also employed in the design of the *xTReD* tag recommendation diversifier [Belém et al. 2013b]. The idea is that a diversified list of tags must cover as many topics related to the target object as possible, and as early in the ranking as possible. As in Belém et al. [2013b], here we adopt an explicit definition (i.e., topic related) of diversity in tag recommendation. We note that by increasing topic coverage, one also indirectly reduces the topic redundancy (or repetition) among the recommended items, which also contributes to raise the topic-related dissimilarity-based novelty of the recommendations [Clarke et al. 2008].

Figure 1 summarizes the preceding discussion by presenting the alternative definitions of novelty and diversity, highlighting those explored in this work. Next, we formally define our target recommendation problem. In this definition, and throughout

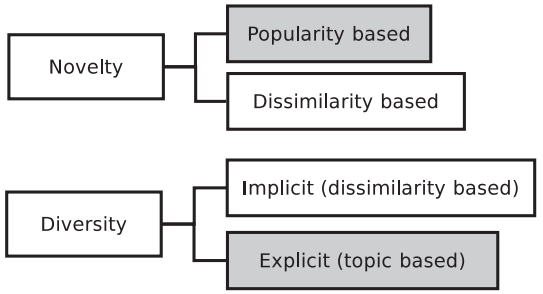


Fig. 1. Different perspectives related to novelty and diversity in tag recommendation. Shaded boxes correspond to the perspectives explored in this work.

the rest of this article, we use the term *diversity* to refer to the explicit (topic-related) diversity (and, indirectly, the topic-related perspective of novelty) and the term *novelty* to refer to the aforementioned popularity-based perspective of novelty.

2.2. Tag Recommendation Task

The tag recommendation problem that we address here can be stated as follows: *Given a set of initial tags I_o priorly assigned to the target object o, and a set of textual features $F_o = \{F_o^1, \ldots, F_o^n\}$, where F_o^i contains the terms in textual feature i of o, produce a sorted list of candidate tags C_o ($C_o \cap I_o = \emptyset$) so that the relevance, novelty, and diversity objectives are maximized, and recommend the k candidates in the top positions of C_o.* Note that in its general form, this problem includes the recommendation to objects with no tags (i.e., $I_o = \emptyset$). Yet, here we focus on scenarios when some initial tags have been assigned to the target object o (i.e., $I_o \neq \emptyset$), leaving the more general case to future work.[2]

Our general strategy to solve this problem comprises the following steps: (1) we incorporate attributes capturing the novelty and diversity of a candidate tag into a state-of-the-art relevance-driven tag recommendation strategy; (2) we use this *implicit* strategy to produce relevance estimates and to rank candidate tags; and (3) the candidate tags are reranked to increase novelty and diversity, by better covering the topics of the target object and bringing more novel (i.e., unpopular) candidates into the top recommendations. Based on such steps, we propose two specific approaches. The first one consists of executing only steps (1) and (2), thus capturing diversity and novelty at the attribute level. The second approach consists of all three steps, addressing relevance, diversity, and novelty jointly, at both attribute and objective levels. We focus on recommending tags for a target object, aiming at improving the quality of the tags in the objects, and indirectly the effectiveness of services that use tags as a data source. We leave the task of tackling novelty and diversity for personalized tag recommendation as future work.

As will be presented in the next sections, our proposed solutions exploit several metrics of tag relevance, diversity, and novelty as inputs to an L2R algorithm. To compute such metrics and learn the recommendation functions as well as to evaluate them, we exploit a training set $\mathcal{D} = \{\langle I_d, F_d \rangle\}$, where I_d ($I_d \neq \emptyset$) contains all tags assigned to object d, and F_d contains the term sets of the other textual features associated with d. There is also a test set $\mathcal{O}$, which is a collection of tuples $\{\langle I_o, F_o, Y_o \rangle\}$, where both I_o and Y_o are sets of tags associated with object o. However, whereas tags in I_o are known (and

[2]The scenarios when $I_o = \emptyset$ are examples of *cold start*. Our new solutions can be applied in such scenarios, as they do not rely only on previously assigned tags to recommend new tags. However, their effectiveness in cold start requires further investigation [Martins et al. 2013].

given as input to the recommenders), tags in Y_o are assumed to be unknown and are taken as the relevant recommendations to the target object o (i.e., the *gold standard*). This split of the tags of each test object is done to enable an automatic assessment of the recommendations, as performed in various previous studies [Rendle and Schmidt-Thie 2010; Yin et al. 2013] and further discussed in Section 5.2. Similarly, there is a validation set $\mathcal{V}$ used for tuning parameters and learning recommendation functions (see Section 5.2). Thus, each object v in $\mathcal{V}$ also has its tag set split into input tags (I_v) and gold standard (Y_v).

3. TAG RECOMMENDATION METRICS

In this section, we introduce the metrics used to capture the relevance, novelty, and diversity of a candidate tag, which are used as tag attributes by the recommendation methods. The relevance and novelty metrics, presented in Sections 3.1 and 3.2, respectively, have been previously proposed [Belém et al. 2013a, 2014; Lipczak and Milios 2011; Vargas and Castells 2011]. The two diversity metrics introduced in Section 3.3 are novel contributions of this work.

3.1. Relevance Metrics

The relevance metrics used in this article, presented in Table II, are grouped into the following four categories based on the aspect they try to capture regarding the tag recommendation task: tag co-occurrence patterns, descriptive and discriminative capacities, and predictability.

Metrics related to *co-occurrence patterns* estimate the relevance of a candidate tag by the frequency and confidence at which they co-occur with tags previously assigned to the target object. In other words, given the initial set of tags I_o of target object o, tags that are often used jointly with tags in I_o are considered good candidates to be recommended. *Descriptive* metrics, in turn, estimate the relevance of a candidate tag c based on how closely it relates to the textual content of the target object, whereas *discriminative* metrics estimate the relevance of a candidate c by its capacity to distinguish the target object o from the others. Finally, *predictability* metrics indicate the likelihood that a term (a candidate) will be used as a tag.

3.2. Novelty Metric

Vargas and Castells [2011] proposed to estimate the novelty of an item in a list of recommendations as the probability that it has not been previously observed. Thus, the lower the popularity of an item, the more novel it is. Bringing this definition to the context of tag recommendation, we note that the *IFF* metric (Equation (9) in Table II) does capture this aspect exactly, as it favors candidates that occur less frequently in the training set. Thus, although in Belém et al. [2014] *IFF* was employed to recommend tags that can better discriminate an object from the others, an aspect that is related to the relevance of the tag to the target object, the same metric was used in Belém et al. [2013a] to increase the novelty of the recommendations—that is, to recommend possibly relevant tags that, because they occur very rarely in the training set, would hardly be recommended by traditional methods.Similarly, here we use *IFF* as a novelty metric as well.

3.3. Diversity Metrics

Considering that users often associate tags to Web content with organization and categorization purposes [Golder and Huberman 2005], tags that are more related to the topics (e.g., categories) of the target object are good candidates for recommendation. The new metrics that we propose here—topic coverage and topic similarity—exploit this idea.

Table II. Metrics Used to Estimate the Relevance of a Candidate Tag c as a Recommendation for an Object o [Belém et al. 2014; Lipczak and Milios 2011]

	Name	Equation/Description
Tag Co-occurrence	*Sum*	Let X be a set of tags and c a candidate tag. $X \rightarrow c$ is an association rule and $\theta(X \rightarrow c)$ is its *confidence*. *Sum* is calculated as $$Sum(c, I_o, \ell) = \sum_{X \subseteq I_o} \theta(X \rightarrow c), \quad (X \rightarrow c) \in \mathcal{R}, |X| \leq \ell, \tag{1}$$ where $\mathcal{R}$ is a set of association rules computed offline over the training set $\mathcal{D}$, and ℓ is the size limit for the antecedent X. As in Belém et al. [2014], we use the *Apriori* algorithm to generate these rules.
	Sum^+	$$Sum^+(c, I_o, k_x, k_c, k_r) = \sum_{x \in I_o} \theta(x \rightarrow c) \times Stab(x, k_x) \times Stab(c, k_c) \times Rank(c, x, k_r), \tag{2}$$ where $Stab(x, k_x)$ is defined in Equation (10), and k_x, k_c and k_r are tuning parameters. $Rank(c, x, k_r)$ is equal to $k_r/(k_r + p(c, x))$, where $p(c, x)$ is theposition of c in the ranking of candidates according to the confidence of the corresponding association rule (whose antecedent is x).
	$Vote$	$$Vote(c, I_o) = \sum_{x \in I_o} j, \text{ where } j = \begin{cases} 1 & \text{if } (x \rightarrow c) \in \mathcal{R} \\ 0 & otherwise \end{cases} \tag{3}$$
	$Vote^+$	$$Vote^+(c, I_o, k_x, k_c, k_r) = \sum_{x \in I_o} j \times Stab(x, k_x) \times Stab(c, k_c) \times Rank(c, x, k_r),$$ $$\text{where } j = \begin{cases} 1, & if \quad x \rightarrow c \in \mathcal{R} \\ 0, & otherwise \end{cases} \tag{4}$$
Descriptive Power	*Term Spread (TS)*	$$TS(c, o) = \sum_{F_o^i \in F_o} j, \text{ where } j = \begin{cases} 1 & \text{if } c \in F_o^i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$
	Term Frequency (TF)	$$TF(c, o) = \sum_{F_o^i \in F_o} tf(c, F_o^i), \tag{6}$$ where $tf(c, F_o^i)$ is the number of occurrences of c in textual feature F_o^i of object o.
	Weighted Term Spread (wTS)	Let the *Feature Instance Spread* of a feature F_o^i associated with object o, $FIS(F_o^i)$, be the average TS over all terms in F_o^i. We define the *Average Feature Spread* $AFS(F^i)$ as the average $FIS(F_o^i)$ over all instances of F^i associated with objects in the training set $\mathcal{D}$. The wTS is defined as $$wTS(c, o) = \sum_{F_o^i \in F_o} j, \text{ where } j = \begin{cases} AFS(F^i) & \text{if } c \in F_o^i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$
	Weighted Term Frequency (wTF)	$$wTF(c, o) = \sum_{F_o^i \in F_o} tf(c, F_o^i) \times AFS(F^i) \tag{8}$$

(Continued)

Table II. Continued

	Name	Equation/Description
Discriminative Power	*Inverse Feature Frequency (IFF)*	$IFF(c) = log \frac{\|\mathcal{D}\| + 1}{f_c^{tag} + 1}$, (9) where f_c^{tag} is the number of objectsin the training set $\mathcal{D}$ that contain c *associated as a tag*.
	Stability (Stab)	$Stab(c, k_s) = \frac{k_s}{k_s + \|k_s - log(f_c^{tag})\|}$, (10) where the tuning parameter k_s represents the "ideal frequency" of a term in the data collection.
Term Predictability	*Entropy (H^{tags})*	$H^{tags}(c) = -\sum_{(c \to i) \in \mathcal{R}} \theta(c \to i)\, log\, \theta(c \to i)$ (11)
	Predictability (Pred)	$Pred(c) = \frac{f_c^{tag,F}}{f_c^F}$, (12) where $f_c^{tag,F}$ is the number of objects in the training set $\mathcal{D}$ in which c appears both as a tag and as a term in *any* other textual feature, and f_c^F is the number of objects in which c is a term associated with any of its textual features (except tags).

Before introducing them, we first estimate the probability that a topic z is associated with a tag t, $\Pr(z|t)$, as $\Pr(z|t) = f(t,z)/f(t)$, where $f(t,z)$ is the number of objects in which z appears as a topic and t appears as a tag, and $f(t)$ is the number of objects containing tag t, both computed in the training set $\mathcal{D}$. We also estimate the probability that a topic z is associated with an object o, $\Pr(z|o)$, as $1/n_o$, where n_o is the number of categories associated with object o, if categories are available. Otherwise, we use the result produced by the LDA algorithm (described in Section 5.2.1) to estimate $\Pr(z|o)$.

Let Z_o be the set of topics associated with object o. We define the *topic coverage* of a candidate tag c for o, $TC(c,o)$, as the fraction of topics of o covered by c—that is,

$$TC(c,o) = \frac{1}{|Z_o|} \sum_{z \in Z_o} J(c,z) \text{ where } J(c,z) = \begin{cases} 1, & \text{if } \Pr(z|c) > \Pr(z) \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

where $|Z_o|$ is the number of topics associated with object o, and $\Pr(z)$ is the prior probability of topic z—that is, the fraction of all objects in the training set $\mathcal{D}$ that have topic z associated with them. We consider that candidate c covers a topic z (c is highly related to z) if the probability of topic z given c is higher than the (prior) probability of z.

Multiple topics may be associated with a given object or tag, although the strength of the semantic association between them may vary across different topics. Yet, the topic coverage metric does not capture such variability. Thus, we propose the topic similarity metric, which measures the cosine similarity between the distribution of topics of the candidate tag and the distribution of topics of the target object, and thus takes the strength of the semantic association between topic and object (or tag) into account. We estimate the strength of the association between topic z and object o by the probability of the topic given the object $\Pr(z|o)$. Similarly, the strength of the association between z and candidate tag c is estimated by $\Pr(z|c)$. The *topic similarity*of a candidate tag c with relation to a target object o, $TSim(c,o)$, is then defined as

$$TSim(c,o) = \frac{\sum_{z \in Z_o} \Pr(z|o) \times \Pr(z|c)}{\sqrt{\sum_{z \in Z_o} (\Pr(z|o))^2} \times \sqrt{\sum_{z \in Z_o} (\Pr(z|c))^2}}. \quad (14)$$

4. TAG RECOMMENDATION STRATEGIES

This section describes the tag recommendation methods analyzed in this article. All approaches employ the *RF* L2R method to build a ranking function that assigns scores to candidate tags according to some criterion (probability of relevance, in the present case), as it was found to be the best L2R algorithm, out of a large set of methods, for this task [Canuto et al. 2013]. Moreover, all of them exploit (subsets of) the tag quality metrics introduced in Section 3 as tag attributes, extracting candidate tags from (1) co-occurrence patterns with tags already in the target object o (i.e., tags in I_o), learned using association rules,[3] and (2) other textual features in o, namely title and description.

We briefly describe our previous relevance-driven method, referred to here as simply *RF*, and the *xTReD* diversifier in Sections 4.1 and 4.2, respectively. We introduce our twonew approaches to jointly address relevance, diversity, and novelty—RF_t and *xTReND*—in Section 4.3.

4.1. Relevance-Driven Tag Recommendation Strategy

We adopt the state-of-the-art *RF*-based tag recommendation strategy as a relevance-driven baseline [Canuto et al. 2013]. *RF* is an ensemble learning method in which n_t decision trees are trained with distinct subsets of the training set (with size n_b), randomly sampled with replacement to reduce the correlation among them. The decision tree learning happens in a recursive way: first, the most discriminative tag attribute (according to some measure, e.g., Information Gain) is selected as a decision node. The selected training samples are split according to a split value (e.g., the average attribute value), and the process repeats in a top-down fashion. To further reduce the correlation between the decision trees, before each split, a fraction ϕ of the attributes is randomly selected to be considered as candidates for splitting, instead of considering the whole set of attributes. The decision rule is given by averaging the n_t predictions of the trained trees. The crucial aspects that make *RF* a good learner are (*i*) the reduced correlation between the decision trees composing the ensemble and (*ii*) the better-than-random-guess predictions of each tree. To achieve strong decision trees (i.e., trees with better prediction performance than random guessing), each tree is typically grown to its maximum depth, containing n_l terminal nodes, where n_l is a tuning parameter.

To apply *RF* to the tag recommendation context, each candidate tag for each object in the training, test, and validation sets is represented as an attribute vector containing the values of different tag quality metrics. Each candidate tag c in the training and validation sets is also associated with a binary label y_c, indicating if c is relevant ($y_c = 1$) or not ($y_c = 0$). Whereas the training set is a source of data for the computation of tag quality metrics, the validation set provides the examples necessary to learn the model (i.e., a tag score function f). As in Canuto et al. [2013], we use the values of the tag relevance metrics presented in Table II as attributes. We refer to this baseline method as *RF*.

4.2. Explicit Tag Recommendation Diversifier

We previously proposed to address diversity in tag recommendation in an explicit way—that is, by seeking to directly maximize the set of categories covered by the recommended tags [Belém et al. 2013b]. In its general form, maximizing topic coverage is an NP-hard problem [Agrawal et al. 2009]. Fortunately, there is a well-known greedy algorithm for this problem, which achieves an approximation factor of $(1 - 1/e) \approx 0.632$ of the optimal solution [Hochbaum 1997]. This is also the best possible polynomial-time

[3]Association rules are generated using the *Apriori* algorithm [Agrawal and Srikant 1994].

ALGORITHM 1: The *xTReD* Algorithm

xTReD(o, τ)

1: $C_o^\tau \leftarrow rec(o, \tau)$ // relevance-driven recommendations
2: $C_o^S \leftarrow \emptyset$
3: **while** $|C_o^S| < min(\tau, |C_o|)$ **do**
4: $\quad t^* \leftarrow argmax_{t \in C_o^\tau} f(o, t, C_o^S)$
5: $\quad C_o^\tau \leftarrow C_o^\tau \setminus \{t^*\}$
6: $\quad C_o^S \leftarrow C_o^S \cup \{t^*\}$
7: **end while**
8: **return** C_o^S

approximation for the problem, unless NP $\subseteq$ DTIME($n^{O(\log \log n)}$), where n is the number of items to be diversified [Feige 1998; Khuller et al. 1999]. Our method, called *Explicit Tag Recommendation Diversifier*, or *xTReD*, builds upon this greedy approach, as described in Algorithm 1.

Specifically, *xTReD* takes as input an object o and a diversification cutoff τ. In its first step, *xTReD* calls a tag recommendation method *rec* to produce an initial ranking C_o of recommended tags, generated with a relevance-focused objective (line 1). Any relevance-driven tag recommender could be used in this step. As in Belém et al. [2013b], here we use *RF*.

Let C_o^τ be the top τ recommendations in C_o. The goal is to produce a permutation of C_o^τ so as to raise the diversity in the top positions of the ranking of recommended tags, given that those tags are often the ones at which the user looks. A complete permutation of C_o ($\tau = |C_o|$) could be produced. However, we can reduce τ for efficiency reasons and as a means to restrict the search for more diverse tags among the most relevant ones, avoiding severe relevance penalties.

The permutation C_o^S is initialized as an empty set (line 2) and is iteratively constructed (lines 3 through 7). The submodular objective function $f(o, t, C_o^S)$ scores each yet unselected tag $t \in C_o^\tau \setminus C_o^S$ in light of the object o and the tags already in C_o^S, selected in the previous iterations of the algorithm (line 4). The highest scored tag, t^*, is then removed from C_o^τ (line 5) and added to C_o^S (line 6). Finally, the produced diverse ranking C_o^S is returned (line 8).

To instantiate the objective function $f(o, t, C_o^S)$ in Algorithm 1, *xTReD* builds upon a state-of-the-art framework for diversifying search results, called *xQuAD* [Santos et al. 2010]. The xQuAD framework instantiates the aforementioned function to score the documents retrieved for a given query proportionally to these documents' coverage, and novelty in light of the multiple possible information needs underlying this query [Santos et al. 2010; Santos and Ounis 2011]. In the context of *xTReD*, instead of a ranking of documents for a query, we seek to diversify a ranking of tags for a given object. More precisely, *xTReD* includes a new instantiation of the objective function $f(o, t, C_o^S)$ such that

$$f\left(o, t, C_o^S\right) = (1 - \lambda) \times \Pr(t|o) + \lambda \times \sum_{z \in Z_o} \Pr(z|o) \Pr(t|o, z) \prod_{t' \in C_o^S} (1 - \Pr(t'|o, z)), \qquad (15)$$

where Z_o is a set of topics associated with the object o, and $0 \leq \lambda \leq 1$ is a tuning parameter used to balance the trade-off between promoting relevance or diversity. The greater the value of λ, the more importance is given to diversity. The idea is to promote tags that are simultaneously highly related to at least one of the topics of the target object and little related to the topics of the tags already selected as recommendation

(captured by the product over $t' \in C_o^S$), hence increasing the coverage of topics in the top positions of the list of recommendations.

When $\lambda = 0$, Equation (3) reduces to $\Pr(t|o)$, which results in a pure relevance-driven tag recommendation, as produced by a nondiversification baseline. In our experiments in Section 6, we define $\Pr(t|o) = 1/r_t$, where r_t is the position of the tag t in the ranking produced by the initial ranker *rec*. To estimate the second half of Equation (15), we infer the distribution $\Pr(z|o)$ of topics $z \in Z_o$ for an object o from the available training data or using the LDA algorithm, as discussed in Section 3.3. Finally, to estimate how much a given tag t covers the topic z of the object o, we approximate the probability $\Pr(t|o, z)$ as $\Pr(t|o, z) \approx \Pr(t|o) \times \Pr(z|t)$, where $\Pr(z|t)$ is an estimate of the probability that tag t is related to topic z, defined in Section 3.3.

4.3. Our Novel Relevance-, Novelty-, and Diversity-Driven Tag Recommenders

We propose two new tag recommendation strategies that jointly exploit relevance, novelty, and diversity. The first one extends *RF* to include the new topic coverage and topic similarity metrics (Equation (13) and Equation (14), respectively), which capture explicit diversity as tag attributes. It also includes *IFF* as an attribute, capturing aspects related to both relevance (i.e., discriminative capacity) and novelty (i.e., rarity). Like *RF*, our new method, referred to as RF_t, still has the objective of maximizing relevance of the recommendations, capturing novelty and diversity indirectly at the attribute level.

Our second approach, *xTReND*, builds upon *xTReD* and RF_t. Although it uses the same general algorithm described in the previous section (Algorithm 1), it differs from *xTReD* in two core components. First, it employs RF_t as the basic recommender (line 1) and thus already captures relevance, novelty, and diversity at the attribute level. Second, it uses a new instantiation of the objective function that also captures the same three aspects. The new objective function is defined as

$$f\left(o, t, C_o^S\right) = (1 - \alpha - \beta) \times \Pr(t|o) + \alpha \times \mathit{IFF}(t) + \beta \times \sum_{z \in Z_o} \Pr(z|o) \Pr(t|o, z) \prod_{t' \in C_o^S} (1 - \Pr(t'|o, z)), \tag{16}$$

where *IFF* is the novelty metric defined in Equation (9). The tuning parameters α and β ($0 \leq \alpha, \beta \leq 1$) are used to balance the trade-off between promoting relevance or novelty or diversity. The higher the values of α and β, the more weight is given to novelty and diversity, respectively.

Thus, *xTReND* captures relevance, (popularity-based) novelty, and explicit diversity at both attribute and objective levels. Its design is motivated by (1) the effectiveness of the *RF* L2R method for tag recommendation [Canuto et al. 2013], and particularly of its extension RF_t proposed in this work, (2) the superiority of capturing diversity explicitly,as opposed to only implicitly, for tag recommendation [Belém et al. 2013b], and (3) the absence of a previous approach that directly includes popularity-based novelty, in addition to explicit topic diversity and relevance, as part of the goal to be maximized.

To better distinguish *xTReND* from our prior diversifier *xTReD*, Figure 2 illustrates the general structure of these methods and their expected results. In the figure, the rectangles represent the ranked list of recommended tags. We use different colors to represent different topics related to the target object.[4] Focusing first on *xTReD*

[4]For the sake of simplicity, we assume in this example that a single topic (color) is associated to each tag (rectangle). In reality, multiple topics may be associated to the same tag t, and the strength of the semantic association between them is given by $\Pr(z|t)$.

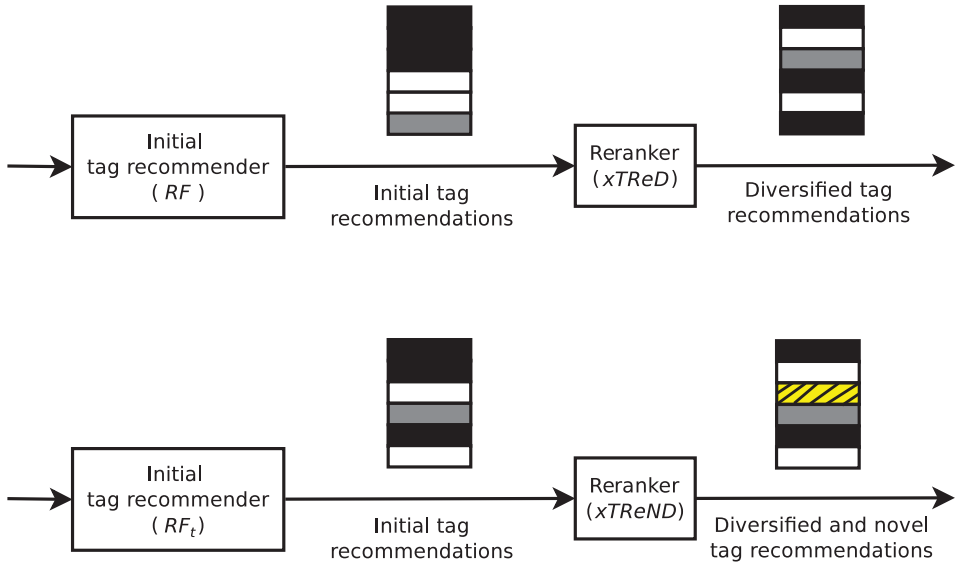


Fig. 2. Illustration of *xTReD* and *xTReND*: general structure and expected results. The rectangles represent the ranked list of recommended tags, and each color represents a different topic related to the target object.

Table III. Example of the Reranking Step of *xTReND* for the Movie *X-Men: The Last Stand*: Statistics of Top Candidate Tags (Candidates Are Sorted by Relevance)

Candidate Tag	Relevance Estimate	Novelty Estimate	Topic Probability Estimates: $\Pr(z\|t)$		
(t)	$\Pr(t\|o)$	*IFF(t)*	*Fantasy*	*Thriller*	*Sci-Fi*
dvd	1.00	1.72	0.04	0.10	0.04
genetics	0.50	6.07	0.00	0.17	0.30
biology	0.33	6.07	0.00	0.23	0.27
comics	0.25	4.21	0.09	0.10	0.11
mckellen	0.20	5.87	0.09	0.09	0.09
marvel	0.17	5.56	0.09	0.16	0.16
mutant	0.14	6.19	0.00	0.09	0.23
super-hero	0.13	5.01	0.07	0.17	0.17
based	0.11	2.26	0.05	0.08	0.07
ummarti2006	0.10	4.69	0.05	0.13	0.05

(top diagram in Figure 2), the initial tag recommender, *RF*, which is driven only by relevance, prioritizes one topic (represented by the black color) over the others, whereas *xTReD* rearranges the results so as to allow tags related to different topics to appear earlier in the ranking. In contrast, *xTReND* (bottom diagram) uses RF_t as the initial recommender, which already introduces some diversification and novelty to the results, compared to *RF*. Like *xTReD*, the *xTReND* reranker also promotes tags related to different topics to earlier positions of the ranking, Additionally, *xTReND* is also able to bring a tag related to a novel topic (represented by the yellow rectangle with diagonal lines) to the object's top recommendations.

We further illustrate the reranking step performed by *xTReND* by focusing on one of the real examples mentioned in Section 1, namely the recommendations for the movie *X-Men: The Last Stand*. For simplicity, consider that only the top $\tau = 10$ candidate tags will be reranked. Table III shows, for each of the top 10 most relevant tag candidates, their estimated values of relevance and novelty, and how much they are related to the

Table IV. Example of the Reranking Step of *xTReND* for the Movie *X-Men: The Last Stand*: $f(o, t, C_o^S)$ Scores for Each Candidate Tag in Each Iteration (Candidates Are Shown in the Order Selected by the Method)

Candidate Tag	Iteration									
	1	2	3	4	5	6	7	8	9	10
genetics	0.177	-	-	-	-	-	-	-	-	-
dvd	0.163	0.158	-	-	-	-	-	-	-	-
biology	0.141	0.135	0.131	-	-	-	-	-	-	-
mckellen	0.093	0.092	0.091	0.090	-	-	-	-	-	-
marvel	0.091	0.089	0.088	0.087	0.087	-	-	-	-	-
mutant	0.088	0.087	0.086	0.085	0.085	0.084	-	-	-	-
comics	0.087	0.085	0.084	0.083	0.083	0.082	0.082	-	-	-
super-hero	0.077	0.075	0.074	0.073	0.073	0.073	0.073	0.072	-	-
ummarti2006	0.063	0.063	0.062	0.062	0.062	0.061	0.061	0.061	0.061	-
based	0.039	0.039	0.038	0.038	0.038	0.038	0.038	0.038	0.037	0.037

three topics (i.e., genres) of the movie. Note that the first column of the table presents the candidate tags sorted by relevance. Table IV shows the $f(o, t, C_o^S)$ scores calculated in each iteration of the methods. Entries containing "-" indicate tags that were already selected in previous iterations. Note that the first column of Table IV presents the list of candidate tags in the order they were selected by *xTReND*—that is, the list of candidates *after* reranking.

In the first iteration of the reranking, no tags have been selected yet ($C_o^S = \emptyset$). Tag "genetics" has the highest score, probably because it presents the highest probability to be related to one of the topics of the movie (Sci-Fi), and also presents good relevance and novelty estimates. Since no movie genre of the considered object has been covered yet, all genres are equally good choices to be covered first.[5] Thus, the algorithm appends tag "genetics" to the new, reranked list of tag recommendations. Next, in the second iteration, tag "dvd" is selected, despite being little related to any of the three genres. This choice is due the high relevance estimate given by the initial recommender, RF_t (see Table III). Following, tag "biology" is selected in the third iteration. This tag is relatively well connected to topic Thriller, which was not yet well covered by the previously selected candidates, according to the statistics of tag occurrences in genres of our MovieLens dataset. Tag "mckellen," referring to one of the main actors, is the next one selected. This tag is somewhat related to all three genres of this movie, since the actor starred other movies of these genres (e.g., other X-Men movies and *The Lord of the Rings*), not to mention that it is also highly novel and specific. Next, tag "marvel" is appended. In comparison with "comics," which appeared first in the relevance-driven ranking (see Table III), "marvel" is more novel and specific, as well as more strongly related to the topics of the considered movie (according to the topic probability estimates). Thus, "marvel" is ranked higher than "comics" after the reranking. The other tags are appended similarly, considering the best trade-off between relevance, novelty, and topic diversity.

5. EXPERIMENTAL SETUP

5.1. Datasets

Our evaluation was performed using five datasets, containing *title*, *tags*, and *description* associated with objects from Bibsonomy, LastFM, MovieLens, YahooVideo, and YouTube. The MovieLens dataset is publicly available.[6] The Bibsonomy, LastFM,

[5]Recall that we are assuming a uniform distribution of the topics (i.e., genres) related to the movie.
[6]http://www.grouplens.org/taxonomy/term/14.

Table V. Overview of Our Datasets

Dataset	Total Number of Objects	Sample Size (after Filtering)
Bibsonomy	543,872	150,000
LastFM	2,758,992	35,975
MovieLens	10,000	6,500
YahooVideo	160,228	140,000
YouTube	9,000,000	150,000

YahooVideo, and YouTube datasets[7] are the same used in Belem et al. [2013a, 2014], and they are also available for public use.[8]

We used different sources of category information for our datasets. Specifically, we used the available categories for YouTube videos as well as the genres associated with each movie in MovieLens. We also collected the musical styles associated with the artists in the LastFM dataset from the AllMusic site,[9] and used them as artist categories. The Bibsonomy and YahooVideo datasets do not contain categories and thus were evaluated using latent topics only.

We removed *stopwords* and applied the Porter stemming algorithm[10] to avoid trivial recommendations (e.g., plurals and other variations of the same word). We also discarded objects with fewer than two tags. Table V shows the total number of objects in our original datasets as well as the sizes of the samples used in our evaluation (after filtering).

5.2. Evaluation Methodology

As in most tag recommendation studies [Heymann et al. 2008; Garg and Weber 2008; Rendle and Schmidt-Thie 2010; Belém et al. 2013a, 2014; Lipczak and Milios 2011; Chen and Shin 2013; Lops et al. 2013], we adopted an *automatic* evaluation approach: half of the tags, randomly selected from all tags previously assigned to an object, are used as the gold standard (i.e., as the relevant tags for that object). These tags are *not* used by the tag recommendation methods, thus being removed from $\mathcal{I}_o$, which is used as input by those methods.

Alternative evaluation approaches would rely on manual evaluation by either the users who created the test objects or by external users. An extensive discussion about the advantages and drawbacks of each evaluation strategy for different situations can be found in Belém et al. [2014]. In summary, we argue that, although desirable, a manual evaluation by the users who created the test objects is very hard to perform, especially if one is covering various systems and methods (as we do). Alternatively, relying on external users to evaluate the recommendations is also expensive and does not scale to the size of our datasets. Moreover, external evaluators might introduce biases to the evaluation, as they have different backgrounds and perceptions of the target content when compared to the real users of the system. The impact of such biases on the conclusions is unknown. Instead, the automatic evaluation simulates a manual evaluation by real users who created the object and/or added tags to it, who are ideal evaluators.

The experiments were performed using a fivefold cross-validation procedure. In other words, the objects were randomly distributed into five equal-size portions. Three portions were treated as the training set, which was used to compute all tag quality metrics

[7]https://figshare.com/articles/data_tar_gz/2067183.

[8]All four datasets are available at https://figshare.com/articles/data_tar_gz/2067183. The Bibsonomy dataset was originally obtained from http://www.kde.cs.uni-kassel.de/bibsonomy/dumps.

[9]http://www.allmusic.com.

[10]http://snowball.tartarus.org/algorithms/porter/stemmer.html.

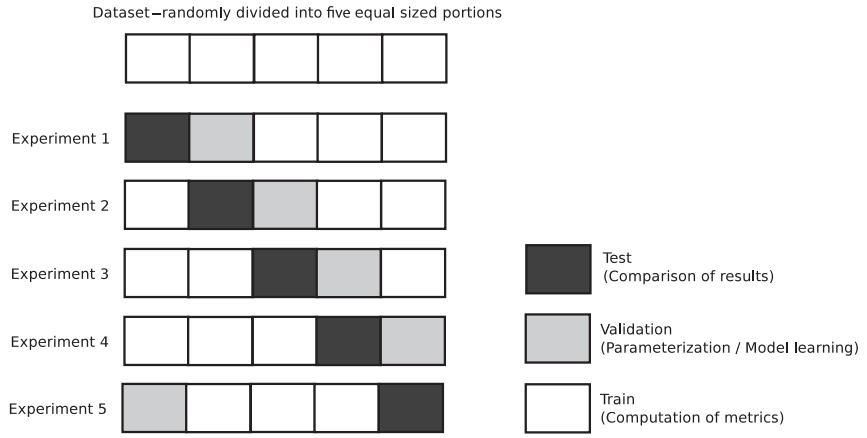


Fig. 3. Illustration of the fivefold cross-validation procedure.

(i.e., tag attributes) exploited by the tag recommendation methods. A fourth portion was used as the validation set, which in turn was used to learn the solutions (i.e., generate the ranking function by RF and RF_t) and to tune parameters of the methods (e.g., λ, α, β, and RF parameters). The last portion was used for testing. We repeat this procedure five times, alternating the roles of each portion of the dataset, as illustrated in Figure 3. For each configuration of training, validation, and test sets, we repeat the experiments five times using different seeds for the random number generator.

As discussed in Section 3.3, we estimate how related a tag t is to a topic z, which is necessary to evaluate topic diversity, by the probability of z given t. We experimented with two sources of topics for objects: (1) an explicit taxonomy represented by categories, obtained from our datasets (see Section 5.1), and (2) implicit topics generated by an unsupervised clustering technique. Specifically, we used LDA [Blei 2012], a probabilistic approach to generate and assign topics for each object based on terms (tags) contained in it. The use of LDA allows us to evaluate our approach in datasets where each object has only one category, as in YouTube, or no explicit category at all, and to compare results across scenarios with different levels of generalization of categories. Next, we further describe the LDA method (Section 5.2.1) and introduce our main evaluation metrics (Section 5.2.2).

5.2.1. Latent Dirichlet Allocation. LDA is a probabilistic model that is based on the assumption that a document can be represented as a mixture of different topics [Blei 2012], whereas a topic is defined as a distribution of words from a fixed vocabulary. Given several topics and the distribution of words for each topic, LDA can be described by a generative process that explains how the content of a given document arises. Specifically, the method "generates" words for a given document as follows:

(1) Randomly choose a topic distribution;
(2) For each word to be generated in the document:
    (a) Randomly draw a topic from the distribution in (1);
    (b) Randomly draw a word from the word distribution corresponding to the selected topic.

This process assumes that each document exhibits topics in different proportions (step 1); each topic associated with a document is drawn from a per-document distribution (step 2a); and each word in the document is drawn from one of its topics (step 2b). In our case, documents refer to objects o in our datasets, and words refer to

tags in $\mathcal{I}_o$ (i.e., previously assigned tags that were not included in the gold standard).[11] In general, the goal of LDA is to exploit the *observed* terms (tags) in objects to infer their *hidden* topic structure (distribution). This can be thought of as "reversing" the aforementioned generative process.

Formally, given $\Pr(z|o)$, the probability distribution of topics for object o, and $\Pr(t|z_i)$, the distribution of tags for a latent topic z_i, the probability $\Pr(t_i|o)$ of a tag t_i appearing in an object o is defined as

$$\Pr(t_i|o) = \sum_{j=1}^{n_Z} \Pr(t_i|z_i = j) \Pr(z_i = j|o), \tag{17}$$

where $\Pr(t_i|z_i = j)$ is the probability of tag t_i appearing in topic j, and $\Pr(z_i = j|o)$ is the probability of topic j being associated with object o. The number of latent topics n_Z is a parameter that allows us to adjust the level of generalization/specificity of topics. The larger the number of topics, the more specific the generated topics.

LDA estimates the distribution of tags given a topic $\Pr(t|z)$ as well as the distribution of topics in an object $\Pr(z|o)$ from a set of unlabeled objects (training set) assuming a *prior* Dirichlet distribution and a fixed number n_Z of topics. A possible approach to infer these probabilities is to use *Gibbs sampling* [Blei 2012], a sampling method performed in m iterations of the two-step method described earlier. This is the method adopted in *pLDA*,[12] which is the implementation of LDA [Liu et al. 2011] that we used to generate topics for each object in our datasets. We discuss how we set the values of n_Z and m in Section 5.3.

5.2.2. Evaluation Metrics. We evaluate the tag recommendation methods in terms of the relevance, novelty, and diversity of their results. We use NDCG@k [Baeza-Yates and Ribeiro-Neto 1999] to assess the relevance of a ranked list of tags. Specifically, let Y_o be the set of relevant tags for object o, C the sorted list of recommendations generated by the method being evaluated, C^k the top k elements in C, and C_i the i^{th} element in C. In addition, let $disc(i) = 1/log(1 + i)$ be a rank discount function that provides a weight for the i^{th} position of the ranking. The *discounted cumulative gain* in the first k recommendations of the ranking, DCG@k, is defined as

$$DCG@k(C, Y_o) = \sum_{i=1}^{k} rel(i) \times disc(i), \tag{18}$$

where $rel(i)$ is equal to 1 if the i^{th} element returned in C is relevant (i.e., $C_i \in Y_o$) and 0 otherwise. The *normalized discounted cumulative gain* in the first k recommendations, NDCG@k, is defined as

$$NDCG@k(C, Y_o) = \frac{DCG@k(C, Y_o)}{IdealDCG@k}, \tag{19}$$

where $IdealDCG@k$ is the value obtained for DCG@k when there are only relevant candidates at the top-k (or fewer) positions.

To assess the diversity of a list of recommended tags, we traditionally use three metrics for evaluating search result diversification methods [Clarke et al. 2011; Santos et al. 2010; Vargas et al. 2012; Dang and Croft 2012]. Two of them—α-NDCG and ERR-IA—are the primary evaluation metrics used in the diversity task of the TREC Web

[11] Initial experiments showed that using terms extracted from other textual features of the object did not improve results, but we intend to further exploit this direction in the future.
[12] http://code.google.com/p/plda.

track [Clarke et al. 2012]. They are cascade metrics that penalize redundancy (and thus also capture the topic-related novelty discussed in Section 2) by modeling the behavior of a user who stops inspecting the ranking once a relevant tag is observed [Vargas and Castells 2011]. Whereas α-NDCG incorporates a notion of the *expected gain* attained by each ranked tag, ERR-IA measures the *expected retrieval performance* with respect to multiple topics.

Specifically, to define α-NDCG@k, we first define α-DCG@k as

$$\alpha\text{-}DCG@k(C, o) = \sum_{i=1}^{k} disc(i) \times \sum_{z \in Z_o} J(C_i, z)(1-\alpha)^{r(i,z,C)}, \tag{20}$$

where *disc* is the same discount function used by *DCG*, and $J(C_i, z)$ is equal to 1 if the i^{th} candidate returned in C is related to topic z and 0 otherwise. Function $r(i, z, C)$ outputs the number of candidates in C recommended before the i^{th} position that are related to topic z—that is,

$$r(i, z, C) = \sum_{j=1}^{i-1} J(C_j, z). \tag{21}$$

The normalized α-DCG@k, α-NDCG@k, is defined as

$$\alpha\text{-}NDCG@k(C, o) = \frac{\alpha\text{-}DCG@k(C, o)}{IADCG@k}, \tag{22}$$

where IADCG@k is the value obtained for α-DCG@k when there is no redundancy—that is, all topics associated with the object appear only once in the ranking.

ERR-IA@k, as implemented for the task of the TREC Web track [Clarke et al. 2012], is defined similarly to α-NDCG@k. The only difference is that the discount function $disc(i)$ is replaced by $disc_{ERR}(i) = 1/i$.

In addition to α-NDCG and ERR-IA, we also assess the diversity of the recommended tags using the (sub)topic recall—*S-Recall* [Zhai et al. 2003]. This metric quantifies the fraction of unique topics associated with the object that are covered by the top ranked tags.

To assess novelty, we use the average inverse popularity in the top k recommendations [Belém et al. 2013a], *AIP@k*, which is the average *IFF* (Equation (9)) of the top k recommended tags, weighted by the rank-aware discount function *disc*, defined earlier. The *AIP@k* of a list C of recommended tags is defined as

$$AIP@k(C) = \frac{1}{K} \sum_{i=1}^{k} disc(i) \times IFF(C_i), \tag{23}$$

where $K = \sum_{i=1}^{k} disc(i) \times IFF_{max}$ is a normalization constant. IFF_{max} is the maximum value of *IFF* possible, which occurs when the tag is not present in the training set.

All diversity metrics use the probability of a topic z given a tag t, $\Pr(z|t)$, to estimate whether a recommended tag is related to a given topic of the object. We consider that a tag t is related to topic z if $\Pr(z|t) > \Pr(z)$ (recall Section 3.3). All metrics are computed over the top k tags in the recommendation list, with $k = 5$.

We evaluate novelty and diversity orthogonally to relevance. Thus, a tag considered irrelevant might contribute with higher novelty or diversity. Alternatively, we could embed relevance in the diversity/novelty metrics such that only relevant tags could contribute to raise these aspects. We opted for an orthogonal assessment of novelty/diversity because, unlike in previous diversification efforts in other contexts [Clarke et al. 2012], our data about tag relevance and object topics is sparse—that

Table VI. Best Parameter Values for Each Tag Recommendation Method

Method	Parameter	Dataset				
		Bibsonomy	LastFM	MovieLens	YahooVideo	YouTube
RF and RF_t	n_t	1	1	1	1	1
	ϕ	0.3	0.3	0.3	0.3	0.3
	n_b	300	300	300	300	300
	n_l	1000	1000	1000	1000	1000
xTReD (w/ categories)	λ	—	1	1	—	0.9
xTReD (w/ latent topics)	λ	0.7	0.9	0.8	0.8	0.7
xTReND (w/ categories)	α	—	0.001	0.01	—	0.01
	β	—	0.95	0.9	—	0.8
xTReND (w/ latent topics)	α	0.005	0.001	0.005	0.005	0.001
	β	0.7	0.9	0.8	0.8	0.7
LDA	n_Z	10	10	19	100	5
	m	150	150	150	150	150
xTReD and *xTReND* (w/ categories/latent topics)	τ	25	25	25	25	25

is, we do not have a complete sample of all relevant tags[13] spread across the topics that they cover for each object. Instead, we estimate how related the tags are to a topic using training data and use this estimate to compute the diversity metrics.

We make a final note regarding our experimental setup. One might argue that the diversity and novelty improvements obtained by our methods over the baselines are expected because (1) the diversifier exploits the same source of topics used to evaluate diversity, and (2) both result evaluation and tag attribute computation exploit tag popularity in the dataset to estimate novelty. However, we argue that this is a valid approach because topic information is commonly available in objects in the form of categories or can be automatically generated by clustering strategies, such as LDA. Popularity information can be also computed and is indeed correlated with novelty and discriminative power. The surprising aspect is the possibility of obtaining large gains in diversity and novelty with little loss (if any) in relevance, as will be discussed in Section 6.

5.3. Parameterization

We ran a series of experiments using data from the validation set to determine the best values for the parameters of each method in each dataset. These values are presented in Table VI.

For the reranking strategies *xTReD* and *xTReND*, we set the number of positions of the ranking to be diversified $\tau = 25$ for efficiency reasons and because the tags in the top positions are much more likely to be selected (and visualized by the user) than lower-ranked tags. Our objective with *xTReD* is to maximize diversity without harming relevance, whereas with *xTReND* we aim to maximize both novelty and diversity without diminishing relevance. Thus, we performed a grid search to find the best values for λ, α, and β (the tuning parameters), as well as for the number of topics n_Z generated by LDA, such that diversity and novelty (in the case of *xTReND*) are maximized without hurting relevance by more than a factor of ϵ%. We varied λ and β in 0, 0.05, 0.1, 0.2, ..., 0.9, and 0.95, and varied α in 0, 0.001, 0.005, 0.01 and 0.1.[14] For each dataset, we also experimented with the following values of n_Z: 5, 10, 100, and the number of predefined categories present in the dataset. We selected the best parameter values by

[13]Note that a tag may be relevant even if it is not in the gold standard.
[14]Values larger than 0.1 were very detrimental to relevance.

setting $\epsilon = 4\%$. Finally, for LDA, we set the number of iterations of the Gibbs sampling at $m = 150$, as suggested by the pLDA tool.

According to our validation experiments, our *RF*-based tag recommenders are very insensitive to parameterization. Indeed, for both *RF* and RF_t, the results obtained with different numbers of trees per bag ($n_t = 1, 5, 50$) are statistically tied (with 95% confidence). We chose $n_t = 1$ due to the lower cost. We also fixed the fraction of attributes selected in each node as ϕ=0.3 after verifying that other values (e.g., 0.25, 0.5, and 0.75) led to the same results. Different sizes of the bootstrap sample n_b also led to the same results, and we set $n_b = 300$. The only parameter that (slightly) impacted results is the number of leaves n_l: the best result was obtained with the largest value tested ($n_l =$ 1,000).

6. EXPERIMENTAL RESULTS

We now discuss representative results of the analyzed tag recommendation methods, namely the state-of-the-art *RF* and *xTReD* as well as our new RF_t and *xTReND* strategies. All results are averages of 25 executions (five folds, five random seeds) along with corresponding 95% confidence intervals.

Recall that our present investigation is driven by the following key research questions, presented in Section 1:

(I) *Q1: Do our new topic -related attributes contribute to produce better tag recommenders?*
Q2: Is our new explicit diversifier and novelty promoter xTReND effective if compared to the state-of-the-art xTReD diversifier?
Q3: How does xTReND perform compared to RF_t, which also captures relevance, diversity, and novelty?
Q4: Is the use of latent topics a viable alternative to our solutions when the target application does not possess an explicit category system to organize content?
Q5: To which extent can we effectively promote novelty and diversity without harming relevance in tag recommendation?

We address *Q*1 by comparing our new RF_t method, which incorporates diversity and novelty at the attribute level, with the relevance-driven *RF* method. We tackle *Q*2 by comparing the new *xTReND* diversifier, which captures all three aspects at both attribute and objective levels, against the previous *xTReD*. We then address *Q*3 by directly comparing our two new strategies. As *Q*4 covers an orthogonal/transversal aspect concerning all previous questions, we tackle it in the context of each individual comparison, analyzing results for all previous questions with explicit categories *and* latent topics. All of these comparisons, which cover various datasets, are presented in Section 6.1. We then tackle *Q*5 by exploring the trade-offs among relevance, novelty, and diversity in Section 6.2. We summarize our findings, providing answers to each question, in Section 6.3.

6.1. Comparing Alternative Tag Recommendation Strategies (Q1-Q4)

Tables VII and VIII show average NDCG (relevance), AIP (novelty), α-NDCG, *ERR-IA*, and *S-Recall* (diversity) results for all methods and datasets for two evaluation scenarios: (1) using the predefined categories available in the datasets as topics and (2) exploiting latent topics. These results were computed over the top $k = 5$ recommended tags and produced with all methods parameterized according tothe best parameter values obtained in the *validation* set (as shown in Table VI). Note that Table VII shows results only for the three datasets where predefined categories are available (LastFM, MovieLens, and YouTube).

Table VII. Relevance, Novelty, and Diversity of the Top $k = 5$ Recommended Tags by All Methods

Dataset	Method	NDCG	AIP	α-NDCG	ERR-IA	S-Recall
LastFM	*RF*	0.483 ± 0.003	0.325 ± 0.003	0.404 ± 0.006	0.365 ± 0.005	0.583 ± 0.009
	RF_t	**0.508 ± 0.005**	0.328 ± 0.003	0.546 ± 0.011	0.492 ± 0.009	0.738 ± 0.014
	xTReD	0.472 ± 0.003	0.353 ± 0.001	0.579 ± 0.008	**0.532 ± 0.006**	0.729 ± 0.011
	xTReND	0.504 ± 0.004	**0.365 ± 0.003**	**0.591 ± 0.011**	**0.530 ± 0.009**	**0.780 ± 0.014**
MovieLens	*RF*	0.415 ± 0.005	0.515 ± 0.005	0.272 ± 0.010	0.220 ± 0.009	0.446 ± 0.011
	RF_t	**0.428 ± 0.004**	0.509 ± 0.004	0.354 ± 0.019	0.285 ± 0.016	0.559 ± 0.017
	xTReD	0.409 ± 0.005	0.523 ± 0.004	0.383 ± 0.006	0.316 ± 0.007	0.575 ± 0.014
	xTReND	0.415 ± 0.005	**0.593 ± 0.005**	**0.437 ± 0.010**	**0.352 ± 0.010**	**0.664 ± 0.013**
YouTube	*RF*	**0.553 ± 0.002**	0.610 ± 0.001	0.749 ± 0.003	0.717 ± 0.003	0.949 ± 0.001
	RF_t	**0.555 ± 0.002**	0.610 ± 0.001	0.798 ± 0.002	0.761 ± 0.002	0.973 ± 0.001
	xTReD	0.535 ± 0.002	0.607 ± 0.001	**0.838 ± 0.002**	**0.813 ± 0.003**	0.980 ± 0.001
	xTReND	0.536 ± 0.002	**0.651 ± 0.001**	**0.837 ± 0.002**	0.807 ± 0.002	**0.985 ± 0.001**

Note: Best average results and statistical ties according to a two-sided t-test with $p < 0.05$ are shown in bold. Evaluation scenario: using predefined categories as topics.

In the following, we discuss these results by focusing first on how our new RF_t method compares against the state-of-the-art *RF*. These two methods have the same relevance-driven objective function and differ at the attribute level: RF_t adds new topic-related attributes capturing explicit diversity. This discussion, which tackles Q1, is presented in Section 6.1.1. We then approach Q2 by comparing our new *xTReND* method against the previous *xTReD* diversifier in Section 6.1.2. Finally, we address Q3, comparing *xTReND* and RF_t, in Section 6.1.3. The treatment of Q4 is included in all of those analyses.

6.1.1. Do Topic-Related Attributes Produce Better Tag Recommendations? We start by comparing the *RF*-based strategies, whose objective functions are focused on relevance only.[15] Considering the use of categories as source of topics, Table VII shows that our new RF_t strategy greatly outperforms the state-of-the-art *RF* strategy in terms of all three diversity metrics in all datasets. The improvements in α-NDCG, *ERR-IA*, and *S-Recall* reach up to 35%, 35%, and 28%, respectively. Corresponding *average* gains, computed across all datasets, are 24%, 23%, and 18%, respectively. We note that the increases in all three diversity metrics are smaller on YouTube, because objects in this dataset (videos) are associated with only one category, which reduces the room for improvements from the use of topic-related attributes.

Although both strategies have relevance as the only objective to be maximized, RF_t obtains such great improvements in diversity over *RF* by exploiting attributes that help to promote tags that are highly related to the topics of the target object and thus have higher chances to cover these topics. Such gains in diversity are accompanied by some (more modest) improvements also in terms of relevance of the recommendations: the average NDCG of RF_t is up to 5% higher. We note that these gains come with no significant additional cost since the new topic-related attributes are easy to compute. Indeed, all probabilities required to compute these attributes can be calculated offline.

Regarding novelty of the recommendations, the average AIP results of both *RF* and RF_t are statistically tied, except in the MovieLens dataset, although the difference is under 2% in this case. One possible explanation for the slightly smaller average AIP obtained with RF_t in this dataset is that MovieLens genres are semantically broader than the categories of the other datasets. As a consequence, tags in this dataset that

[15]Although RF_t also exploits diversity attributes, its objective function is based only on relevance.

Table VIII. Relevance, Novelty, and Diversity of the Top $k = 5$ Recommended Tags by All Methods

Dataset	Method	NDCG	AIP	α-NDCG	ERR-IA	S-Recall
Bibsonomy	*RF*	**0.454 ± 0.001**	0.554 ± 0.001	0.574 ± 0.004	0.479 ± 0.004	0.781 ± 0.003
	RF_t	**0.455 ± 0.001**	0.555 ± 0.001	0.580 ± 0.005	0.482 ± 0.004	0.789 ± 0.004
	xTReD	0.443 ± 0.001	0.553 ± 0.001	0.668 ± 0.003	**0.561 ± 0.003**	0.878 ± 0.002
	xTReND	0.444 ± 0.001	**0.569 ± 0.001**	**0.673 ± 0.004**	**0.564 ± 0.004**	**0.883 ± 0.003**
LastFM	*RF*	0.483 ± 0.001	0.325 ± 0.001	0.570 ± 0.007	0.535 ± 0.008	0.807 ± 0.006
	RF_t	**0.490 ± 0.001**	0.324 ± 0.001	0.587 ± 0.008	0.552 ± 0.008	0.824 ± 0.007
	xTReD	0.468 ± 0.001	0.326 ± 0.001	0.716 ± 0.006	0.678 ± 0.006	0.956 ± 0.003
	xTReND	0.473 ± 0.002	**0.333 ± 0.001**	**0.725 ± 0.007**	**0.688 ± 0.007**	**0.960 ± 0.003**
MovieLens	*RF*	0.415 ± 0.002	0.515 ± 0.002	0.452 ± 0.009	0.378 ± 0.009	0.656 ± 0.010
	RF_t	**0.426 ± 0.003**	0.506 ± 0.001	0.431 ± 0.012	0.381 ± 0.010	0.653 ± 0.016
	xTReD	0.411 ± 0.002	0.515 ± 0.001	**0.570 ± 0.007**	**0.472 ± 0.007**	**0.816 ± 0.008**
	xTReND	0.418 ± 0.003	**0.531 ± 0.001**	0.542 ± 0.014	**0.482 ± 0.012**	0.799 ± 0.014
YahooVideo	*RF*	**0.809 ± 0.001**	0.433 ± 0.001	0.509 ± 0.001	0.341 ± 0.002	0.586 ± 0.001
	RF_t	**0.810 ± 0.001**	0.433 ± 0.001	0.507 ± 0.001	0.339 ± 0.002	0.585 ± 0.002
	xTReD	0.788 ± 0.001	0.439 ± 0.001	0.561 ± 0.001	**0.382 ± 0.002**	0.623 ± 0.001
	xTReND	0.779 ± 0.002	**0.463 ± 0.001**	**0.568 ± 0.005**	**0.385 ± 0.006**	**0.628 ± 0.004**
YouTube	*RF*	**0.553 ± 0.001**	**0.610 ± 0.001**	0.556 ± 0.005	0.507 ± 0.005	0.834 ± 0.004
	RF_t	**0.553 ± 0.001**	**0.610 ± 0.001**	0.556 ± 0.005	0.507 ± 0.005	0.834 ± 0.004
	xTReD	0.539 ± 0.001	0.608 ± 0.001	**0.691 ± 0.002**	**0.645 ± 0.003**	**0.955 ± 0.001**
	xTReND	0.540 ± 0.001	**0.609 ± 0.001**	0.684 ± 0.004	0.638 ± 0.004	**0.952 ± 0.002**

Note: Best average results and statistical ties according to a two-sided t-test with $p < 0.05$ are shown in bold. Evaluation scenario: using latent topics (LDA).

are more related to the topics (and thus are promoted by RF_t) tend to be more general and occur more often, thus having lower AIP, if compared to tags in the other datasets.

If LDA topics are used (Table VIII), the gains of RF_t in relevance and diversity over *RF* are much more modest (if any), probably because the topics are generated in an unsupervised way, exploiting only the previously assigned tags. Some of these tags might be too general (i.e., "seen" and "based") or even too noisy (i.e., unrelated to the object's content) and thus might not be very appropriate for topic inference. Yet, we do observe some statistically significant improvements in average NDCG (e.g., in MovieLens) as well as in each diversity metric (e.g., LastFM and Bibsonomy). Such improvements reach 3% in average NDCG and in average α-NDCG.

6.1.2. Is Our New xTReND Effective When Compared to the State-of-the-Art xTReD? We now compare *xTReND*, our new diversifier with novelty promotion, with the state-of-the-art *xTReD* diversifier. In common, they address relevance and diversity at the objective function level, although only *xTReND* directly exploits popularity-based novelty. Moreover, only *xTReND* includes the new topic-related attributes.

Table VII shows that when categories are used as topics, *xTReND* outperforms *xTReD* with gains in AIP (novelty) of 8% on average and maximum gains of 13%. The corresponding gains when LDA topics are used are 3% and 6%, respectively, according to Table VIII. These gains are due to the promotion of tags with higher *IFF*. The surprising aspect is that such gains are achieved with no harm to diversity or relevance in most cases. Indeed, our results show that for most datasets and scenarios, *xTReND* produces at least the same diversity as *xTReD*, although in some cases there are large improvements.

For instance, note the increase in 14% of average α-NDCG if categories are used, and in 2% of average ERR-IA when LDA topics are used, both in the MovieLens dataset. Indeed, if categories are used as source of topics, the diversity of the results produced

by *xTReND* is at least as good as that of *xTReD*, although often better, *in all cases*, which indicates that it is possible to promote more specific tags that also are highly related to the topics of the object. The few exceptions when the novelty promoted by *xTReND* hurts diversity occur when LDA topics are used (e.g., α-NDCG and S-Recall on MovieLens). However, the differences in such cases are under 5% and are probably due to the higher focus given by *xTReD* to diversity when compared to *xTReND*, which also promotes novelty.

Similarly, we note that the improvements in average AIP (novelty) obtained with *xTReND* over *xTReD* also come with no detrimental impact on relevance. Instead, we do observe some significant improvements in average NDCG, with gains reaching 7% (e.g., LastFM when categories are used).

6.1.3. How Does xTReND Compare to RF_t? We now compare our two new solutions: RF_t, which captures relevance, novelty, and diversity aspects at the attribute level only, and *xTReND*, which addresses all three aspects at both attribute and objective function levels.

If categories are used as source of topics, Table VII shows that *xTReND* outperforms RF_t in terms of both diversity and novelty with gains of 11% in average AIP, 12% in average α-NDCG, 12% in average *ERR-IA*, and 9% in average *S-Recall*, all computed on average across all datasets. The maximum improvements on these metrics on any dataset reach 16%, 24%, 24%, and 19%, respectively.

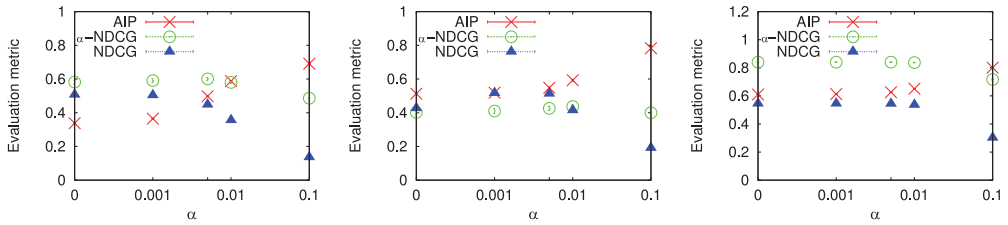
According to Table VIII, the results are similar if LDA topics are used: the gains in average AIP, α-NDCG, *ERR-IA*, and *S-Recall* are, on average, 3.4%, 20%, 22%, and 14%, respectively, reaching 7%, 26%, 26%, and 22% (also respectively). We note that such gains come with only a small impact (if any) on relevance: compared to RF_t, the average NDCG results produced by *xTReND* is at most 4% lower. Thus, it is possible to provide further gains in diversity and novelty by exploiting these aspects at the objective function level.

6.2. Trade-Offs among Relevance, Novelty, and Diversity (Q5)

Finally, we tackle research question *Q*5 and analyze the trade-offs among relevance, novelty, and diversity by quantifying how each aspect is affected as we favor one over the others. Ultimately, we want to assess the extent to which one can improve novelty and/or diversity without significantly hurting relevance. To this end, we focus on our best method, *xTReND*, which explicitly captures all three aspects, and analyze its sensitivity to parameters α and β, the weights given to novelty and diversity, respectively.

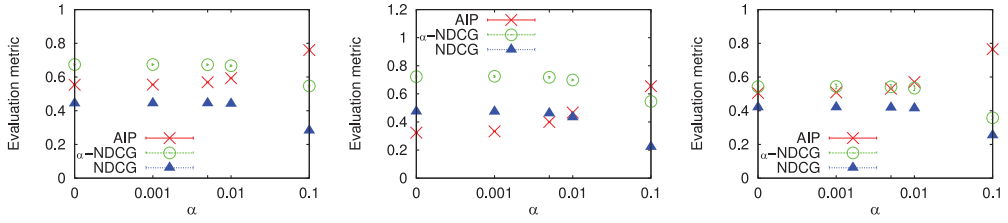
We vary α and β in the same ranges of values used for parameterizing the method (see Section 5.3), and we evaluate the relevance, diversity, and novelty of the recommendations produced by *xTReND* in the *test sets*. We perform experiments in each evaluation scenario, and when using latent topics, we also analyze the impact of varying the number of topics n_Z (see discussion later). As we vary α (or β), we compare the results produced by *xTReND* with (1) the results produced by *xTReND* when the parameter being varied is set to 0 but all other parameters are fixed at their best values (as shown in Table VI) and (2) the results obtained when $\alpha = \beta = 0$—that is, the results produced by RF_t. The first comparison allows us to assess whether favoring one factor impacts the other compared to the case when the latter is maximized (i.e., corresponding weight is set at the best value). The second comparison allows us to assess the extent to which relevance is degraded as we favor novelty or diversity, since as shown in Tables VII and VIII, RF_t produces the best results in terms of relevance in all datasets.

Figures 4 and 5 show the impact of parameter α on the average AIP (novelty), NDCG (relevance), and α-NDCG (diversity) results in both evaluation scenarios, whereas

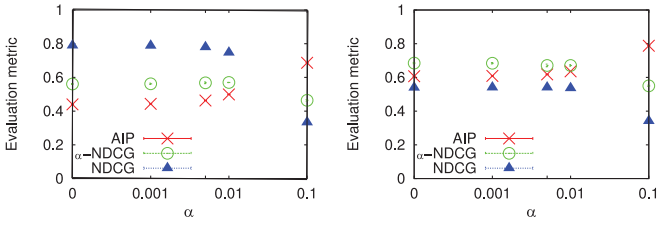


(a) LastFM (b) MovieLens (c) YouTube

Fig. 4. Impact of varying parameter α on average NDCG (relevance), AIP (novelty), and α-NDCG (diversity), computed over the top $k = 5$ recommended tags. Evaluation scenario: using predefined categories as topics.

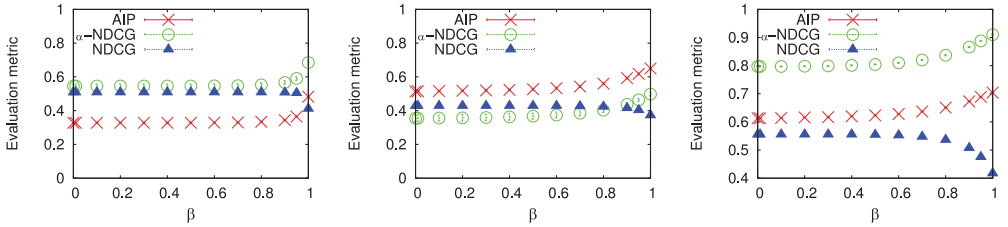


(a) Bibsonomy (b) LastFM (c) MovieLens



(d) YahooVideo (e) YouTube

Fig. 5. Impact of varying parameter α on average NDCG (relevance), AIP (novelty), and α-NDCG (diversity), computed over the top $k = 5$ recommended tags. Evaluation scenario: using latent topics (LDA).



(a) LastFM (b) MovieLens (c) YouTube

Fig. 6. Impact of varying parameter β on average NDCG (relevance), AIP (novelty), and α-NDCG (diversity), computed over the top $k = 5$ recommended tags. Evaluation scenario: using predefined categories as topics.

Figures 6 and 7 show the impact of parameter β on the same metrics. All figures show the impact of one parameter when all other parameters are kept fixed at their best values. Results for the other diversity evaluation metrics are similar to those of α-NDCG and thus are omitted. We note that Figures 4 and 6 show results only for the three datasets where predefined categories are available (LastFM, MovieLens, and

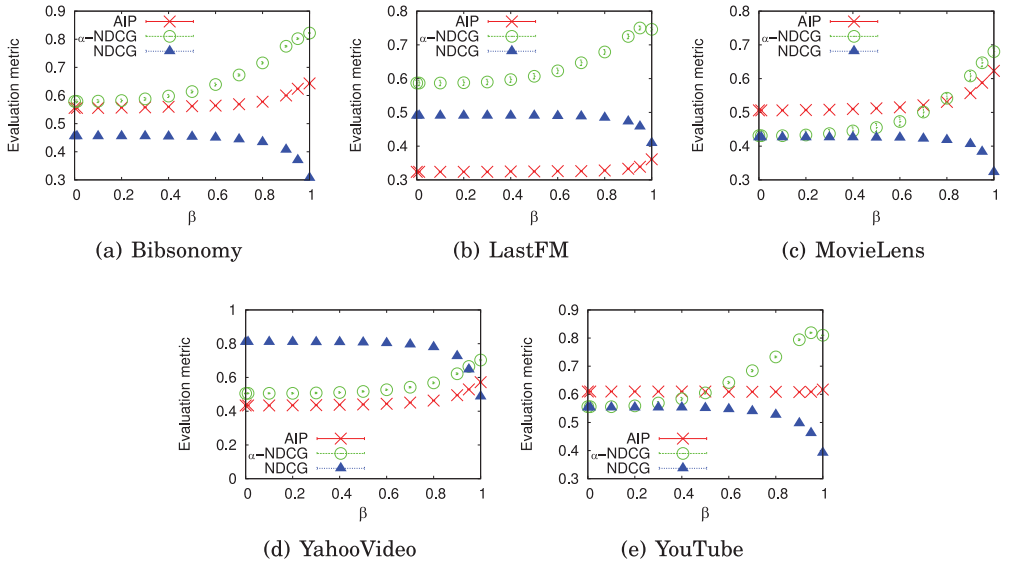


Fig. 7. Impact of varying parameter β on average NDCG (relevance), AIP (novelty), and α-NDCG (diversity), computed over the top $k = 5$ recommended tags. Evaluation scenario: using latent topics (LDA).

YouTube). We also note that all figures report average results computed over the top $k = 5$ recommended tags, along with corresponding 95% confidence intervals, although some intervals are not visible, as they are smaller than the symbols used.

Focusing first on the impact of α, Figures 4 and 5 show that average AIP results always increase as we increase the values of α, which is expected. However, values of α beyond a certain threshold, which depends on the dataset, are harmful to both relevance and diversity. Such large values of α lead to recommending very rare tags, which may be noisy and usually present low information about the topics to which they belong. For example, setting α to the maximum value tested ($\alpha = 0.1$) causes an increase in average AIP of as much as 105%. However, such improvements come at the cost of a decrease in average NDCG, compared to the initial scenario of $\alpha = 0$, which varies from 36% to 73%. Similarly, the drop in average α-NDCG (diversity) varies from 3% to 35%. Compared to the results produced by RF_t, the increase in average AIP is even higher (up to 111%), but so is the decrease in average NDCG, which varies from 52% to 74%. Similarly, the reduction in average α-NDCG varies from 1% to 17%. The exception is the MovieLens dataset when evaluated with categories, a scenario which provides 13% improvements in α-NDCG when compared to RF_t.

We now turn to the impact of β on the results. Figures 6 and 7 show that, as expected, the average α-NDCG results (diversity) always increase with β but so do the average AIP results (novelty). This indicates that tags that are highly related to the topics of the target object (diversity) also present a good level of specificity (novelty). However, very large values of β may hurt relevance by promoting tags related to the topics of the target object but less related to the object in particular. For example, compared to the case when $\beta = 0$ and the other parameters are set at their best values, increasing β to 1 leads to improvements in average α-NDCG and *AIP* of as much as 40% and 47%, respectively. But it also causes a quite dramatic decrease in average NDCG (from 13% to 25%). The differences are even more striking when we compare these results against those produced by RF_t: whereas the improvements in average α-NDCG and AIP reach

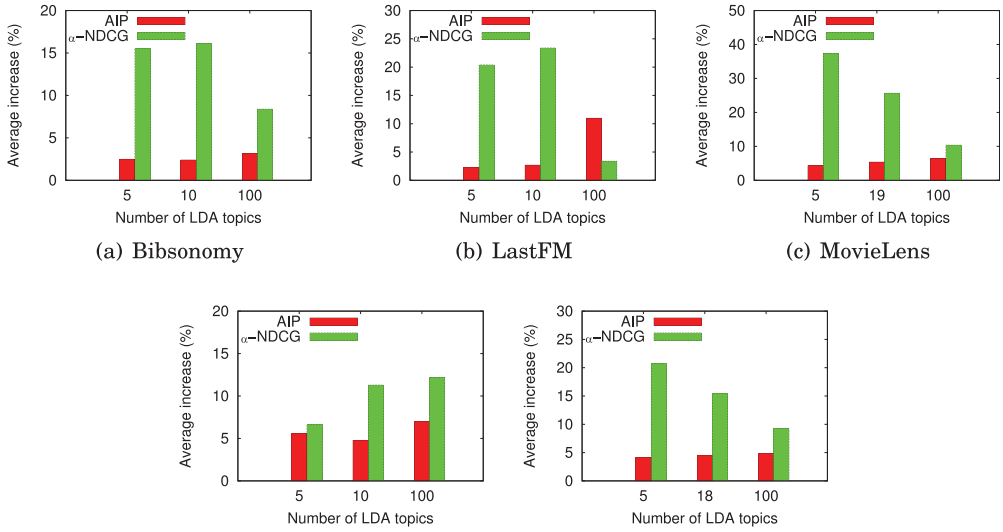


Fig. 8. Impact of varying the number of LDA topics on average AIP (novelty) and α-NDCG (diversity) computed over top $k = 5$ recommended tags: average increase over no diversification and novelty promotion.

58% and 47%, respectively, the impact in relevance may be quite detrimental, with losses of as much as 40% in average NDCG.

Yet, as discussed in Section 6.1.3, it is possible to obtain improvements in both diversity and novelty if we allow a small degradation in relevance (up to $\epsilon = 4\%$). To further analyze the trade-offs among novelty, diversity, and relevance, here we consider a more restrictive scenario when the maximum degradation in average NDCG allowed is only $\epsilon = 1\%$. Results indicate that *xTReND* is still capable of producing gains in novelty and diversity under such constraints. When categories are exploited, the gains in α-NDCG (diversity) reach 8%, 14%, and 2% for the LastFM, MovieLens, and YouTube datasets, respectively, when compared to results produced by RF_t. The corresponding gains in average AIP (novelty) reach 11%, 7%, and 3%, respectively. When latent topics are exploited, there are gains of 10%, 10%, 16%, 4%, and 13% in average α-NDCG for the Bibsonomy, LastFM, MovieLens, YahooVideo, and YouTube datasets. The corresponding gains in average AIP are 2%, 8%, 3.4%, 3%, and 1%. Thus, even if we severely restrict any possible degradation in relevance, *xTReND* can still achieve expressive improvements in diversity and novelty, particularly considering that simultaneously maximizing these three (often conflicting) objectives is quite challenging.

The aforementioned results using LDA topics were obtained by setting the number of latent topics n_Z at the best values obtained in the validation set for each dataset (Table VI). The larger the value of n_Z, the more specific are the generated topics, which may impact the diversity and the novelty of the recommended tags. Thus, as a final set of experiments, we evaluate how the novelty, diversity, and relevance of the results produced by *xTReND* using latent topics is affected as we vary the number of topics used in 5, 10 and 100. In these experiments, α and β are set at their best values. Figure 8 shows improvements in average α-NDCG (diversity) and average AIP (novelty)[16] obtained

[16]Here we analyze diversity and novelty gains instead of absolute values of the evaluation metrics, because the absolute values are not directly comparable for different values of n_Z.

over the initial recommendations produced by RF_t for different numbers of LDA topics. For every dataset, we find that the number of topics does not impact the relevance of the results. In other words, average NDCG results are statistically tied across all values of n_Z tested, thus being omitted from the graphs. In contrast, the improvements in average AIP, compared to RF_t, slightly increase as a larger number of topics is used (the improvements increase by as much as 11%). Such increase in average AIP occurs because to cover more specific topics, the diversifier promotes tags that are probably more specific as well, and thus with higher *IFF* values. However, α-NDCG gains tend to be higher for a smaller number of topics, as they are are easier to cover compared to when a larger number of more specific topics is used. The exception is YahooVideo, in which the improvements in α-NDCG are larger for the higher value of n_Z. We conjecture that this might be due to the larger number of collaboratively created tags present in that application, allowing a higher variability of tags and thus latent topics.

6.3. Summary of Our Findings

We now summarize our main findings with respect to research questions Q1 through Q5. Our experimental results revealed the following:

(1) The use of our new topic-related metrics at the attribute level by an L2R-based tag recommendation approach such as *RF* does contribute to produce better tag recommendations, particularly if predefined categories are used as topics, allowing substantial gains in diversity as well as some (modest) gains in relevance (Q1).
(2) Our new diversifier with novelty promotion, *xTReND*, outperforms the state-of-the-art diversifier *xTReD*, as it allows a higher level of novelty in the recommended tags while keeping at least the same relevance and diversity (and often also improving them) in most cases (Q2).
(3) *xTReND* provides a better trade-off among the three objectives (relevance, novelty, and diversity), being the best alternative between our two new solutions (Q3).
(4) Although more modest, the improvements of our new methods over the baselines are still significant if LDA topics are used, implying that such unsupervised topic inference strategy can be used to extend the applicability of our solutions to scenarios where predefined categories are not available (Q4).
(5) Although relevance, novelty, and diversity of recommendations may seem to be conflicting objectives, it is possible to effectively increase novelty and diversity with only a slight impact on relevance (Q5).

7. RELATED WORK

7.1. Relevance-Driven Tag Recommendation

Most previous tag recommendation techniques focus on relevance, exploiting co-occurrence patterns computed over a history of tag assignments [Canuto et al. 2013; Belém et al. 2014; Sigurbjörnsson and Zwol 2008; Garg and Weber 2008; Heymann et al. 2008; Wu et al. 2009; Krestel et al. 2009; Menezes et al. 2010], words extracted from multiple textual features of the target object [Lipczak et al. 2009; Lipczak and Milios 2011; Wang et al. 2009; Zhang et al. 2009; Lu et al. 2009; Katakis et al. 2008], and metrics of tag relevance to filter out irrelevant terms or give more importance to the relevant ones [Canuto et al. 2013; Belém et al. 2014; Lipczak and Milios 2011; Wang et al. 2009; Katakis et al. 2008]. In Belém et al. [2014], we proposed several heuristic methods that jointly exploit these three dimensions and showed that they outperform previous approaches in various datasets. Moreover, several existing tag recommendation strategies treat the problem as a multiple candidate tag ranking problem, recommending tags that are in the top positions of the generated ranking. This motivated us, as well as other authors [Cao et al. 2009; Wu et al. 2009], to exploit L2R-based strategies to

automatically "learn" tag ranking functions based on a set of metrics that estimate tags' relevance. In Canuto et al. [2013], we performed a comparative study of the use of L2R techniques for tag recommendation, analyzing eight different L2R techniques: *RF*, MART, λ-MART, ListNet, AdaRank, Genetic Programming (GP), RankSVM, and RankBoost. We found that there is a clear winner group of methods, composed by the decision tree–based strategies *RF*, MART, and λ-MART, with a slight advantage of *RF* over the other two methods.

Other dimensions of the problem (e.g., personalization) have also been tackled, often using the folksonomy (i.e., the history of tag assignments of the users [Rendle and Schmidt-Thie 2010]). Collaborative filtering and *FolkRank* [Jäschke et al. 2007], as well as pairwise interactions tensor factorization (PITF) [Rendle and Schmidt-Thie 2010], and the graph-based ranking proposed in Guan et al. [2009] fall into this category. Feng and Wang [2012] modeled the folksonomy as a heterogeneous graph containing tags, users, and objects as nodes, and employed an optimization strategy, OptRank, to learn the weights of the edges that connect these nodes. Yin et al. [2011], in turn, considered the temporal aspect of tagging systems (i.e., the variation of user interests over time). Finally, we performed a large experimental study comparing personalized with object-centered tag recommendation [Belém et al. 2014]. We do not address personalization but rather focus on recommending useful tags for a target object, a task that has been referred to as object-centered tag recommendation [Belém et al. 2014]. However, we note that our approaches can be easily extended to include other user-related metrics as components of both RF_t and *xTReND*, thus producing personalized recommendations. We leave that as future work.

In common, all of the aforementioned efforts focused on tag relevance as the single criterion of tag *quality*. Yet, other aspects, such as novelty and diversity, may also be important. In the following, we first discuss previous attempts to address novelty and diversity in various IR services and then focus on prior efforts in the particular domain of tag recommendation.

7.2. Novelty and Diversity in Information Retrieval

Result diversification is a problem that has been addressed in various contexts, particularly Web search [Santos et al. 2015]. In this context, two main families of diversification approaches have emerged to tackle query ambiguity [Santos et al. 2012].

Implicit approaches seek to promote diversity by scoring a given search result proportionally to its difference to the results ranked ahead of it, such as in terms of these results' textual dissimilarity [Carbonell and Goldstein 1998] or the divergence of their language models [Zhai et al. 2003]. In contrast, *explicit* approaches seek to diversify the search results on the basis of their coverage of some property of the user's query, such as multiple query categories [Agrawal et al. 2009] or multiple query reformulations [Santos et al. 2010; Dang and Croft 2012]. Considering that categories may be absent or noisy (e.g., vague or with nonuniform granularity) in some applications, Yu et al. [2014] proposed the use of latent topics generated by LDA as an alternative to categories and query intents in the problem of query result diversification in e-commerce sites. Dang and Croft [2012], in turn, proposed an election-based approach to diversify search results, considering the proportion/popularity of each topic of a query. We note that explicit approaches for diversification represent the current state of the art [Santos et al. 2012] and provide the basis for our approach.

In the general context of (item) recommendation, previous work mostly focused on implicit approaches to promote novelty and diversity. Celma and Herrera [2008], as well as Vargas and Castells [2011] and Hurley and Zhang [2011], evaluated novelty and diversity in terms of popularity and dissimilarity of items, based on the idea that novel and diverse items must be different from all items that have been seen or consumed. As

discussed in Section 2.1, novelty was estimated by the inverse of the popularity of the items (popularity-based novelty) as well as by the average distance (or dissimilarity) of an item to other items in a given context (e.g., the application as a whole or a specific user), referred to as distance-based novelty. Diversity, in turn, was estimated as the average *pairwise* distance between recommended items [Nehring and Puppe 2002; Vargas and Castells 2011]. Note that distance-based novelty and diversity, as previously defined, are closely related but different concepts: the former is taken from the perspective of all other items in a given context, whereas the latter is evaluated within the list of recommended items.

Hurley and Zhang [2011] formulated the trade-off between diversity and relevance as a binary optimization problem, with an input control parameter that, similarly to our methods, allows an explicit tuning of this trade-off. Zhang et al. [2012] introduced Auralist, a music recommendation framework that promotes diversity, novelty, and serendipity (a concept similar to the distance-based novelty [Vargas and Castells 2011]). They showed that the inclusion of novelty, diversity, and serendipity does improve user satisfaction, although it may slightly impact relevance of the recommendations. Lathia et al. [2010], in turn, defined novelty and diversity under a temporal perspective—that is, novel and diverse items should be different from those that have been seen or recommended in the past. Instead of aggregating relevance, novelty, and diversity as a single objective, Ribeiro et al. [2012] exploited a multiobjective Pareto optimization algorithm to jointly address these three recommendation quality criteria. The solution in this case is a set of "nondominated" recommendation functions instead of a single function.

Küçüktunç et al. [2013] and Shi [2013] addressed the problem of producing diversified and novel recommendations on graphs. Specifically, Küçüktunç et al. [2013] modeled the problem as the task of returning a set of items that extend the history of interests of a user in some items. The only data they assumed as available is the graphitself (e.g., a social network or a product co-purchasing graph), not relying on predefined topics. Their proposed diversity metric—expanded relevance—penalizes recommended items that are close to each other in the graph and thus present expanded sets (sets of neighbors in the graph) with high intersection and low coverage of the relevant results. Finally, they presented a greedy diversification algorithm called *BestCoverage*, which optimizes the expanded relevance of the result set. In a different direction, Shi [2013] proposed a method based on a first-order Markovian graph with transition probabilities between user-item pairs. The author proposed a "cost flow" concept, in which items with lower costs are recommended to a user.

Szpektor et al. [2013] addressed the problem of diversifying question recommendations in question-answering applications. According to the authors, showing the users only the main topics in which they had previously expressed interest is not the best strategy to encourage user participation in answering questions. They found that promoting both diversity and novelty provided significant improvements in the number of answers; the daily session length; and other activities, such as voting.

7.3. Novelty and Diversity in Tag Recommendation

Despite previous studies tackling diversity and novelty in item recommendation in general, to our knowledge, the only previous attempt to explore such aspects in the specific context of tag recommendation was performed in two of our previous studies [Belém et al. 2013a, 2013b]. In the first study [Belém et al. 2013a], we extended our previous relevance-driven Genetic Programming–based method [Belém et al. 2014] to include metrics related to both novelty and diversity as attributes and as part of the objective function. This approach, called GP_{rnd}, captures novelty and diversity of a list of recommended tags implicitly by introducing metrics that assess the semantic

Table IX. Comparison of our New Solutions (RF_t and *xTReND*) with Previous Tag Recommendation Methods

	Attribute Level				Objective Level				
	Relevance	Novelty (popularity)	Diversity (implicit)	Diversity (explicit)	Relevance	Novelty (popularity)	Diversity (implicit)	Diversity (explicit)	Use of LDA
[Sigurbjörnsson and Zwol 2008]	✓				✓				
[Lipczak and Milios 2011]	✓				✓				
[Wang et al. 2009]	✓				✓				
RF [Canuto et al. 2013]	✓				✓				
GP_{rnd} [Belém et al. 2013a]	✓	✓	✓		✓	✓	✓		
xTReD [Belém et al. 2013b]	✓			✓	✓			✓	
RF_t	✓	✓		✓	✓				✓
xTReND	✓	✓		✓	✓	✓		✓	✓

distance between different tags (diversity) and the inverse of the popularity of the tag in the application (novelty).

In contrast, in Belém et al. [2013b], we adopted an explicit notion of diversity by exploiting the multiple categories associated with an object. This approach, called *xTReD*, reranks the list of tag recommendations produced by any relevance-based method (we used *RF* in Belém et al. [2013b]), bringing tags that better cover the multiple topics (categories) of the target object to higher positions of the ranking. We note that by doing so, *xTReD* also avoids redundancy (i.e., repetition of topics in close positions of the ranking). This indirectly helps the novelty of the recommendations, estimated from a perspective of dissimilarity between topics covered by the recommended tags.[17] Our experimental evaluation, comparing both explicit and implicit approaches, indicated that the former is more effective, improving diversity, and indirectly the (topic level) novelty of the recommendations with no harm on relevance.

We summarize how our new methods, RF_t and *xTReND*, differ from prior work on (object-centered) tag recommendation in Table IX.[18] To our knowledge, no previous tag recommendation strategy has jointly addressed relevance, *popularity-based* novelty, and *explicit topic* diversity as an objective to be maximized. Moreover, although category information may be unavailable in some Web 2.0 applications, no other alternative sources of topics for objects (e.g., unsupervised clusters) were previously exploited in tag recommendation diversifiers. More broadly, despite the various studies discussed in the previous section, to our knowledge, we are also the first to jointly exploit the same three aspects in IR services in general. We note that Table IX does not include personalization, which is also an important and orthogonal aspect of tag recommendation. As mentioned in Section 7.1, introducing this aspect into our solutions is a subject for future work.

[17]This notion of novelty is very similar to the aforementioned distance-based novelty, although evaluated at the topic level.

[18]We make two observations about Table IX. Although *RF* uses the *IFF* metric, a popularity-based novelty metric, as a tag attribute, its use is motivated by the goal of recommending tags with higher discriminative power, an aspect that is related to relevance. Thus, *RF* is essentially a relevance-driven method. Second, we omit the dissimilarity-based novelty aspect from the table, since despite prior efforts to formally define it [Vargas and Castells 2011], no existing tag recommender explicitly captures this aspect. Yet, as mentioned in Section 2.1, this aspect is strongly related to diversity, being thus indirectly captured by those methods that tackle such an aspect.

8. CONCLUSIONS AND FUTURE WORK

In this work, we addressed the problem of recommending relevant, novel, and diverse tags in Web 2.0 applications. Besides, relevance, novelty, and diversity are important aspects for the effectiveness and utility of tag recommendations. Whereas diversity promotion helps to better cover the multiple possible topics (e.g., categories) of objects, avoiding redundancy, novelty may help to promote more rare, specific, and complementary tags to increase the completeness of the content descriptions and thus the findability of objects and recall of tag-based search.

We proposed two new strategies to tackle the tag recommendation problem, covering all three aspects of the problem (relevance, novelty, and diversity) at different levels. Our first strategy, RF_t, captures topic diversity as well as novelty at the attribute level while aiming to maximize relevance as an objective function. The second method, *xTReND*, reranks the recommendations provided by any tag recommender to directly promote relevance, novelty, and topic diversity. Here we use RF_t as basic recommender applied before the reranking, thus building a solution that addresses the problem at both attribute and objective levels.

We compared our new strategies with two state-of-the-art tag recommendation methods: a relevance-driven *RF*-based method and our previous diversifier (*xTReD*). Our evaluation, using real data from five popular Web 2.0 applications, shows that the use of our new topic-related metrics at the attribute level (as performed by RF_t) does contribute to produce better tag recommendations, particularly if predefined categories are used as topics, allowing substantial gains in diversity (up to 35%) as well as some (modest) gains in relevance (up to 5% in NDCG). We also found that our new diversifier with novelty promotion, *xTReND*, outperforms the state-of-the-art diversifier (*xTReD*), as it allows a higher level of novelty (gains of up to 13% in AIP) in the recommended tags while keeping at least the same relevance and diversity (and often also improving them) in most cases. Overall, our new method, *xTReND*, is the best alternative considering the trade-offs among relevance, novelty, and diversity. Although more modest, the improvements of our new methods over the baselines are still significant if LDA topics are used, implying that such unsupervised topic inference strategy can be used to extend the applicability of our solutions to applications where predefined categories are not available. Finally, although relevance, novelty, and diversity of recommendations may seem to b conflicting objectives, it is possible to effectively increase novelty and diversity with only a slight impact on relevance.

As future work, we intend to deal with other aspects of the problem, such as personalization and the novelty of recommendations in the perspective of specific users. We also plan to employ other multiclustering strategies as an alternative to LDA to inferthe object topics, which is useful when no category information is available.

REFERENCES

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. 5–14.

R. Agrawal and R. Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*. 487–499.

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

F. Belém, E. Martins, J. Almeida, and M. Gonçalves. 2013a. Exploiting relevance, novelty and diversity in tag recommendation. In *Advances in Information Retrieval*, Vol. 7814. Springer, Berlin, Germany, 380–391.

F. Belém, E. Martins, J. Almeida, and M. Gonçalves. 2014. Personalized and object-centered tag recommendation methods for Web 2.0 applications. *Information Processing & Management* 50, 4, 524–553.

F. Belém, R. Santos, J. Almeida, and M. Gonçalves. 2013b. Topic diversity in tag recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. 141–148.

D. Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55, 4, 77–84.

S. Canuto, F. Belém, J. Almeida, and M. Gonçalves. 2013. A comparative study of learning-to-rank techniques for tag recommendation. *Journal of Information and Data Management* 4, 453–468.

H. Cao, M. Xie, L. Xue, C. Liu, F. Teng, and Y. Huang. 2009. Social tag prediction based on supervised ranking model. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*. 35–48.

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. 335–336.

Ò. Celma and P. Herrera. 2008. A new approach to evaluating novel recommendations. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'08)*. 179–186.

X. Chen and H. Shin. 2013. Tag recommendation by machine learning with textual and social features. *Journal of Intelligent Information Systems* 40, 2, 261–282.

C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. 75–84.

C. Clarke, N. Craswell, and E. Voorhees. 2012. Overview of the TREC 2012 Web track. In *Proceedings of the 21st Text Retrieval Conference (TREC'12)*. 1–8.

C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. 659–666.

V. Dang and W. Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. 65–74.

U. Feige. 1998. A threshold of $\ln(n)$ for approximating set cover. *Journal of the ACM* 45, 634–652. Issue 4.

W. Feng and J. Wang. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*. 1276–1284.

F. Figueiredo, F. Belém, H. Pinto, J. Almeida, and M. Gonçalves. 2012. Assessing the quality of textual features in social media. *Information Processing & Management* 49, 1, 222–247.

N. Garg and I. Weber. 2008. Personalized, interactive tag recommendation for Flickr. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'08)*. 67–74.

S. Golder and B. Huberman. 2005. The Structure of Collaborative Tagging Systems. Retrieved January 11, 2016, from http://arxiv.org/abs/cs.DL/0508082.

Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. 2009. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 540–547.

P. Heymann, D. Ramage, and H. Garcia-Molina. 2008. Social tag prediction. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. 531–538.

D. Hochbaum (Ed.). 1997. *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Co.

N. Hurley and M. Zhang. 2011. Novelty and diversity in top-n recommendation—analysis and evaluation. *ACM Transactions on Internet Technology* 10, 4, 14:1–14:30.

R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thie, and G. Stum. 2007. Tag recommendations in folksonomies. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 506–514.

I. Katakis, G. Tsoumakas, and I. Vlahavas. 2008. Multilabel text classification for automated tag suggestion. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

S. Khuller, A. Moss, and J. Naor. 1999. The budgeted maximum coverage problem. *Information Processing Letters* 70, 1, 39–45.

R. Krestel, P. Fankhauser, and W. Nejdl. 2009. Latent Dirichlet allocation for tag recommendation. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*. ACM, New York, NY, 61–68.

O. Küçüktunç, E. Saule, K. Kaya, and Ü. Çatalyürek. 2013. Diversified recommendation on graphs: Pitfalls, measures, and algorithms. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 715–726.

N. Lathia, S. Hailes, L. Capra, and X. Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. 210–217.

X. Li, L. Guo, and Y. E. Zhao. 2008. Tag-based social interest discovery. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. 675–684.

M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. 2009. Tag sources for recommendation in collaborative tagging systems. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*.

M. Lipczak and E. Milios. 2011. Efficient tag recommendation for real-life data. *ACM Transactions on Intelligent Systems and Technology* 3, 1, Article No. 2.

Z. Liu, Y. Zhang, E. Chang, and M. Sun. 2011. PLDA+: Parallel latent Dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology* 2, 3, 26:1–26:18.

P. Lops, M. de Gemmis, G. Semeraro, C. Musto, and F. Narducci. 2013. Content-based and collaborative techniques for tag recommendation: An empirical evaluation. *Journal of Intelligent Information Systems* 40, 1, 41–61.

Y. Lu, S. Yu, T. Chang, and J. Hsu. 2009. A content-based method to enhance tag recommendation. In *Proceedings of the 21st International Conference on Artificial Intelligence*. 2064–2069.

B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. 2009. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 641–650.

E. Martins, F. Belém, J. Almeida, and M. Gonçalves. 2013. Measuring and addressing the impact of cold start on associative tag recommenders. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia'13)*. 325–332.

G. Menezes, J. Almeida, F. Belém, M. Gonçalves, A. Lacerda, E. Moura, G. Pappa, A. Veloso, and N. Ziviani. 2010. Demand-driven tag recommendation. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*.

K. Nehring and C. Puppe. 2002. A theory of diversity. *Econometrica* 70, 3, 1155–1198.

S. Rendle and L. Schmidt-Thie. 2010. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 81–90.

M. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys'12)*. 19–26.

R. Santos, C. Macdonald, and I. Ounis. 2010. Exploiting query reformulations for Web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. 881–890.

R. Santos, C. Macdonald, and I. Ounis. 2012. On the role of novelty for search result diversification. *Information Retrieval* 15, 5, 478–502.

R. Santos, C. Macdonald, and I. Ounis. 2015. Search result diversification. *Foundations and Trends in Information Retrieval* 9, 1, 1–90.

R. Santos and I. Ounis. 2011. Diversifying for multiple information needs. In *Proceedings of the 1st International Workshop on Diversity in Document Retrieval*. 37–41.

L. Shi. 2013. Trading-off among accuracy, similarity, diversity, and long-tail: A graph-based recommendation approach. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. 57–64.

B. Sigurbjörnsson and R. Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. 327–336.

Y. Song, L. Zhang, and C. Giles. 2011. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web* 5, 1, 1–31.

I. Szpektor, Y. Maarek, and D. Pelleg. 2013. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 1249–1260.

S. Vargas and P. Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*. 109–116.

S. Vargas, P. Castells, and D. Vallet. 2012. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. 75–84.

J. Wang, L. Hong, and B. D. Davison. 2009. Tag recommendation using keywords and association rules. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-497/.

L. Wu, L. Yang, N. Yu, and X. Hua. 2009. Learning to tag. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 361–370.

D. Yin, S. Guo, B. Chidlovskii, B. Davison, C. Archambeau, and G. Bouchard. 2013. Connecting comments and tags: Improved modeling of social tagging systems. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*. 547–556.

D. Yin, L. Hong, Z. Xue, and B. Davison. 2011. Temporal dynamics of user interests in tagging systems. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. 1279–1285.

J. Yu, S. Mohan, D. (Pew) Putthividhya, and W. Wong. 2014. Latent Dirichlet allocation based diversified retrieval for e-commerce search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM'14)*. 463–472.

C. Zhai, W. Cohen, and J. Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*. 10–17.

N. Zhang, Y. Zhang, and J. Tang. 2009. A tag recommendation system based on contents. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*.

Y. Zhang, D. Séaghdha, D. Quercia, and T. Jambor. 2012. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. 13–22.

Received November 2014; revised March 2015; accepted July 2015