Topic Diversity in Tag Recommendation*

Fabiano M. Belém, Rodrygo L. T. Santos, Jussara M. Almeida, Marcos A. Gonçalves
Computer Science Department
Universidade Federal de Minas Gerais, Brazil
{fmuniz, rodrygo, jussara, mgoncalv}@dcc.ufmg.br

ABSTRACT

Tag recommendation approaches have historically focused on maximizing the relevance of the recommended tags for a given object, such as a movie or a song. Nevertheless, different users may be interested in the same object for different reasons—for instance, the Star Wars movies may appeal to both adventure as well as to fantasy movie fans. In this situation, a sensible strategy is to provide a user with diverse recommendations of how to tag the object. In this paper, we address the problem of recommending relevant and diverse tags as a ranking problem. In particular, we propose a novel tag recommendation approach that explicitly takes into account the possible topics (e.g., categories) underlying an object in order to promote tags with high coverage and low redundancy with respect to these topics. We thoroughly evaluate our proposed approach using data collected from two popular Web 2.0 applications, namely, LastFM and MovieLens. Our experimental results attest the effectiveness of our approach at promoting more relevant and diverse tags in contrast to state-of-the-art relevance-based methods as well as a recently proposed method that takes both relevance and diversity into account.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Algorithms, Experimentation

Keywords

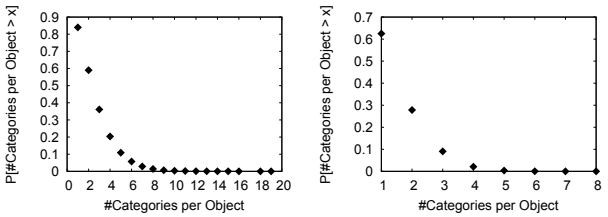
Tag Recommendation, Relevance, Diversity

*This work is supported by the INWeb (grant 57.3871/2008-6) and by the authors grants from CNPq and FAPEMIG.

RecSys'13, October 12–16, 2013, Hong Kong, China.

http://dx.doi.org/10.1145/2507157.2507184.

1. INTRODUCTION



(a) Artist styles in AllMusic (b) Genres in MovieLens

Figure 1: Complementary cumulative distribution of the number of categories per object.

The Web 2.0 is characterized by an unprecedented amount of user-generated content. This content usually comprises a main *object* (e.g., a video) 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 [5]. Among all textual features, tags have attracted special attention as they offer an effective data source for Information Retrieval (IR) services such as search [17] and classification [13], and may capture user interests reasonably well [17]. 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 improving the quality of the generated tags and, indirectly, of the IR services that rely on them.

Tag recommendation methods have historically focused mostly on maximizing the *relevance* of the recommended tags [5, 18, 27]. 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 and effectiveness [4, 25]. For example, a list of synonyms that well describe the object's content is arguably relevant, but also redundant and less useful than a more diversified list covering more aspects related to the object. As argued in [4], objects on the Web 2.0 may be *multifaceted*, being related to various aspects and topics.

We illustrate this point by showing, in Figure 1, the distributions of the number of styles for different music artists and the number of genres for different movies, computed over datasets collected from AllMusic and MovieLens[1], re-

[1]http://www.allmusic.com and http://MovieLens.umn.edu

spectively. According to Figure 1, 84% of the artists and 63% of the movies are associated with *more* than one category (style or genre). Take for instance the movie "Sister Act", starring Whoopi Goldberg. Its main genre is Comedy, but it also presents elements from the Action and Musical genres. Arguably, it would be appropriate to recommend tags related to all these genres for this movie. According to Figure 1, the distribution is even more biased towards larger numbers of categories for music artists.

While *diversity* is an important aspect in tag recommendation, our previous work [4] is, to our knowledge, the only one to address aspects other than relevance for tag recommendation. In [4], we defined the diversity of a list of recommended tags *implicitly*, as the average *semantic distance* [4, 13, 19] between each pair of tags in the list, such that a set of synonyms or semantically similar words has low diversity. However, this approach has the disadvantage of being more subject to data sparsity, because the estimates of the semantic similarities are based on tag co-occurrences [4].

We here propose an alternative approach to address diversity in tag recommendation. We define diversity *explicitly*, in terms of the capacity of the recommended tags to cover different aspects or topics of the target object, an approach inspired by diversification algorithms proposed in the context of search [22]. To this end, we exploit an explicit taxonomy represented by categories, commonly available in Web 2.0 applications, as topics for objects.

In order to recommend relevant and diverse tags, we model tag recommendation as a ranking problem. That is, we aim to produce a ranking function f that assigns scores to candidate tags based on: (1) relevance estimates generated by a baseline state-of-the-art tag recommendation strategy, and (2) our diversification approach, which rescores candidate tags such that they better cover the topics of the target object. This allows us to sort candidate tags according to their joint relevance and diversity estimates.

Our method is a re-ranking strategy, and thus can be applied over the recommendations produced by any relevance-based strategy. In particular, we apply it over a state-of-the-art Genetic Programming (GP) based tag recommender [5], here referred to as GP_r, and over a novel strategy that uses *Random Forests* (RF), a learning-to-rank technique [6] that has not been applied to tag recommendation yet. Our approach, called explicit Tag Recommendation Diversifier or $xTReD$, modifies the original scores produced by GP_r (or by RF), bringing tags that better contribute to covering the topics of the target object to higher positions of the ranking.

We evaluate our method on datasets collected from 2 popular applications - LastFM and MovieLens - comparing it against GP_r and GP_{rnd} [4], which extends GP_r to include metrics of diversity and novelty. Our goals are: (1) assessing to which extent we can improve diversity without harming relevance, and (2) comparing our explicit diversification strategy against the implicit GP_{rnd} method. Our results show that our method outperforms the best relevance-based strategy on which it is based by up to 45% in diversity without harming relevance. It also outperforms GP_{rnd} by up to 54% in diversity and 10% in relevance, with the best results produced when xTReD is applied jointly with RF. Moreover, among the two relevance-driven strategies, the new RF-based recommender outperforms the state-of-the-art GP_r by up to 7% in relevance.

In sum, the main contributions of this paper are: (1) an alternative definition of diversity for tag recommendation that explicitly captures the multifaceted nature of many Web 2.0 objects; (2) a novel strategy to diversify tag recommendations that leads to significant gains in diversity, with no harm to relevance, over state-of-the-art strategies; (3) an analysis of Random Forests as an alternative learning-to-rank strategy for recommending relevant tags, and (4) an extensive evaluation of all strategies in large real datasets.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 formally states the problem we tackle. Section 4 introduces the tag recommendation methods analyzed here. Our experimental methodology and results are discussed in Sections 5 and 6, while Section 7 offers conclusions and directions for future work.

2. RELATED WORK

The tag recommendation literature is rich in techniques exploiting co-occurrence patterns computed over a history of tag assignments [5, 4, 27], words extracted from multiple textual features of the target object [18], as well as metrics of tag relevance to filter out irrelevant terms or give more importance to the relevant ones [5, 18]. In [5], we proposed several heuristic methods that jointly exploit these three dimensions, and showed that they outperform previous approaches in various datasets. We, as well as other authors [7, 27], also exploited learning-to-rank (L2R) techniques, such as RankSVM [5, 7], RankBoost [27] and Genetic Programming (GP) [5], to "learn" a model that ranks tags based on a set of relevance metrics. Other dimensions of the problem, such as personalization, have also been tackled, often using the user's history of tag assignments [20].

In common, all these previous efforts focus on tag relevance as the single criterion of tag *quality*. However, other aspects such as diversity may also be important. Indeed, result diversification is a problem that has been addressed in other contexts, particularly web search [11]. In this context, two main families of diversification approaches have emerged to tackle query ambiguity [23]. *Implicit approaches* seek to promote diversity by scoring a given search result proportionally to its difference to the results ranked ahead of it, e.g., in terms of these results' textual dissimilarity [8] or the divergence of their language models [28]. 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 [1] or multiple query reformulations [22]. Explicit approaches represent the current state-of-the-art [23], and provide the basis for our approach.

In the general context of (item) recommendation, previous work mostly focused on implicit approaches. Celma *et al.* [9] and Vargas *et al.* [25] 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. Ribeiro *et al.* [21] exploited a multi-objective pareto optimization algorithm to jointly address relevance, novelty and (implicit) diversity. Wartena and Wibbels [26] improve item recommendation by detecting different topics of interest in the user profile and then generating recommendations for each of these topics.

We are aware of only one previous attempt to explore result diversification in the specific context of tag recommendation. Very recently, we extended our previous GP-based method [5] to include metrics related to both diversity and

novelty [4]. However, inspired by [25], we previously adopted an *implicit* definition of diversity, which captures the semantic distance between different tags in a recommendation list. In contrast, building from recent search result diversification efforts [22, 23], we here adopt an alternative definition by *explicitly* representing the multiple topics associated with an object, as discussed in Section 3.

The original GP-based solution proposed in [5] and its recent extension [4] are here taken as baselines for comparison against our novel approach, being described in Section 4.1. We note that, although we currently do not address personalization, our approach can be easily extended to produce personalized recommendations, by taking into account a topic distribution biased towards each user's interests.

3. PROBLEM STATEMENT

The tag recommendation problem we address here can be stated as: *Given a set of initial tags I_o, priorly assigned to the target object o, and a set of textual features $F_o = \{F_o^1, ..., F_o^n\}$, 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 both relevance and diversity objectives are maximized, and recommend the k candidates in the top positions of C_o.* We focus on recommending tags for a target object, aiming at improving the quality of tags in the objects, and indirectly the effectiveness of services that use tags as a data source. We leave the task of tackling diversity in personalized recommendations to the future.

The main focus of this work is on the diversity of the recommended tags. Diversity can be defined implicitly or explicitly. As an example of the former, Vargas *et al.* [25] defined diversity as the average pairwise difference between the items in the recommendation list. Inspired by them, we have previously defined diversity in the specific context of tag recommendation as the average semantic distance between the recommended tags, such that a list of synonyms or semantically related words present low diversity [4].

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 [22, 23], being useful to increase the chances that at least one document will satisfy the information need of the target user. This is also the perspective we employ here, thus 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.

Given our focus, we approach the tag recommendation task as a *ranking problem*. That is, we aim at developing a *ranking function* that assigns scores to each candidate tag c, allowing us to sort candidates so that those that represent more relevant and diverse recommendations for the object o appear in higher positions in C_o. This is achieved in two steps: (1) a state-of-the-art relevance-driven tag recommendation strategy is used to produce relevance estimates, and (2) our diversifier approach rescores candidate tags such that they better cover the topics of the target object.

4. TAG RECOMMENDATION METHODS

This section describes the analyzed tag recommendation methods. All approaches employ a learning-to-rank (L2R) technique, for example, Genetic Programming (GP) or Random Forest (RF), to build a ranking function that assigns scores to candidate tags according to some criterion (e.g., probability of relevance). The two state-of-the-art baselines use GP as L2R technique, and are described in Section 4.1. We present our new relevance-driven RF-based strategy as well as our new diversifier approach in Section 4.2. All methods extract candidate tags from (1) co-ocurrence patterns with tags already in the target object o (tags in I_o), and (2) other textual features in o, namely, title and description.

Table 1: Metrics of tag relevance.

Relevance aspect	Metrics
Tag co-occurrences	Sum, Sum^+, $Vote$, $Vote^+$ [24]
Discriminative power	Stability ($Stab$) [24], Inverse Feature Frequency (IFF) [13]
Descriptive power	Term Spread (TS) [13], Term Freq. (TF) [2] weighted TS (wTS), weighted TF (wTF) [5]
Predictability	$Entropy$ [5]

4.1 State-of-the-Art Baselines

Our baseline methods are based on Genetic Programming (GP), a framework inspired by the biological mechanisms of genetic inheritance and evolution of individuals in a population [3]. GP is a non-linear method that has been applied to various IR tasks. We were the first to use it for recommending tags [4, 5], having obtained competitive (or superior) results over other approaches.

GP implements a global search mechanism by evolving a population of individuals over multiple generations. Each individual, representing a possible solution for the target problem, is modeled as a tree composed of terminals (leaves) and functions (inner nodes), related to the target problem. In each generation, each individual is evaluated by a *fitness* function, defined based on quality metrics related to the problem at hand. Only individuals with the highest fitness values are selected, according to some selection method, to evolve the population. An initial randomly generated population is evolved, generation after generation, through *crossover* and *mutation* operations. The individual with the best fitness value, usually part of the last generation, is chosen as the final solution for the problem.

GP directly optimizes a target (fitness) function, and allows for easy extensions to include more problem-related features (terminals) and to address other aspects of the target problem (by, for instance, adapting the fitness function). In particular, as reported in [5], GP outperforms other learning-to-rank techniques (RankSVM) in some scenarios, being statistically tied with it in many others. Thus, GP provides a promising framework to develop effective tag recommendation solutions.

In our first GP-based baseline, here referred to as GP_r [5], an individual is a tag ranking function with a tree representation built from nodes defined as follows. The sum ($+$), subtraction ($-$), multiplication ($\times$), division ($/$) and natural logarithm (ln) operations are used as inner nodes. Protected division and logarithm are implemented such that these operations return 0 (default) when their inputs are outside their domains. Terminals are either constants, uniformly distributed between 0 and 1, or variables, selected from a set including (the values of) the metrics listed in Table 1. These metrics capture the following aspects related to the *relevance* of a candidate tag c to a target object o: (1) co-occurrence patterns, whose strengths are expressed by the

confidences of association rules between previously assigned tags (tags in I_o) and c, where the rules are extracted from the training set, (2) the power of c to discriminate o from other objects, (3) the power of c to describe o's content, and (4) the predictability of c as a tag in the object collection. For illustration purposes, a tree representation of function $Vote + wTS/0.5$ has operation '+' as root, with variable $Vote$ and operation '/' as its children. The inner node '/' has variable wTS and constant 0.5 as children.

The fitness of an individual represents the quality of the recommendations produced by the corresponding ranking function. It is computed as the Normalized Discounted Cumulative Gain ($NDCG$) [2] of the top-k tags in the recommendation list produced by the function for a sample of objects in a validation set, with k=5.

Only relevance metrics are used by GP_r (as terminals and fitness function). Since our approach diversifies GP_r results, comparing both methods allows us to assess to which extent we can exploit a tradeoff between relevance and diversity.

GP_r was recently extended to include attributes related to the diversity and novelty of a list of tags [4]. This extension, here called GP_{rnd} (our second baseline), adds the Average Distance to other Candidates (ADC) metric to the list of terminals. ADC captures the diversity that a candidate c brings to a recommendation list C_o by the average semantic distance between c and each other candidate tag in C_o. The ADC of c with respect to C_o is defined as $ADC(c, C_o) = \frac{1}{|C_o|}\sum_{t \in C_o, t \neq c} dist(c, t)$, where $dist(c_1, c_2)$ measures the dissimilarity between candidates c_1 and c_2 as the relative difference between the sets of objects O_1 and O_2 in which they appear as tags. That is, $dist(c_1, c_2) = \frac{|O_1 \setminus O_2|}{|O_1 \cup O_2|}$. If both sets are empty, $dist(c_1, c_2)$ is set to 1, the maximum value. GP_{rnd} also adopts a different fitness function, aiming at jointly maximizing relevance, diversity and novelty. For each object o in the validation set, the fitness of a candidate ranking function f is computed as:

$$fitness(C_o) = \alpha \times AIP@k(C_o) + \beta \times AILD@k(C_o) + (1 - \alpha - \beta) \times NDCG@k(C_o), \quad (1)$$

where C_o is the recommendation list produced by f for o, and $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ are tuning parameters. The fitness function includes $NDCG@k$ to capture the relevance of the recommended tags in C_o, as well as $AIP@k$ and $AILD@k$, which are related to the novelty and the diversity of these tags, respectively.

$AIP@k$, or Average Inverse Popularity over the top k positions of the ranking, was adapted from [25] to be a normalized average of the IFF values of the top-k tags in C_o. IFF, a variation of the traditional IDF metric [2], favors tags that do not occur too often. It was used to estimate the novelty of a tag under the assumption that the lower the popularity of a tag, the higher its novelty [4].

$AILD@k$, or Average IntraList Distance in the top k positions [25], captures the diversity of the tags in C_o. Specifically, given the normalization constant $K' = (k^2 - k)/2$ and $dist(c_i, c_j)$ as defined above, $AILD@k$ of C_o is computed as $AILD(C_o) = \frac{1}{K'}\sum_{i=1}^{k}\sum_{j=i+1}^{k} dist(c_i, c_j)$. Note that both $AILD@k$ and ADC metrics capture the notion of diversity in a list of tags implicitly, by means of the $dist(c_i, c_j)$ function: a list with tags that are more semantically *dissimilar* is considered more diverse.

Although GP_{rnd} captures novelty, in addition to relevance and diversity (our goal), a comparison against our solution is fair since, as noted in [4], a simplified version of the method with the novelty component turned off (α=0) does not outperform the complete solution in any of the three aspects.

4.2 Our New Solutions

We now describe the new tag recommendation strategies proposed here. We start by describing a relevance-driven tag recommender that exploits Random Forests to learn a tag ranking function, being thus an alternative to GP_r (Section 4.2.1). We then introduce our topic diversifier, which can be jointly applied with any relevance-driven strategy to produce relevant and diversified recommendations (Section 4.2.2).

4.2.1 Random Forests Based Tag Recommender

Random Forests (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 in order to reduce the correlation among them. The decision tree learning happens in a recursive way: first, the most discriminative feature (according to some measure, such as Information Gain) is selected as a decision node. The selected training samples are split according to a split value (such as the average attribute value), and the process repeats in a top-down fashion. In order to further reduce the correlation between the decision trees, before each split a fraction ϕ of the features 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 decision tree is typically grown to its maximum depth, containing n_l terminal nodes, where n_l is a tuning parameter.

In order to be applied to our context, each tree outputs a relevance score to be assigned to each candidate tag c in a (test) object. The scores given by each tree to c are averaged and used to produce the final ranking of candidates. RF has achieved promising results in other contexts (e.g, document ranking [14]). It is used here as an alternative to GP_r to produce a relevance based ranking of candidate tags. We use the RF implementation available in the RankLib tool[2], and exploit the same attributes used by GP_r (Table 1) to build the trees.

4.2.2 Explicit Tag Recommendation Diversifier

Unlike GP_{rnd}, we propose to address tag diversity in an explicit way, by seeking to directly maximize the set of topics (or categories) covered by the recommended tags. In its general form, maximizing topic coverage is an NP-hard problem [1]. 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 [15]. This is also the best possible polynomial-time approximation for the problem, unless NP $\subseteq$ DTIME($n^{O(\log\log n)}$), where n is the number of items to be diversified [12, 16]. Our method, called explicit Tag Recommendation Diversifier, or $xTReD$, builds upon this greedy approach, as described in Algorithm 1.

[2] http://people.cs.umass.edu/~vdang/ranklib.html

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 *based only on relevance* (line 1). Any relevance-driven tag recommender could be used in this step. We here evaluate: (1) GP_r, to make our approach comparable to the GP_{rnd} baseline, and (2) our new relevance-driven method *RF*.

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 $\quad t^* \leftarrow \arg\max_{t \in C_o^\tau} f(o, t, C_o^S)$
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

Algorithm 1: The xTReD algorithm.

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 that the user looks at. A complete permutation of C_o (τ=$|C_o|$) 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 3-7). The sub-modular 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, we build upon a state-of-the-art framework for diversifying search results, called xQuAD [22]. The xQuAD framework instantiates the aforementioned function in order 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 [22]. In our adaptation, instead of a ranking of documents for a query, we seek to diversify a ranking of tags for a given object. More precisely, we propose a novel instantiation of the objective function $f(o, t, C_o^S)$, such that:

$$f(o, t, C_o^S) = (1 - \lambda) p(t|o) + \lambda \sum_{z \in Z} p(z|o) p(t|o, z) \prod_{t' \in C_o^S} (1 - p(t'|o, z)), \quad (2)$$

where Z 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 which 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. Besides addressing an explicit notion of diversity, our strategy has the advantage of making the weight given to diversity adjustable at recommendation time, unlike GP_{rnd}, which needs to train a new model for each new value of α, and is hence more time-consuming.

When $\lambda = 0$, Equation (2) reduces to $p(t|o)$, which results in a pure relevance-driven tag recommendation, as produced by a non-diversification baseline. In our experiments in Section 6, we define $p(t|o) = 1/r_t$, where r_t is the position of the tag t in the ranking produced by the initial ranker *rec*. In order to estimate the second half of Equation (2), we infer the distribution $p(z|o)$ of topics $z \in Z$ for an object o from the available training data, as we will discuss in Section 5.2. Finally, to estimate how much a given tag t covers the topic z of the object o, we approximate the probability $p(t|o, z)$ as $p(t|o, z) \approx p(t|o) \times p(z|t)$, where $p(z|t)$ is an estimate of the probability that tag t is related to topic z. In turn, $p(z|t)$ is defined as $p(z|t) = f(t, z)/f(t)$, 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 in the training set.

5. EXPERIMENTAL SETUP

This section describes the setup used in our experimental evaluation of the tag recommendation methods, including datasets (Section 5.1), evaluation methodology (Section 5.2), and method parameterization (Section 5.3).

5.1 Datasets

Our evaluation is performed using two datasets, containing *title*, *tags*, *description* and *categories* associated with objects from MovieLens and LastFM. The MovieLens dataset[3] contains 100,000 tags applied to 10,000 movies. The genres associated with each movie were used as categories. The LastFM dataset[4] is the same used in [4, 5]. We collected the musical styles associated with 35,975 artists in this dataset from the AllMusic site[5], and used them as artist categories.

We removed *stopwords* and applied the Porter stemming algorithm[6] to avoid trivial recommendations (e.g., plurals and other variations of the same word). We also discarded objects with fewer than 2 tags. Our processed datasets comprise a sample of the textual features of 35,975 LastFM artists and 6,500 MovieLens movies.

5.2 Evaluation Methodology

As in most tag recommendation studies [4, 5, 24, 18], we adopted an automatic evaluation approach: half of the tags, randomly selected from all tags previously assigned to an object, are used as *gold standard*, i.e., as the relevant tags for that object. These tags are *not* used by the tag recommenders, being thus removed from $\mathcal{I}_o$, which is used as input by the recommenders. A manual evaluation of tags is an expensive process in terms of time and human effort, besides being subjective, and thus is left for future work.

As in [5], the experiments were performed using a 5-fold cross-validation. That is, the objects are distributed into five equal-sized portions. Three portions are treated as the training set, used to compute all features exploited by both GP and RF. A fourth portion is used as the validation set,

[3] http://www.grouplens.org/taxonomy/term/14
[4] http://www.vod.dcc.ufmg.br/recc/
[5] http://www.allmusic.com
[6] http://snowball.tartarus.org/algorithms/porter/stemmer.html

which, in turn, is used to "learn" the solutions (i.e., generate the ranking function by both techniques) and to tune parameters of the methods (e.g., λ, GP and RF parameters). The last portion is used for testing.

In order to estimate how related a tag t is to a topic, which is necessary to evaluate topic diversity and used by the diversifier algorithm, we estimate the probability of a topic given a tag using training data. That is, the level of relationship of a tag t to a topic z is given by $p(z|t)$, defined in Section 4.2. The topics correspond to categories obtained from our datasets, as described in Section 5.1.

We evaluate the tag recommendation methods in terms of the relevance and the diversity of their results. We use $NDCG@k$ to assess the relevance of a ranked list of tags. To assess the diversity of this list, we used three metrics traditionally used for evaluating search result diversification methods [10]. Two of them—ERR-IA and α-NDCG—are the primary evaluation metrics used in the diversity task of the TREC Web track [11]. They are cascade metrics that penalize redundancy by modeling the behavior of a user who stops inspecting the ranking once a relevant tag is observed [25]. While ERR-IA measures the *expected retrieval performance* with respect to multiple topics, α-NDCG incorporates a notion of the *expected gain* attained by each ranked tag. We also report (sub)topic recall—*S-Recall* [28], which quantifies the fraction of unique topics associated with the object that are covered by the top ranked tags. All diversity metrics use the probability of a topic z given a tag t, $p(z|t)$, to estimate whether a recommended tag is related to a given topic of the object. They are computed over the top-k tags in the recommendation list, with k=5 as in [4, 5].

Following [4], we evaluate diversity orthogonally to relevance. Thus, a tag considered irrelevant might contribute with higher diversity. Alternatively, we could embed relevance in the diversity metric such that only relevant tags could contribute to raise diversity. We opted for an orthogonal assessment of diversity because, unlike in previous diversification efforts in other contexts [11], our data about tag relevance and object topics is sparse, that is, we do not have a complete sample of all relevant tags spread across the topics they cover for each object. Instead, we estimate how related the tags are to a topic using training data, and use this estimation in the diversity metrics. One might argue that tags considered irrelevant[7] should not contribute to raise diversity, despite being related to a topic of the object. We note however that we tuned all methods to maximize the average diversity across all objects, *without harming relevance* (on average). After this tuning, we observed that a tag considered irrelevant contributed to amplify diversity for only a small fraction of the objects (less than 4%). Thus, we did indeed filter out the vast majority of such cases.

Note that we can evaluate the diversity of recommendations for objects with a single category by verifying if the recommended tags are related to that single topic. However, the redundancy is high when recommending tags from the same topic, which could cause a loss in α-$NDCG$.

5.3 Parameterization

We ran a series of experiments using data from the validation set to determine the best parameters for each method. For both GP_r and GP_{rnd} we fixed the parameters of the evolutionary process at the same values reported in [5]. We set the population size at 200, the maximum tree depth at 8, and the maximum number of generations at 200. We also employed *tournament selection* of individuals to evolve the population, i.e., we randomly select, with replacement, 2 individuals from the population and choose the one with highest fitness. Moreover, crossover and mutation operations were performed at rates 0.6 and 0.1, respectively.

For GP_{rnd}, as in [4], we set α=β, varying both at the same time[8], in the [0,0.6] interval. These parameters capture the tradeoff between relevance and the combination of novelty and diversity. Larger values of α (or β) lead to great losses in relevance. The value that lead to the best diversity[9] without harming relevance is 0.25 and 0.1 for LastFM and MovieLens, respectively.

According to validation experiments, our RF-based tag recommender is very insensitive to parameterization. The results obtained with different numbers of trees (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) lead to the same results. Different sizes for the bootstrap sample n_b also led to the same results, and we set n_b=300. The only parameter that (slightly) impacts results is the number of leaves n_l: the best result is obtained with the largest value tested (n_l=1000).

For our diversifier $xTReD$, we set the number of positions of the ranking to be diversified τ=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 also to maximize diversity without harming relevance. Thus, we performed a grid search to find the best values for λ, the trade-off between relevance and diversity, as will be discussed in Section 6.1.

6. EXPERIMENTAL RESULTS

We now discuss the results of the analyzed tag recommendation methods, namely, the baselines GP_r and GP_{rnd}, and our new strategies RF and $xTReD$. Whenever necessary, we refer to our diversifier as $xTReD_{rec}$, where rec is the initial ranker used (GP_r or RF) All results are averages of 25 executions (5 folds, 5 random seeds). Our goal is to answer the following key research questions:

1. Can we effectively diversify tag recommendations without harming relevance? We explore the tradeoff between relevance and diversity in Section 6.1.
2. Is our explicit diversifier effective in contrast to a state-of-the-art implicit approach? In Section 6.2, we compare $xTReD$ against GP_{rnd} and against the relevance-driven strategies it builds upon (either GP_r or RF).
3. Is Random Forests an effective learning-to-rank technique when applied to tag recommendation? In Section 6.2, we also compare our new RF-based method against the state-of-the-art GR_r recommender.

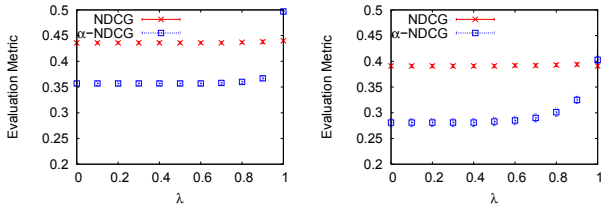
6.1 The Relevance and Diversity Tradeoff

We here analyze the sensitivity of our method to its key parameter λ, by searching for the best parameter value in

[7] Note that a tag may be relevant even if it is not in the gold standard.

[8] The results obtained following this approach are not worse than the best results when we set α=0 (thus removing the novelty component AIP) and varied only β.

[9] The best results were chosen in terms of α-$NDCG$, but the best parameter values are the same for the other metrics.



(a) LastFM (b) MovieLens

Figure 2: Impact of parameter λ in $xTReD$ effectiveness.

the *validation set*. Figure 2 shows average $NDCG$ and α-$NDCG$ results for $xTReD_{GP_r}$, along with 95% confidence intervals[10], as functions of λ. Results for $xTReD_{RF}$ and for other diversity evaluation metrics are similar (omitted).

When λ=0, $xTReD$ outputs the same recommendation list as the baseline it uses as initial ranker (GP_r or RF). The closer λ gets to 1, the more the baseline's recommendation list is rearranged, favoring diversity, which reaches its maximum at λ=1, for both datasets. The gains of $xTReD_{GP_r}$ in α-$NDCG$ over the results for λ=0 (i.e., GP_r) reach 39% in LastFM and 44% in MovieLens for λ=1. Corresponding gains for $xTReD_{RF}$ over RF are 46% and 40%.

Such impressive gains in diversity come with no statistically significant loss in relevance (NDCG). Thus, for both initial rankers, the maximum diversification level (λ=1) does not hurt relevance. This is possibly due to: (1) tags promoted by $xTReD$ may present high relevance as they tend to be related to the topics (categories) of the object, and (2) $xTReD$ promotes relevant tags even if λ=1, due to the probability $p(t|o,z) \approx p(t|o) \times p(z|t)$. Indeed, recall that $p(t|o,z)$ is part of the diversity component of Equation (2), representing the probability that tag t is related to *both* topic z and object o. Thus, the diversifier also promotes tags with high chance of being related to the object. In contrast, the benefits introduced by GP_{rnd} are much more modest (if any), as discussed in the next section.

Thus, recalling our first research question, we found that it is possible to significantly increase diversity without harming relevance with our xTReD method by adjusting its parameter λ. Best results are obtained with λ=1.

6.2 Test Results

We now turn to our second and third research questions, on the effectiveness[11] of our explicit diversification approach and of the RF relevance-driven method. Towards answering them, we perform a series of experiments with objects in the *test set*, comparing our $xTReD$ approach (with both RF and GP_r as initial rankers), the state-of-the-art GP_{rnd} implicit diversifier, and the two relevance-driven methods (GP_r and RF). All methods are parameterized with the best parameter values found in the validation set.

In the following, we fix the maximum allowed degradation in relevance at 3%, and assess the gains in diversity, in terms of all metrics, achieved by $xTReD$ and GP_{rnd} over the relevance-driven method they build upon. We choose the 3% threshold to favor the GP_{rnd} baseline since this threshold leads to the best tradeoff between relevance and diversity for that method. We note however that our $xTReD$ methods produce statistically tied relevance results even if we allow maximum diversification (as discussed in Section 6.1).

[10] The intervals are not visible in some points as they are smaller than the symbols used.

[11] Regarding the efficiency (execution time) of our methods, the training stage of the L2R-based methods (GP_r and RF) represents the only significant additional cost, but this stage can be performed entirely offline [5], thus not impacting the recommendation time.

Table 2: Results of the analyzed methods in test set. Best results (and their statistical ties) in bold.

	Method	*NDCG*	*α-NDCG*	*ERR-IA*	*S-Recall*
LastFM	GP_r	0.436	0.357	0.329	0.499
	GP_{rnd}	0.423 ▽	0.369 ○	0.339 ○	0.511 ○
	$xTReD_{GP_r}$	0.439 ○△	0.498 △△	0.458 △△	0.643 △△
	RF	**0.466** △△	0.391 △△	0.353 △△	0.563 △△
	$xTReD_{RF}$	**0.464** △△	**0.567** △△	**0.517** △△	**0.717** △△
MovieLens	GP_r	0.389	0.280	0.235	0.462
	GP_{rnd}	0.387 ○	0.282 ○	0.238 ○	0.459 ○
	$xTReD_{GP_r}$	0.389 ○○	0.406 △△	0.339 △△	0.633 △△
	RF	**0.404** △△	0.302 △△	0.250 △△	0.509 △△
	$xTReD_{RF}$	**0.402** △△	**0.424** △△	**0.353** △△	**0.656** △△

Table 2 shows average $NDCG$, α-$NDCG$, ERR-IA and S-$Recall$ results for all methods, for both datasets. It also shows statistical significance indicators (according to a two-sided t-test with $p < 0.05$): the first indicator symbol refers to the comparison of the corresponding method against GP_r, while the second refers to the comparison against GP_{rnd}. Symbols △, ▽ and ○ indicate significant improvement, decrease and statistical tie, respectively, with 95% confidence.

We start by comparing the state-of-the-art baselines. We find that GP_{rnd} produces only small improvements, on average, in explicit diversity (up to 3.4% in α-$NDCG$, 3.1% in ERR-IA and 2.5% in S-$Recall$). These average results are consistent with [4], which reports similar gains in terms of implicit diversity, captured by the $AILD$ metric (Section 4.1). However, due to higher variability, the average gains of GP_{rnd} in the explicit diversity metrics adopted are *not* statistically significant, as they are tied with GP_r results with 95% confidence. This is because $AILD$ promotes infrequent tags, which tend to be more dissimilar to any tag since there is little information about their co-occurrence (the source of information for $AILD$ to estimate the semantic differences between tags). On the other hand, infrequent tags also carry little information about their related topics, being less appropriate to cover the object's topics.

We now turn to our diversifier $xTReD$. Following the discussion in Section 6.1 (which was based on results of the validation set), we find that $xTReD_{GP_r}$ greatly outperforms GP_r, achieving gains of up to 45%, 44% and 37%, on average, in α-$NDCG$, ERR-IA and S-$Recall$, respectively, in the test sets. The corresponding gains of $xTReD_{RF}$ over RF reach 45%, 47% and 29%, respectively. Such gains in diversity come with no statistically significant losses in relevance in neither dataset. One might argue that our diversity improvements are expected because the diversifier exploits the same source of topics used to evaluate diversity. However, this is a valid approach because this information is commonly available in objects in the form of categories. The surprising aspect is the possibility of obtaining very large gains with no loss in relevance, as discussed in Section 6.1.

We next compare our $xTReD$ methods against the state-of-the-art GP_{rnd} baseline, which provides implicit diversification. We find that $xTReD_{GP_r}$ outperforms GP_{rnd} by as much as 44%, 43% and 38% in average α-$NDCG$, ERR-IA and S-$Recall$, respectively. The gains are larger in MovieLens, but are also quite impressive in LastFM, ranging from 26% to 35%. In terms of relevance, $xTReD_{GP_r}$ produces results with average $NDCG$ at least as good as those of GP_{rnd}, with even some modest gains (3%) in LastFM. Com-

paring $xTReD_{RF}$ against GP_{rnd}, we find that the former provides even further improvements in all diversity and relevance metrics, with average gains in α-$NDCG$, ERR-IA and S-$Recall$ and $NDCG$ reaching 54%, 53%, 43% and 10%, respectively. In sum, answering our second question, our explicit $xTReD$ diversifier, particularly when using RF as the initial ranker, is much more effective than the alternative GP_{rnd} method in improving the diversity of recommendations produced by the relevance-driven strategies.

Finally, tackling our third research question, we find that our new RF method outperforms the state-of-the-art GP_r strategy, with statistically significant gains in all (relevance and diversity) metrics, in both datasets. The average improvements in $NDCG$, α-$NDCG$, ERR-IA and S-$Recall$ reach 7%, 10%, 7% and 13%, respectively. Thus, the use of RF is a competitive alternative to GP_r, producing gains in terms not only of relevance but also diversity, even though both strategies are focused on relevance only and exploit the same set of attributes. This happens because since RF recommends more relevant tags in higher positions of the ranking (higher $NDCG$), these recommended tags have higher probability of being related to the topics of the target object, thus presenting a higher topic diversity. This also impacts the effectiveness of our diversifier, which is higher when RF is exploited as initial ranker, as discussed above.

In sum, we found that: (1) although relevance and diversity of recommendations may seem conflicting objectives, it is possible to effectively increase diversity without harming relevance with our $xTReD$ diversifier, which also has the advantage of being flexible (adjustable) at recommendation time, (2) our explicit diversifier is much more effective at reducing redundancy and covering different topics of the target object than the state-of-the-art GP_{rnd} baseline, the only previous tag recommender that exploits diversity (implicitly), and (3) our new relevance-driven RF tag recommender outperforms the alternative, state-of-the-art GP_r in terms of not only relevance but also diversity.

7. CONCLUSIONS AND FUTURE WORK

We have addressed the problem of recommending tags for Web 2.0 objects, which are usually multifaceted, by proposing two new methods: a diversification approach ($xTReD$), which exploits an explicit taxonomy to decrease redundancy and increase the number of topics of the target object covered by the recommended tags, and a relevance-driven Random Forests (RF) based tag recommender. Our evaluation of the methods on two datasets showed that it is possible to increase the average diversity of a list of recommended tags in up to 45% with no significant impact on relevance. We also found that our explicit diversifier is much more effective than GP_{rnd}, a state-of-the-art implicit diversification strategy and the only alternative to $xTReD$ in the literature, reaching gains in both diversity and relevance of 54% and 10%, respectively. Our approach also has the advantage of being adjustable at recommendation time, unlike GP_{rnd}, which requires the user to choose the weight given to the relevance/diversity tradeoff at model training time. Finally, we also found that our new RF-based recommender outperforms a state-of-the-art relevance driven Genetic Programming based method not only in terms of relevance (gains of up to 7%) but also diversity (gains of up to 10%).

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 multi-clustering strategies as an alternative to infer the object topics, which can be useful when no category information is available.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

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

[3] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone. *Genetic Programming – An Introduction On The Automatic Evolution Of Computer Programs And Its Applications*. Morgan Kaufmann, 1998.

[4] F. Belém, E. Martins, J. Almeida, and M. Gonçalves. Exploiting relevance, novelty and diversity in tag recommendation. In *ECIR*, pages 380–391, 2013.

[5] F. Belém, E. Martins, T. Pontes, J. Almeida, and M. Gonçalves. Associative tag recommendation exploiting multiple textual features. In *SIGIR*, pages 1033–1042, 2011.

[6] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[7] H. Cao, M. Xie, L. Xue, C. Liu, F. Teng, and Y. Huang. Social tag prediction based on supervised ranking model. In *ECML PKDD*, pages 35–48, 2009.

[8] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[9] O. Celma and P. Herrera. A new approach to evaluating novel recommendations. In *RecSys*, pages 179–186, 2008.

[10] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*, pages 75–84, 2011.

[11] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 Web track. In *TREC*, 2012.

[12] U. Feige. A threshold of $\ln(n)$ for approximating set cover. *J. ACM*, 45:634–652, 1998.

[13] F. Figueiredo, F. Belém, H. Pinto, J. Almeida, and M. Gonçalves. Assessing the quality of textual features in social media. *IP&M*, pages 222–247, 2012.

[14] P. Geurts and G. Louppe. Learning to rank with extremely randomized trees. *JMLR*, 14:49–61, 2011.

[15] D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., 1997.

[16] S. Khuller, A. Moss, and J. S. Naor. The budgeted maximum coverage problem. *Inf. Proc. Lett.*, 70:39–45, 1999.

[17] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW*, 2008.

[18] M. Lipczak and E. Milios. Efficient tag recommendation for real-life data. *TIST*, 3(1):2:1–2:21, Oct. 2011.

[19] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *WWW*, pages 641–650, 2009.

[20] S. Rendle and L. Schmidt-Thie. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, pages 81–90, 2010.

[21] M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *RecSys*, pages 19–26, 2012.

[22] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW*, pages 881–890, 2010.

[23] R. L. T. Santos, C. Macdonald, and I. Ounis. On the role of novelty for search result diversification. *Inf. Retr.*, 15(5):478–502, 2012.

[24] B. Sigurbjörnsson and R. Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, pages 327–336, 2008.

[25] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*, pages 109–116, 2011.

[26] C. Wartena and M. Wibbels. Improving tag-based recommendation by topic diversification. In *ECIR*, pages 43–54, 2011.

[27] L. Wu, L. Yang, N. Yu, and X. Hua. Learning to tag. In *WWW*, pages 361–370, 2009.

[28] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17, 2003.