Exploiting Novelty and Diversity in Tag Recommendation[⋆]

Fabiano Belém, Eder Martins, Jussara Almeida, and Marcos Gonçalves

Computer Science Department, Federal University of Minas Gerais, Brazil
{fmuniz,ederfm,jussara,mgoncalv}@dcc.ufmg.br

Abstract. The design and evaluation of tag recommendation methods have focused only on relevance. However, other aspects such as novelty and diversity may be as important to evaluate the usefulness of the recommendations. In this work, we define these two aspects in the context of tag recommendation and propose a novel recommendation strategy that considers them jointly with relevance. This strategy extends a state-of-the-art method based on Genetic Programming to include novelty and diversity metrics both as attributes and as part of the objective function. We evaluate the proposed strategy using data collected from 3 popular Web 2.0 applications: LastFM, YouTube and YahooVideo. Our experiments show that our strategy outperforms the state-of-the-art alternative in terms of novelty and diversity, without harming relevance.

Keywords: Tag Recommendation, Relevance, Novelty, Diversity.

1 Introduction

Many Web 2.0 applications became very popular mainly due to the strong incentives to users creating and sharing their own content as well as to the establishment of online communities and social networks. User generated content is typically composed by a main *object*, which can be stored in various media types (e.g., text, audio, video, image), and by several sources of data associated with the object, here referred to as object features. There are several types of object features. For instance, *content features* are sources of data that can be extracted from the object itself, such as the color histogram of an image, whereas *textual features* are blocks of text often assigned by the users to the object, such as tags, a title, a description, and comments.

Tags, out of all textual features, deserve special attention as they often drive content organization, providing good descriptions and reflecting the users' interests [5]. Moreover, tags offer a valuable data source for information retrieval services, particularly for rich media (image, audio, video) objects. This is because, as pointed out by [4], the size of object collections and the rate at which new content is uploaded to popular applications as well as the (typically poor)

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

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 380–391, 2013.

quality of user generated content (particularly rich media content) bring great challenges to existing multimedia information retrieval techniques. As a matter of fact, recent studies demonstrated that tags are among the best textual features to support information retrieval services such as automatic object classification [5] and content recommendation [7].

In this context, tag recommendation strategies aim at suggesting *relevant* and *useful* keywords to users, helping them in the task of assigning tags to content. Ultimately, these mechanisms aim at improving the quality not only of the generated tags, by making them more complete and accurate and reducing their noise (e.g., misspellings, unrelated words) but also of the information services that rely on tags as data source.

Research on recommendation systems has, historically, focused mostly on maximizing the *relevance* of the recommended items [11]. In tag recommendation, specifically, relevance can be defined from an object centered perspective and from a personalized one. In the former, a relevant term describes well the content of the target object [3], whereas in the latter a relevant term describes correctly the target content *and* satisfies the target user's interests [6].

However, relevance only may not be enough to guarantee recommendation usefulness and effectiveness [13,11]. For instance, consider a list of recommended tags given to user u to describe object o, in which all terms are related to o's content, but all of them are synonyms. Moreover, suppose that these terms have already been used by u to describe o. In both cases, although the recommended tags have maximum relevance, they are less interesting and useful than a more diverse list of terms that brings novel information for the given object and user. This is particularly important because multimedia objects on the Web 2.0 may be *multifaceted*, that is, they may be related to various aspects and topics. Take, for instance, the case of a video about the use of genetic algorithms to control robots. Tags related to robotics, artificial intelligence and even genetics could be adequate. Thus, newer and more diverse tags may help better capture these various facets of the object. Therefore, not only relevance but also *novelty* and *diversity* are key aspects for tag recommendation.

In the general context of item recommendation, Vargas *et al.* [11] define novelty as how different an item is from other items observed in a given context (e.g., the application as a whole, a user, a group of users, etc). The diversity of a list of recommended items, in turn, refers to how much different each item is from the others in the list [11]. Inspired by this previous effort, we here address novelty and diversity in the specific context of tag recommendation, leaving the context of specific users for the future, when we will provide personalized recommendations. We analyze tag novelty in the application context. That is, we consider a tag as novel if it is not observed very often in the application, which can be estimated by the inverse of the popularity of the tag in the object collection. We also estimate the diversity of a list of recommended tags by the average semantic distance [10,5] between each pair of tags in the list, such that a set of synonyms or semantically similar words has low diversity. Thus, novelty and diversity are distinct but related concepts: a list of recommended tags is diverse if each item in the list is novel with respect to the others.

Specifically, we here address the problem of recommending relevant, novel and diversified tags. The problem is stated as a multiple term candidate ranking problem where the ranking is a combined function of relevance, novelty and diversity. In other words, we aim at producing a ranking function f which assigns scores to candidate terms based on various tag quality metrics, thus allowing us to sort them according to their joint relevance, novelty and diversity estimates.

This is the first work to explore novelty and diversity for tag recommendation. In particular, we here apply metrics that have never been used for tag recommendation, and we develop a novel tag recommendation strategy. This strategy extends a state-of-the-art method based on Genetic Programming (GP) [3], here referred to as GP_{rel}, which is focused on maximizing the relevance of recommended tags, to include novelty and diversity metrics both as attributes and as part of the objective function to be optimized. We refer to this new strategy as $GP_{rel+nov+div}$. We here choose a GP based approach because: (1) GP is very flexible, allowing the easy introduction of new metrics and new objective functions, including multiobjective functions, as in the present case; (2) it is a very effective machine learning technique, presenting results that are as good as the traditional RankSVM method in tag recommendation [3]; and (3) it has been applied to other *ranking* problems [1], presenting a good theoretical foundation.

We evaluated our method using data collected from the popular LastFM, YahooVideo and YouTube applications, comparing it against the state-of-the-art method GP_{rel} (our baseline). Our results show that the new $GP_{rel+nov+div}$ strategy produces gains over the baseline in up to 14% in novelty with no detrimental impact on relevance. The gains in diversity are more modest (2.5%), however $GP_{rel+nov+div}$ is a promising and flexible approach, which can be extended by the inclusion of new metrics and new objective functions to capture other novelty, diversity and relevance aspects of the problem.

In sum, the main contributions of this paper are: (1) definition and explicit exploration of novelty and diversity in the context of tag recommendation; (2) proposal of a new tag recommendation strategy that jointly explores relevance, novelty and diversity; and (3) a thorough experimental evaluation of the proposed strategy, comparing it against a state-of-the-art method, considering relevance, novelty and diversity as evaluation criteria.

The rest of this paper is organized as follows. Section 2 discusses related work, and Section 3 formally defines the problem addressed here. Section 4 presents the metrics used by the analyzed tag recommendation strategies, which are described in Section 5. Our experimental evaluation is discussed in Section 6. Section 7 concludes the paper, presenting some directions for future work.

2 Related Work

With the focus only on relevance, most of the existing tag recommendation strategies exploit a combination of the following dimensions: (i) co-occurence of terms with tags previously assigned to the object; (ii) terms extracted from multiple textual features, such as title and description; and (iii) relevance metrics, such as *Term Frequency* (TF) [2], to filter out irrelevant terms or give more importance to the relevant ones [3,10,9]. Based on these three dimensions, a few studies exploit *learning to rank* (L2R) techniques [12,3] to "learn" a model that allows to rank tags based on a set of relevance metrics. RankSVM [3], RankBoost [12] and Genetic Programming [3] are examples of L2R techniques already explored in tag recommendation.

Despite the importance and benefits of considering novelty and diversity for recommendation purposes, these aspects have been little explored in general recommendation systems [13,11]. For instance, Zhou *et al.* [13] measure novelty as the *Inverse User Frequency* (IUF), defined as the log of the inverse of the number of users who like the item. Vargas *et al.* [11] evaluate novelty not only in terms of popularity, but also in terms of the intra-list dissimilarity among recommended items. Lathia *et al.* define novelty and diversity under a temporal perspective [8], that is, novel items should be different from what was seen or recommended in the past. Although these previous studies addressed novelty and diversity in content recommendation, to the best of our knowledge, no previously proposed tag recommendation method explicitly considers these aspects.

3 Problem Definition

The *novelty* of an item (e.g., a tag, a movie or any type of element being recommended) can be defined as how different this item is from all other items observed in a given context [11]. This context can be, for instance, the items that have been observed by a single user or by a group of users, or even all items in the application. 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.

The *diversity* of a list of recommended items, in turn, refers to how different these items are among each other. Thus, novelty and diversity, though different, are related concepts, given that in a diverse set of recommended items, each item is novel with respect to the others [11]. Note that novelty and diversity should not be taken independently from relevance, because a non-relavant random item tend to be novel, although it does not represent an adequate recommendation.

We here define the novelty of a tag in the context of the application, estimating a tag's novelty 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 less "novel" and more "obvious" recommendation. According to this definition, noisy terms such as typos may be considered highly novel. However, our methods will jointly exploit the aspects novelty and relevance, avoiding noise. Besides that, the weight given to each aspect can be adjusted, as we will see in Section 5.

We also estimate the diversity of a list of tags by the average semantic distance between each pair of tags in the list. The metric used to estimate the semantic distance between a pair of tags is defined in Section 4.3.

The task of recommending tags for a target object o is defined as follows. Given a set of tags I_o that have been priorly assigned to o and a set of textual features F_o associated with o (e.g., o's title), generate a set of candidate tags C_o ($C_o \cap I_o = \emptyset$) ranked based on the relevance, novelty and diversity of each tag in C_o for o, and recommend the k terms more highly ranked in C_o.

In this context, many tag recommendation methods, and in particular those analyzed here, exploit co-occurrence patterns by mining relations among tags assigned to the same object in an object collection. The process of learning such patterns is defined as follows. There is a training set $D = \{(I_d, F_d)\}$, 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 O, which is a collection of objects $\{(I_o, F_o, Y_o)\}$, where both I_o and Y_o are sets of tags previously assigned to object o. While tags in I_o are known and given as input to the recommender, tags in Y_o are assumed to be unknown and taken as the relevant recommendations for o (i.e., *gold standard*). As in previous studies [6,3], we split the tags of each test object into these two subsets simply to facilitate an automatic assessment of the recommendations, as further discussed in Section 6.2. Similarly, there might be also a validation set V used for tuning parameters and "learning" the recommendation functions. Thus, each object $v \in V$ also has its tag set split into input tags I_v and gold standard Y_v.

Given our focus, we here treat the tag recommendation task as a *ranking problem*. That is, we aim at developing a *ranking function* which assigns scores to each candidate term c in C_o, allowing us to sort them so that terms that represent more relevant, novel and diverse recommendations for object o appear in higher positions. The ranking function $f(R(c), N(c), D(c, C))$ is a function of the relevance $R(c)$ and of the novelty $N(c)$ of given candidate term c, as well as of the diversity $D(c, C)$ of c with respect to a list of candidates C. Two issues that must be addressed to define f are: how to define $N(c)$ and $D(c, C)$ in the tag recommendation context, and how to effectivelly combine them with $R(c)$ to build function f. We address these issues in the next two sections.

4 Tag Recommendation Metrics

We here present the relevance, novelty, and diversity metrics used by the analyzed tag recommendation methods. Some of these metrics have been previously proposed for the broader context of item recommendation [11]. We here adapt them to the specific context of tag recommendation. Moreover, unlike previous work, these metrics are here used not only to evaluate the effectiveness of recommendations, but also as part of the (objective function of the) methods.

4.1 Relevance Metrics

The relevance metrics used here are categorized into three groups based on the aspect they try to capture regarding the tag recommendation task. The categories are: *tag co-occurrence patterns*, *descriptive* and *discriminative* capacities.

Metrics related to *co-occurrence patterns* estimate the relevance of tags that 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. These co-occurrence patterns are based on association rules, that is, implications of type $X \rightarrow c$, where the antecedent X is a set of tags and the consequent c is a candidate tag for recommendation. The importance of an association rule is given by its support σ and confidence θ. Given a rule $X \rightarrow c$, its support $\sigma(X \rightarrow c)$ is the number of times X co-occurred with c in the training set D, whereas its confidence $\theta(X \rightarrow c)$ is the conditional probability that c is assigned as tag to an object $d \in D$, given that all tags in X are also associated with d.

We here consider four metrics related to tag co-occurrence patterns previously proposed in [10]. They are Sum, $Sum{+}$, $Vote$ and $Vote{+}$. Given a candidate tag c for a target object o, $Sum(c, o)$ is the sum of the confidences of all rules whose antecedent contains terms in I_o and whose consequent is c, whereas $Vote(c, o)$ is the number of such rules. $Sum{+}$ and $Vote{+}$ are weighted versions of Sum and $Vote$, respectively, using the Stability ($Stab$) metric [10] as weight.

$Stab$ gives more importance to terms with intermediate frequencies in the collection, thus penalizing terms that are either too common and general or very rare and specific, which represent poor recommendations as they have poor discriminative capacity. The Stability of a candidate c is defined as $Stab(c, k_s) = \frac{k_s}{k_s + |k_s - log(f_c^{tag})|}$, where k_s is the "ideal" or "most stable" frequency of a term (parameter adjusted to the collection) and f_c^{tag} is the frequency of c as tag in the training set D.

$Sum{+}$ and $Vote{+}$ are then defined as:

$$Sum+(c, o, k_x, k_c, k_r) = \sum_{x \in I_o} \theta(x \rightarrow c) \times Stab(x, k_x) \times Stab(c, k_c) \times Rank(c, o, k_r) \qquad (1)$$

$$Vote+(c, o, k_x, k_c, k_r) = \sum_{x \in I_o} I(x \rightarrow c) \times Stab(x, k_x) \times Stab(c, k_c) \times Rank(c, o, k_r) \qquad (2)$$

where k_x, k_c and k_r are tuning parameters, and $Rank(c, o, k_r)$ is equal to $\frac{k_r}{(k_r + p(c,o))}$, where $p(c, o)$ is the position of c in the ranking of candidates according to the confidence of the corresponding association rule. Moreover $I(x \rightarrow c)$ is equal to 1 if rule $x \rightarrow c$ belongs to R, the set of rules computed offline over the training set D, and 0 otherwise.

Descriptive capacity metrics estimate the relevance of a candidate tag c based on how closely it relates to the textual content of the target object. A widely used metric is *Term Frequency* (TF) which is the number of occurrences of c in all textual features (except tags) of object o. In contrast, the *Term Spread* (TS) [5] of a candidate c is the number of textual features of o (except tags) that contain c. Thus, unlike TF, TS takes the structure of the object, composed by different textual features, into account.

Belém *et al.* [3] proposed weighted versions of TF and TS, referred to as wTF and wTS, which weight the occurrence of each term based on the average descriptive capacity of the textual feature in which it appears. The average descriptive capacity of a feature is estimated by the *Average Feature Spread* (AFS) heuristic [5]. Let the *Feature Instance Spread* of a textual feature $\mathcal{F}_o^i$

associated with object o, $FIS(\mathcal{F}_o^i)$, be the average TS over all terms in $\mathcal{F}_o^i$. $AFS(\mathcal{F}^i)$ is defined as the average $FIS(\mathcal{F}_o^i)$ over all instances of $\mathcal{F}^i$ associated with objects in the training set $\mathcal{D}$.

Discriminative capacity metrics estimate the relevance of a candidate c by its capacity to distinguish an object from the others, which is important to discriminate objects into different categories or levels of relevance. In addition to the aforementioned *Stab* metric, we also consider the *Inverse Feature Frequency* (IFF) and the entropy metrics [3].

The IFF metric is an adaptation of the traditional Inverse Document Frequency (IDF) that considers the term frequency in a specific textual feature (tags, in the present case). Given the number of objects in the training set $|D|$, the IFF of candidate c is given by $IFF(c) = \log \frac{|D|+1}{f_c^{tag}+1}$, where f_c^{tag} is the frequency of c as tag in D. The value 1 is added to both numerator and denominator to deal with new terms that do not appear as tags in the training set. As discussed in [3], this metric may privilege terms that do not appear as tags in the training set. However, other relevance metrics (e.g, TF) will be considered in the final recommendation function. Thus, their relative weight can be adjusted.

Finally, the entropy of term c in the tags feature is defined as $H^{tags}(c) = -\sum_{(c\rightarrow i)\in R} \theta(c \rightarrow i) \log \theta(c \rightarrow i)$. If a term occurs consistently with certain tags, it is more predictable, thus having lower entropy. Terms that occur indiscriminately with other tags are less predictable, having higher entropy. It is better to recommend more consistent and predictable terms (i.e., with lower entropy).

4.2 Novelty Metric

Vargas *et al.* [11] proposed to estimate the novelty of an item based on its popularity, that is, the novelty of an item is related to 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 does capture exactly the aspect proposed by Vargas *et al.*, as it favors candidates that occur less frequently in the training set. Thus, we here propose to use IFF as a novelty metric.

Note that, although Belém *et al.* [3] have previously used IFF to recommend tags, their purpose was recommending tags that can better discriminate an object from the others, an aspect that is related to relevance. Here, IFF is also used to raise 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.

4.3 Diversity Metric

Another desired property of a list of recommended items is diversity, that is, each item in the list should represent a different piece of content from the others. In the context of tag recommendation, we want to avoid redundant recommendations, such as synonyms and semantically similar terms[1], aiming at capturing different concepts (i.e., facets) related to the target object.

Like in [11], we here estimate the diversity of a candidate term c with respect to a list C_o of candidates for recommendation for target object o as the average semantic distance between c and each other term in C_o. Thus, we define the *Average Distance to other Candidates* (*ADC*) as $ADC(c, C_o) = \frac{1}{|C_o|} \sum_{t \in C_o, t \neq c} dist(c, t)$, where $dist(c, t)$ measures the dissimilarity between candidate terms c and t.

There are various ways of estimating the dissimilarity between two terms. We here estimate the dissimilarity between terms t_1 and t_2 by the relative difference between the sets of objects O_1 and O_2 in which they appear as tag, i.e., $dist(t_1, t_2) = \frac{|O_1 - O_2|}{|O_1 \cup O_2|}$. If both sets are empty, we set $dist(t_1, t_2)$ equal to the maximum value, i.e., 1. Note that by measuring the dissimilarity between two terms in this way, we are basically using the set of objects in which each term appears as tag to represent its possible meanings. Thus, terms that appear in very different sets of objects most probably have very different meanings.

Once again, we emphasize that taking only diversity, or novelty, into account does not necessarily lead to appropriate recommendations. They must be considered jointly with relevance for the sake of effective recommendations.

5 Tag Recommendation Strategies

We now describe the analyzed tag recommendation strategies, including the baseline and our new strategy, which extends the baseline to include new metrics that capture both novelty and diversity as well as a new objective function that jointly considers relevance, novelty and diversity.

5.1 State-of-the-Art Baseline

Our baseline is the state-of-the-art method proposed in [3], based on Genetic Programming (GP). We refer to this strategy as GP_{rel}, since it exploits only relevance. GP_{rel} generates a set of candidate terms C_o for recommendation to object o containing: (1) terms that co-occur with tags previously assigned to o (i.e., tags in I_o), and (2) terms extracted from other textual features associated with o, namely, its *title* and *description*.

Given a target object o, GP_{rel} computes a list L_m of tag relevance metrics (defined in Section 4.1) for each candidate term in C_o. In the learning phase, a binary label is assigned to each candidate, indicating whether it is relevant to the object, for objects in the training set D. Through an evolutionary process that explores operations such as mutation and crossover, GP_{rel} learns a function f that maximizes a given objective function, which captures the relevance of a set of recommendations for the object. We here use the *nDCG* at the top k

[1] Particularly synonym terms that have different roots, otherwise applying stemming would be enough to remove redundancy.

positions of the ranking as objective function[2]. Function f is then used to rank and recommend candidate terms for unseen objects in the test set O (test phase).

GP_{rel} can be easily extended to exploit new metrics and objective functions, including functions which combine multiple objectives, as we discuss next.

5.2 Our New Strategy

Our new strategy, called $GP_{rel+nov+div}$ exploits the same set of candidate terms of GP_{rel}. However, it introduces new metrics as features in the list L_m and as part of the objective function. Specifically, we include *Average Distance to other Candidates* (*ADC*), defined in Section 4.3, in L_m and (indirectly) in the objective function. Moreover, unlike in GP_{rel}, which exploits IFF only as a relevance metric in L_m, in $GP_{rel+nov+div}$ we also have it as part of the objective to be optimized, which changes the search space for recommendation functions.

Specifically, in order to add the novelty of a list of recommended terms C to the objective function of $GP_{rel+nov+div}$, we employed the metric *Average Inverse Popularity* over the top k positions of the ranking, $AIP@k$, adapting it from [11] to our context. $AIP@k$ is defined here as a normalized average of the IFF values of the first k recommended terms. Let $disc(i) = 1/log(1+i)$ be a rank discount function that provides a weight for the i^{th} position of the ranking. $AIP@k$ of list C is defined as: $AIP@k(C) = \frac{1}{K}\sum_{i=1}^{k} disc(i) \times IFF(c_i)$, where c_i is the i^{th} term in C and $K = \sum_{i=1}^{k} disc(i) \times IFF_{max}$ is the normalization constant.

We introduce diversity to the objective function by using the *Average IntraList Distance* in the top k positions of the list of recommended terms C (AILD@k) [11], defined as $AILD@k(C) = \frac{1}{K'}\sum_{i=1}^{k}\sum_{j=i+1}^{k} dist(c_i, c_j)$. $K' = (k^2 - k)/2$ is a normalization constant, and $dist(c_i, c_j)$ is as defined in Section 4.3.

Finally, we define the new objective function (*Fitness*) as a convex linear combination of the three aspects (relevance, novelty and diversity) as $Fit(C) = \alpha AIP@k(C) + \beta AILD@k(C) + (1 - \alpha - \beta)nDCG@k(C)$, where $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ are tuning parameters to weight the evaluation metrics.

6 Experimental Evaluation

6.1 Data Collections

The tag recommendation methods were evaluated in three datasets, containing *title*, *tags* and *description* associated with objects from three applications: LastFM, YouTube and YahooVideo. These datasets include the textual features associated with 2.758.992 artists in LastFM, 160.228 videos of YahooVideo and more than 9 million videos of YouTube. For the experiments, we sampled 150,000 objects from each collection, removed *stopwords* and performed stemming with the Porter algorithm[3] to avoid trivial recommendations such as plural and other variations of a same word.

[2] Results for $P@k$ as objective function are similar.

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

Table 1. Tuning of parameter $\lambda = \alpha = \beta$ for $GP_{rel+nov+div}$. Best results (best tradeoff between relevance, novelty and diversity) in bold.

	LastFM			**YahooVideo**			**YouTube**		
$\lambda = \alpha = \beta$	$nDCG$@5	AIP@5	$AILD$@5	$nDCG$@5	AIP@5	$AILD$@5	$nDCG$@5	AIP@5	$AILD$@5
0.00	0.429	0.293	0.892	0.754	0.423	0.892	0.510	0.613	0.973
0.20	0.427	0.315	0.902	0.755	0.444	0.903	0.509	0.627	0.972
0.25	**0.422**	**0.334**	**0.910**	0.758	0.438	0.901	0.512	0.635	0.973
0.40	0.369	0.537	0.948	0.753	0.450	0.912	0.509	0.650	0.974
0.50	0.330	0.617	0.957	**0.749**	**0.465**	**0.914**	**0.503**	**0.664**	**0.975**
0.60	0.238	0.760	0.974	0.705	0.519	0.933	0.495	0.676	0.975

6.2 Methodology

Similarly to most studies in tag recommendation [3,10,9], we adopted an automatic approach for evaluation: we used a subset of the object's tag as a *gold standard*, i.e., the relevant tags for that object. These tags are not considered for the calculation of the metrics. The remainder subset of tags (I_o) is used as input for the recommenders. More specifically, we fixed half of the tags of each object (randomly selected) as gold standard and half as input. This methodology was adopted because the manual evaluation of tags is an expensive process in terms of time and human effort, besides being subjective. Thus, we leave the manual evaluation of the strategies for future work. The experiments were performed using 5-fold cross-validation with the validation fold being used for parameter tuning. The reported results are averages of the 5 test folds.

As evaluation metrics we used $nDCG@k$, a traditional IR relevance metric [2], as well as $AIP@k$ and $AILD@k$, adapted from [11] as described in Section 5.2 to evaluate novelty and diversity, respectively. We computed these metrics for the top 5 terms in the ranked list of recommendations (i.e., $k = 5$).

6.3 Representative Results

We now describe the parameterization of each strategy and discuss the main results of our evaluation of both recommendation methods considered. The results are averages of 25 runs (5 *folds* × 5 seeds). We also compute 95% confidence intervals, omitted in Table 1 for space reasons. In any case, with this confidence level, the deviations of the results are in general inferior to 2% of the averages.

Parameterization. We first performed a series of experiments to determine the best values for the parameters of the analyzed methods, using a **validation set**. We fixed the parameters of the GP framework as in [3], focusing on parameters α and β, which control the weight given to novelty, diversity and relevance in the objective function. To that end, we started by giving the same weight for novelty and diversity, and testing different tradeoffs between the the sum (novelty + diversity) and relevance, i.e., we set $\alpha = \beta = \lambda$. We tested values of λ in the interval [0,0.6], since the relevance started dropping a lot after $\lambda = 0.6$.

Table 1 shows the results. In general, the higher the value of λ, the higher the values of novelty (AIP) and diversity ($AILD$) of the recommendations, but also the higher the reduction in relevance ($nDCG$). This was expected since relevance and novelty/diversity may be seen as conflicting objectives. For instance,

Table 2. Average results and 95% confidence intervals. Best results and statistical ties in bold.

Collection	*Method*	*nDCG*@5	*AIP*@5	*AILD*@5
LastFM	GP_{rel}	**0.429** ± 0.002	0.293 ± 0.006	0.892 ± 0.002
	$GP_{rel+nov+div}$	0.422 ± 0.004	**0.334** ± 0.015	**0.910** ± 0.005
YahooVideo	GP_{rel}	**0.755** ± 0.005	0.423 ± 0.005	0.892 ± 0.004
	$GP_{rel+nov+div}$	**0.749** ± 0.007	**0.465** ± 0.010	**0.914** ± 0.004
YouTube	GP_{rel}	**0.510** ± 0.004	0.613 ± 0.004	**0.973** ± 0.002
	$GP_{rel+nov+div}$	**0.503** ± 0.004	**0.664** ± 0.006	**0.975** ± 0.002

random recommendations may present high novelty and diverstiy although very low relevance. Our goal is to maximize novelty and diversity without compromising relevance. Thus, for each dataset, we chose the λ value that produced the higher gains in *AIP* and *AILD*, causing a loss of at most 2% in *nDCG* with regards to the results of the GP_{rel} baseline (indicated as λ=0 in the Table). For instance, in LastFM, with λ=0.25, it is possible to obtain gains of at least 14% in *AIP* and 2% in *AILD*, loosing only 1.5% in *nDCG*.

Given that the simultaneous optimization of the three aspects, i.e., novelty, diversity and relevance, may be hard to accomplish, we also tested a combination of only two objectives, i.e., relevance and novelty as well as relevance and diversity. That is, we first fixed $\alpha = 0$ varying β, and then fixed $\beta = 0$ varying α. However, none of these strategies outperformed the original $GP_{rel+nov+div}$ in terms of relevance, novelty and diversity, thus we chose the best values of the parameters found in Table 1 to be used in the experiments with the test set.

Evaluation of the Recommendation Strategies. Having the parameters defined in the validation set, we used them to perform experiments in the test set to compare the strategies. Results are shown in Table 2. We start with a general observation regarding the behavior of the strategies in the different datasets: the value of relevance tends to be higher in YahooVideo, followed by Youtube and LastFM. This may be explained by several factors. In YahooVideo and LastFM, tags are collaboratively created (any user can assign tags to a content), while only the video uploader can do this in YouTube. Moreover, the average number of tags per object is larger in YahooVideo than in LastFM, favouring the methods in the former due to the higher availability of data in the training set. Also, it is difficult to extract relevant candidate terms from the textual features of LastFM because, in general, there is little intersection among the contents of the title, description, and tags associated with a same object in that applications [3]. This makes the distinction between relevant and irrelevant terms by the several relevance metrics that are based on frequency and *spreading* of the terms in the textual features much harder.

Comparing our new strategy $GP_{rel+nov+div}$ with the state-of-the-art GP_{rel}, we obtained gains in *AIP* (novelty) of 14% in LastFM, 8% in YouTube and 10% in YahooVideo, losing at most 1.6% in *nDCG*. Thus, it is possible to obtain novel recommendations while maintaining similar levels of relevance with the new proposed objective function, althought relevance and novelty may be conflicting objectives. However, it is more difficult to improve diversity, since the *AILD* results are already very high in GP_{rel}: our gains are below 2.5%. This happens because the data is sparse, making the values of distance between tags typically large, with small differences between them, given that there is little information about tag co-occurrences. Notice also that the gains in novelty and diversity are higher in LastFM and YahooVideo, where tags are collaboratively created, exhibiting, thus, higher variability.

7 Conclusions and Future Work

We here defined novelty and diversity for tag recommendation and proposed a new recommendation strategy that considers both aspects jointly with relevance. Our strategy produces gains of up to 14% in novelty without harming relevance, over a state-of-the-art strategy that only exploits relevance. The corresponding gains in diversity are more modest (up to 2.5%). However, we note that the GP framework is flexible, allowing the inclusion of new attributes and objective functions that capture other aspects of the problem. Thus, as future work, we plan to explore new metrics and objective functions (e.g., temporal novelty and topic diversity), and perform human judgment of recommendations.

References

1. Almeida, H., Gonçalves, M., Cristo, M., Calado, P.: A combined component approach for finding collection-adapted ranking functions based on genetic programming. In: SIGIR (2007)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley (1999)
3. Belém, F., Martins, E., Pontes, T., Almeida, J., Gonçalves, M.: Associative tag recommendation exploiting multiple textual features. In: SIGIR (2011)
4. Boll, S.: MultiTube–Where Web 2.0 And Multimedia Could Meet. IEEE MultiMedia 14(1) (2007)
5. Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M.: Assessing the quality of textual features in social media. IP&M (2012)
6. Garg, N., Weber, I.: Personalized, interactive tag recommendation for flickr. In: RecSys (2008)
7. Guy, I., Zwerdling, N., Ronen, I., Carmel, D., Uziel, E.: Social media recommendation based on people and tags. In: SIGIR (2010)
8. Lathia, N., Hailes, S., Capra, L., Amatriain, X.: Temporal diversity in recommender systems. In: SIGIR (2010)
9. Lipczak, M., Hu, Y., Kollet, Y., Milios, E.: Tag sources for recommendation in collaborative tagging systems. In: ECML PKDD (2009)
10. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW (2008)
11. Vargas, S., Castells, P.: Rank and relevance in novelty and diversity metrics for recommender systems. In: RecSys (2011)
12. Wu, L., Yang, L., Yu, N., Hua, X.: Learning to tag. In: WWW (2009)
13. Zhou, T., Kuscsik, Z., Liu, J., Medo, M., Wakeling, J., Zhang, Y.: Solving the apparent diversity-accuracy dilemma of recommender systems. National Academy of Sciences of the United States of America 107(10) (2010)