More Efficient Tagging Systems with Tag Seeding

Vikas Kumar[1], Daniel Kluver[1], Loren Terveen[1], John Riedl[1]
[1]GroupLens Research - University of Minnesota, Twin Cities, USA
vikas,kluver,terveen@cs.umn.edu

Abstract

Tags are a useful mechanism for users to find, organize, and understand items on a web site. A tagging system evolves to reflect a user community's understanding of an information space. However, the evolution can be problematic: specifically, the ratio of tags to items can decrease over time, to the extent that tags no longer effectively discriminate among items. The primary reason is over application of existing popular tags to most of the items. This makes it difficult for a user to search and navigate using specific tags. We address this problem of disproportionate application of existing popular tags by eliciting applications of those tags that are under applied *but* highly relevant. In this paper, we introduce two metrics to identify such under applied relevant tags from the system. We then design a controlled study to elicit more applications of these tags, to show how the declining discriminating power of tags is counteracted by interventions in tag applications using these carefully chosen under applied tags, namely *seeded tags*.

We found that users were able to apply these underutilized tags, with the result that about eleven months of lost navigational and search efficiency of the tagging system was attained back. We also monitored subsequent usage of these tags in the system and found that the *seeded* tags does attract more users in navigation (such as clicks and searches) than before our intervention, suggesting that *tag seeding* has a persistent effect over the system.

1 Introduction

Tagging systems allow users to search, navigate and explore using a shared vocabulary. Systems like StackOverflow,del.icio.us and Flickr support such annotations on a massive scale, enabling tags to evolve as a popular means of categorizing and organizing content. A collaboratively generated shared vocabulary is shown to be more effective [4] capturing users' diverse perspectives on the information space.

However, a tagging vocabulary can evolve into a suboptimal state; specifically, it can become less effective for tag's basic functionality of navigation and search. This is due in part to a design feature common to nearly all tagging systems i.e. in choosing tags to display, the system favors more popular tags. These tags being more visible, tend to get applied even more often [3, 8] i.e. the rich gets richer. In contrast, other tags that might be usefully applied, but not popular enough become less visible, thus even less likely to be applied. Over time, whole folksonomy can become less diverse and less useful for discriminating the items in an information space.

The evolution of tags and their declining discriminating power is best explained by Chi et al. [2] using information theory concepts. Chi et al. demonstrate how over utilization of existing tags affect organizing and categorizing feature of tags which result in decline of navigational and search efficiency.

In this paper, we address this problem of tagging efficiency and propose solutions and best practices to *reverse* its declining trend in the system. We derive new algorithms based on the concept of *Tag Seeding* proposed by Peters et al. [11] to facilitate an opportunity for those less visible underutilized tags.

Tag Seeding suggests existence of underutilized tags in the system that require special attention and careful guidance to grow among other prosperous, popular tags. However, it is more important to identify and differentiate underutilized *relevant* tags from irrelevant futile tags. In this paper, we address this critical issue of identifying true *candidate* seeding tags (i.e. tags that are underutilized, yet relevant and highly applicable). We introduce *two novel metrics* to recognize these tags: *Opportunity Gap* and *Potential*. *Opportunity Gap* aims to find tags which are relevant but lack exposure/visibility in the system; whereas *Potential* metric identifies tags that have significant capability (or potential) to apply more than their existing applications.

To analyze if *seeded tags* are capable enough to add significant discriminating power to the shared vocabulary we guide our work with the research question:

RQ1: Does tag seeding *improve* the navigational and search efficiency of tagging systems?

We formulate our second research question to further understand the effectiveness of derived algorithms:

RQ2: Which algorithm performs better at *tag seeding*?

Moreover, we establish our third guiding research question to understand if the tags given the *initial boost* (seeding) can sustain their growth in the system:

RQ3: Does seeding tags lead to *lasting changes* in their visibility and usage?

The outline for rest of the paper is as follows: We discuss prior work exploring tagging systems in detail in *Related Work*. *Background Concepts* provides a brief explanation of information theory concepts followed by an overview of MovieLens dataset. In *Methods & Metrics* section we formalize the *Opportunity Gap* and *Potential* metrics and explain the derived algorithms followed by an outline of the experimental setup. The key findings and observations exploring the research questions are discussed in *Results* section. We discuss the implications and limitations of our results and conclude with general discussion in *Conclusion & Future Work*.

2 Related Work

Collaborative Tagging Systems allow people to access and organize online information quickly and efficiently. Tagging is emphasized to be independent of any guidelines and that users should perform the process of free labeling [9, 6] . Clay Shirky indicates that the absence of *controlled* vocabulary makes users think collectively to categorize contents making the system more efficient than traditional ontological and hierarchical based system. However, this free form indexing of contents comes with a cost of an unsystematic and unstructured vocabulary [7].

Prior work investigates this evolving vocabulary and tagging behavior, often referred in study of *folksonomy* (a term coined by Vander Wal in 2007 to describe tags generated by community of diverse users). Golder & Huberman [5] attempt to understand the tagging patterns and argue that over a period of time the growth of tags stabilize and reflect a structure or an order in the system. In an attempt to refine the structure of social tagging system researchers have proposed many theoretical and practical approaches. The theoretical approaches include *semantic enrichment* [**?**] whereas a set of practical approaches was discussed in a work of Peters et al. [11] defined as *Tag Gardening*. Peters et al. make an analogy of *tagging system for a garden* and tags as "savaged wildly grown plants in the garden". The ideology is that a tagging system can be taken care in a way similar to a garden where tags (or plants) can grow reasonably well naturally with careful guidance. For instance: the gardening activities-*fertilizing & Landscaping* suggest merging synonyms like *"Sci fi"* or *"sci-fi"* and creating visual distinctions in homonyms such as *"jaguar (animal)"* and *"jaguar(car)"* respectively; *weeding* to remove or hide inaccurate or idiosyncratic tags; and *seeding* to provide a careful guidance and close attention to under grown tags. Motivated from the concept of *tag seeding* suggested by Peters et al. we formalize the concept (of seeding) by proposing new metrics and algorithms to identify candidate seeding tags in this paper.

However, to identify the relevant underutilized tags and understand direct effectiveness of seeding, it is critical to understand the evolution of tagging systems. The influence of *existing tags* in the system have been studied by Fu et al. [3] who conducted a controlled study to understand the semantic imitations in social tagging systems. He found that users' new tag applications are influenced by presence of existing tags resulting in *semantically similar and less diverse* set of tags. In related work, Cattuto et al. [1] was able to predict user tags with high accuracy by considering tags already applied to an item. Community influences are further explored by Sen et al[8] showing that users' tagging pattern depend on the set of tags shown to the user.

Chi et al. [2] explore tagging system in terms of its *efficiency* to clearly understand the evolution of shared vocabulary. A tagging system efficiency lies in its ability to enable users to discover items quickly using tags. However, using information theory concepts Chi shows that the efficiency is set to *decline* after a period of time and that tags reaches a point where diversity of tags stabilizes and loses specificity to search and navigate efficiently over the constantly growing item size. Taking cue from this work we analyze our research dataset (MovieLens) and notice similar phenomenon of declining efficiency. We discuss the information theory concepts and observations from the dataset in the following section.

3 Background concepts

In this section we provide an overview of information theory concepts used in the paper to describe some of the metrics and evaluation techniques used by Chi et al, in context of MovieLens. We evaluate each metric in context of MovieLens to show how the efficiency of system has been consistently declining over time.

In MovieLens, a single tag application is referred to an event when a user u_i annotates a movie m_i with a tag t_i. Considering tag application events, we can easily define probability of a movie $p(m_i)$, probability of a tag $p(t_i)$ etc. For example, $p(m_i)$ is ratio ratio of applications for movie m_i to all applications. Using these notations following we describe following information theory concepts:

3.1 Entropy

Entropy is a measure of uncertainty or unpredictability of information content. The entropy of a system increases with increase in total number of events or if it follows a more uniform distribution.

In the domain of tagging, the entropy of tags measures the diversity of tags (*diversity* in this context is solely based on distinct number of tags present in a system i.e. more distinct tags imply more diverse set and vice-versa). For example, consider a single tag applied to each movie; the diversity (or entropy) of tags is 0 bits; whereas if there is a set of unique tags applied uniquely to each movie then diversity (or entropy) is high. We expect tagging systems to have a constant increase in entropy to maintain a good ratio of tags to increasing number of items. A decreasing diversity will reflect disproportional usage of existing tags in the system.
Similar to entropy of tags, entropy of movies measures how diverse (or distinct) movies exist in a system and is expected to keep increasing in the system due to constant addition of new movies.

3.2 Conditional Entropy

This quantity measures the amount of randomness or uncertainty in tags T (or movies M) given that we have knowledge about the movies M (or tags T). The conditional entropy of a movie given a tag measures the uncertainty to identify a movie given a tag i.e. how hard it is for an user to search for a movies given a tag. For example, given a tag that applies to every movie, it will be difficult and thus uncertain to search specific movie in the system.

3.3 Mutual Information

The mutual dependence between tags T with tags $t_1, t_2, ...t_n$ & movies M with events $m_1, m_2...m_n$ is given by mutual information. In this context, the tags (T) and movies (M) related by applications are two discrete random variables and mutual information $I(T; M)$ explains how much tags can explain about the movies. In a tagging system high dependence between tags and movies would suggest tags having high discriminating power for the movies. The higher the mutual information, more easy it is for a tag to explain about a movie. However, if tags are applied independently of movies i.e. every tag is applied to every movie then there will be *zero* mutual information and hence *zero* discriminating power.

4 Dataset: MovieLens Tags

We present the analyses and observations of metrics described in previous section in context of MovieLens. MovieLens is a movie recommendation site which also provides a social tagging platform. A user can apply new or existing tags to each movie or use tags to search and navigate other movies. MovieLens has 30,187 unique tags for 14,495 movies applied by 7194 distinct users. To understand the growth of tags over time we analyze the activity logs of over a period of 24 months starting January 2010 until December, 2012.

Tagging in MovieLens follows the long tail distribution with 1% of tags constituting 46.23% of all applications. Observing MovieLens characterized by stabilizing tag diversity (tag entropy) over last several months (top right of Figure 1) with diversity of movies consistently increasing as shown in Figure 2 reflect how number of tags to movies ratio has grown over time resulting in less discriminating power for tags due to over utilization of existing tags.

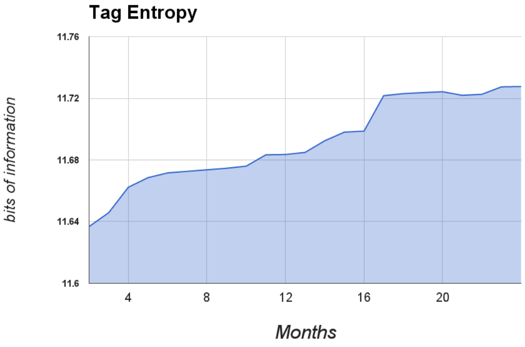


Figure 1: Entropy of tags in the MovieLens plotted against time.

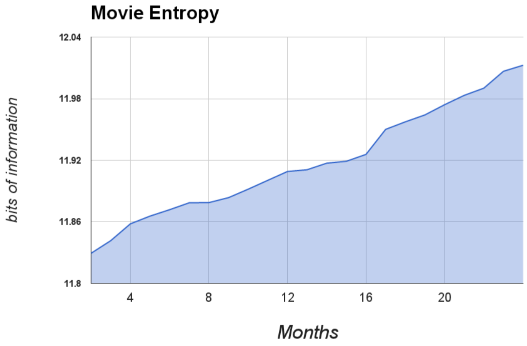


Figure 2: Entropy of movies in the MovieLens plotted against time.

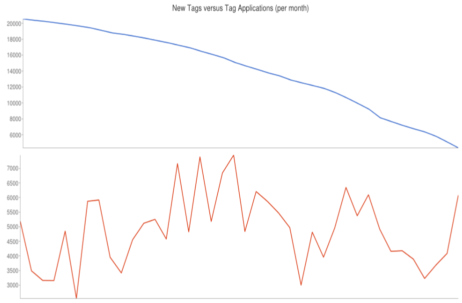


Figure 3: Number of new tags each month(Upper plot). Total Applications per month(Lower plot).

To clearly understand the phenomenon, we explore another parameter i.e. the frequency of new tags introduced in the system compared with the frequency of tag applications per month respectively. The plot in Figure 3 shows the *declining* diversity of tags (i.e. declining number of new tags per month). However the number of applications per month is randomly ordered and can be said to have neither decreasing nor increasing over time suggesting exhausting tag vocabulary. Moreover, the observation made by Chi et al, that the mutual information between tags and items decreases is also exhibited in MovieLens. *Figure 4* shows the declining mutual information over time in MovieLens predicting the decreasing mutual dependence of tags and movies suggesting the decreasing navigational and search efficiency of the system.

4.1 Tag Genome

This section discusses about a tag parameter known as Tag Genome from previous work [10] in MovieLens. The parameter and its usage in the context of this paper is described as follows:

Tag Genome quantifies and predicts the relationship between each tag $t \epsilon T$ (set of tags) and $m \epsilon M$ (set of movies) using machine learning techniques, denoted as $rel(t, m)$. The value suggests how strongly a tag t applies to a movie m with range between 0 *(least relevant)* and 1*(most relevant)*.

Genome considers current tag applications, movie synopsis, reviews, etc to determine the relevance value. The tag

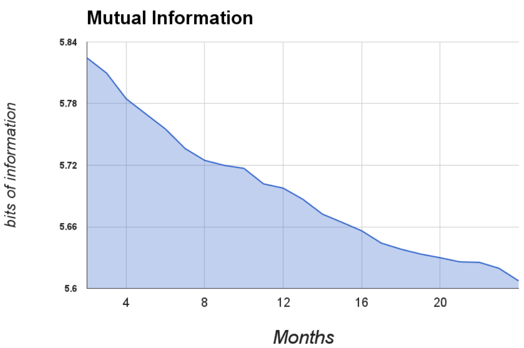


Figure 4: Mutual Information of tag and movie in MovieLens plotted against time.

set T is selected carefully from the system discarding any idiosyncratic tags such as "to watch list", "my dvd collection" etc or cast names.

In this work, we use the tag genome in two ways:

1. Identify *most relevant movies* for a tag.

2. Generate an ideal set of relevant movies for a tag with $rel(t, m) > 0.65$ [1] to be used in the *Potential* metric (details in next section)

Limitation: Tag Genome metric may not be available for other tagging systems in its present form as it is for MovieLens and hence we mention this as one of our limitations of this research. However Vig et al [10] in his work have mentioned about the adoption of this technique for other similar systems.

5 Methods & Metrics

In this section, we introduce two novel metrics to identify candidate seeding tags. We also discuss other techniques (such as Baseline) we use to compare and evaluate the metrics in the proceeding subsection of algorithms.

5.1 Metrics

The goal: to explore whether *seeding* improves efficiency of system and helps increase in application of underutilized tags, it is important to select tags that are in general capable to have more applications though under applied in the current state of the system. For example, the tag *"hitler"* is not applied often, but that is because it is applicable only to few movies related to World War II or German history. Instead, we want tags that can apply to many more existing movies such as *"entirely dialogue"* or *"talky"* that are relevant to a much larger set of movies than their current set of applications.

Based on such observed nature of tags, we describe two novel metrics: *Opportunity Gap* and *Potential*.

Metric 1- Opportunity Gap: As we discussed earlier, popular existing tags have higher visibility in the system causing an imbalanced applications for such tags compared to other existing tags. *Opportunity Gap* aims to address this imbalance by identifying tags that are relevant but lack opportunity being under applied and less visible in the system. Specifically it identifies tags having higher probability of user interaction for the number of times it is shown (or visible) to users, compared to other tags. We analyze different user interactions and quantifies in a single quantity to call it as *Demand*. As the name suggests, *Demand* captures the demand of a tag based on users interaction with a tag. We include tag search and tag clicks as the user interactions to compute the *Demand* of a tag i.e. if a tag receives more number of clicks and searches then it has demand in the system. Similarity we define *Supply* of a tag based on presence of a tag i.e. its applications and visibility in the system. More applications or more visibility suggests more supply in the system.

Based on the *Supply* and *Demand* of a tag *Opportunity Gap* aims to identify tags that have higher demand even though having less supply in the system. Infact it is this gap in supply and demand that we identify as an opportunity for the tag to grow by seeding them into system explicitly. Quantitatively, we define *Opportunity Gap* as the ratio of *demand* to the *supply* of a tag in the system. We define *Demand* as a quantity to determine the *usage* (or user demand) of a tag in users navigation and searches in the system.

$$OpportunityGap = \frac{Demand}{Supply}$$

Where *Demand* is defined as:

$$log(tag\ clicks\ +\ search)$$

and *Supply* as:

$$tag\ views * log(tag\ applications + C)$$

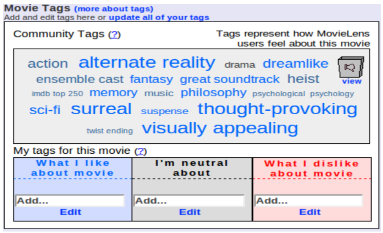
We define each of the parameters below: ***tag clicks & search***: provides implicit feedback about tag's relevance in the system. Users can navigate MovieLens content by searching or clicking on tags i.e. more clicks and searches for a tag implies more relevance. The log of cumulative tag clicks and search is taken to normalize the observed long tail distribution of the values.

tag views: counts the definite number of tag views (or appearance) during user navigation. A tag is shown to user at different venues in the system such as tag cloud (Figure: 5), tag search results and movie search pages.

And, ***tag applications:*** captures the number of annotations a tag has with all items. More applications imply more popularity leading to more exposure for the tag (supply). The *log* of the parameter is taken to normalize the observed long tail distribution. The constant $C = 10^{-6}$ is used to avoid divide by zero error i.e. when tag is applied only once.

The MovieLens system further provides an explicit way to capture user ratings for a tag. Users record their feedback for a tag using a "thumbs up" or "thumbs down".

[1] The value is chosen based on the empirical analysis and manual observation of the ideal set for tags.



Movie Tags (more about tags)
Add and edit tags here or update all of your tags

Community Tags (?)

Tags represent how MovieLens users feel about this movie

My tags for this movie (?)

Figure 5: *Tag Expression* : tag cloud on the movie details page

Using these voting options, we define *tag rating* as a ratio of distinct "thumbs up" votes to "total votes" i.e. :

$$tagrating = \frac{C(thumbs\ up) + 1}{C(thumbs\ up) + C(thumbs\ down) + 2}$$

where, $C(x)$ is count of x, and the constants 1 and 2 are Laplacian correction terms.

We include this explicit rating with *Opportunity Gap* as shown below to weight those tags higher having better user ratings. Finally, we formulate *Opportunity Gap (OG)* as:

$$OG = tagrating * \frac{log(tag\ clicks\ +\ search)}{tag\ views * log(tag\ applications + C)}$$

The top 20 tags selected by *Opportunity Gap* are shown in Appendix *I-Table 1.* We find that it does a good job of identifying opportunity lacking tags such as "bollywood", "mentor" etc. However in the process it is prone to select few non-generic tags such as *"dr. seuss"* (16*th* tag in the list). It is applied to specific popular kids movies such as- *"Dr. Seuss, The Lorax"*, *"The Cat in the Hat"* etc. Tags in such popular movies leads to higher search and clicks for the tag (or *Demand*) compared to the number of times it is applied in the system (*Supply*). But note that applications are limited for this tag and cannot be applied to any more movies. We understand that such tag is relevant for the system but *not necessarily applicable for seeding.* For such cases, it is mandatory to determine the *applicability* of a tag. *Potential* metric in next section attempts to address this specific problem.

Metric 2- Potential: *Potential* is an information theoretic approach to measure the relative ability of a tag to apply more. It determines how *well* a tag is currently applied to the movies in the system i.e. how well the tag's *existing* applications (say $M_{current}$) can tell about tag's *ideally*[2] applicable movies (say M_{ideal}). For example tags such as *"mentor"*, have relatively small number of existing applications of 24 compared to its ideal applications approximating 134 *movies*). M_{ideal} is computed based on the predicted relevance values of a tag and movies using tag genome (Section:IV.1). To precisely understand the concept numerically we define M_{ideal} as a vector of all movies with each point being a boolean variable depicting whether the movie applies ideally to the tag or not. Similarly, $M_{current}$ is a vector of all movies with each point depicting if the movie is currently applied to the tag or not. Using these variables we define *Potential* as conditional entropy:

$$H(M_{ideal}|M_{current}) = H(M_{ideal}) - I(M_{ideal}; M_{current})$$

where, $H(M_{ideal})$ represents the entropy or uncertainty of these movies among all set of movies in the system. $I(M_{ideal}; M_{current})$ represents mutual information shared by both set of the movies and $H(M_{ideal}|M_{current})$ represents the conditional entropy. The conditional entropy quantifies the amount of randomness of uncertainty in one set i.e. M_{ideal} given that we have knowledge about $M_{current}$. The conditional entropy allows us to understand if the current applications are able to explain well (less randomness) about the ideal applications. If the value is small then the randomness or uncertainty about the ideal set is reduced considerably given the tag's current applications i.e. current applications are able to cover most of the ideal applications in the system. Whereas, if the value is large then current applications presumably fail to contain movies that should have been ideally applied.

We show the top 20 tags selected by *Potential* in Appendix *I-Table 1.* We observe that *Potential* due to its nature is able to select applicable tags and successfully discarding tags like *"dr. seuss"* due to their low applicability. However due to its high bias towards applicability of tags makes the metric to select popular tags like *"action"*, *"drama"* etc as relevant *seeding* tags. Such tags though applied well, are still the most common tags applicable to large number of movies (*primarily* due to existence of large number of movies with similar genre of Action, Drama etc). However, *these tags don't require seeding* as they are sufficient popular to be highly visible in the system.

We discuss more derived metrics and corresponding algorithms we explored to compare our metrics in the following section.

5.2 Algorithms

We derive algorithms to select *top N* tags for *seeding* based on the *two metrics.* Due to lack of any performance benchmark, we also propose a *Baseline* algorithm, to compare and evaluate other algorithms. The five algorithms chosen are as follows:

Baseline: As a simple baseline, we select a random set of tags for seeding. However, to make sure that this random selection is somewhat plausible i.e. should contain tags that are neither too obscure nor too popular. To achieve this we randomly select from the universal set of tags that are moderately applied (10 to 50 percentile of tag applications) by a moderate number of users (10 to 50 percentile of distinct users).

Opportunity Gap: The *top N* tags are selected in decreasing order of their *Opportunity Gap* value.

Potential: Same as *Opportunity Gap*, we select *top N* tags in decreasing order of their *Potential* value.

[2] The ideal applicable movies for tag is determined by *tag genome* (discussed in previous section)

Hybrid: *Potential* & *Opportunity Gap.* The basic metrics have some limitations such as *Opportunity Gap* penalizes relatively popular tags (very high supply) whereas *Potential* promotes highly applied *popular* tags (most of the movies in the system are applied with these tags). Achieving a balance somewhere in middle is required to have tags with higher *Opportunity Gap* and relatively higher *Potential* too. Empirically we realize that simply multiplying both the absolute values is able to identify better set of tags which we call the *Hybrid* set.

Personalized: The algorithms described so far identify *global* set of candidate seeding tags. However, individual users because of specific interest are likely to be not aware of certain genre of tags. For example: users with interest in *thrillers* and *action* movies may not find *romantic* or *romantic comedy* related tags easy to apply, no matter how efficiently *Potential* or *Opportunity Gap* suggests. Hence, we explore a simple approach to personalize seeding tags based on the interests of individual users. The steps to personalize are:

1. Identify tags ($U_{current}$) currently applied by user U in system.

2. Determine the most similar tags (U_{sim}) for the set $U_{current}$. The similarities between tags are evaluated by normalized mutual information[3] defined as:

   $$Sim(t_1, t_2) = I(t_1; t_2)/H(t_2)$$

   where $I(t_1; t_2)$ is the mutual information between the two tags and $H(t_2)$ is the entropy.

3. Select top 50 most similar tags from $U_{top50} = top50(U_{sim50}$.

4. Determine the *seeding* value (using *Hybrid* algorithm) for each tag in U_{top50}. Reorder them to form U'_{top50} in decreasing order of seeding value.

5. Finally, select *top N* tags from the reordered set U'_{top50} to have the final personalized seeded set U_{final}.

This approach is simple and at the least facilitates us to find the similar personalized seeding tags based on users' current applications. In case user does not have any current tag application we do not assign that user in our experiment to Personalized approach. We discuss about the experiment in the following section.

5.3 Experiment

We aim to investigate each of the research questions with an explicit user intervention for *seeded tags.* We designed a web based interface and invited users to apply chosen tags to a set of movies as shown in Appendix *I*: Figure 6. In following paragraphs we discuss about key aspects of the interface and the experimental setup.

Tags: The design objective of the experiment is to gather new tag applications respectively for each seeded tags. Users are assigned mutually exclusively to each of the five condition (corresponding to five algorithms). The top 20 tags are selected from each algorithm forming five set of tags for each condition. To be precise, every user in a condition is shown the same set of top 20 tags except the Personalized conditions- where we select top 20 unique tags for each of the user assigned to this condition.

As a part of the study, each user in experiment is required to apply a set of movies for the tags chosen. Each participant is asked to apply the *tag shown* to a set of six movies. A user can apply the tag to *all*, *some* or *none* of the movies on the page based on user expected relevance of tag to the movies. We explain the movies selection process in following section.

Movies: There are six movies pre-selected for each tag shown to the user. To apply tags to a movie, users can simply select a movie (by clicking on the *movie tile*). Each *tile* contains details to help user recollect about the movie or relate to the theme of the movie. It consists of *Album art*, *Release Year*, *Synopsis*, *Genre*, *Actors*, *Directors* and *tags* that already exist for the movie.

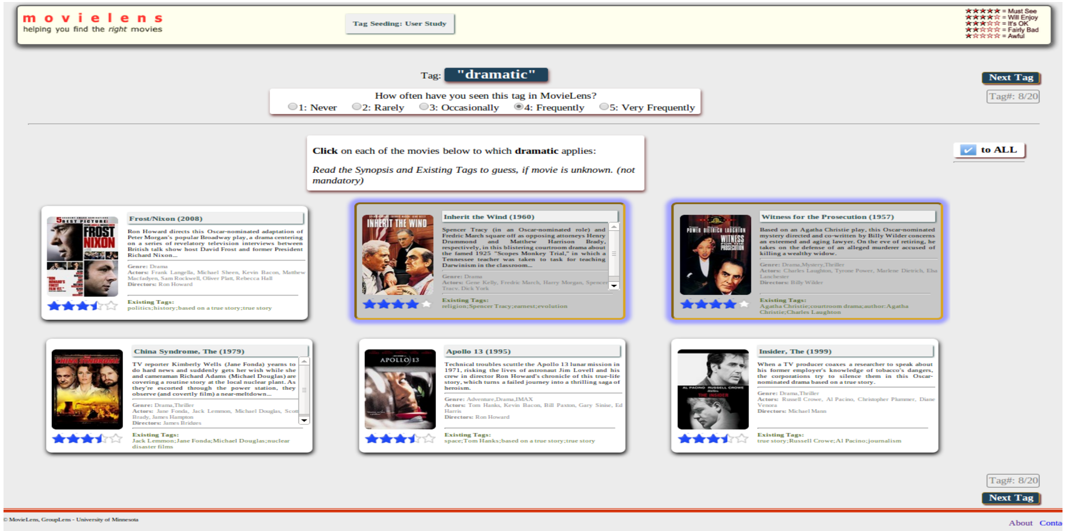
The set of applicable movies for a tag is determined based on the *relevance value* of tag to movies defined by *tag genome* (see section *Datasets*). Movies with higher relevance value and not yet applied are selected. We reorder this set by popularity of movie (in system) to determine the top six movies for a tag. There exist two rationales for reordering movies by popularity instead of the relevance value: (a) Applying tags to popular movies will help *seeded tags* to gain more exposure compared to applications to a relatively unknown or rarely navigable movie, (b) In pre-experiment trials with small group of users we found that it was hard to apply tags to movies thought to be relevant but *rarely known* in general.

Recycling Movie Set: We keep movies recycling to avoid over-application of tag and movie pairs from multiple users in same condition. Movies are replaced if the pairs together get more than 9 distinct user tag applications. This number is determined on the median of applications for movies in MovieLens.

Participants: The participants for the study are chosen (*without replacement*) from a sample of users who are active since 2009. We choose a sample of 1611 such users and randomly assigned each user to a condition. Users were communicated via email to participate in the study.

Survey: Users at the end of experiment were requested to participate in a survey with 17 Likert-Scale based questions. These questions were designed to determine the *experience of users with MovieLens and their expertise with tags and movies.* It was critical to make sure that the tag applications for each condition are unbiased by the set of assigned users expertise and experience; to facilitate an unbiased comparison of algorithms. We use *Kruskal-Wallace Rank Sum test* to determine if the users responses are different in each condition. In our analysis of survey responses,

[3] We find that normalized mutual information works well in measuring the similarity of these tags than other similarity measure such as cosine or Jaccard



Tag: "dramatic"

How often have you seen this tag in MovieLens?
○1: Never ○2: Rarely ○3: Occasionally ●4: Frequently ○5: Very Frequently

Click on each of the movies below to which **dramatic** applies:

Read the Synopsis and Existing Tags to guess, if movie is unknown. (not mandatory)

Frost/Nixon (2008)
Ron Howard directs this Oscar-nominated adaptation of Peter Morgan's popular Broadway play, a drama centering on a series of revelatory television interviews between British talk show host David Frost and former President Richard Nixon...
Genre: Drama
Actors: Frank Langella, Michael Sheen, Kevin Bacon, Matthew Macfadyen, Sam Rockwell, Oliver Platt, Rebecca Hall
Directors: Ron Howard
Existing Tags: politics;history;based on a true story;true story

Inherit the Wind (1960)
Spencer Tracy (in an Oscar-nominated role) and Fredric March square off as opposing attorneys Henry Drummond and Matthew Harrison Brady, respectively, in this blistering courtroom drama about the famed 1925 "Scopes Monkey Trial," in which a Tennessee teacher was taken to task for teaching Darwinism in the classroom...
Genre: Drama
Actors: Gene Kelly, Fredric March, Harry Morgan, Spencer Tracy, Dick York
Existing Tags: religion;Spencer Tracy;earnest;evolution

Witness for the Prosecution (1957)
Based on an Agatha Christie play, this Oscar-nominated mystery directed and co-written by Billy Wilder concerns an esteemed and aging lawyer. On the eve of retiring, he takes on the defense of an alleged murderer accused of killing a wealthy widow.
Genre: Drama,Mystery,Thriller
Actors: Charles Laughton, Tyrone Power, Marlene Dietrich, Elsa Lanchester
Directors: Billy Wilder
Existing Tags: Agatha Christie;courtroom drama;author:Agatha Christie;Charles Laughton

China Syndrome, The (1979)
TV reporter Kimberly Wells (Jane Fonda) yearns to do hard news and suddenly gets her wish while she and cameraman Richard Adams (Michael Douglas) are covering a routine story at the local nuclear plant. As they're escorted through the power station, they observe (and covertly film) a near-meltdown...
Genre: Drama,Thriller
Actors: Jane Fonda, Jack Lemmon, Michael Douglas, Scott Brady, James Hampton
Directors: James Bridges
Existing Tags: Jack Lemmon;Jane Fonda;Michael Douglas;nuclear disaster films

Apollo 13 (1995)
Technical troubles scuttle the Apollo 13 lunar mission in 1971, risking the lives of astronaut Jim Lovell and his crew in director Ron Howard's chronicle of this true-life story, which turns a failed journey into a thrilling saga of heroism.
Genre: Adventure,Drama,IMAX
Actors: Tom Hanks, Kevin Bacon, Bill Paxton, Gary Sinise, Ed Harris
Directors: Ron Howard
Existing Tags: space;Tom Hanks;based on a true story;true story

Insider, The (1999)
When a TV producer coaxes a researcher to speak about his former employer's knowledge of tobacco's dangers, the corporations try to silence them in this Oscar-nominated drama based on a true story.
Genre: Drama,Thriller
Actors: Russell Crowe, Al Pacino, Christopher Plummer, Diane Venora
Directors: Michael Mann
Existing Tags: true story;Russell Crowe;Al Pacino;journalism

© MovieLens, GroupLens - University of Minnesota

Figure 6: Experimental Interface

we did not found any differences in the users set. Due to the space limitation and scope of paper we plan to not discuss the survey any further in *Results* section.

Experiment Timeline: The experiment was conducted for 15 days starting *May 1st, 2013* until *May 15th, 2013.* A total of 1611 email invitations were sent with 320 users per condition. In order to complete the experiment, users were required to provide their inputs for all 20 tags shown.

6 Results & Discussions

We now discuss the results of our experiment and their implications, in the light of our research questions.

Experiment Response Summary: Out of 1611 emails sent, a total of 279 users (17.4%) responded in which 191 users (12%) completed the experiment and 88 users did not. The number of users per condition and mean applications per user in each condition respectively is shown in *Table 2.* A total of 10,899 new tag applications are generated from set of seeded 321 tags; applying to 1392 distinct movies during the experiment. Based on these responses we evaluate and discuss our first research question:

Table 2: Number of users and applications per user for each condition.

Algorithm	number of users	Applications per user
BaseLine	42	65.26
Opportunity Gap	44	48.32
Potential	34	60.35
Hybrid	31	48.35
Personalized	40	61.18

***RQ1**:Does tag seeding improve the navigational and search efficiency of tagging systems?*

In our previous section, we mentioned how Mutual Information of tags and movies can determine the search and navigational efficiency of tags. But we noted in *Figure 4* that this value has been consistently declining over time. Here we seek to observe if the tags carefully chosen if applied to the current system helps in reversing the lost information from the system. To understand this we calculate the mutual information of the system if in case the new tag applications were added to the system. In *Table 3* the amount of mutual information that system *gains* is shown. In contrast to an existing trend of decreasing mutual information, the tag seeding experiment suggests a gain in mutual information; thus able to improve the *discriminating power* of tags for movies.

Among all algorithms we record *Opportunity Gap* leading with the best gain of 0.069 bits, followed closely by *Baseline* algorithms with 0.064 bits. We suspect *Baseline* to be relatively better than others based on the highest number of applications it has achieved (65.26 application per user). The tags chosen for *Baseline* on an average are relatively well known to user compared to tags chosen by other algorithms resulting in easy application for users. The *Hybrid* and *Potential* algorithms had smaller, but still *substantial effects* on the information gains of 0.042 and 0.038 bits respectively. The Personalized algorithm had the smallest gain at only 0.017 bits gained. The point to keep in mind here is that the number of tags for personalized algorithm relatively very large than other algorithms. This is due to assignment of unique personalized set of tags for each user. However, in our next analysis we will notice how information gain per user is the best for Personalized algorithm. A *key observation* in all these algorithms is that although the gain of $\frac{1}{15}$*th* of a bit may seem small, but looking at past we are able to add enough information to attain the efficiency level it had *eleven months* before the experiment.

***RQ2**:Which algorithm performs better at tag seeding?*

Unfortunately the statistics on information theory values is a poorly studied field, therefore we know of no way to directly measure if the differences we observed between al-

Table 1: Top twenty tags by algorithm

Baseline	Opportunity Gap	Potential	Hybrid	Personalized
black and white	bollywood	original	original	science fiction
christianity	jesus	mentor	melancholic	super hero
dark	treasure	comedy	noir thriller	superheroes
ensemble cast	noir thriller	storytelling	mentor	wizards
father-son relationship	stranded	criterion	masterpiece	suspenseful
franchise	mars	loneliness	supernatural	noir thriller
high school	dark hero	oscar (best directing)	excellent script	dancing
inspirational	cheerleading	dramatic	stylish	fight scenes
mental illness	arnold	action	splatter	brutality
mystery	screwball comedy	relationships	women	lone hero
nazis	christian	weird	musicians	modern fantasy
nudity (topless - notable)	neo-noir	visually appealing	teen movie	fantasy world
oscar (best cinematography)	entirely dialogue	great acting	suspenseful	comedy
pixar	conspiracy theory	cult classic	brutality	teen movie
police	dragon	horror	neo-noir	action
politics	dr. seuss	family	dancing	excellent script
stupid	mountain climbing	great ending	great ending	fairy tales
tense	short-term memory loss	talky	unlikely friendships	supernatural
thriller	spielberg	story	detective	women
violent	marx brothers	cerebral	screwball comedy	dark hero

Table 3: Measured efficiency and efficiency gain for each condition. Existing system efficiency = 5.607

Algorithm	**Efficiency**	**Efficiency Gain**
BaseLine	5.671	0.064
Opportunity Gap	5.676	0.069
Potential	5.645	0.038
Hybrid	5.649	0.042
Personalized	5.624	0.017

gorithms are statistically significant. A direct comparison of gains may not be suitable since each condition (or algorithm) in the study have *varying number of user responses* and number of tag applications. We will try, however to gain an understanding of these differences by considering the per user information gain.

Per user efficiency measures the gain in information due to the applications *per user*. This is at best a crude estimate of the total effect on system by single users' tag applications. This will allow us to measure the information gain independent of the number of users responded in an algorithm during the study.

The *mean* and *median* of information gains per user for each algorithm is shown in *Table 4* with Box and Whisker charts in *Figure 7*.

We realize that the measure of action of each user independently shows few of the algorithms (*Opportunity Gap, Hybrid, Personalized*) having a *positive gain* whereas *Baseline* and *Potential* causing information loss (or negative gain). This should *not* be taken to mean that the *Baseline* and *Potential* algorithms are damaging the system, rather it means that information gain by these algorithms is dependent on the action of *many users working together*. The *key point* is that since tagging systems are built with mass action of many users we do not consider relying on multiple users to be a negative result. However, in this case of per user, it is easier to use statistical non parametric methods to compare algorithms independently. We find that the non parametric *pairwise Wilcoxon* test suggests that each pairwise difference between algorithms is indeed *significant*

Table 4: Efficiency bit gain per user

Algorithm	**Mean Per User**	**Median Per User**
	Efficiency Gain	**Efficiency Gain**
Baseline	-0.000240	-0.000356
Opportunity Gap	0.000737	0.000617
Potential	-0.000456	-0.000519
Hybrid	0.000382	0.000363
Personalized	0.000155	0.0000968

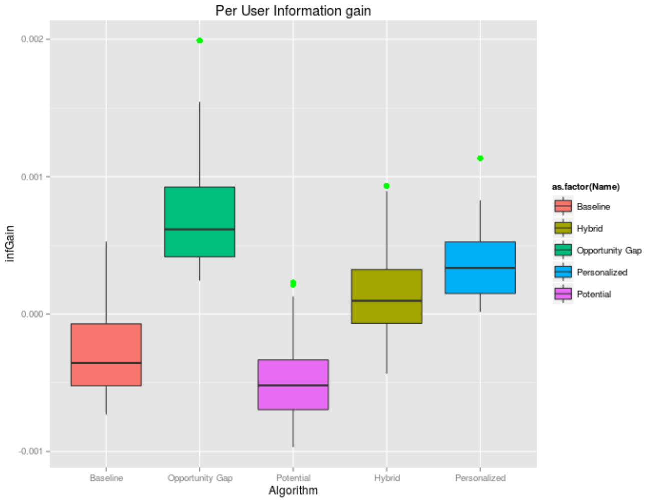


Figure 7: Box and whisker plot for the per user information gains

($p << 0.05$).

Implications: The *findings* for both research questions show that *tag seeding* is able to *improve* the quality of tagging systems. The crucial point is that *"within three weeks of the experiment we were able to counteract eleven months of the decline in the efficiency in MovieLens"*. Though it is hard to answer *RQ*2 accurately given the measures, we can say that the *Opportunity Gap* appears to be the *best* at selecting *candidate seeding tags* accompanied with best per user efficiency gain. While the *Potential* and *Baseline* algorithms did not perform as well, it is important to note that both these algorithms have a positive effect on total gain. Unlike *Opportunity Gap* these methods do not require analysis of system logs thus simplifying the metric for systems where logs may not be easily accessible.

7 Post Experiment System Analysis

Although seeded tags are able to show improvement in efficiency of the system, it is important to analyze if these tags being under applied or unknown to users affect their tagging behavior when exposed with more tag applications. We seek to observe if there is any significant change in user interactions to seeded tags compared to other already existing *un*-seeded tags- i.e. our third and final research question: ***RQ3***:*Does seeding tags lead to lasting changes in their visibility and usage?*

To answer this question, we analyze the system for respective user interactions. We primarily focus on four important interactions with the tag i.e. *Tag Clicks, Tag Search, Tag Applications and Tag Views.* Out of these interactions *Tag Clicks* and *Tag Search* reflect users' preference or interest in the tags. Whereas, more applications capture tags popularity and relevance with other similar items as seen by user. In the same way, *Views* which is number of times a tag is shown in system can tell us about visibility of the tag in system. With these user interactions we investigate using a simple approach to answer the research question.

Approach: The objective is to measure if the seeded tags are able to capture user's attention post the experiment. This is important for the tags which were once hidden and obscure to realize their potential among other popular tags. To achieve this we are required to improve the exposure of these tags and thus we added the 10,899 new tag applications from the experiment back to the system, increasing total number of applications for 321 seeded tags; hence increasing the visibility. After the tag applications are added, we monitor user interactions for a period of time. We specifically measure average number of each user interactions per day over that period. Note that we conducted our user controlled study between *May 1st, 2013* until *May 15th, 2013* and we added the applications on *July 15th, 2013* into the system. We started our post experiment interaction monitoring from *July 18th, 2013* until *September 23rd, 2013.* However to understand if the changes in user interactions for these tags are significant enough we analyzed the interaction parameters for same number of days before the experiment from *Jan 18th, 2013* until *March, 26th 2013.* Note that we did consider and analyzed logs from other multiple dates and number of days before and after the experiment but didn't found any significance change in the result.

With our parameters (user interactions) we compare the change in user activities before and after the experiment for seeded and unseeded tags. This helps us determine if the change in activities was significant compared to natural growth of tags in system over a period of time i.e. it may happen that tags in system achieve a certain bare minimum number of applications as time passes.

Regarding users, it is important to note that users who participated in the experiment may have had an influence over the tags shown during the experiment, and hence can impact the interaction parameters in after experiment analysis. We remove these users from the set of users to monitor and evaluate the results for the gain in percentage of user activities for seeded and unseeded tags. *Table 5* shows the gain in percentage change in user activities before and after the experiment.

We further recalculate the *Opportunity Gap* and *Potential* metric for the tags before and after the experiment. We expect both of these parameters to show decrease in values due to more applications and visibility of these tags.

Table 5: Percent change for *non-participating* users. (**Bold** values shows significant difference with $p << 0.05$)

Parameters	**Seeded Tags**	**Unseeded Tags**
Tag Clicks	15.61%	14.85%
Tag Search	14.61%	12.69%
Tag Applications	36.21%	31.82%
Tag Visibility	**73.42%**	**35.46%**

Results and Implications: The proportion test (or chi-squared test) is used in order to find if the percent change for seeded and unseeded tags are significant (at level $\alpha = 0.05$). We observe that the tag views have grown significantly after the experiment which is not a surprise given that we added around $10,000$ applications thus increasing the visibility. Though we didn't found any significant increase in usage of seeded tags but a positive increase in percent with marginally higher than unseeded tags shows us that these tags have been able to sustain at least the same momentum as unseeded tags. We do not consider this as a negative result since the tags were not able to gain even this momentum before the experiment. In case of the two defined metric, we see that *Potential* does show a significant change in post experiment suggesting that the coverage of seeded tags for the ideal movies is better than before. *Opportunity Gap*...........

8 Conclusion & Future Scope

We are able to show that seeding the under-applied relevant tags is able to help *improve the navigational efficiency of a* *tagging system*. Based on the results, we see that every algorithm we investigated had a positive effect on the information value of the system. Though the gain seems to be very small, it is worth noting that the information has been constantly declining for many years. Just within three weeks of running this experiment, the *Opportunity Gap* collected enough tag applications to counteract *"eleven months"* of declined efficiency in MovieLens.

Our results do not let us state conclusively that one of our algorithms is better than the others. However, we can say that the *Opportunity Gap* algorithm appears to be the best as selecting seeding tags. We attempt to answer this question using per user gain but more future work would be required to understand the effectiveness per algorithm. Nonetheless, we have begun the process of understanding the space of different approaches for choosing seeding tags, and we hope future work will continue exploring different techniques.

While the *Potential* and *Baseline* algorithms did not perform as well compared to Opportunity Gap algorithm, it is important to note that both these algorithms had a positive effect on total value and neither of them require any knowledge of historic behavior and can be computed without search or view logs of the tagging system making them more suitable for dataset with no usage logs.

In the post experiment system analysis we are able to show that the tags given a minimal opportunity are able to grow among other popular tags and *sustain* their growth in the system. Increased visibility and applications for each user reflect that users have started using these tags which otherwise could have remain hidden thus under-applied in the system.

Our results are overall very encouraging for the future of tagging systems. We have shown that it is quite possible to reverse the loss of efficiency in a tagging system. Furthermore, we have given several algorithms that can be used to guide future work in *tag seeding* as a solution to efficiency loss.

As future work we hope to explore *implicit* tag seeding in a live system such as MovieLens. Implicit tag seeding can be as simple as posting questions like "Do you think any of the tags below applies to the movie?" during user navigation of movies. Another approach would be to include a feature such as "Tag of the week" and ask users to apply this tag in the system.

9 Acknowledgments

The authors thank Ting-Yu Wang for his contribution in developing the user interface. Authors would also like to thank Shilad Sen and the rest of GroupLens for their feedback on the ideas presented in this paper. We also thank the MovieLens community for their responses, feedback and suggestions. This paper is funded by National Science Foundation grant IIS 09-64695.

References

[1] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104:1461–1464, 2007.

[2] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, 2008.

[3] W.-T. Fu, T. Kannampallil, R. Kang, and J. He. Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(3):12, 2010.

[4] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[5] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of information science*, 32(2):198–208, 2006.

[6] A. Mathes. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication*, 47(10):1–13, 2004.

[7] I. Peters. Against folksonomies: Indexing blogs and podcasts for corporate knowledge management. In *ONLINE INFORMATION-INTERNATIONAL MEETING-*, page 93. LEARNED INFORMATION LTD, 2006.

[8] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190. ACM, 2006.

[9] C. Shirky. Ontology is overrated: Categories, links, and tags. *Clay Shirky's Writings About the Internet*, 2005.

[10] J. Vig, S. Sen, and J. Riedl. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 93–102. ACM, 2011.

[11] K. Weller and I. Peters. Seeding, weeding, fertilizing. different tag gardening activities for folksonomy maintenance and enrichment. In *Triple-I Conference, Proceedings of I-Semantics 2008, Graz, Austria*, pages 110–117, 2008.