Mining Tags Using Social Endorsement Networks

Theodoros Lappas
Dept of Computer Science
Univ. of California - Riverside
Riverside, CA 92521
tlappas@cs.ucr.edu

Kunal Punera
Yahoo! Research
701 1st Avenue
Sunnyvale, CA 94089
kpunera@yahoo-inc.com

Tamas Sarlos
Yahoo! Research
701 1st Avenue
Sunnyvale, CA 94089
stamas@yahoo-inc.com

ABSTRACT

Entities on social systems, such as users on Twitter, and images on Flickr, are at the core of many interesting applications: they can be ranked in search results, recommended to users, or used in contextual advertising. Such applications assume knowledge of an entity's nature and characteristic attributes. An effective way to encode such knowledge is in the form of *tags*. An untagged entity is practically inaccessible, since it is hard to retrieve or interact with. To address this, some platforms allow users to manually tag entities. However, while such tags can be informative, they can oftentimes be inadequate, trivial, ambiguous, or even plain false. Numerous automated tagging methods have been proposed to address these issues. However, most of them require pre-existing high-quality tags or descriptive texts for *every* entity that needs to be tagged. In our work, we propose a method based on *social endorsements* that is free from such constraints.

Virtually every major social networking platform allows users to endorse entities that they find appealing. Examples include "following" Twitter users or "favoriting" Flickr photos. These endorsements are abundant and directly capture the preferences of users. In this paper, we pose and solve the problem of using the underlying social endorsement network to extract useful tags for entities in a social system. Our work leverages techniques from topic modeling to capture the interests of users and then uses them to extract relevant and descriptive tags for the entities they endorse. We perform an extensive evaluation of our proposed approach on real large-scale datasets from both Twitter and Flickr, and show that it significantly outperforms meaningful and competitive baselines.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering, Search Process

General Terms

Algorithms, Experimentation

Keywords

Tag extraction, Endorsement networks, Multimodal data mining

SIGIR'11, July 24–28, 2011, Beijing, China.

1. INTRODUCTION

The "Web 2.0" incarnation of the World Wide Web [26], with its emphasis on user generated content (UGC) and online social interactions, has transformed the way information is accessed by users. The focus of users' attention is now not on webpages managed by a central service provider, but on *entities* that are often uploaded and maintained by other users on social systems like Twitter, Facebook, Flickr, and YouTube. These entities can represent anything that is of interest to users: products, brands, celebrities, pictures, videos, blog posts etc. The lack of a centralized organizing or cataloging authority has both positive implications (e.g. an amazing explosion in content creation by web users) and potentially negative ones as well (e.g. challenges in discovering and recommending relevant content for users.) *Social Tagging* has emerged as an effective way to alleviate some of these challenges. Social Tagging refers to the practice of publicly labeling or categorizing entities in a shared, online environment [31]. The idea is simple: select meaningful tags (keywords) to encode the characteristic attributes of an item, thus making it easily accessible to relevant applications. As cataloged in studies on Flickr [23] and del.icio.us [14], tags can sometimes be a means of information discovery by themselves. However, they typically serve as valuable meta-data that can be used to aid information access mechanisms like recommendation services [7, 16], image search [10], enterprise search [12], and web search [3, 17].

USER-GENERATED TAGS. Many social platforms allow users to manually tag entities. Even though these user-generated tags can sometimes be informative, they also present numerous shortcomings. First, as noted by Mathes [24], user-generated tags suffer from problems stemming from an uncontrolled vocabulary, leading to ambiguity and lack of canonicalization. In addition, tags can often be trivial or inadequate, failing to capture all the attributes of entities. Finally, tags can sometimes be false or misleading, submitted either purposefully (e.g. to manipulate the accessibility and reputation of an entity) or simply due to human error.

In addition to the above issues, user-generated tags face the issue of *availability*. Selecting a compact set of high-quality tags for an entity can be a non-trivial and time-consuming task, which requires adequate knowledge of an entity's nature and attributes. This makes tagging less attractive to users, leading to a shortage or even absence of tags for many of the entities. Studies on the "long-tail" of information needs, such as those by Anderson [1] and Goel et al. [13], assert that users would benefit from a system where a large fraction of entities have been assigned high quality tags. In order to address this, some social platforms have been designed explicitly around the Social Tagging paradigm. Examples include the ESP game of Von Ahn et al. [32] and the study by Heymann et al. [17] on del.icio.us. These efforts have demonstrated that useful tags can be obtained for a large fraction of entities. However, in social platforms that do

not offer a tagging feature (e.g. Twitter) or when tags are not the central focus (e.g. Flickr), the fraction of entities with useful tags is minimal [23]. Hence, there is a great need of automated support for entity-tagging on such social systems; this is the focus of our work.

AUTOMATED TAGGING METHODS. Automatically extracting tags from text "associated" with entities has been the focus of multiple studies [4, 8, 18, 19, 30]. Past works on social tagging have mostly experimented with the del.icio.us system, where entities (documents) can be robustly represented by their own textual content. However, in scenarios where there is little or no text associated with an entity, such approaches are inapplicable. Consider, for example, a photo on Flickr with no comments or other descriptive text attached to it. One of the primary contributions of our approach is its ability to assign meaningful tags, even to such "cryptic" entities.

Another family of approaches assumes that the social system already has some tags associated with most entities, albeit noisy in nature [7, 18, 19, 30]. These approaches then fall into one of two categories: 1) recommend tags by studying similarities and correlations among entities, and 2) heuristically refine an entity's noisy set of tags. An example of the latter is the work by Krestel et al. [19], who use topic models for tag refinement. Other approaches rely on the historical tagging behavior of users to recommend tags for new documents [4, 8]. All these approaches assume the presence of tags, as well as robust features that allow the comparison and grouping of entities based on notions of similarity. As a result, they have been shown to perform well on social systems like del.icio.us that are built around social tagging of feature-rich entities, i.e. documents. However, we do not expect these approaches to perform well when applied to social systems like Twitter or Flickr that are the focus of our work and include entities with complex non-textual representations and sparse or non-existent tag-sets. In fact, we empirically demonstrate this by comparing our approach to the model of Krestel et al. [19] in the experiments section.

MINING SOCIAL ENDORSEMENT NETWORKS. The relative scarcity of pre-existing tags (and other text content) and their ambiguous nature motivate us to look for an alternative source of information. In this paper we demonstrate that *Social Endorsement Networks* [22] can be used to extract high-quality tags from the noisy and inadequate text associated with entities in social systems like Twitter and Flickr.

Most social systems allow users the option to endorse entities that they find interesting and/or appealing. An endorsement can be expressed in different ways. On Twitter, users have the option to *follow* other users and get updates on their tweets. On Facebook, users can *like* various types of items such as pictures or videos. On Flickr users can *favorite* pictures that they find appealing. In contrast to the often ambiguous and trying task of tagging, endorsements are an intuitive and concise method of expression. *We do not need users to describe an entity or tell us why they like it; we simply need them to tell us that they find it appealing via a single click.* Hence, chiefly due to their inherent simplicity and central role in user activity, endorsements are abundant in social systems. In this work our goal is to develop methods to use these pervasive social endorsements networks to extract meaningful tags from text associated with entities. We still have to deal with the ambiguity associated with the intent behind each endorsement. However, the general abundance of endorsements allows us to leverages techniques from statistical topic modeling to capture user interests and use them to extract meaningful tags for the entities they endorse. At a very high level, the intuition behind our approach is that entities that are frequently co-endorsed by users with the same interests are more likely to have the same or similar tags.

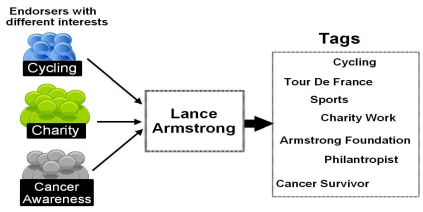


Figure 1: Twitter user Lance Armstrong attracts endorsements due to his various appealing aspects: Pro Cyclist, Philanthropist, Cancer Survivor. Our approach mines these aspects and translates them into meaningful tags.

OUR APPROACH. Our approach works in two phases. The first phase figures out why users choose to endorse the entities that they do. Intuitively, each entity can be mapped to a one or more topics. These topics capture the different aspects that make it appealing to different users. Consider a photograph of a beautiful *landscape* from *Hawaii*. This photo can attract endorsements from nature-lovers who appreciate the beauty of the scenery, or from people who have a particular interest in the island of Hawaii. Another example is given in Figure 1, showing the different aspects of Twitter-user and world-famous cyclist Lance Armstrong.

We uncover the observed endorsements using a generative-model approach based on Latent Dirichlet allocation (LDA) [5]. This allows us to reconstruct the "aspect-space" that the entities reside in. Clearly, not *all* endorsements are due to such coherent topical reasons, and our approach does not require them to be. A user may endorse an entity even if it does not really appeal to her interests, simply because one of her friends did. Such endorsements cannot be translated into meaningful tags; even if we could somehow assign a tag that captures this scenario (e.g. one that encodes the relationship between the user and her friend), it would be of little use to other users or to information retrieval and data mining applications. Thus, while our approach does not explicitly distinguish among endorsements, it is driven by those generated due to aspects of general interest.

The product of the first phase is the set of learned aspects, along with the correlation of each entity with each aspect. In the second phase, these aspects are fleshed out using the text associated with the entities. In practice, this phase translates each aspect into a probability distribution over a vocabulary of terms. Note that we do not require the presence of text for every entity in the network. Instead, we only need enough text to allow us to converge to the distribution of each aspect. We can then use the correlation of an entity with each aspect (learned in the first phase) to extract tags. In the experiments section, we show how one can obtain text for entities in social platforms such as Twitter and Flickr. We also demonstrate how our approach is able to extract high quality tags even when noisy or no text is given as input for some of the entities.

CONTRIBUTIONS.

1. We propose the use of social endorsement networks to extract high quality tags for entities embedded in social systems. To the best of our knowledge, this is the first such application of endorsement networks.
2. We propose an approach that has no requirement for pre-existing tags and can produce high-quality results even when the input consists of *noisy* text that is available for *only a portion* of the entities.
3. We conduct extensive experiments on large scale datasets from two very different social systems -Twitter and Flickr- to demonstrate the efficacy of our approach. Our results indicate that our approach is highly effective at finding high quality tags: in certain cases performing 100% better than the closest baseline.

ORGANIZATION.

The rest of the paper is organized as follows. In Section 2 we motivate the use of social endorsement networks and present our approach. Next, we experimentally demonstrate the merits of our method over a broad set of strong baselines and an extensive evaluation in Section 3. In order to distinguish our work from related existing works, we delay the discussion of related work until Section 4. Finally, we conclude in Section 5.

2. MINING TAGS FROM ENDORSEMENTS

In this section we begin with a formal discussion and motivation of the problem considered in this paper. In Section 2.2, we explore possible approaches and discuss their shortcomings. Finally we present our proposed approach in Section 2.3.

2.1 Motivation

The goal of this work is to devise a method of obtaining high-quality tags for entities embedded in social systems, in order to enhance their accessibility and make them visible to applications such us search and recommendation systems. The scope of the problem is best explained through some examples.

Flickr: On Flickr, the entities to be tagged are pictures. Users have the option to tag pictures with keywords of their own selection. While these manually assigned tags are sometimes informative, they can often be trivial or inadequate. In addition, some pictures are tagged with too few or too many tags. Our study of the pictures uploaded in the group FlickrCentral (the largest user-group on Flickr) showed that 53% of the pictures had either less than 2 or more than 30 tags.

Twitter: The entities of interest on the Twitter platform are the users themselves. Our primary focus is on users with a significant number of followers, since they are more likely to be of interest to other users. Twitter is a particularly challenging scenario, since it does not allow users to tag each other.

One of the primary challenges of automated tag extraction is finding an appropriate source of informative text associated with the entities. This is easier in cases where the entities are documents themselves, as in the case of the del.icio.us platform. However, for entities embedded in social platforms, finding good text can be a non-trivial task. On Twitter, we can consider the tweets of each entity as a source of text. However, as we show in the experiments section, the ambiguous and often irrelevant nature of the tweets often leads to poor results. Another source of text that we explore is the set of results retrieved by issuing the user's name as a query to a major web search engine. However, here we have to deal with issues like the availability of a nickname instead of the user's actual name, ambiguity of results due to common names, and noisy or irrelevant web content. On Flickr, we can use the comments submitted by the users for each picture. Again, we are faced with the issue of dealing with irrelevant or trivial comments or, for some pictures, the complete lack thereof. For the rest of this section, we assume access to a corpus, albeit incomplete and noisy, that associates text with *some* of the entities. We further comment on the specifics of the text we use for each social system in the experiments section.

NOTATION. First, we establish some notation common to the rest of this section. Let $\mathcal{E}$ denote the set of entities we are interested in tagging, with d_e representing the "bag-of-terms" associated with each entity, with terms coming from vocabulary $\mathcal{V}$. Let $n_{t|e}$ denote the number of times term t occurs in d_e, $n_{\cdot|e}$ the total occurrences of terms in d_e, and $n_{t|\cdot}$ the total number of unique entities that have associated text containing t. We will introduce additional notation when needed later in the section.

2.2 Baselines for the Tag Extraction Problem

The earlier examples show that, in order to be effective, an approach has to be able to work with *noisy text* and also tackle cases when text is *unavailable* for some of the entities. To better motivate our approach, we begin by discussing the shortcomings of some intuitive baselines. We then go on to give details of our approach and describe how it overcomes these shortcomings.

2.2.1 *Baseline 1:* TFIDF

A simple method for obtaining tags from the text associated with entities is to treat such text as documents in a information retrieval setting and to use the popular TF-IDF method for scoring words [2]. In terms of our problem, given an entity e from a collection $\mathcal{E}$ and a term t, using the notation from Section 2.1, the measure is defined as follows:

$$TFIDF(t,e) = \frac{n_{t|e}}{n_{\cdot|e}} \log \frac{|E|}{n_{t|\cdot}} \quad (1)$$

While this method has been shown to be highly effective in various information retrieval tasks, it suffers from a number of drawbacks. First, as shown in our experiments, this method often results in extremely obvious or obscure tags. For example, in Twitter, for the entity *Al Gore*, the former vice president of USA, the top two tags obtained by TFIDF were "al gore" and "futurama". The first is trivially the entity's name. The second is a popular TV show that sometimes features Al Gore as a gimmick, hardly one of the entity's most characteristic attributes. Further, if there exist multiple reasons that make an entity appealing in a social system, then it is likely that TFIDF will identify redundant tags from the dominant aspect before identifying tags from any of the others. Finally, this approach depends heavily on the text available for each entity: uninformative text leads to irrelevant tags, and the utter lack of text renders the method inapplicable. All these issues emerged in our experiments, where TFIDF was used as a baseline. In order to address these problems, we clearly need a more sophisticated approach.

2.2.2 *Baseline 2:* TEXTASPECTS

A key insight from the above discussion is that an entity can appeal to users of a social system for multiple distinct reasons.

Examples: *Some Twitter users might follow Lance Armstrong's due to his status as a world-class cyclist, while others choose to follow him for his significant charity work. On Flickr, a picture of the Eiffel tower in Paris may be favorited by people who like Paris in particular, as well by those who care about architecture in general.*

We refer to these different reasons for an entity's appeal as *Aspects*. Using such aspects, we can identify non-redundant tags that are representative of the different facets of an entity and avoid some of the issues that ailed the TFIDF approach. The problem of tag extraction can now be broken down into two distinct steps.

1. Identify the aspects that make each entity appealing and compute a term-based representation of each aspect.

2. Extract high-quality tags, based on each entity's aspects of appeal and each aspect's term-based representation.

Next, we consider an intuitive solution to sub-problem 1 and discuss its shortcomings. Then, we give our own solution to this sub-problem in Section 2.3. A general solution to sub-problem 2 is given in Section 2.4.

A reasonable assumption is that (some of) the aspects of an entity's appeal are represented in the text associated with it.

Examples: *On Twitter, Lance Armstrong's tweets concern his cycling career, as well as his charity work. Hence, both of these central aspects of his appeal are captured in his tweets. Similarly, on Flickr, the comments associated with a picture of the Eiffel Tower may address admiration for Paris, as well as for impressive architecture feats.*

We refer to such aspects that are latent within text as TEXTASPECTS. One principled way to extract TEXTASPECTS is by modeling the text associated with entities using the Latent Dirichlet Allocation (LDA) approach [5]. We use this approach as a baseline.

The LDA generative model states that each word in a document is generated with the document first picking a topic, and then picking a word associated with the topic. The topics are picked from a document-specific distribution. In our *Text Generation Process*, the text associated with an entity is assumed to be generated by the above process, with the topics representing the aspects of an entity's appeal. Let $\mathcal{K}$ be the set of aspects, α and β be the Dirichlet smoothing parameters, Θ the set of aspect-entity distributions, and Φ the set of term-aspect distributions. In the generation process, t_i and z_i indicate a word in a d_e and the topic that generates it, respectively. Given this notation, the text generation process is as follows:

Text Generation Process

1. For all aspects k, sample $\phi_k \sim Dir(\beta)$
2. For all entities e, sample $\theta_e \sim Dir(\alpha)$
3. For each term-slot in d_e
   (a) Sample an aspect $z_i \sim Mult(\theta_e)$
   (b) Sample a term $t_i = w \sim Mult(\phi_{z_i})$

The parameter-sets Θ and Φ can be learned by determining an assignment of term-slots to aspects $\mathbf{z}$, following the collapsed Gibbs sampling method [35]. The core of the method relies on sampling a new value for the aspect z_i that generated the term w at term-slot t_i by using the aspect assignments of all other term-slots (i.e.$\mathbf{z}_{-\mathbf{i}}$). Formally,

$$P(z_i = k|t_i = w, \mathbf{t}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{i}}) \propto \frac{n_{k|e,-i} + \alpha}{|d_{e,-i}| + \alpha|\mathcal{K}|} \times \frac{n_{t|k,-i} + \beta}{n_{\cdot|k,-i} + \beta|\mathcal{V}|} \quad (2)$$

where $n_{k|e,-i}$ is the number of times aspect k is observed for entity e, $n_{t|k,-i}$ is the number of times term t is sampled from aspect k, $|d_{e,-i}|$ is number of term occurrences associated with e, and $n_{\cdot|k,-i}$ is the total number of terms generated from aspect k, with all these quantities computed over all slots except the i^{th} one.

The learned distributions $\Theta : \{\theta_{ek} = p(k|e), \forall e \in \mathcal{E}, k \in \mathcal{K}\}$ and $\Phi : \{\phi_{kt} = p(t|k), \forall k \in \mathcal{K}, t \in \mathcal{V}\}$ are called the **aspect-entity** and **term-aspect** distributions, respectively. θ_e gives us the probability that the appeal of a given entity e is due to aspect k and ϕ_k represents the probability of a term showing up given an aspect k. Given this knowledge, techniques such as the one we describe in Section 2.4 can be then used to extract tags.

While TEXTASPECTS solves some of the problems with redundancy and coverage of tags, it faces a different issue: the mined TEXTASPECTS are just a *proxy* to the true aspects of an entity's appeal, since they only capture the facets of an entity that are present in the associated text. Let us consider the following examples.

Twitter: While Lance Armstrong's tweets may cover the cycling and philanthropy aspects of his appeal to users, they might not capture his appeal to others who identify with and follow Armstrong because he is a famous American from Texas, or because he is a cancer survivor.

Flickr: On Flickr, a picture of Paris taken by a famous photographer may appeal to people who follow this particular photographer's work. It might be the case, however, that this aspect is not adequately represented in the comments associated with the photograph, and therefore will be erroneously overlooked.

The above examples demonstrate that TEXTASPECTS is still limited by the available text. If the available text is inadequate, some of the entity's appealing aspects will be missed. If the text is uninformative, the produced aspects will lead to inferior tags. Finally, if no text is available at all for an entity, then the approach is not applicable at all. The above analysis motivates us to find a way to disconnect the process of learning the appealing aspects of an entity (i.e., the aspect-entity distribution) from the often inadequate and noisy entity-text.

2.3 Our Approach: ENDORSEASPECTS

In the previous section we discussed the shortcomings of learning aspects directly from text, and motivated the need for an alternative rich and easily accessible data source. In particular, we highlighted the need for an alternative way to estimate the aspect-entity distribution $\theta_{ek}(= p(k|e))$, i.e. the probability that a given entity is appealing due to a specific aspect k. In this section we give an approach for learning this aspect-entity distribution by mining the underlying *social endorsement graph*. Social endorsements are ideal for our purpose since they directly encode the appeal of the entity to the users. In addition, endorsement functionality available in virtually every major social networking platform, ensuring an abundance of endorsement data for almost all entities. On Twitter, the endorsement graph consists of the underlying follower-graph. On Flickr, users endorse pictures by including them to their Favorites. In the context of such a graph, we compute the probability θ_{ek} as follows:

$$\theta_{ek} = p(k|e) = \frac{p(e,k)}{p(e)} \propto \sum_{u \in \mathcal{U}} p(u)p(k|u)p(e|k), \quad (3)$$

where $p(u)$ is the probability that the user u makes an endorsement. This probability can be assumed to be the same for all users in $\mathcal{U}$. In domains where there is reasonable evidence that some users have a stronger propensity to endorsements than others, it can be learned from a training corpus.

Conceptually, $p(k|u)$ is the probability that user u makes an endorsement due to aspect k (**aspect-user** distribution) and $p(e|k)$ is the probability that entity e gets endorsed due to (its association with) aspect k (**entity-aspect** distribution). Hence, our approach computes the required term-aspect distribution by first breaking down social endorsement network into the the aspect-user and entity-aspect distributions.

Examples: *A user who is interested in Cycling is likely to endorse entities based on this aspect. Moreover, most users are likely to have many interests, and will likely have a distribution over these interests ($p(k|u)$). Further, a famous Cyclist such as Lance Armstrong is more likely to be endorsed due to his status in the context of this aspect ($p(e|k)$).*

Since aspects are learned directly through endorsements, our approach gives the appealing aspects even for entities that have no usable text attached to them. Provided that we have enough text from other entities to learn the term distribution of each aspect, we can then still generate high-quality tags. Next, we show how we can learn the aspect-user and entity-aspect distributions by mining the endorsement network.

2.3.1 Learning Appealing Aspects from Endorsements

We learn the appealing aspects from the endorsement network by modeling it using an LDA-like generative process. Informally, the *E*ndorsement Generation Process states that each endorsement made by a user is generated with the user first picking an aspect

from his personal distribution of interests, and then picking an entity to endorse based on that aspect. Formally, let $\mathcal{U}$ denote the set of users who endorse entities picked from the set $\mathcal{E}$. For each user u, the set of entities endorsed by her is denoted by $\mathcal{E}(u) \subseteq \mathcal{E}$. Let the set $\mathcal{K}$ be the set of aspects, α' and β' be the Dirichlet smoothing parameters, $\Psi : \{\psi_{uk} = p(k|u), \forall u \in \mathcal{U}, k \in \mathcal{K}\}$ be the set of aspect-user distributions, and $\Xi : \{\xi_{ke} = p(e|k), \forall k \in \mathcal{K}, e \in \mathcal{E}\}$ be the set of entity-aspect distributions. In the generation process, e_i denotes an entity in the endorsed set $\mathcal{E}(u)$ of a user u. Then, y_i denotes the aspect e_i is sampled from. Given this notation, the endorsement generation process is as follows:

Endorsement Generation Process

1. For all aspects k, sample $\xi_k \sim Dir(\beta')$
2. For all entities u, sample $\psi_u \sim Dir(\alpha')$
3. For each endorsement-slot in $\mathcal{E}(u)$
   (a) Sample an aspect $y_i \sim Mult(\psi_u)$
   (b) Sample a entity $e_i = l \sim Mult(\xi_{y_i})$

As in the standard LDA model, the distribution-sets Ψ and Ξ can be learned via the collapsed Gibbs sampling method [35]. This requires an expression that gives the probability that one particular endorsement from a user u to entity $e_i = l$ was due to aspect y_i. This can be derived by following the steps outlined by Griffiths and Steyvers [15]. Due to paucity of space, we skip the derivation and give here the final sampling equation.

$$P(y_i = k|e_i = l, \mathbf{e_{-i}}, \mathbf{y_{-i}}) \propto \frac{n_{k|u,-i} + \alpha'}{|\mathcal{E}(u)| + \alpha'|\mathcal{K}|} \times \frac{n_{e|k,-i} + \beta'}{n_{\cdot|k,-i} + \beta'|\mathcal{E}|} \quad (4)$$

where $n_{k|u,-i}$ is the number of times aspect k is observed for user u, $n_{e|k,-i}$ is the number of times entity e is sampled from aspect k, $|\mathcal{E}(u)|$ is number of entities endorsed by user u, and $n_{\cdot|k,-i}$ is the total number of entities generated from aspect k, with all these quantities computed over all endorsement-slots except the i^{th} one.

The learned aspect-user distributions ψ_u and entity-aspect distributions ξ_k give us the probabilities $p(k|u)$ and $p(e|k)$ that we need to compute Eq. (3) and obtain θ_e, i.e. the distribution over the set of aspects for each entity. We can then plug θ_e into the *Text Generation Process* from Section 2.2.2 and learn the term distribution ϕ_k of each aspect k. The next section describes the details of this learning process. The corresponding plate notation is given in Figure 2. The Endorsement Generation Process follows standard LDA principles. For the Text Generation Process, the θ distribution is now known, which is why it is shown shaded in the plate notation.

2.3.2 *Efficiently Learning the Term-Aspect Distributions*

In the TEXTASPECTS approach discussed in Section 2.2.2, the aspect-entity distribution θ_e of each entity e and the term-aspect distribution ϕ_k of each aspect are learned simultaneously by sampling Eq. 2. In the case of ENDORSEASPECTS, however, θ_e is obtained via an independent mining of the endorsement graph and is thus already known. Therefore, we can re-write Eq. (2) as follows:

$$P(z_i = k|t_i = w, \mathbf{t_{-i}}, \mathbf{z_{-i}}) \propto p(k|e) \times \frac{n_{t|k,-i} + \beta}{n_{\cdot|k,-i} + \beta|\mathcal{V}|} \quad (5)$$

where $p(k|e) = \theta_{ek}$ is already known. Via Gibbs sampling, we can now sample aspects from this equation and infer ϕ_k for each aspect k accordingly.

In their paper on efficient topic model inference [35] Yao et al. introduce SparseLDA, an algorithm for the efficient evaluation of Gibbs sampling distributions. The method is based on the observation that parts of the computations are independent to each other and can thus be cached to reduce the computational cost. Given Eq. (5) we can apply the SparseLDA sampling process to our approach by noticing that the right-hand side of the equation can be re-written as the product of the two following factors:

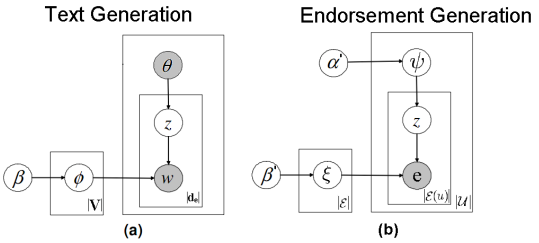


Figure 2: Text and Endorsement Generation in ENDORSEASPECTS.

$$r = \sum_k \frac{p(k|e)\beta}{n_{\cdot|k,-i} + \beta|\mathcal{V}|} \quad , \quad q = \sum_k \frac{p(k|e)n_{t|k,-i}}{n_{\cdot|k,-i} + \beta|\mathcal{V}|}$$

The sampled value $x \sim U(0, r+q)$ can now fall within two buckets, depending on whether $x < r$ or $r < x < q$. The numerator of r is now a constant and can be precomputed for each aspect-entity combination. The denominator can be cached for each aspect and updated by simply subtracting and adding the values of re-sampled aspects. For q, we can cache $\frac{p(k|e)n_{t|k}}{n_{\cdot|k}+\beta|\mathcal{V}|}$ for every aspect k, by remembering the entities for which $p(k|e) \neq 0$. A single multiplication for each aspect k with $n_{t|k} \neq 0$ is then enough to compute q for a given term t.

2.4 From Aspects to Tags

After the appealing aspects of each entity have been extracted and mapped to meaningful term distributions, the final stage consists of assigning useful tags to each entity. We accomplish this via use of the popular *Information Gain measure*. In our context, given a term t and an entity e, we define the measure as follows:

$$\begin{aligned} IG(t|e) = &-p(e)\log p(e) \\ &+p(t)[p(e|t)\log p(e|t) + p(\neg e|t)\log p(\neg e|t)] \\ &+p(\neg t)[p(e|\neg t)\log p(e|\neg t) + p(\neg e|\neg t)\log p(\neg e|\neg t)] \end{aligned} \quad (6)$$

This equation measures the reduction in the entropy associated with entity e, incurred by the presence or absence of term t. The entropy itself is the same for all terms, thus it can be omitted. We compute $p(t)$ as follows:

$$p(t) = \sum_k \sum_e p(e)p(k|e)p(t|k) \quad (7)$$

$p(e) = 1/|\mathcal{E}|$ is assumed to be the same for all entities, $p(k|e)$ is as computed by Eq. (3) and $p(t|k)$ is computed as part of the term distributions learned by the second stage of our approach.

$p(e|t)$ is computed as follows:

$$p(e|t) = p(e,t)/p(t) = \sum_k p(e)p(k|e)p(t|k)/p(t) \quad (8)$$

After computing the information gain scores, we can tag an entity with a fixed number of top-scoring terms, or simply report all terms with a score over a given threshold. Assigned tags need to be good discriminators, in order to be characteristic of the entity. At the same time, tags need to be frequently used terms that are familiar to the users. This is necessary for applications like search, where an entity needs to be matched to user queries based on its tags. We justify our choice of Information Gain by showing in our experiments that it reports terms that satisfy both of these requirements.

3. EXPERIMENTS

In this section, we illustrate the efficacy of our approach through a rigorous experimental evaluation on real data. We begin with a description of the datasets used in the experiments, as well as the baselines that we compare our approach to. We then discuss each experiment in detail.

3.1 Datasets

Each dataset we use in our evaluation consists of a set of users, a set of endorsable entities, the underlying endorsement graph and a piece of textual information associated with each entity.

Twitter: Twitter is a social networking and microblogging website that allows users to broadcast short messages (tweets) to other users that explicitly choose to "follow" them. This act of following a user is interpreted as an endorsement. For our experiments, we used the publicly available crawl of the Twitter graph published by Kwak et al. [21][1]. The crawl consists of 41.7 million user profiles and 1.47 billion social relations. In our experiments, we consider every user with at least 2,000 followers to be an endorsable entity, giving us a total of about 38,400 entities. In order to obtain the text associated with each entity, we harvested all their tweets from July-December of 2009, the same year the graph was crawled.

An inspection of the tweets revealed that the content was often not only sparse, but also very noisy. For example, while an actor may sometimes tweet about his work and projects, the majority of his posts are not related to movies or acting. Hence, we create a second dataset of text associated with entities that was more conducive for the purposes of tagging. We start by making the following useful observation about twitter users: highly-followed entities are often public figures or authorities in some topic of broad interest. Such users use their real name when registering, since they use Twitter as a medium to advertise themselves. First, we extract the real full name of each user. We then submit the name as a query to a major search engine and aggregate the text of the top-5 returned webpages. As we show in our experiments, the data from the search engine turns out to be more useful than the tweets for extracting tags.

Flickr: Flickr is an image hosting website, where users can endorse an entity (image) by including it in their 'Favorites'. Flickr users also have the option to create and join groups, where people can post images and comments. For our experiments, we used the data for all users and images that are affiliated with FlickrCentral, the largest group on Flickr. FlickrCentral consists of users with diverse interests; we collected data for the 370,144 members of the group. After removing images with less than 10 endorsements, we were left with a set of 138,225 distinct images. For each image, we collected the comments attached to it, and, where available, the set of user-assigned tags and descriptive note attached by the image's owner. This comprised the text data associated with each entity. The data was extracted from a complete snapshot of Flickr from 2008.

3.2 Evaluated Approaches

ENDORSEASPECTS: This refers to our own approach of mining aspects from the social endorsement network as described in Section 2.3. The mined aspects are then used to extract tags using the approach outlined in Section 2.4. For Twitter, the number of aspects was set to 400; this was determined via experiments on a validation set of tags of wefollow.com, described in Section 3.3.2. For Flickr, the number of aspects was set to 1000, tuned on a validation subset of images. We implemented our approach by modifying Mallet [2], which also handles hyper-parameter optimizations.

TEXTASPECTS: This baseline refers to the approach of mining aspects from the text associated with each entity (see Section 2.2.2). This can be considered an implementation of an approach proposed by Krestel et al. [19]. As above, the approach in Section 2.4 is used to mine the tags. The number of aspects used were set and hyper-parameter optimization was performed in the same way as for ENDORSEASPECTS.

NN-TEXTASPECTS: For this baseline, the text associated with an entity is enriched with the text associated with its 5 nearest neighbors, selected based on the Jaccard similarity of their endorsement sets. Conceptually, this enriches the text of each entity by taking into consideration the (local) information encoded in the endorsement graph. The TEXTASPECTS approach is then applied on this enriched text representation.

TFIDF: This baseline for mining tags is described in Section 2.2.1.

3.3 Evaluation on Twitter

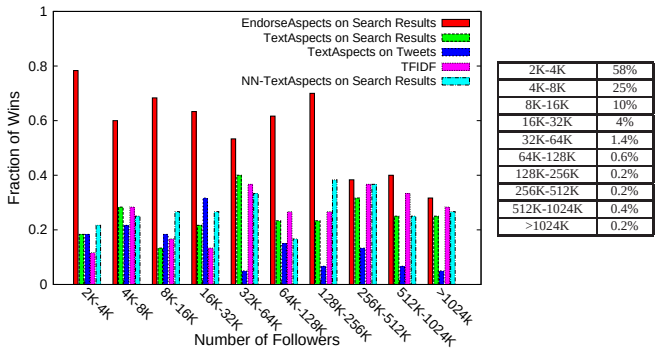
3.3.1 Twitter Evaluation via Manual Labeling

In this section we describe our evaluation of the quality of the tag-sets reported by the different approaches for the Twitter dataset. The study was conducted as follows: first, we separated the entities into 10 groups based on their number of followers. The first group includes all entities with a number of followers in $[2000, 4000)$, the second group those in $[4000, 8000)$, and so on. The purpose of this grouping was to evaluate how the approaches perform for entities of different popularity, as indicated by the number of their followers. We then randomly selected 30 entities from each range, which were then shown to two annotators, along with the tag-sets recommended by each approach. The annotators were asked to pick the approach with the best tag-set; they could also pick multiple winners or no winners at all. For each range, we report the average fraction of wins for each approach, taken over the judgments of both annotators. We quantify the inter-rater agreement by reporting the average Kappa statistic of agreement between each pair of annotators on the wins for each approach; this value was 0.455. To see that this signifies a robust agreement between annotators note that ENDORSEASPECTS outperforms baseline approaches by a wide margin, and hence large values of the Kappa statistic are difficult to obtain as marginal distributions of annotator preferences over approaches are very biased. Another interesting way to visualize results is to group entities by a measure of their activity (number of tweets); we report results this way taking care to avoid biases when estimating the performance of approaches in various ranges.

The fraction of wins of each approach over different groups of users are shown in Figures 3 and 4. As can be seen, our approach (ENDORSEASPECTS) consistently outperforms all others in all specified ranges. For all entities with up to 256K followers, ENDORSEASPECTS is significantly better than the competition. In fact, for some ranges (e.g. 2K-4K and 4K-8K), the achieved score is more than double that of the second best. This gap is reduced for the last three ranges of followers. It is important to observe that this is due to the scores of ENDORSEASPECTS dropping, since the scores of the other methods remain in the same levels as in the other ranges. Considering the nature of ENDORSEASPECTS and its dependence on the underlying endorsement graph, this reduction can be easily explained: users with over 256K followers are very famous public figures, attracting interest from people of different interests. Britney Spears and Ellen Degeneres are both examples of such individuals. Their initial source of fame may have been their achievements in a

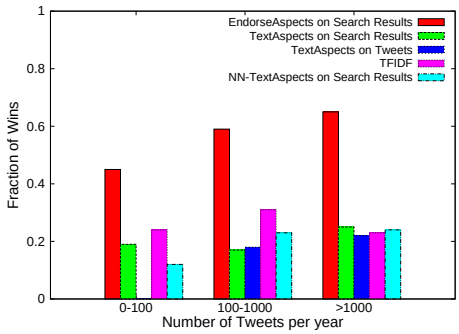
[1] http://an.kaist.ac.kr/traces/WWW2010.html

[2] `http://mallet.cs.umass.edu/`



2K-4K	58%
4K-8K	25%
8K-16K	10%
16K-32K	4%
32K-64K	1.4%
64K-128K	0.6%
128K-256K	0.2%
256K-512K	0.2%
512K-1024K	0.4%
>1024K	0.2%

Figure 3: Fraction of wins for each approach with entities grouped by number of followers and the percentage of entities in each group.



0-100	82.2%
100-1000	14.5%
> 1000	3.3%

Figure 4: Fraction of wins for each approach with entities grouped by activity level and the percentage of entities in each group.

particular area (music and acting, respectively). However, their popularity has grown to a point that they attract followers who may not even have an interest in these areas. This throws off ENDORSEASPECTS, since it cannot easily determine the aspects that attract the entity's endorsements. It is important to note that such entities make up for only small portion of the endorsable population. Moreover, these entities often have name recognition, which reduces the necessity of finding tags. As can be seen in Figure 3, 99.2% percent of the entities have less than 256K followers, and these are the entities that our approach excels at. When grouping entities by activity level as well ENDORSEASPECTS outperforms all competing approaches over all ranges by wide margins (Figure 4).

Overall, our results show three points worth noting. First, it is very clear from the preceding results that using endorsement network information lets us obtain much better tags that simply using text associated with entities (even when applying sophisticated approaches like TEXTASPECTS). Second, we can answer whether it is endorsement information alone that helps or the way ENDORSEASPECTS exploits it. As explained before NN-TEXTASPECTS was designed to enrich the text representation of entities using local information from the endorsement graph. However, we notice that even though NN-TEXTASPECTS improves upon TEXTASPECTS it does come close to ENDORSEASPECTS in terms of performance. This indicates that using text in any form at the first stage of constructing the aspects tends to lead the system astray (in spite of using neighborhood information) and ENDORSEASPECTS is able to exploit the endorsement network information to obtain much cleaner signal. A third point worth noting is that TEXTASPECTS performs much better when run on Search Results as opposed to when run on Tweets, showing that the reasons why users follow entities and hence appropriate tags for them are not very correlated to the tweets produces by these entities.

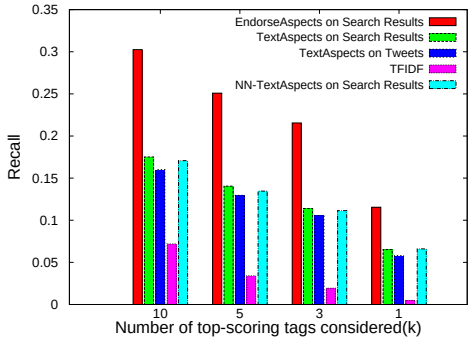


Figure 5: Recall on the wefollow tag-prediction task.

Tagging Entities with Few Followers and No Associated Text: The results so far are reported on entities with at least 2000 followers as we believe that these are of the most interest to Twitter users. However, for some interesting entities - such as new ones - their number of followers may be less than our threshold or there may not be available any text associated with them (they may not tweet much or may resolve to ambiguous search results). Here we report experimental results on such entities.

We conducted this experiment on entities with at least 1000 followers, leading to 146000 entities. Of these, web text was obtained for only those with more than 2000 followers. The goal in this experiment was to then obtain tags for those entities with number of followers between 1000 and 2000 for whom we had no associated text. ENDORSEASPECTS does not require associated text for *every* entity in the corpus, however, the all baselines evaluated in the previous study do, and hence, are no longer applicable. Therefore, we modify the NN-TEXTASPECTS approach to use the 5 nearest neighbors *with* associated text. For the manual evaluation, we selected 50 entities from the 1000-2000 followers range, uniformly at random. For each entity, the winning approaches were selected by three annotators. The average fraction of wins for ENDORSEASPECTS and modified NN-TEXTASPECTS were **78.6%** and **14%**, respectively. The value of Kappa statistic averaged over annotators pairs and approaches was 0.52. These results show that ENDORSEASPECTS works extremely well for entities that have few followers and no text associated with them. These experiments also show that the impressive performance of ENDORSEASPECTS reported in Figures 3 and 4 is not due to any bias in selecting entities by restricting the number of followers to be at least 2000.

3.3.2 *Evaluation with WeFollow Tags*

In this experiment we evaluate each approach on the task of predicting tags that Twitter users have assigned to themselves. Even though Twitter does not provide a tagging feature, users can tag themselves by visiting wefollow.com. We found 16,000 entities from our experimental setup that had tagged themselves with at least one tag. An inspection of the data revealed that user-assigned tags can be noisy or irrelevant: single-letter tags, tags with non-letter characters, humorous or inaccurate self descriptions (e.g. "god", "great", "lover", "geek"), misspellings are just some of the factors that make this evaluation particularly challenging. To ameliorate the effects of such factors, we stem the tags, remove stop-words and eliminate tags with less than ten occurrences in the corpus. This brings the total to 32099 (non-distinct) tags for the entities. This experiment measures *Recall@k*: for each approach, we report the percentage of the wefollow tags of each entities found in the top-k tags ($k \in \{1, 3, 5, 10\}$) recommended for it.

As can be seen from Figure 5, ENDORSEASPECTS clearly outperforms the others, for all values of k. Despite the difficulty of pre-

Aspect label	Top-5 Entities				
UK TV Personalities	Chris Moyles	Jonathan Ross	Phillip Schofield	Fearne Cotton	Stephen Fry
Technology People	Leo Laporte	Kevin Rose	Alexal Brecht	John C. Dvorak	Patrick Norton
NBA basketball	Lamar Odom	Derek Fisher	Dwight Howard	Shaquille O'Neal	Baron Davis
US Teen Celebs	Selena Gomez	Miley Cyrus	David Henrie	Mitchel Musso	Emily Osment
Cycling	Lance Armstrong	Levi Leipheimer	George Hincapie	Johan Bruyneel	Dave Zabriskie
Twilight	Peter Facinelli	Twilight Fans	Rachelle Lefevre	Rob Pattinson News	Taylor Lautner
Political Commentators	Rachel Maddow MSNBC	MirHossein Mousavi	Claire McCaskill	TehranBureau.com	HuffingtonPost
Brazilian Celebs	Rafinha Bastos	Marco Luque	Felipe Andreoli	Luciano Huck	Oscar Filho
Country Music	Taylor Swift	Carrie Underwood	Kellie Pickler	Keith Urban	Brad Paisley
Rock Music	Pitchfork	David Lynch	NME Magazine	Sonic Youth	Liam Gallagher

Table 1: Top-5 entities (by $p(e|k)$) for 10 selected aspects

dicting the self-assigned tags, ENDORSEASPECTS achieved a recall as high as 30%. None of the other three TEXTASPECTS based approaches stands out, while TFIDF was consistently last. TFIDF tends to report tags that are more specific to the entity, while users prefer more generic tags to encode themselves and their different characteristics. For example, a famous painter is more likely to choose to tag himself with terms like "artist" and "painter", rather than the name of his wife or his hometown.

Toward a more thorough analysis of the results, we examine the tags that hurt the recall of ENDORSEASPECTS the most. Let *success(t)* be the number of times the wefollow tag t was correctly predicted by ENDORSEASPECTS and *fail(t)* be the number of times it was missed (for k=10). First, we select the 20 wefollow tags with the highest *fail(t)-success(t)* difference. For each selected tag t, we report the seven tags that were most commonly reported for entities tagged with t (the number seven was chosen solely due to space constraints). The results are shown in Table 2. The 1st column holds the tags ENDORSEASPECTS misses. The 2nd column holds the percentage of missed predictions due to each tag. The 3rd column holds the seven most common alternately reported tags.

As shown in the table, these 20 tags alone are responsible for almost 30% of all missed predictions. Note that the reported terms come from the same domain are clearly connected to the missed tag. Another observation is that ENDORSEASPECTS often returns terms that represent more specialized facets of the missed tag. (e.g. see results for "music" and "sports") The consistently high relevance of the tags demonstrates that, even if a tag is missed, the domain and nature of the entity are still correctly identified. This leads to the assignment of appropriate tags that capture the entity's attributes.

3.3.3 Evaluation of Aspects

In this section we evaluate the aspects computed by our approach. Aspect quality is crucial, since it directly affects the tag extraction process in our model. Moreover, good quality of aspects opens the door to using them as browsable pages or for recommendation systems. The experiment is conducted as follows. First, we pick the 50 aspects with the highest $p(k)$ value, which represents the probability that an endorsement made due to aspect k and can be computed as follows $p(k) = \sum_{u \in \mathcal{U}} p(u) * p(k|u)$. Here $\mathcal{U}$ is the set of all users and $p(u)$ encodes the propensity of a user u to make an endorsement. We consider this probability to be equal for all users. $p(k|u)$ is the probability that user u makes an endorsement after picking aspect k from his distribution of interests, as computed by our model.

For each aspect k, we manually assigned a descriptive label, based on the nature of the entities that were given a high probability with respect to k (i.e. $p(e|k)$) by our model. For example, if the entities with high $p(e|k)$ values tend to be basketball players from the NBA, the label "NBA basketball players" was given to the aspect. If the labeling process revealed aspects with the same labels we keep the aspect with the highest $p(k)$ value. This results in 32 distinct aspects.

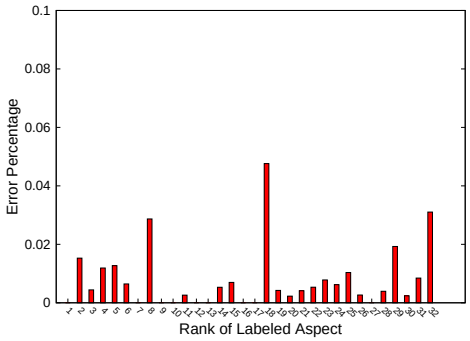


Figure 6: Results of evaluation of aspects on Twitter

Next, we picked the 5 entities with the highest $p(e|k)$ value for each aspect and presented the entire set to three annotators in a randomized order. The annotators were asked to mark from among the 32 labels those that they found appropriate for each entity. The performance of ENDORSEASPECTS for each aspect k was then measured via the following error function:

$$Error(k) = \frac{\sum_{e_1,e_2} I(h(e_1,e_2,k) \neq 0 \text{ and } h(e_1,e_2,k) \neq r(e_1,e_2,k))}{I(h(e_1,e_2,k) \neq 0)} \quad (9)$$

where $h(e_1, e_2, k)$ is equal to 1 if the annotator marked e_1 as relevant to k and e_2 as non-relevant, -1 if she marked e_1 as non-relevant to k and e_2 as relevant and 0 if the marked them both as relevant. Similarly, $r(e_1, e_2, k)$ is 1 if $p(e_1|k) > p(e_2|k)$, -1 if $p(e_1|k) < p(e_2|k)$ and 0 if $p(e_1|k) = p(e_2|k)$, as computed by ENDORSEASPECTS. Conceptually, the error function iterates over all possible entity pairs and adds a penalty point when ENDORSEASPECTS contradicts the annotator. Observe that a penalty can only be incurred for a pair of entities if k was marked by the annotator as relevant for one entity and non-relevant for the other. Thus, the overall error is computed as the percentage of such pairs that actually incurred a penalty. The results are shown in Figure 6. The x-axis holds the selected aspects, in descending order of $p(k)$ from left to right. The y-axis holds the $Error$ values.

The reported $Error$ is consistently low for all aspects. The highest observed value was lower than 5%, while for 9 of the aspects the error was equal to zero. This shows that human annotators agree to a high degree with the assignments of entities to aspects found by ENDORSEASPECTS. As qualitative evidence, we present the top-5 entities (by $p(e|k)$) for 10 of the top-50 aspects. As demonstrated clearly by the error measurements of the previous experiment, any subset of the aspects would produce convincing anecdotal evidence. The aspects were selected purely to maximize the verifiability of the results by the reader. The 1st column of Table 1 holds the labels of the selected aspects, while the selected entities are shown in the 2nd column. The table shows that the reported entities are both famous

wefollow Tag	Fraction of Errors	Reported tags						
social_media	6.3%	marketing	news	business	video	success	internet	music
music	4%	album	band	record	track	news	hip_hop	tour
media	2.7%	obama	politics	republican	elections	marketing	news	music
business	2.1%	marketing	news	social_media	video	blog	media	management
web	1.7%	social_media	marketing	news	media	technology	video	company
sports	1.4%	league	draft	player	basketball	nba	espn	nfl
author	1.3%	marketing	social_media	news	media	business	spiritual	management
technology	1%	marketing	news	social_media	blog	video	media	music
comedy	1%	news	business	home	social_media	starcraft	videos_channel	video
radio	0.9%	music	hiphop	news	obama	rap	mixtape	artist
artist	0.9%	jewelry	handmade	items	art	dollars	shopping	vintage
nonprofit	0.8%	poverty	africa	featured_sponsor	qstring	donate	global	nytimes
actor	0.8%	tv_episode	episode	hollywood	film	movie	trailer	dvd
startup	0.7%	marketing	technology	media	business	company	management	news
geek	0.7%	news	social_media	blog	marketing	video	media	playstation
advertising	0.7%	marketing	social_media	news	business	video	management	media
journalist	0.6%	obama	bbc	social_media	london	politics	british	media
podcast	0.6%	social_media	media	blog	marketing	comment	news	business
consulting	0.6%	marketing	social_media	business	media	success	industry	sales
innovation	0.5%	social_media	marketing	media	management	industry	business	technology
	29.5%							

Table 2: Tags reported most frequently by ENDORSEASPECTS for entities tagged with selected wefollow Tags

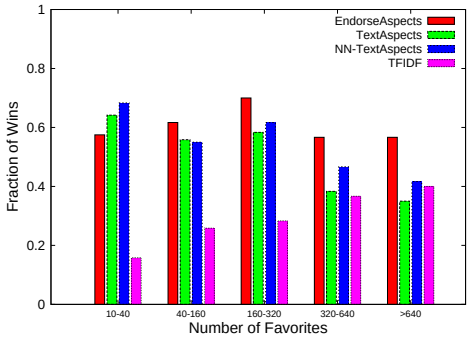


Figure 7: Results of evaluation via manual labeling on Flickr

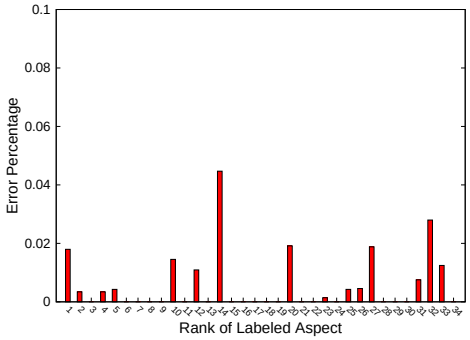


Figure 8: Results of evaluation of aspects on Flickr

individuals (or organizations) and highly relevant to the respective aspects. Just as importantly, the automatically discovered aspects are cohesive and represent interesting concepts.

3.4 Evaluation on Flickr

3.4.1 Flickr Evaluation via Manual Labeling

In this section we describe our evaluation of the quality of the tag-sets reported by the various approaches for the images in the Flickr dataset. The evaluation is conducted in a manner similar to that of the Twitter dataset, described in Section 3.3.1: images are put into 7 groups based on number of endorsements, 30 entities per range are selected, two annotators rate tags for each image. The results are shown in Figure 7. ENDORSEASPECTS consistently reported competitive results, achieving the best score for 4 of the 7 ranges. TEXTASPECTS and NN-TEXTASPECTS were also competitive, especially for images that fall within the first ranges.

Tagging in the Absence of Text: While indicative of the competitiveness of our approach, this experiment does not demonstrate one of its primary advantages: the ability to assign tags to an entity, even when there is no text associated with it. Theoretically, the NN-TEXTASPECTS approach also shares the same advantage, since it can borrow text from the entity's neighbors. To evaluate both methods on this task, we removed all the textual information attached to all the images that were evaluated in the previous study. We then re-applied both approaches to the corpus and generate new sets of tags for each image. A manual inspection of the results verified that the negative effect of removing text on ENDORSEASPECTS was negligible, while NN-TEXTASPECTS failed to produce meaningful results. The outcome of this trial, together with the results of the original user study, indicate the following: the local enhancement offered by the nearest-neighbor approach may help the discovery of appropriate tags, but it fails to produce high-quality results when no text is attached to the entity itself.

3.4.2 Evaluation of Aspects

In this section we evaluate the aspects computed by our approach for the Flickr Dataset. The methodology is the same as for the Twitter dataset (Section 3.3.3). After picking and labeling the 50 aspects with the highest $p(k)$ values, we eliminate those with duplicate labels. This produces a set of 34 distinct aspects. The 5 images with the highest $p(e|k)$ value for each aspect were then presented to a human annotator in randomized order, along with the labels of selected aspects. The annotator was asked to mark the labels that she found appropriate for each image. As we did for the Twitter dataset, we used the error function defined in Eq. (9) to evaluate the results. The produced error values are shown in Figure 8. The x-axis holds the selected aspects, in descending order of $p(k)$ from left to right. The y-axis holds the error percentages. As in the case of Twitter, the observed error percentage was consistently low for all aspects. The highest observed value was lower than 5%. Further, no error was observed for 19 of the 34 aspects. This shows that the computed aspects are cohesive and well-defined.

4. RELATED WORK

Here we review some relevant past works and point out the differences compared to our own work.

SOCIAL TAGGING. With the emergence of Web 2.0, social tagging has attracted much attention. Several approaches [19, 18, 29] assume pre-existing tags. The problem is then formalized as one of re-ranking or propagating existing tags across the entities. This is achieved by utilizing various similarity measures or nearest-neighbor methods [6]. The major shortcoming of such methods is the requirement of a significant number of high-quality tags. However, several major social networking platforms (e.g. Twitter) do not even provide a social tagging feature. Our approach has no requirement for pre-existing tags and thus overcomes this drawback. Chirita et al. [8] mine tags from raw text. However, their approach (which is customized for Web pages) focuses exclusively and depends heavily on text, disregarding the underlying social endorsement graph. In fact, our work is the first to utilize social endorsement data toward improving entity tagging.

ENDORSEMENT NETWORKS. Different facets of social endorsement networks have been considered in the past. Kunegis et al. [20] study a type of endorsement graph from Slashdot.org. In a recent paper [22], Lappas and Gunopulos mine groups of frequently co-endorsed entities toward an interactive recommendation system. As part of their work, the authors apply a very basic group-tagging mechanism that assumes the existence of a piece of text for each entity. Each group is then tagged based on the overlap of the term-sets of its members.

TOPIC MODELING. In their seminal work on topic modeling Blei et al. [5] introduce Latent Dirichlet Allocation (LDA), an unsupervised method to infer latent topics from unlabeled data. The method has been extensively used in several domains, including that of tag recommendation. The approaches by Xu et al. [34] and Krestel et al. [19] are examples of such work, although they both assume pre-existing tags. Si and Sun [28] present an LDA-based approach for tagging documents, making it applicable only to textual entities. Jianshu et al. [33] apply LDA on tweets, toward the identification of topic-sensitive influential Twitter users. We also present experimental results on Twitter data, even though the focus of our work is radically different. In addition, we evaluate the LDA-on-Tweets approach as the baseline (TEXTASPECTS) in our experiments. In our work, we use LDA on the underlying social endorsement graph to mine the topics that drive users to endorse the various entities. While LDA has been applied to graph corpora before [27, 9, 11, 25] in various settings, ours is the first work to apply it to social endorsement graphs toward tag extraction.

5. SUMMARY AND FUTURE WORK

In this work we demonstrated that social endorsement networks can be used to extract high-quality tags from noisy text associated with entities embedded in social systems. Our proposed approach models the reasons why entities appeal to users of social systems and then uses these aspects of appeal to extract tags. Finally, we evaluated our approach against intuitive and competitive baselines on large-scale real-world datasets from Twitter and Flickr. In future work, we would like to evaluate the impact of our extracted tags on tasks such as recommendation systems, contextual advertising, and expert finding.

6. REFERENCES

[1] C. Anderson. *The long tail: Why the future of business is selling less of more*. Hyperion Books, 2008.

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

[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07*.

[4] P. Basile, D. Gendarmi, F. Lanubile, and G. Semeraro. Recommending smart tags in a social bookmarking system. In *(SemNet 2007)*, 2007.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] A. Budura, S. Michel, P. Cudré-Mauroux, and K. Aberer. Neighborhood-based tag prediction. In *ESWC '09*.

[7] A. Byde, H. Wan, and S. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. In *ICWSM '07*.

[8] P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic gen. of pers. annotation tags for the web. In *WWW '07*.

[9] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS '01*.

[10] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *MULTIMEDIA '06*.

[11] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML '07*.

[12] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *WWW '06*.

[13] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. In *WSDM '10*.

[14] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. Aug. 2005.

[15] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl. 1):5228–5235, April 2004.

[16] C. Hayes and P. Avesani. Using tags and clustering to identify topic-relevant blogs. March 2007.

[17] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08*.

[18] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *PKDD 2007*.

[19] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *RecSys '09*.

[20] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *WWW '09*.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10*.

[22] T. Lappas and D. Gunopulos. Interactive recommendations in social endorsement networks. In *RecSys '10*.

[23] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06*.

[24] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Tech. Rep., UIUC, 2004.

[25] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *KDD '08*.

[26] T. O'Reilly. What is web 2.0. design patterns and business models for the next generation of software, September 2005.

[27] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. UAI '04.

[28] X. Si and M. Sun. Tag-LDA for Scalable Real-time Tag Recommendation. *Journal of Computational Information Sys.*, 2009.

[29] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*.

[30] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *SIGIR '08*.

[31] J. Trant. Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, 10(1), 2009.

[32] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[33] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10*.

[34] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized lda. In *MM '09*.

[35] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD '09*.