Abstract

Tagging has gained a lot of ground as a lightweight annotation system, proven to work well in large scale deployments. This paper explores the world of collaborative tagging, and highlights some of the issues one is faced with when trying to utilize the tag meaning for knowledge extraction, with a special focus on using the tag meaning to create a user interest profile.

KEYWORDS: tags, tagging, folksonomy, collaborative tagging, collaborative filtering

1 Introduction

Content in the Web 1.0 era was mostly static, and user participation in content creation and refinement played only a minor role. After that, the phenomena labeled Web 2.0 emerged a few years back, and suddenly user created content started getting vast amounts of coverage even in mainstream media. Web 2.0 is characterized by creativity, collaboration and sharing with other users. In this environment, a steady stream of new content in various formats is constantly made available and ready for consumption. To facilitate better findability and linking, many sites started using a lightweight classification system known as tagging. A tag is a label applied to a resource, a freely chooseable keyword which relates to the resource. Contrary to classical keywords, tags are tiny pieces of functionality - typically dynamic searches linking to other resources with the same tag - and therefore do not exist outside the computer world.

Due to the unstructured nature of tags, mapping them to a purer form of conventional, structured metadata is not without challenges. For some systems, such mapping is not needed, but in many other cases refined data would be much preferred. This paper describes these challenges and ways to overcome them, with a special focus on utilizing the tag metadata to create a user interest profile.

The paper is organized in the following way: Chapter 2 describes tagging, its background and the group phenomena behind collaborative tagging and the incentives for users to tag. Chapter 3 explores the challenges of extracting meaning from the tags. Chapter 4 deals with the specific case of creating a user interest profile from extracted tag meaning. Also, an algorithm capable of adjusting to diverse tagging habits while still catching the essential user interests is proposed. The final conclusions are found in Chapter 5.

2 Background

Tagging is the process of annotating content with simple metadata in the form of a flat list of tags, unstructured keywords. Although tagging content for oneself for later reference is useful as such, tagging web content in a collaborative fashion can unleash a lot of further potential realized by these added annotations. In the following sub chapters we will discuss swarm intelligence, finding content using “the long tail” and the emergence of folksonomies, which are all examples of the benefits that can be achieved from such collaboration.

Collaborative tagging and filtering is commonly utilized in Web 2.0 applications, but the problem to be solved is not a new one. It’s one that face most content publishers at one point or the other, namely that of content categorization and discoverability. Because of increasing amounts of user generated content the need for filtering is apparent.

2.1 Folksonomies

The term folksonomy is used to describe collaborative tagging in pursuit of categorizing content. [1] defines folksonomies as “a folk taxonomy of important and emerging concepts within the user group”. Folksonomies are typically Internet based, with user participation as a crucial ingredient. The goal is to improve the discoverability of the content, for yourself and for others. Systems commonly enable viewing the tags of other persons who have tagged the content and facilitate navigation from resource to resource using tag relationships.

Seeing others’ tags, in turn, will many times aid in finding related content and making connections one didn’t see before. [2] shows that in terms of path length, for a graph describing the tag relationships the distance remained relatively small although the nodes in the graph increased. In their experiment the average path length between resources was 3.5, which results in potential discovery from any resource to another within an average 3.5 clicks. The short distance suggests that tag based discovery can be very efficient indeed.

2.2 Social bookmarking

Social bookmarking sites like del.icio.us¹, with over a million registered users, have proven in practice that tagging in all its simplicity is quite a working concept. It is argued that

¹http://del.icio.us/
tagging works well because it strikes a suitable balance between the individual and the community; entering tags is low effort for the user, but it benefits both the user himself and the greater community [3].

The tags that are entered by users are generally more reliable than the keywords provided by the HTML page author using the meta HTML element. The study in [3] indicate that the more popular a website is, the less does its meta-data reflect the tags that site visitors would use. Sometimes it’s obvious that author added keywords are there only to lure traffic to the site, but it’s also possible the user has a different incentive and different perspective on the page in question.

2.3 The long tail

The concept of “the long tail” is tightly linked to the Web 2.0 trend. It describes how marginal niche products can still in total make up significant cash flows in the new (Web 2.0) economy. The problem in earlier days, it is said, was that not enough people were able to find the service or product in question. Niches are per definition not widely known when they emerge, so how would users know what to search for? One way to overcome this is tag based discovery. [3], however, found that users tend to focus their bookmarking and tagging activity on popular pages and less on unpopular ones, which would contradict the usefulness of tags for this purpose.

2.4 Collective and Swarm Intelligence

2.4.1 Background

The collective intelligence of a larger crowd can be far greater than that of a few single experts. Sir Francis Galton is one of the people commonly accredited with this discovery. At a live stock fair he stumbled upon an interesting competition where people were to guess the weight of an ox, it is told. Although no single member of the self proclaimed live stock experts guessed the weight correctly, the average value of the guesses was not more than a pound off [4].

Economist Adam Smith’s invisible hand which guarantees fair prices under fair competition, and the ability of ants to find the shortest path from one place to another are other examples of swarm intelligence in practice.

While swarm intelligence and collective filtering doesn’t require a great deal of building on earlier findings, collective intelligence, improving upon earlier knowledge through collaboration is the central part.

2.4.2 Crowdsourcing

Wikipedia[2] is good example of what collective intelligence can accomplish. With the contribution of lots and lots of users, the gathered knowledge grows far beyond that of any single contributor. One should note that it’s not about imitation, but comes down to everybody doing their part. This somewhat controversial way of “letting the crowd do the work” is commonly referred to as crowdsourcing. To distinguish between the commercial side of crowdsourcing and projects that are more about common good, the Wikipedia model is sometimes referred to as peer production instead.

The Amazon Mechanical Turk[3] is a crowdsourcing service where requesters can provide small mechanical tasks that are best done by humans. Workers then do the work and get a (small, often cents) payment for their work.

As a final example, Google Image Labeler[4] utilize crowd sourcing to get images labeled. Participants play a game where one should tag random images, getting a score based on how well the tags match those entered by the automatically assigned game partner. In the process Google gets fairly good tags suggested that they can use for indexing the image material.

2.5 Tags for Article Discovery

Swarm intelligence can be utilized as inspiration for creating artificial intelligence applications of various kinds. A typical usage scenario for the web is collective filtering: using this group phenomena for filtering out the top content from a wide range of commercial and user generated content. The generated “intelligence” is utilized by services where many people do small parts to filter out and classify various web resources. In social bookmarking the user mass determines the content popularity by bookmarking (and tagging).

The web formed by all the interconnected blogs of the Web is commonly referred to as the blogosphere. Blogs have taken advantage of tags since early on, and it could be claimed these connections played an important role in making the blogosphere what it is today - a network where links between blogs are an essential ingredient. Technorati[5], which specializes in user generated content, especially in finding new interesting blog articles, lets blog owners register to claim their blog, which they can tag with relevant keywords. For common users, the service offers users the ability to keep informed by providing feeds for the tag (keyword) they want.

2.6 Tags for Multimedia Discovery

Modern search engines are good at finding exact knowledge, while searches for something less well defined is substantially more difficult. To improve the situation especially sites with heavy multimedia usage rely greatly on tags. The web’s number one video sharing site YouTube[6] lets the person uploading set tags to increase discoverability, but rely on other mechanisms like video view counts for popularity rating. Flickr[7], which is one of the web’s largest photo sharing sites, is heavily dependent on tags for findability. In fact it was one of the pioneers that moved tagging into main stream. A further example of tag utilization in multimedia discovery is Last.fm[8], which is an Internet radio and music community right in the spirit of Web 2.0. It features recommendations based on past listening history as well as some social features, part of which involve tags. For instance, tagging music

---

2http://www.wikipedia.org/
3http://www.mturk.com/
4http://images.google.com/imagelabeler/
5http://technorati.com/
6http://www.youtube.com/
7http://www.flickr.com/
8http://www.last.fm/
This is a tag cloud - a list of tags where size reflects popularity, sort: alphabetically | by size

.net 2008 3d advertising ajax api apple architecture art article articles au audio blog blogging blogs book books business community computing cool css culture design development diy download ebooks economics education email english education environment fashion fic finance firefox flash flickr food free freeware fun funny gallery games google graphics green hardware health history home howto html humor illustration images imported inspiration interesting internet iphone java javascript jobs language library lifehacks linux mac magazine management maps marketing math media mobile money movies music news obama online opensource osx photo photography photos photoshop php politics portfolio productivity programming python rails recipe recipes reference

Figure 1: A Tag Cloud

with its genre is popular, resulting in the ability for anyone to listen to user compiled lists of recommended music by a certain genre.

For many hard core taggers, the tags play a big role in finding new content. Users often navigate through the content by clicking on related tags or directly from tags in a so called “tag cloud” which visualize the overall tag popularities.

### Challenges of Tag Refinement

#### 3.1 Introduction

When many people engage in the loose, small-scale collaboration we know as collaborative tagging, it’s inevitable that tag usage and tag conventions vary a lot. Different lexical forms, plural and singular noun forms, alternative spellings, synonyms and polysemy (words with multiple meanings) are some of the challenges facing anyone wanting to efficiently utilize the tag meaning.

Due to the inherent unstructured nature of a tag, different people will use different tags for different things, and use various lexical forms and different words altogether. However, if users are presented with tag suggestions based on what other people have used, the tags for the resources tend to converge [5]. There is a lot to be won by combining the same concepts under the same word root - for instance “movie” and “movies” should reasonably be interchangeably usable within the system. Refinement through tag analysis and clever tag suggestions are therefore needed to maximize the convergence and thus the usefulness of the tags.

There is also the question of precision level. If the tag “beagle” is used, it won’t get linked from the “dog” tag automatically. Taking these kinds of lexical relation could improve the general usefulness of tags quite a bit. [3] argues that tagging is (currently) more useful for broad categorization than with specific categorization.

An interesting aspect of tag utilization is processing an existing user’s tag cloud to create a user profile describing his/her interests; cf. Chapter 4. The user profile interests would then be used to form more accurate and interesting search results and better linking for the user - all small steps of providing enjoyable knowledge discovery.

#### 3.2 Tag Classifications

After analyzing what types of tags people use, [5] divides them into five different categories: content-based tags (e.g. videos, open source), context-based tags (e.g. Los Angeles), attribute tags (e.g. dad’s blog), subjective tags (e.g. funny, cool), and organizational tags (e.g. todo, to-read). Especially tags in the organizational category are clearly of little value for others than one self. The usefulness of the other categories varies, but generally content and context based tags are the most useful ones.

Since automatic lexical and other processing of tags after they have been entered is complex and error prone, it is essential for the tagging application to offer good tag suggestions. In the suggestion process, the system can improve tag uniformity in a number of ways. For instance, it can suggest tags known to co-occur with the first entered tag, it can use stemming, and take advantage of thesauri.

#### 3.3 Co-occurrence

When we have the resources and their corresponding tags, we can easily form absolute and relative co-occurrence frequencies between a tag pair. The absolute co-occurrence says how many times two tags were used together in total, and the relative co-occurrence frequency states how often tag1 was used together with tag2 compared to the total number of times tag1 was used in total.

#### 3.4 Stemming

To obtain the word root form from a tag, a process known as stemming is applied. In the stemming plurals will be reduced to singular form (“blog” to “blogs”), and derived words can be combined into one (“blog” into “blogging”). The Porter stemmer algorithm is the predominantly used stemming algorithm for the English language. Tools and libraries are freely available to perform the stemming.
3.5 Thesauri

WordNet\textsuperscript{9} is one of the more complete semantic lexicons for the English language. Using WordNet, normalized words can be classified according to their lexical class. Of the lexical classes, verbs, adjectives and adverbs are for the most part not tags that other people can take advantage of. Tag words that have several potential lexical classes should be included in an initial analysis, and co-occurrence can be used to determine if that tag is relevant for the given context.

Wikipedia can also be used to extract some semantic relationships. The dbpedia\textsuperscript{10} project focuses on extracting structured data from Wikipedia and making the data available for the Semantic Web in RDF format.

4 Building a User Interest Profile

4.1 Background and Dataset

For audio-visual search or browsing, being able to utilize known user interests when listing matches and suggesting related material has interesting prospects. By carefully analyzing the tags a user has used, we can create a user interest profile from the data. To overcome the cold start problem, utilizing existing user data from other services is possible.

Using public APIs to access the del.icio.us data, we obtained a dataset which contained the 100 last tagged resources of 1035 users, as well as a list of all their (public) tags along with the total tag usage rates.

4.2 Considerations

When tagging, one usually types in the tags in a text field. Different systems use different techniques to separate the tags from each other. del.icio.us currently allows only one word tags, so multi-word tags are typically written together as one, sometimes with underscores or in camel case. Many other systems allow multi word tags; separating tags with commas is indeed a fairly intuitive and workable solution.

As we seek to map the tag to a certain meaning, the different kind of word forms etc. are standing in the way of that. We will therefore utilize stemming, and lexical analysis through WordNet. We argue that - especially when using the tag meaning to create a user interest profile - nouns are what’s important. They usually denote the subject matter, the substance of the resource which in turn describes ones interest.

4.2.1 Subjective Tags

After some experimentation and reasoning we found that for creating a user interest profile, some types of tags do not have to be taken into consideration. Let’s consider a user who tags a nice article about python programming. He might tag that “article, nice, python, toread”. In it self, only the content-based and context-based tag types describe a characteristic relevant for user interest - in this case, “python”. The other tags describe properties of the tagged resource. Analyzing the content of the resource could yield some additional interest data, but that is beyond the scope of this experiment. We find it unlikely that adjectives would infer any helpful data about the subject as few resources are tagged with negative opinions. An uninteresting resource would simply not get tagged since there is no reason to come back to it.

Even positive adjectives are not that commonly used. For our test data of roughly 280 000 tags, the tag “cool” was used only 871 times, which is about 0.3% of the tags. The tag “bad” was only used eight times in total. We subsequently find the subjective tags are superfluous with regards to forming the interest cloud.

4.2.2 Organizational Tags

Although the reasoning for subjective tags also partly apply to organizational tags, they could potentially be used for weighting of the content-based tags. An initial idea was to give content-tags used in combination with organizational tags like for instance “to-read”, “todo” a higher rating since it describes what the user is likely to want more information about in the near future. However, organizational tags were not used widely enough - some very rough numbers indicate somewhere well below 0.5% in total for our dataset - and had marginal impact for the people that used them. We therefore decided to focus on other algorithms at this point.

4.2.3 Tagging Habits

As stated earlier, tagging habits vary a lot between people. Below are some statistics about the tagging habits of del.icio.us users in our dataset (1035 users, the latest 100 tagged resources per user):

- 1.55% of the users have over 1000 tags, 2.2% had over 900 tags, 14.2% had over 500 tags, 50% had over 178 tags and 5.1 have very few tags are not used in the analysis.
- The average number of tags that users assign to a resource (bookmark) was 3.1 ± 2.4 tags/resource, the maximum was 20.5 tags/resource.
- The average number of total number of tags of a user was 273 ± 234 tags/user and the maximum was 2046 tags/user.

As we can see from the numbers, the standard deviation is quite large with regards to the total number of tags used per user. After analysing a few different user types, it’s clear that without taking the tagging habits of the user into account, we will not get satisfactory results.

4.2.4 Time Sensitivity

Since our dataset contains both the recent tags as well as numbers for overall all-time usage, we could look at short term and long term trends to see which tags are of rising interest for the user, making the assumption that the most recent tags describe the current user interests. Taking time into account is left for future experiments, since that particular data isn’t always available.

4.2.5 Tag Usage Frequency

We make the basic assumption that frequently used tags are core interests of the user. We also assert that each person
has a limited number of real interests. If we didn’t make this assumption, the analysis results would easily become too watered down to be useful. This is especially true when analyzing users who evidently tag for discoverability, using multiple similar tags or whatever alternative tag they imagine could be used by others. Tag spam is another (unrelated) problem in collaborative tagging systems, but for user interest profiling that is not an issue as it simply means spammers would not get a good interest profile about themselves.

We define the absolute tag frequency as the total number of times a tag has been used by a user. The relative frequency is the number a tag has been used divided by to how many times any tag has been applied to a resource.

Now that we have established the need to get only a select amount of tags to represent the user tag profile, there is inevitably the question of how many tags is appropriate. By looking at the most frequently used tags for some sample users, we come to the conclusion that using some fixed number of tags would not give good results due to the extremely varying tagging habits people can have. For instance, a user with few tags would get assigned interests much too easily. Using a fixed absolute frequency on the other hand would give heavy taggers who use many tags per resource interests they primarily don’t really have.

To get more even numbers for all users, we have therefor adopted an algorithm based on the a cumulative relative tag frequency.

### 4.3 Algorithm

To extract the user interest profile from the remote data we exercise the following procedure:

1. Fetch and store tags and resource URLs.
2. Run the tags through a stemmer. Since the stemmer can’t handle other languages than English, we ignore all non-ascii tags.
3. Pass the word root the stemmer provided us with to WordNet, to get some lexical data for the word. Unfortunately, lexical analysis of tags consisting of two words written together won’t succeed. (These are therefore discarded.)
4. Choose a cutoff frequency. Having experimented with the cutoff frequency, we have found 30% to be a fairly useful value.
5. Select the tags which have a) known noun word roots and b) whose relative frequency cumulatively add up to the chosen cutoff frequency. Many times considering only the relative frequency will give unjust results, we therefore also include the tags which have the same absolute frequency as the absolute frequency of the last tags to be included by the cumulative relative tag criteria.
6. Create the user interest profile by listing the selected tags along with their frequencies, co-occurring tags and co-occurrence frequencies.

For the tags to be considered to co-occur we require them to be used together more than once. Otherwise the co-occurrence would very easily grow beyond usefulness. In case the co-occurrence list grows too large, higher values could be tested.

For doing the lexical analysis we used the “MIT Java Wordnet Interface” library to categorize the used tags and the stemmer included in the library for lexical classification.

## 5 Conclusions

Tagging has proven itself to be a powerful form of lightweight annotation that can be used in many situations. While combining the tagging power of the masses can yield high rewards, utilizing the tagging data in refined forms pose the implementer with some challenges.

Users want to get good, accurate search results, and especially when searching audio visual material, one often has quite a fuzzy search criteria. Here, knowledge about the user’s interests could be used to improve the perceived accuracy of the search result listings. In this paper we have described an algorithm by which a tag listing of the user’s primary interests can be obtained by analyzing the tagging habits of the user on a social bookmarking site where the user is a member. For a new player in the audio-visual search market, this helps to alleviate the cold start problem where established players already have access to a lot more data (through analyzed search history) than the user even realize.

### References


---

1<http://www.mit.edu/~markaf/projects/wordnet/>