Constructing Evolutionary Taxonomy of Collaborative Tagging Systems

Junjie Yao  Yuxin Huang  Bin Cui
Department of Computer Science & Key Laboratory of High Confidence Software Technologies (Ministry of Education), Peking University {junjie.yao, yuxin, bin.cui}@pku.edu.cn

ABSTRACT

Collaborative tagging systems allow users to label online resources. The tags are generally correlated and evolving according to the change of web contents, and the popularity of tags represent evolution of social interests. Taxonomy is a promising solution to organize the data in tagging systems. In this demonstration, we propose to construct the evolutionary taxonomy which incorporates the correlation and evolution of tags, as user generated tags grow and change temporally. We demonstrate that our approach is intuitive and efficient in tag organization which exploits the evolving characteristic of collaborative tagging systems.

Categories and Subject Descriptors
H.2.8 [Information Systems]: [DATABASE MANAGEMENT], Database applications

General Terms
Algorithms, Experimentation

Keywords
Tagging System, Evolutionary, Taxonomy

1. INTRODUCTION

Collaborative tagging systems have gained significant popularity for their revolutionary ways of organizing online resources. Tagging behavior of users does not require any strict expertise training. Each user can provide tags to a certain online object, according to his/her own knowledge. Tagging is especially suitable for online corpora, in which the distributions or types of contents are diverse and change dynamically. Many tagging websites have been established, such as Del.icio.us for web page bookmarking, Youtube for video sharing, Flickr for photo sharing, etc.

Currently, most tagging sites present tags in a tag cloud style. This kind of flat list is easy to implement, but cannot capture the rich semantic correlation among tags. The taxonomy is a hierarchical classification structure used to organize online resources better, which can be used to flexibly describe the corpora content with varying granularity. There are some attempts to automatically construct tag taxonomies [3]. Figure 1 illustrates a part of tag taxonomy extracted from the tag space of Del.icio.us.

![Figure 1: An example of taxonomy (Del.icio.us)](image)

Different from previous small corpora, the tagged resources in current online tagging systems are large and dynamic [2]. Representing users’ interest, tags always change and evolve temporally. The previous unsupervised tag taxonomy construction did not consider the temporal information. Here we investigate how to incorporate the temporal smoothing and the taxonomy construction into an unified process, and propose the evolutionary taxonomy construction mechanism to demonstrate the taxonomy evolution.

2. SYSTEM OVERVIEW

Some researchers recently proposed to exploit tags’ statistics information to automatically extract tag taxonomy from tag space [3]. The existing taxonomy extraction algorithms are able to generate hierarchical concept structure from a static tag space. However, it cannot capture the evolving and dynamic nature of the tag space [1]. In this section, we present our solution for the evolutionary taxonomy construction which exploits the evolving tags.

2.1 Data models

In collaborative tagging systems, a tagging action involves user, tag, resource, and time. Users generally provide multiple tags for an object. We define "co-occurrence" as two tags used by the same user to annotate the same object. This co-occurrence relations among tags generally indicate some semantic aspects, e.g., homonymy, synonymy, subsumption and so on. The tag space can be represented by an undirected weighted graph, denote as $G = (V,E)$ where $V$ is the...
set of tags and $E$ is the set of tag correlations. There exists an edge between two tags $u$ and $v$ if and only if there exists “co-occurrence” relationship between them with its weight $w(u, v)$ determined by the strength of co-occurrence. We choose an association rule based way to extract taxonomies from tag co-occurrence graph. The association rule methods hold an assumption that strong co-occurrence tag pairs indicate a “type-of” subsumption relationship [3]. An example of association rule graph is shown in Fig 3, where the larger dot indicates more general concept.

![Diagram](image)

**Figure 3:** An example of association rule graph

Formally, let $t_1, t_2, \ldots, t_m$ be $m$ consecutive temporal intervals, and $G_1, G_2, \ldots, G_m$ be the co-occurrence graph of tags identified for each of the intervals $t_1, t_2, \ldots, t_m$. We can build a sequence of taxonomies, $T_1, T_2, \ldots, T_m$, where $T_i$ is the taxonomy corresponding to $t_i$, such that $T_i$ satisfies the current tag characteristics represented by $G_i$ and does not deviate too much from the previous tag taxonomies.

### 2.2 Taxonomy extraction with evolution

To capture the changing oftemporal taxonomies, the straightforward approach generates temporal taxonomies at each consecutive time interval. Although such approach can well represent the snapshot of current tag space, independently extracted taxonomy in each time interval may be vulnerable to the noisy tags or transient tag bursts and deviate significantly from the taxonomies of preceding intervals.

We design a novel evolutionary taxonomy fusion approach, where a flexible smoothing strategy is adopted in the evolutionary taxonomy generation. The taxonomy is generated online by fusing the tag information among current interval’s graph $G_i$ with historical taxonomies and graphs. In this approach, the tag’s frequency and correlations at interval $t_i$ are fused with $T_{i-1}$ and $G_{i-1}$. If the current chosen edge for $T_i$ is reflected consistently in $T_{i-1}$, we just add the edge into $T_i$.

When it is different from $T_{i-1}$, we exploit the information in $G_{i-1}$ and $G_i$, and select a suitable edge which is adjacent to the it with the highest accumulated weight defined as $arg\max_u(u, v)\lambda + (1 - \lambda) * w(u, v)_{i-1}$, where $\lambda$ is the parameter to control the portion of the temporal taxonomy and the historical ones. In this unified process, the historical information is embedded into the taxonomy extraction process smoothly.

### 3. RESULT DEMONSTRATION

We collected a large corpora from Del.icio.us to demonstrate the performance of taxonomy construction and evolution. The dataset contains more than 200,000 users’ 270 million tagging actions dated from 2007.01 to 2008.10. Figure 2 shows an example of extracted evolutionary taxonomies. The first taxonomy $a$ is simply generated using the association rule graph of tags in Jan 2008. Under “movies”, there are some common topics, e.g., “hollywood”, “actors”, “netflix” (a DVD rental site) and some temporal hot tags, e.g., newly released movie “juno”. The taxonomy $b$ is the snapshot taxonomy of Feb 2008, where the newly released movie “rambo” (Rambo 4) and “oscar” appear. The taxonomy $c$ shows hot and new topics. We generate the taxonomy $c$ with fusion of the historical information. By integrating the previous information, the tag “netflix” remains in taxonomy $c$ as its accumulated frequency is higher than “rambo”. “oscar” remains in $c$ as it is a very hot topic in Feb due to the annual OSCARS movie event.

Our solutions to the evolutionary taxonomy problems provide more stable and smooth tag evolution. We can utilize it to better analyze and study the taxonomy evolution patterns in the dynamic tag space. The constructed taxonomy can reflect the evolution of user interests and web contents, and facilitate the organization of online resources.

### 4. ACKNOWLEDGMENTS

This research was supported by NSFC under Grant No. 60603045, NSFC China-Australia Special Fund under Grant No. 60710053.

### 5. REFERENCES

