# EMERGENT ONTOLOGY DISCOVERED FROM FOLKSONOMIES

by

© Feng Wu

A Thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

**Master of Science**

**Computer Science Department**

Memorial University of Newfoundland

**January 2015**

St. John's Newfoundland and Labrador

# ABSTRACT

Collaborative tagging websites or systems allow users to associate freely-determined keywords (tags) with a particular resource. The collection of users' tags and resources is referred to as a folksonomy. Unlike traditional forms of metadata, the meaning of, and relationships between, tags are not rigorously defined, limiting the usefulness of tag-based metadata. We propose a novel approach to enrich tagging systems by constructing a tag ontology that captures semantic relationships among tags. We first consider regularities that can be exploited in a folksonomy. Then, we show how user-level tag vocabulary can be used for tag meaning disambiguation. Following this, we introduce a distance model to calculate the relatedness of two sets of resources within a folksonomy, and use this to develop a method for discovering tag relations. A series of experiments we conducted demonstrate the effectiveness of the method. We conclude the thesis with example use cases where our method can be applied to improve folksonomy data organization and queries.

# ACKNOWLEDGEMENTS

My thesis supervisor Dr. Jeffrey Parsons has provided me great insights and inspirations, without which this thesis is impossible to come to life. It is very fortunate for me to have the opportunity to work with and be under the guidance of Jeff, who shares the same passion as I do in the field of data management and modeling, and whose advices always enlighten me when I am out of ideas. I would like to thank him who made my journey to the master degree at Memorial University meaningful and colorful.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

Web 2.0 has enabled the creation of large volumes of user-generated content (UGC). On one hand, the flexibility of UGC allows users to contribute data with few restrictions. On the other hand, unlike content created by information professionals, UGC is less organized, less structured and normally lacks metadata. As a consequence, it is difficult to make effective use of much UGC.

Metadata is indispensible for information organization and retrieval. Traditionally metadata is curated by dedicated professionals. The library and information science discipline has developed and utilized schemes for information categorization and classification for centuries. However, given the scale of available UGC today, it is unimaginable that information professionals could provide us a comparable magnitude of metadata.

Another characteristic of UGC shared online is that the authors or information professionals can hardly foresee how the data might be used, which raises the problem

that professionally created metadata may be disconnected from context in which the end users consume the UGC, hence reducing the value of the metadata.

User-created metadata can help improve the usability of UGC [1]. User-created metadata takes various forms. It can be implicit such as searching keywords, purchase history, or browsing histories. It can also be explicit in the formats of reviews, comments, personal taxonomies, or tags.

This thesis focuses on user-created metadata in the form of ***collaborative tagging***. Collaborative tagging websites or systems allow users to associate freely determined keywords (tags) with a particular resource. Collaborative tagging websites exist to tag an enormous variety of resources such as products, photographs, URLs, podcasts, computer games, music and videos. The hashtags employed by various social media such as Twitter also form tag-resource relations. The dataset arising from all users' tags and resources is commonly referred to as a ***folksonomy*** [2].

Caution should be exercised to distinguish author-generated tagging and user-generated folksonomy. Some tagging services, such as YouTube and Flickr, are author-generated tagging where only the content submitter has the right to tag the content. On the other hand folksonomy, according to the author who coined the term, must hold three qualities [2]:

- *Result of personal free tagging of information and objects for one's own retrieval*
- *Tagging in a social environment*
- *Act of tagging is done by the person consuming the information*

The novel approach proposed in this thesis requires that different users mark the same tag on the same resources. This behavior implies agreement on the meaning of the tag from different users, which will be explained in more detail in the following chapters. Many websites are social-tagging powered or social-tagging enabled. Delicious.com, CiteULike, and LibraryThing, among others, are the best known examples. However, the existence of user-generated tags in other digital forms predates the adoption of tagging by web services. User-created directory trees, bookmark folder systems, contact books and email labels are taxonomy systems in which user-generated tags are employed by information consumers for information classification and retrieval.

Tagging as a method of generating metadata is compatible with our cognition process. We remember concepts by learning their features, which act as an indexing for knowledge retrieval [3, 4]. For example, to find a tool to drive a nail, we may use a hammer. But if there is no hammer around, we may consider some other durable and heavy objects that have a surface to hit the nail, such as a rock. Essentially, we retrieve concepts by considering their relevant features.

Likewise, when we encounter some resources online, we tag them with key words that denote their features. These key words act as a query point later when we need to retrieve the tagged resources again. Unlike our cognition process, however, we share the features we recognized as tags. So that other users or systems could know what we have learned about the concepts and make use of the knowledge. In this sense, tagging can be deemed as mimicking cognition process for concept feature identification. The key words used for tagging a particular resource well summarize the relevant features from the perspective of the tagger.

Nevertheless, compared to the human cognition process, there is a major functionality missing in the state of the art tagging websites or systems. Our brain can organize perceived features into feature networks, or feature ontologies. In turn we can reason about new features based on existing ones. For instance, if someone describes something as “chewy”, then we could automatically infer that the thing is probably “edible” for the reason that chewy things are mostly edible things as well. The ability of inference makes it easier for human to understand and make use of the world around us.

Ontology is the organization of concepts for a particular domain, which describes the concepts and their relations [5]. The feature ontologies in our brain offer us navigation

paths for information retrieval. Currently online tagging systems are unable to provide similar hierarchical tag structures for users to navigate through the tagging data space. Most tagging systems present a “frequency weighted list”, also known as “tag cloud”, to the users for finding information. Sinclair et al. [6] designed an experiment to test the usefulness of tag clouds for information retrieval. They concluded that tag clouds are insufficient navigational tools for folksonomy based datasets.

Tag ontologies can potentially alleviate the major disadvantage of tagging systems – the lack of structural relations among tags compared to other classification approaches such as taxonomies. Emergent tag ontologies elicited from folksonomies bring the best of both worlds: end users have the freedom to create metadata in the form of tags in an uncontrolled manner, while the collectively created folksonomy enabled us to infer the meaning of, and the relationship between, tags.

Emergent tag ontologies can improve information retrieval in the following ways. First, resources that are not explicitly tagged by a given tag, but instead are tagged by some other tags that are identified as sub concepts of the given tag, can be included in the query result of the given tag, which in turn improves the search recall of the query. As an example, resources that are tagged as “owl” but without the tag “bird” can be retrieved with the query “bird” if the tag “owl” is identified as a sub concept of the tag

“bird”. Second, instead of using tag clouds, ontologies can be employed as navigational path for end users browsing the tagging system dataset. Users can start browsing with more abstract concepts and subsequently find more precise concepts by following the concept relations of the ontology.

The main contribution of this thesis is threefold: First, we derived a novel algorithm that clusters resources that are being marked with a same tag, and within each cluster the resources express same or similar semantic meaning of the tag. Second, we designed a distance measurement model to gauge the semantic distance between different resource clusters. In this way, not only we could evaluate the effectiveness of the clustering algorithm, but also it provides us information of relativeness of resource clusters belonging to different tags. Finally, we use a subsumption model grounded by instanced based conceptual modeling theory to find semantic related resource clusters from different tags. As the meaning of tags are expressed by assigning to relevant resources, the relations between tags are also revealed when the relations of underlying resources are discovered. With tag relations at hand, we can build a emergent tag ontology on top of the folksonomy as a whole.

In the next chapter we summarize existing methods for eliciting semantic relations between tags. Subsequently, we discuss the regularities presented in the folksonomy

that are yet to be explored, which leads to our proposed novel approach. In chapter 4 we describe how we utilized user level tag vocabulary for tag meaning disambiguation. In chapter 5 we introduce a distance measurement model to calculate the relatedness of two sets of resources in a folksonomy. Equipped with this model, we propose a method for tag relation discovery in chapter 6. We demonstrate some use cases in chapter 7. Chapter 8 presents conclusions and discusses further research directions.

# Chapter 2. Related Work

## 2.1 Methods for discovering tag relations

### Overview

Ways of enhancing tagging systems have been discussed by researchers from a diversity of perspectives and goals, such as tag conceptualization, tag recommendation, natural language processing and categorization, and information retrieval. In summary, existing works aiming to improve tagging systems can mainly be split into three groups:

1. Establishing relations among free form text tags using occurrence statistics of folksonomy datasets
2. Mapping free form text tags to external vocabularies and semantic sources
3. Modifying the tag format so that users annotate with semantic tags instead of plain text

### Establishing relations among free form text tags using occurrence statistics

Heymann and Garcia-Molina [7] designed an algorithm to automatically build a hierarchy of tags from folksonomy datasets. This approach explores the inherent

hierarchical relationship in a similarity graph. Firstly, similarities between tags are calculated. For each tag, all resources that have been assigned with the tag form a vector. The similarity between two tags is represented by the cosine similarity between the two underlying vectors. A tag similarity graph is plotted with tags as vertices and edges reflecting the degree of similarity between two tags. After obtaining the tag similarity graph, the algorithm repeatedly picks the most central (the vertex that has the most edges) tags to form a tree structure. After each pick, the similarities between the picked tag and other tree node tags are calculated and the picked tag is positioned as child node of the most similar tree node. If all the calculated similarities below a certain threshold, the picked tag is positioned under the root node. The authors cautioned that the variations between different tag datasets could impact the attempt to elicit semantic relationship among tags. A dataset which is of low density (users tag few resources), low overlap between users (resources are tagged by few users), and with some special tag distributions is more difficult to use for semantic inference. A similar approach can be found in [8] where tags are clustered in different granularity based on tuning the relatedness for clustering.

Mika [9] restructured the folksonomy tripartite hypergraph, where the vertices represent sets of actors *A*, tags *T*, and resources *R* into three bipartite graphs AT, TR, and AR. The bipartite graph is then transformed to calculate similarity measures. To do this,

taking AT graph for example, the graph is firstly transformed into a matrix B where $b_{ij}$ represents the affiliation between actor $a_i$ with the tag $t_j$. From matrix B we can eventually obtain two matrices depicting relations between actors and tags respectively. A tag relation network based on overlapping actors is represented by another matrix T = B'B. And an actor relation network based on tag overlapping is obtained as matrix A = BB'. The author then use a cohesiveness measure for clustering based on the calculated similarities to construct synonym sets for each vertex. Broader/narrower tag relations can be extracted by examining superset/subset relations calculated from overlapping actors or resources with which the tags are linked. In practice, near-perfect overlaps are a good approximation of superset/subset relations. Plus, the hierarchy extracted from actors-tags graph is based on sub-community relationships. The author evaluated the two hierarchies generated by AT graph and TR graph, and concluded that the hierarchy based on AT graph yields more easily interpretable results.

Markines et al. [10] summarized approaches for assessing similarities between tags and/or resources. First, the authors presented how the graphs can be constructed. While analyzing the TR graph mentioned above, the weight of the edges can be determined either by simply 0 or 1, as the number of votes from users using the tag on the resource, or as the number of averaged votes from users where the more the user tags the less the weight of a single tag attributed to the user. The authors also discussed

the semantic importance of AR graph, and suggested to add a “user tag” for TR graph analysis. Secondly, the authors listed methods for calculating similarities on the graphs as following:

- Matching: Similarity scores are calculated based on the numbers of elements in the intersection of the two vectors.
- Overlap: The number of elements in the intersection of the two vectors is divided by the number of elements in the smaller vector.
- Jaccard: The number of elements in the intersection of the two vectors is divided by the number of elements in the union of the two vectors.
- Dice: The number of elements in the intersection of the two vectors times 2 is divided by the total number of elements in both vectors.
- Cosine: The number of elements in the intersection of the two vectors is divided by the square root of the product of the numbers of elements in both vectors.
- Mutual Information

Although all the above methods return a value for the similarity measurement, the rankings between each evaluated tag pair are more accurate in representing the semantic closeness of concepts. The authors employed WordNet and calculated the

Jiang-Conrath distance between tag pairs as the baseline for comparing of different similarity measures for tags. The result indicated that the mutual information measurement is more accurate in representing tag closeness. Another similar evaluation on tag similarity and relatedness can be found in [11].

Song, Qiu, and Farooq [12] presented a method to build a hierarchical tag structure based on the specificity/abstractness of tags. To this end the algorithm ranks the tags by two measures: either consider their numbers of appearance and entropy of tags (spanness of tags across different topics), or the relative occurrence of tags (tag $t_i$ appears more than tag $t_j$, and in most of the cases when $t_j$ is present then $t_i$ is also present). The second step of the algorithm builds a hierarchy tree based on the ranked tags by adding the current highest ranked tag to the leaf node of the tree where the two tags have the highest relatedness score.

To summarize, research in this direction utilizes measurement of similarity or closeness among tags in one way or another. Albeit most resulting ontologies are rated positively in evaluations, it is often not clear whether the chosen measurement of similarity has any semantic root, which in turn renders the choice rather a haphazard one. In addition, ontologies constructed this way are often in the form of a top-down tree structure, where any node has strictly one parent node. It might not be semantically accurate to

model the relations between tags with this restriction. Most of all, research in this direction concentrates mostly on analyzing relations between two of three tagging elements (users, resources, and tags [9]). Our proposed approach differs from existing research in three ways:

- We developed a quantitative measurement tool to evaluate the tag relations discovered.
- Tag relations are constructed without predefined structural limitations such as tree structures.
- We take all three tagging elements into consideration at the same time, which reveals hidden regularities that are not obvious when analyzing only two elements at a time.

### Map free form text tags to external vocabularies and semantic sources

Basso, Ferreira, and da Silva [13] proposed a mixed method for mapping text tags to WordNet concepts to improve user level tag navigation and information retrieval. To generate disambiguated mapping, four factors were considered to calculate a semantic similarity measure: (i) Co-occurrence of tags; (ii) the title of the tagged resource; (iii) descriptions of the tagged resource; (iv) other available information. As an example, the authors analyzed a tag vocabulary belonging to a particular user, where the tag “java” is

among it. After querying WordNet the word “java” returns two results, as a programming language or as a beverage. However, the most co-occurring tag with the tag “java” within the user’s tag vocabulary is “prolog”. WordNet returns only one meaning for the word “prolog”. Since WordNet organizes all its concepts in a hierarchical structure where a node’s parent is the more abstract concept, a distance value can be calculated by traversing the common parent nodes between two concepts. The calculation revealed that the concept of “java” the programming language is more related to “prolog”. Hence the tag “java” in the user’s tag vocabulary is deemed as the programming language. Other information such as the title of the tagged resources can be used in a similar way. Later the authors then built a personal tag hierarchy containing all the user’s tag vocabulary in the same way as WordNet organizes its concepts.

Angeletou, Sabou, Specia, Motta, and Specia, Motta [14, 15] presented a two stage approach where the tags are firstly clustered using cosine similarity, and afterwards the tags in the same cluster are mapped to some semantic web ontology to further establish relationship between the tags, and as a side effect, disambiguation. The reason for the clustering is to later help choose the meaning of the tag that is most related to the cluster of the concept if the external sources provide more than one meaning for the tags, similar to the example described above. In the same vein, Laniado, Eynard, and Colombetti [16] proposed a system to map tags to WordNet and subsequently generate

a tag tree according to the taxonomy of WordNet. Tag disambiguation is realized by consulting WordNet with other tags associated with the same resources, so that the entries of the most related meanings are selected.

Van Damme, Hepp, and Siorpaes [17] provided an overview of a holistic approach to generate ontology from folksonomy. The input under consideration includes but is not limited to:

- Statistical analysis of tags
- Implicit social network
- Online lexical resources like WordNet
- Semantic web resources

All in all, several questions remain unanswered during the process of mapping tags to external semantic resources. For a given folksonomy dataset, it is not clear if the vocabulary and the conceptualization of the vocabulary used by the tagging users overlaps with the often used external semantic sources such as WordNet. In contrast to most semantic sources, folksonomies are dynamic, evolving, and domain specific. Even if a mapping is established, it is still possible that the mapping could only partially represent the true semantics that the end users intended to convey by using the term.

Furthermore, grouping synonyms defined by general semantic sources may ignore the substantial distinction between the terms in certain domains.

## Redesign tag format

In the redesigning tag format stream of research, as users have to annotate with more complex data structures rather than simple text strings, tag disambiguation is mostly left to the end users. To enable semantic tagging, the first goal is to devise a proper data structure for representing tags. Several authors discussed the formal conceptualizations or ontologies of tagging [18-21].

Kreiser, Nauerz, Bakalov, Konig-Ries, and Welsch [22] presented a tagging system that allows users to maintain a personal ontology. During the tagging process users can use existing semantic tags or create new ones by specifying the meaning of a tag as well as its relations with other tags. The meanings of tags are stored as Resource Description Framework (RDF) entities. Tanasescu and Streibel [23] implemented a system where tags are kept as plain text but users have the ability to tag tags, and also tag the relations between tags or between tags and resources. In their scheme the disambiguation of tags relies on presenting the relationship between other tags during tag query. Lachica and Karabeg [24] introduced a tagging system which creates a collaborative ontology during the tagging process. In this paper the authors discussed

the unwillingness of users to provide metadata for tags, and inferred that the reason could be that the users participating in tagging practice did not find the semantic metadata supportive for their personal goals of using the tagging service.

The approach of using semantically enriched tags can also be found in [25-27], while the distinctions are that they separated ontology maintenance tasks from the tagging process. These approaches often result in a two tier architecture where the first tier is to facilitate semantic tagging and the second tier is to maintain and manage user generated ontology.

Instead of letting the users bear the burden of properly maintaining tag ontology, other researchers seek ways that external semantic sources can be consulted during the tagging process, including WordNet, Wikipedia [28-30], DBpedia, and OpenCyc [31]. In [32], the authors further attempted to generate lightweight ontology on top of the enriched semantic tags, using WordNet as a bootstrapping tool.

Other authors, e.g. [33], considered the content of the resources being tagged (mainly text based), the tags collaboratively applied to the resources, and the collections of tags used by the users (personomy) and WordNet for term disambiguation and semantic tag recommendation/navigation.

Several semantic tagging services proposed and implemented by the above authors are discontinued at the time of writing this paper. One of the most important factors for this could be that enriched tag format oftentimes requires additional inputs from the end users, either being disambiguation alone, or both disambiguation and identifying semantic relations between the tag being used and other tags. Lachica and Karabeg [24] argued that tagging system users consider the cost of a few additional inputs to surpass the benefit of having a more semantically structured tag vocabulary for all users.

The reason behind this observation might be that, if we look at the user tag vocabulary level, individual end users always curate their tagging vocabulary in a way that the typical folksonomy semantic problems are not present. Homonyms and arbitrary use of synonyms are mostly not an issue in the tagging vocabulary of individual end users. Wetzker, Zimmermann, Bauckhage, and Albayrak [34] observed that, unlike folksonomy as a whole, tag vocabularies employed by single users are characterized as more semantically stable. Heckner, Heilemann, and Wolff [35] discussed that one of the major use cases of collaborative tagging systems is to facilitate saving content for possible later use. In the following chapters, we reason that end users maintain an implicit individual ontology for the tags with which they annotate, so that it is unnecessary for them to explicitly use any more complicated tag formats, which adds no value on their personal goals of using the tagging service.

## 2.2 Instance based conceptual modeling and property precedence

Our approach is inspired by Parsons and Wand's previous work on conceptual modeling. Parsons and Wand [36] argued that the most prevailing conceptual modeling practice, where instances must belong to given classes, is the very root cause of a range of data management problems related to schema integration, schema evolution, and interoperability. Instead, class membership should be inferred based on the properties of the instances. Parsons and Wand [37] further developed the class-independent instance conceptual modeling methodology with a formal definition of how to maintain the semantic relationship between instances, properties, and classes. The core idea behind this methodology, called *property precedence*, is easily applied to the folksonomy dataset. Comparing to conceptual modeling, in folksonomies the resources being tagged can be treated as instances, and tags can be treated as properties. Property precedence in this case can help in finding meaningful relations, and consequently, in building a tag ontology.

Conventionally, for data management, classes are firstly identified, and defined as a collection of properties. In turn relations between classes, including class inheritance, are determined. A database is created according to the class schema and instances are populated and tied to some specific classes. At first, this approach seems reasonable since human beings are used to relate to things as instances of some classes. As an

example, an employee record with an employee ID number X could probably pinpoint a single person of interest. In a slightly more complicated situation where employees can have multiple roles, super/sub class relationships can be established. In this situation this employee can also be an instructor or professor, making it possible that this record of employee possesses other properties which ordinary employee could not have.

However, Parsons and Wand argued that the existence of instances is not and should not be dependent on the existence of classes. Instead, classes exist with the sole purpose that some instances share some interesting properties which could solve some problems at hand, thus instance group memberships are defined as classes based on whether instances possess certain properties.

Hence conventional conceptual modeling approach is use case oriented, for classes and class relations are defined with the problems that the model is intended to solve in mind. When the decoupling of conceptual model and its use case context happens to the datasets, such as during scenarios of data integration, schema evolution, and interoperability, predefined classes no longer represent their underlying instances faithfully, or are unable to connect the new use cases to the existing instances.

To illustrate the decoupling problem, suppose we would like to integrate two datasets. One dataset contains only the employee table, the other dataset contains only the

contractor table. If there are people who are both employee and contractor, an easier integrating solution would be building a new table to store the employee ID number and corresponding contractor ID number. But if the two tables contain same or similar information about the person such as address or email, we can see that this solution may cause data inconsistency. A new conceptual model including a class of person with subclasses of employee and contractor is more desired. Though this solution brings the issue of data migration and possibly redesign of all the client code.

The concept modeling methodology, property precedence, combined with instance based conceptual modeling, is designed to solve the aforementioned decoupling issues. In this conceptual modeling approach, instances are independently maintained without affiliation with any classes. The existence of an instance is expressed with a set of properties belonging to the instance. The definitions of classes, and the associations between classes and instances, are emergent and context related. Class is defined as a group of instances which satisfy certain constraints on a certain set of properties. In this way, class membership can be inferred, rather than designated. A dynamic class membership is preferred when different domains require different semantics on the definition of classes. Furthermore, losing class membership does not lead to the loss of the instance.

Property precedence complemented instance based conceptual model by further detailing the semantics of the three artifacts of the model – instance, property, and class. Any instance must possess properties. Properties are descriptions of instances, and each instance possesses properties with a unique set of values. In addition, the relationship between properties is defined in terms of property precedence. The definition of property precedence is as follows:

> Let $P_1$ and $P_2$ designate two properties. $P_1$ will be said to precede $P_2$ if for every instance *x* possessing $P_2$, *x* also possesses $P_1$.

By applying property precedence, subsumption relations between properties can be established. Thereafter, implicit properties can be discovered if they are the preceding properties of explicit maintained properties of instances. We demonstrate the idea with two properties: “has legs” and “has 4 legs”. It is semantically correct that anything that has 4 legs must have legs. Hence “has legs” precedes the property “has 4 legs”. And for an instance that explicitly possesses the property “has 4 legs”, we can infer the instance implicitly possesses the property “has legs”.

In instance based conceptual modeling, classes are containers of instances that are relevant to the context, which bridges the gap between data and its context. Since a class is specified in terms of specific properties, class inheritance can be inferred based

on the relation between the two sets of class specific properties. Explicitly, for class *a*, if all of its class defining properties $p_{a1}…p_{an}$ are preceding the class defining properties $p_{b1}…p_{bm}$ of class *b*, then class *a* is the superclass of class *b*. It is easy to verify since the defining properties of class *a* precede the defining properties of class *b*, any instances possess properties $p_{b1}…p_{bm}$ must also possess properties $p_{a1}…p_{an}$, resulting that instances that possess the set of properties $p_{b1}…p_{bm}$ are a subset of the instances that possess the set of properties $p_{a1}…p_{an}$.

After defining the semantic relations between instance, property, and class, Parsons and Wand have enabled the reasoning ability of instance based conceptual models to a greater extent. Besides inferring implicit properties based on property precedence, class, superclass, or subclass membership can be inferred based on certain properties. And conversely from class membership certain properties can be inferred.

One prominent difference between internet datasets and datasets with carefully designed schemas is that internet datasets are decoupled with the context of their use cases. Folksonomies are one of the paragons of internet datasets where users consume the data with different purposes in mind. The framework of property precedence and instance based conceptual model suits the characteristics and the elements of folksonomy datasets, where resources being tagged correspond to instances, and tags

correspond to properties. Once the property precedence relationship between tags are discovered, end users can create any domain-related conceptual models with ease, and accessing resources can be greatly improved compared to current methods. Thus, our research goal is to derive property precedence relationship between tags using existing folksonomy datasets.

## 2.3 Subsumption model

As reasoned above, once the property precedence relationships between tags are discovered, we can introduce emergent ontologies on top of folksonomy datasets. By the definition of property precedence, tag $T_1$ precedes tag $T_2$ if every resource being tagged as $T_2$ is also tagged as $T_1$.

Sanderson and Croft [38] has proposed a model for deriving concept hierarchy from text:

> For two terms, *x* and *y*, *x* is said to subsume *y* if the following two conditions hold,
>
> $$P(x|y) = 1, P(y|x) < 1 \quad (1)$$
>
> where $P(a|b)$ denotes the probability that *a* happens given *b* happens.

The model states that term *x* subsumes *y* if the documents in which *y* occurs are a subset of the documents in which *x* occurs. Because *x* subsumes *y* and because it is more frequent, in the hierarchical relations, *x* is the super concept of *y*.

Compared with property precedence, the two models share a similar statistic formula in deciding hierarchy among entities. Schmitz [39] adopted this statistic model on the Flickr tagging dataset. However, the parameters of the original formula were adjusted to account for the qualities of the dataset:

> For two tags *x* and *y*, *x* is said to subsume *y* if the following two conditions hold,
>
> $$P(x|y) >= t,\ P(y|x) < t, \quad (2)$$
>
> and both tag *x* and *y* should be tagged on at least $D_{min}$ number of documents, and both tag *x* and *y* should be used by at least $U_{min}$ number of users. *t* is the co-occurrence threshold.

Schmitz conducted a series of experiments with varied value of t, $D_{min}$ and $U_{min}$, looking for a balance between too many error pairs or too few meaningful subsumption pairs. A useful range of $U_{min}$ was 5 to 20. And a useful range of $D_{min}$ was 5 to 40. Notably, if the co-occurrence threshold *t* is higher than 0.9, although the number of error pairs is

reduced somehow, the number of overall produced subsumption pairs is drastically reduced. On the Flickr dataset used by Schmitz, the author claimed that meaningful tag relations can be produced when value of threshold *t* was between 0.7 and 0.8.

The author also suggested several future research opportunities including:

- Faceted ontologies
- Integrating with domain specific upper model ontologies
- Community feedback and moderation

The ontology induction method proposed in this thesis incorporates the idea of subsumption for discovering super/sub tag pairs. However, our approach also addresses issues of homonymy, synonymy, and other semantic problems of folksonomies.

## 2.4 Summary

Previous research has not fully addressed the lack of semantics of tags. On one hand, plain text tags are easy to use for end users, but it is hard to infer the tags' semantic meanings from the end users' perspective. Some research relied on co-occurrence to determine tag relations, ignoring the meaning of tag text altogether. Other research mapped tag text to external dictionaries or semantic sources, without considering that tags belonging to a folksonomy may convey different semantics than the external

sources. Essentially they treated external sources as tag ontologies and hoped that the relations between tags are reflected in these sources.

On the other hand, rich format tags can store more metadata describing the tags, but they require additional input from end users, with the result that the end users become reluctant to adopt the new formats. Our proposed approach firstly explores the meaning of tags and how such meanings are represented in a folksonomy dataset. After examine the meanings of individual tags we can construct tag ontologies based on semantic relatedness of tags.

# Chapter 3. Folksonomies as data management tools

In the previous chapter, we discussed that tagging system users are reluctant to adopting more structured tag formats which requires users to make an effort to identify the meaning of tags. In this chapter we argue that tagging system users have already exerted effort for maintaining an unambiguous tagging vocabulary in plain text format.

Folksonomies generally consist of at least the following three sets of entities:

- Users are the ones who assign tags to online resources in social tagging systems. In some literature, they are referred as actors, corresponding to the terminology often used in social network analysis.
- Tags are keywords chosen by users to describe online resources of interest. Depending on the systems, tags can be single words, phrases, or combination of symbols, number and alphabets. Other than plain text tags, some tagging systems allow that users can use more complex constructs such as RDF entities as tags. Tags are referred as concepts in some analyses.

- Resources are the objects being tagged by the users in the tagging systems. Different tagging systems are generally designed to cater to a particular kind of objects. Tagging system designed for books, photos, academic papers, documents, URLs are the most popular ones. Resources are also referred as instances, objects, or documents by some authors.

In tagging systems, users interpret resources by assigning most meaningful tags to the resources. Hence the meaning of tags is the key for any analysis on folksonomies. Golder and Huberman [40] identified three major problems which dilute the connections between tags and their intended meanings:

- *Polysemy and homonymy*. Different words share the same spelling, or a word has several meanings. The presence of polysemy and homonymy renders free text tags prone to ambiguity.
- *Synonymy*. The inconsistent usage of synonyms makes it difficult to be sure that relations between tags can be clearly defined. Specifically, the semantic boundaries between a given word and its synonyms vary among different cultures, communities, and domains. For instance, the words App and Application seem interchangeable in most of the contexts, but in certain domains apps are a strict subclass of applications.

- *Basic level variation*. Experiments demonstrate that, when asked to identify dogs and birds, subjects used “dog” and “bird” more than, say, “beagle” or “robin”, and when asked whether an item in a picture is an X, subjects responded more quickly when X was a “basic” level [41]. Basic level denotes the most useful specificity of a person recognizing a concept. In the bird and dog experiment, animal experts demonstrated basic levels that were at levels of greater specificity than non-experts. Dog experts might consider using “beagle” rather than more general term “dog” as description of a picture. Knowledge differences between tagging system users brings basic level variations into folksonomy. Domain experts tag with more specific terms than average users, but the more specific terms are less popular and used only by sub-communities of expert users. Basic level variation is rooted in human cognition and learning processes. Our understanding of concepts evolves when we recognize new features or update existing features of the concepts.

These problems make determining the meaning of tags, and establishing semantic relations among tags in folksonomies especially difficult. The problems of the presence of idiosyncratic tags such as misspelled, acronymic, or compound word tags are also common in collaborative tagging setting. Nevertheless, individually these tags are hardly used by enough users to be considered meaningful. If an idiosyncratic tag, whether in

common language form or not, is used by a large number of users, then the text of the tag may in fact bear a special meaning relevant to certain domains.

## Folksonomy as a collection of personomies

Our approach remedies the above three problems by analyzing so called “personomies”, which are user-level tag vocabularies and usage patterns. Heckner et al. [35] surveyed users of author-created tagging system such as Flickr and YouTube, as well as consumer-created tagging system such as Delicious. The results showed that the motivations behind users of consumer-created tagging system are more inclined toward retrieving the content for later reference, rather than facilitating other users to discover their items. In other words, the users of consumer-created tagging system treat the service more as a personal information management tool. Wetzker et al. [34] observed that unlike folksonomy as a whole, personomies are void of the three major problems mentioned above. For the goal of using tagging systems to save content for possible later use, collaborative tagging system users tend to avoid the use of synonymous tags. Furthermore, polysemy and homonymy are largely absent in the user-level tag vocabularies so that each tag employed by a single user conveys a dedicated meaning.

Our approach takes advantage of the well-maintained user-level tag vocabularies to identify different meanings of a tag. More concretely, based on the previously mentioned observations, we formulate several assumptions on folksonomies:

*Assumption 1*: In most cases, any single tag used by an individual user conveys a dedicated and unaltered meaning, and all resources being tagged with this tag by the same user manifest this unique meaning.

In folksonomies, the distribution of tags for resources follows power law distribution [7]. That is, although there is no restriction on what tags can be assigned to a given resource, a majority of taggers voluntarily assign some of the same tags to that resource. The phenomenon of assigning the same tags on a resource by different users exhibits that taggers generally agree the meaning of the tags they employed, which leads to our second assumption:

*Assumption 2*: For a specific tag that is assigned by multiple users to a single resource in a folksonomy, the meaning of this tag is the same or similar to the majority of the taggers who assigned the tag to the resource. In other words the tag conveys same or similar meanings in each of the user's vocabulary.

The two assumptions are supported by evidence mentioned in various sources [34, 40, 42-44]. Furthermore, we believe that tagging system users are aware of the semantic relations between tags within their personomies, which leads to the third assumption:

*Assumption 3*: Tagging system users tag with the concepts that they are more familiar with. Hence, domain experts tag with more specific concepts than general users.

The third assumption is stated in [40] as basic level variations, which is stated as one of the major problems for folksonomy users reaching consent on the meaning of tags. Tanaka and Taylor [41] concluded from a series of experiments that the basic levels of domain experts in their domain of expertise are often at the levels which subjects with novice knowledge of the domain considered as subordinate levels. The experiments showed that subjects in their domain of expertise can (a) differentiate subordinate level categories as effectively as basic level; (b) identify objects using subordinate level names as frequently as basic level; and (c) categorize with subordinate level as fast as basic level.

The domain of expertise variance of tagging system users can cause basic level variations. However, such variance on tag assignment can help us identify the relations between more abstract concepts and more concrete concepts, since more concrete

concepts are used only by users who possess the domain knowledge. In this case resources that are tagged with the more concrete concepts may or may not be tagged with the more abstract concept, because domain expert users who used more concrete concepts as tags view these as more natural than more abstract concepts.

The first two assumptions are the foundations for the method for tag meaning disambiguation in the next chapter. The third assumption will help us develop the method to discover super-concept/sub-concept relations between tags in chapter 6.

## Dataset

In the following chapters we carry out a series of experiments to test and verify the methods proposed. The dataset for the experiments is acquired from the most popular tagging website Delicous.com. Delicious.com (henceforth referred to as “Delicious”) is a tool to organize web pages. It is a social bookmark manager that allows you to easily add sites you like to your personal collection of links, and to manage and organize your collection with freely assigned keywords for each link that you added. Delicious is not unique as a way to manage bookmarks, but its emphasis on user added keywords as a fundamental organizational tool is distinctive. These keywords, which are referred to as “tags” on the site, allow users to describe and organize content with any vocabulary they choose.

To use the tagging service, when browsing a web page which users would like to add to Delicious, they save the link along with any tags they want to associate with the page. Later the users can retrieve the links that they have saved by browsing the associated tags. Furthermore, browsing any specific tag allows the user to access other users' saved links which are annotated with that tag.

Thanks to the extensive research that has been done with Delicious folksonomy, existing Delicious datasets are made available to researchers online. We have obtained our Delicious dataset from Tagora project [45]. In the dataset there are 5,860,000 tag assignments, which are user, resource, tag triples. There are 1,310,000 distinct resources, 38,745 distinct users, and 192,649 distinct tags.

# Chapter 4. The meanings of a tag

## 4.1 Manifesting set of tags

By exploring the aforementioned assumptions presented in folksonomies, we derived a method to elicit different meanings of a given tag. For any given tag in a personomy tag vocabulary, the meaning of the tag is stable across all the resources being tagged with the tag by the specific user. Hence in the personomy, resources that are tagged with the same tag express some degree of similarity. If a certain number of users assigned the same tag on both of two resources, then these resources are said to be ***tagged similar*** on the give tag. The relation is named ***tagged similarity***. The number of users sharing the tag among the two resources is the ***strength of the tagged similarity.***

Taking into consideration that folksonomies often arise from different context, the strength of tagged similarity relations should be adjusted to adapt to the characteristics of the folksonomy under study to counter noise or spam.

Assumption 2 tells us that users generally agree on the meaning of a tag when the tag is assigned to the same resource by these users. And if some other resources are deemed

tagged similar on a given tag with this resource, then the meaning of the tag is similar on all the resources that are tagged similar. Hence a group of resources can be connected together with tagged similarity relations on a given tag and they manifest a consistent meaning for the tag. In this way we can cluster resources being marked with the same tag into different clusters, with each cluster manifesting a consistent meaning of the tag. We formally define such clusters as:

For a given tag, if (1) a set of resources are annotated by this tag, and (2) for any two resources belonging to the set, there exists one or a number of tagged similarity relations on this tag that connects them, and (3) there is no other resource out of the set that is tagged similar on this tag to any resources in this set, then the set of resources is said to be manifesting a similar meaning of the given tag. The resource set is named ***manifesting set*** on the given tag.

Notably, the assumptions ensure a higher degree of meaning similarity for a given tag within a manifesting set, but it does not verify that meanings of the tag among different manifesting sets are necessarily different in a significant way. Therefore different manifesting sets of a given tag may or may not reflect different linguistic meanings of the term representing the tag. Rather, they reflect the user communities' interests of

associating the tag to a group of resources which are of no interest for other communities.

To better illustrate the idea of manifesting set, consider the case of the tag ‘Apple’. ‘Apple’ conveys at least two meanings, so that it might be used to tag resources which relate to either technology products or a kind of fruit. If ‘Apple’ is used to annotate resources *R1*, *R2*, *R3*, and *R4*, and user *U1* tagged on *R1*, *R2*; *U2* tagged on *R2*, *R3*; *U3* tagged on *R4*. We can see that *R1*, *R2*, and *R3* may manifest a similar meaning of the tag ‘Apple’, while *R4* might manifest a different meaning. Apparently if there are more users who make *R1*, *R2*, and *R3* associative tagged similar without *R4*, then it is highly probable that *R4* is manifesting a different meaning of the tag in contrast to the set of *R1*, *R2*, and *R3*.

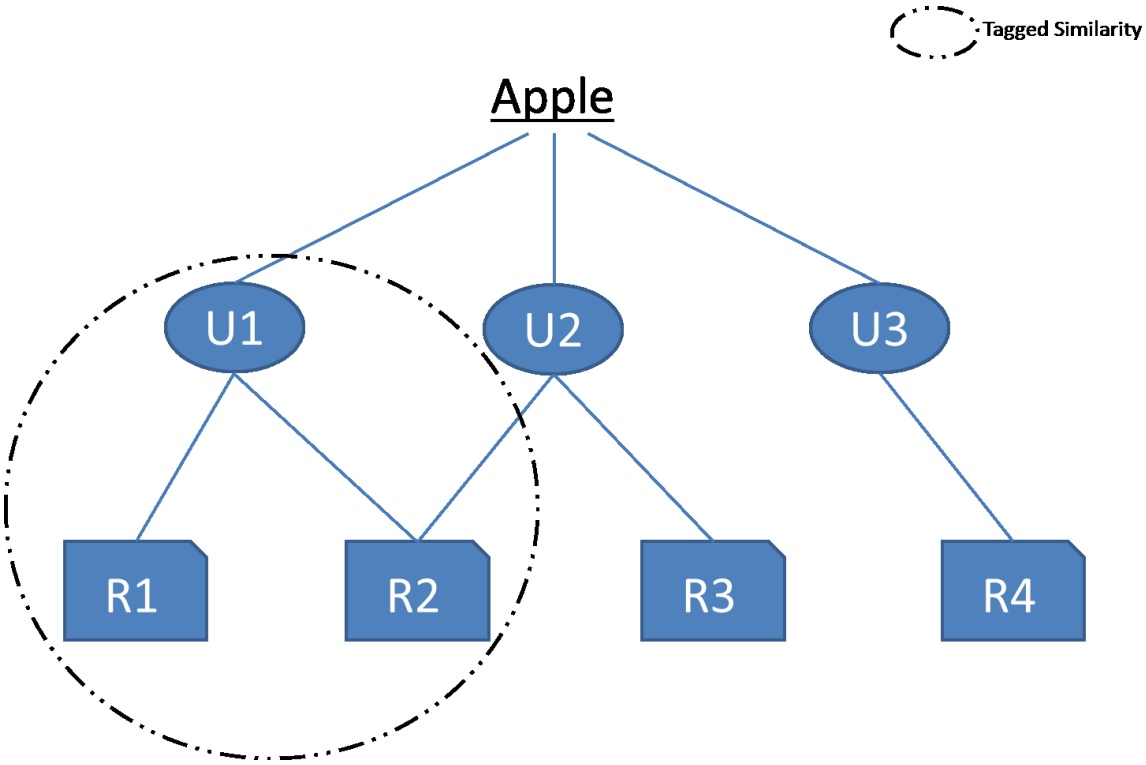


Figure 4-1 Resources R1 and R2 manifest similar meaning

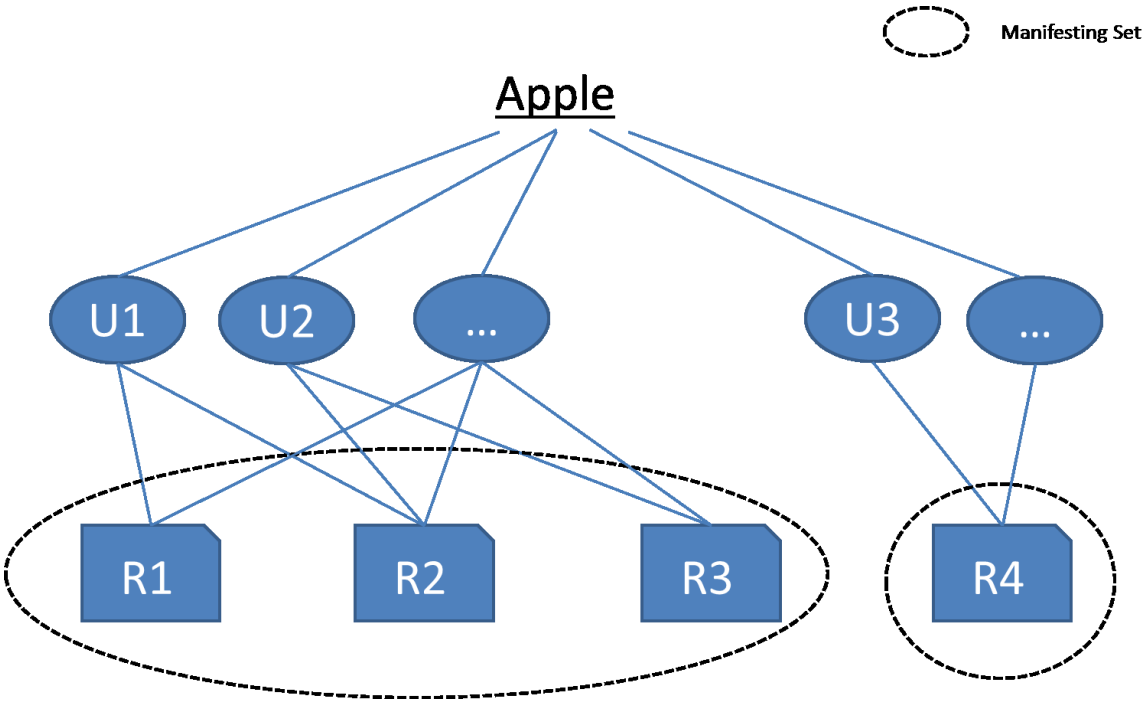


Figure 4-2 Resources R1, R2, and R3 manifest similar meaning

## 4.2 Experiments on the meanings of manifesting sets

In this section we implemented an algorithm for finding manifesting sets of given tags, and queried the most used tags within manifesting sets to illustrate that manifesting sets expresses homogenous meaning on given tags.

## Implementation

Firstly, we implemented an algorithm to cluster resources associated with a given tag into manifesting sets based on the tagged similarity. As explained in the previous chapter, we cluster manifesting sets according to the tagged similarity strength. The algorithm is given in Figure 4-3:

```
1   MSC(List<List<Set<Resource>,Set<User>>> clusters):
2       //Initial input is a list of resources and their corresponding users
3       //So each resource set only contains one resource
4       for(int outer = clusters.size()-1;outer>=0; outer--)
5           for (int inner = outer-1; inner >= 0; inner--)
6               int similarityStrength = clusters.get(inner).get(1)
                                            .intersect(clusters.get(outer).get(1))
                                            .size()
                    // clusters.get(inner).get(1) returns
                    //the set of users associated with the cluster indexed by "inner"
7               if (similarityStrength >= threshold)
8                   clusters.get(inner).merge(clusters.get(outer);
9                   clusters.get(outer).remove();
10                  break;
11      return clusters
```

Figure 4-3 An algorithm for clustering manifest sets

For any given tag which is associated with *n* resources, the complexity of the algorithm is $\Theta(n^2)$, which is comparable with the complexity of single link clustering algorithm [46].

## Examine the homogeneity of manifesting sets

By using the algorithm described in the previous section, we obtained manifesting sets of several tags. To decide whether the resources contained in manifesting sets share similar topics, we count the five most used tags on all resources within each manifesting

set of the given tags, excluding the given tags themselves, to see if the co-occurring tags convey close related meanings. We run the experiments with a minimal tagged similarity strength of 2, which means only when at least two users who tagged two resources with the same given tag do we consider the two resources tagged similar. We use the similarity strength of 2 instead of 1 is to offset idiosyncratic usage of tags. Also we excluded any manifesting sets that contain less than 10 resources so that the meanings conveyed by the manifesting sets reflect the perspectives of at least a number of users.

| | | |
|---|---|---|
| Tag: xp | Manifesting Set 1 | windows, software, computer, reference, tools |
| | Manifesting Set 2 | programming, agile, development, java, software |
| Tag: language | Manifesting Set 1 | programming, ruby, perl, lisp, python |
| | Manifesting Set 2 | chinese, japanese, reference, writing, oriental |
| Tag: health | Manifesting Set 1 | sleep, science, reference, lifehacks, life |
| | Manifesting Set 2 | gesundheit[1], de, med, dgk-webs, impfung[2] |

[1] German word for health.

[2] German word for vaccination.

| | | |
|---|---|---|
| Tag: interview | Manifesting Set 1 | job, jobs, career, work, resume |
| | Manifesting Set 2 | art, illustration, design, painting, portfolio |
| Tag: opera | Manifesting Set 1 | css, web, browser, firefox, design |
| | Manifesting Set 2 | music, entertainment, classical, theater, netradio |
| Tag: fish | Manifesting Set 1 | aquarium, science, reference, aquaria, gallery |
| | Manifesting Set 2 | health, food, seafood, mercury, diet |

**Table 4-1 Most used tags in manifesting sets of given tags**

As shown in Table 4-1, for some tags, different manifesting sets reflect that the tag terms are homonyms. But in other cases the meanings of the tags may or may not be comparable to the linguistic meanings that the terms bear. Specifically some tags are abbreviations or domain specific terms. The meanings of the tags are formed by the usage of the tags on certain resources that the communities of tagging system users consider relevant to the tags. Different communities have different interests, hence judge the relevance differently. For example, since the word “fish” is not polysemous, in this case the two manifesting sets represent different interests of associating the tag

“fish” on different sets of resources. We might infer that one community is composed of aquaculturists, and the other is of gastronomists.

We also plotted the two manifesting sets of the tag “xp” as a graph to visualize the topics of all resources within the manifesting sets to examine the homogeneity of the manifesting sets. In the plot below, we plotted resources that are tagged with the tag “xp” as vertices. Although all of the resources are tagged with the same tag, they may be tagged by different users. So if a number of same users tagged both resources with the tag “xp”, meaning there were tagged similarities between the two resources, an edge was plotted to connect the two vertices.

Furthermore, for each resource that represented as a vertex, we query the folksonomy to find the most used tag that is associated to this resource (Although all the resources are tagged with the term “xp”, this does not necessarily mean that the most used tag on the resources is “xp”. If for some resources, the most used tag is “xp”, we query the secondly most used tag). Figure 4-4 presents the most used tag for each resource near the vertices.

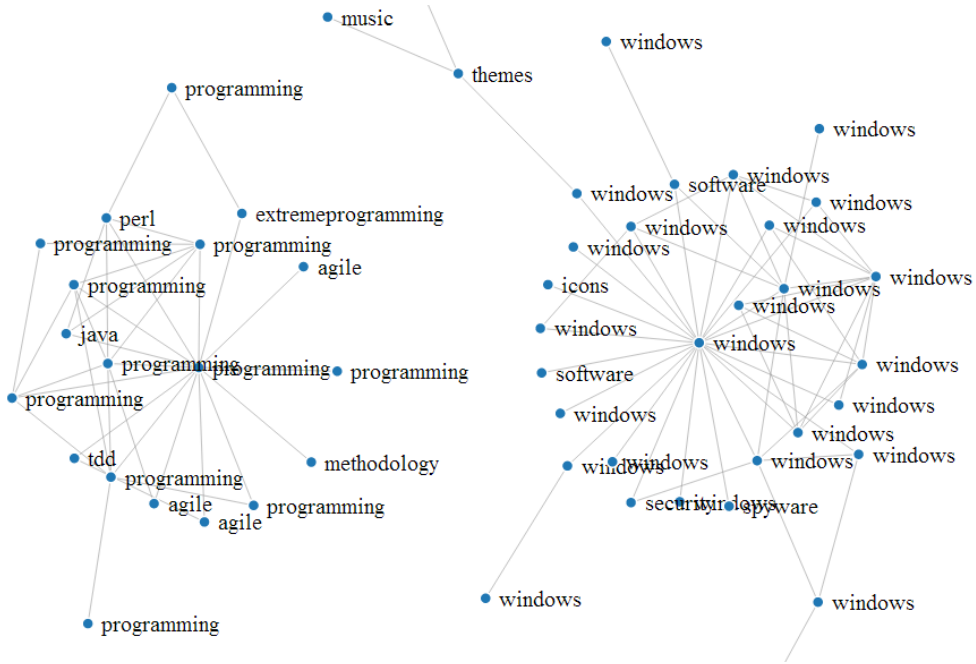


Figure 4-4 The manifesting sets of tag “xp”

In this case we can clearly discern two manifesting sets. The manifesting set at the right may reflect resources related to manifesting set 1 of tag “xp” shown in Table 4-1, with the most tags within the set as “windows, software, computer, reference, tools”, while the manifesting set at the left may represent resources related to manifesting set 2 of tag “xp” as “programming, agile, development, java, software”. The manifesting sets contain resources that are semantic similar in term of the meanings of the given tag.

## 4.3 Conclusion

The results from the experiments on discovering manifesting sets by clustering with tagged similarities provided evidence that manifesting sets of tags convey similar semantic meaning of the tags.

# Chapter 5. Validation of the homogeneity of manifesting sets

## 5.1 Are the resources in manifesting sets quantitatively homogeneous?

Essentially, the process of discovering manifesting sets of a given tag is to cluster resources that are associated to the tag into groups which exhibit similar semantic meanings. Comparable with most clustering algorithms in the domain of data mining or machine learning, the partitions of resources with tagged similarity on given tags require some validation mechanism to demonstrate its correctness [47]. In the previous chapters, we visually described the most frequent co-occurring tags that appear within manifesting sets. In this chapter, we develop a more rigorous method to quantitatively measure the homogeneity of a group of resources in a folksonomy.

Evaluation of clustering algorithms is a challenging task. Several forms of validity criteria are designed to evaluate clustering results in various circumstances. criterion selection depends on the kind of the clustering problem and available information. Generally these clustering validity criteria can be grouped as [48]:

(a) External criteria, where results are evaluated with information that is not made available to the clustering algorithm.

(b) Internal criteria, or structural properties, where results are evaluated using the same information that is available to the clustering algorithm.

Several well-known validation criteria are external criteria, such as F-measure, Entropy, Normalized mutual information, and Purity. However, these validation criteria require human judgment to classify resources into predefined categories as gold standards, which requires a good level of inter-judge agreement [46]. For the case of producing manifesting sets of tags in folksonomies, as the semantic meanings of a given tag may not be equivalent to its linguistic meanings, the task of predefining categories for resources classification is beyond the capability of human experts. Taking the tag "health" for example as shown in Table 4-1, the two manifesting sets of this tag express similar linguistic meaning of the term "health". Nonetheless the two manifesting sets show the different interests of two Delicious.com user communities who are probably different language speakers. It is hard, if not impossible, to predict that the resources that are associated with this tag can be grouped as such, but not resources that are associated with other tags.

Most of the internal criteria utilize the concept of diameters of the clusters and the distances between clusters [49]. In this chapter, we present a method to determine the centrality of manifesting sets using the vector space model.

The vector space model, or term vector model, represents text documents as vectors of identifiers [50]. In folksonomy, the tags that are associated to a resource can be naturally represented as a vector of terms. These vectors convey the topics of the underlying resources. For a group of resources, the individual vectors can be combined to form a ***combined tag vector*** reflecting the semantics of the group. Yet if the resources in the group convey ideas that are hardly related, then the combined tag vector will contain more tags than a group of resources with comparable size but of which the resources are closely related. In a sense the length of the combined tag vector of a group of resources reflects the relatedness of the semantics of the resources.

However, some online resources may involve more topics than others. To illustrate this, we randomly selected 2 resources from the experiment dataset. The tags associated to these two resources are shown in Table 5-1:

| Resource | Tags assigned to the resource | Number of tags |
|---|---|---|
| 1 | bittorrent, p2p, wired, media, internet, article, filesharing, software, technology, interview, news, etc. | 35 |
| 2 | flickr, iphoto, mac, osx, photography, photo, plugin, software, apple, etc. | 12 |

Table 5-1 Two randomly selected resources and their assigned tag counts

In a folksonomy, some resources can contain more topics than others. It could be that the some resources such as articles refer to a greater number of topics. Hence, a group of resources in a manifesting set may refer to a great diversity of topics even though they share some common topics. In this case, the combined tag vector of some manifesting sets might be lengthier than others, although the resources are indeed related to each other on several shared topics.

In order to decide if a group of resources shares a common topic expressed by some tags, all they need to share is a number of common tags. So instead of measuring the length of the combined tag vector, which includes all the tags appearing on each and every resource of the group, to denote the semantics of the group, a subset of the combined tag vector is adequate. This subset should contain sufficient topics that, for each resource in the group, some of its tags are included. We term the subset of the

combined tag vector the ***diameter*** of the cluster of resources. The tags contained in the diameter vector are the common topics shared within the group of resources.

As an example, for the two resources shown in Table 5-1, the combined tag vector may have a length of 35 + 12 - number of common tags. But they share at least one common tag “software”. In this case they are related with the topic “software”. So the diameter of the two resources is 1.

We have discussed that tags in the form of plain text may convey multiple meanings. Therefore, if there is one common tag in a group of resources, the meaning of the tag could be homonymous, rendering the relatedness of the resources less reliable. Research on term sense disambiguation, e.g. [51-53], has found that the co-occurrence of two terms provides enough information to determine the meaning of each term. So we decided that the subset of combined tag vector should contain at least two shared tags for every resource in the group. This implies that the diameter we developed has a minimum value of 2. Figure 5-1 illustrates the concept of diameter.

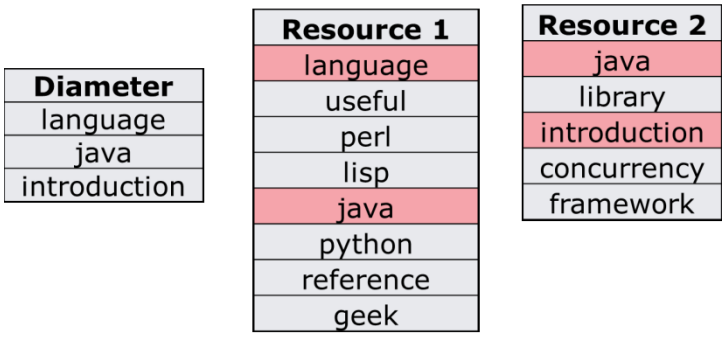


Figure 5-1 Example diameter and its relation with individual resources

In the following section, we present experiment results on using the diameter to measure the relatedness of groups of resources.

## 5.2 Evaluating diameter measure for cluster centrality validation

We implemented the above diameter measure to test the homogeneity of manifesting sets of given tags. In our experiment, we allow that the diameter vector shares two common tags with a minimum of 90% of all resources in the resource group. This is because there are resources on which only a few tags are assigned. And those few tags may be very idiosyncratic or personal, which appear only in few users' tag vocabulary. In the extreme cases, some resources are only assigned with one tag. Hence to allow the diameter to share at least two common tags with these resources, the diameter must contain more idiosyncratic tags. On the other hand, the ultimate purpose of constructing a diameter is to use the length of the diameter vector to denote the

relative relatedness of the group of resources. Therefore it is the vector length difference that is indicative when comparing whether some resources are more similar than others.

We set up test scenarios to validate the correctness of using diameter for cluster centrality validation. First we tried to construct diameter vectors on groups of random selected resources. The results are shown in Table 5-2.

| | Number of resources | Combined tag vector length | Diameter length |
|---|---|---|---|
| First group of randomly selected 100 resources | 100 | 302 | N/A |
| Second group of randomly selected 100 resources | 100 | 299 | N/A |
| Group of randomly selected 200 resources | 200 | 563 | N/A |
| Group of randomly selected 100 resources each of which has at least 2 tags | 100 | 442 | 420 |

**Table 5-2 Applying diameter measure on randomly selected resources**

In the first three groups of randomly selected resources, we are unable to construct the diameter vector. Further investigation revealed that, for the first 100 resources, 31 are

associated with only one tag. As the diameter vector cannot share two tags with at least 90% of resources in these random selected resources groups, we are unable to construct the diameter vector in these cases. For the group of 100 randomly selected resources, each of which has at least 2 tags, the length of the diameter is comparable with the length of the combined tag vector. So to ensure that the diameter vector shares at least two tags of all tags assigned to each resource, the diameter vector must contain almost all tags assigned to the group of resources. It indicates that the resources in the group have few overlaps in the tags assigned to them.

Next, we test diameter on several manifesting sets obtained using the method described in the previous chapter.

| Manifesting sets of tag | Number of resources | Combined tag vector length | Diameter length |
|---|---|---|---|
| “jobs” | 91 | 1646 | 13 |
| “downloads” | 74 | 2309 | 3 |
| “freeware” | 278 | 4777 | 2 |
| “software” | 4399 | 22808 | 11 |

Table 5-3 Diameter measure on several manifesting sets

The results (Table 5-3) show that the lengths of the diameters are an order of magnitude lower than the lengths of the combined tag vectors. In addition, despite the size of the manifesting sets, the lengths of diameter vectors stay relatively constant (and small) in all cases. We expect the length of diameter vector to be small in these cases because resources in a manifesting set should share one common topic.

In the next experiment we randomly separate a manifesting set of the tag “tech” into two portions and test the diameter measure on them.

| | Number of resources | Combined tag vector length | Diameter length |
|---|---|---|---|
| Whole manifesting set | 951 (100%) | 16023 | 5 |
| First portion of the manifesting set | 218 (22.9%) | 5554 | 6 |
| Second portion of the manifesting set | 733 (77.1%) | 13929 | 5 |

Table 5-4 Diameter measure on portions of manifesting set

The length of the diameter vector of the whole manifesting set is small as expected. For each portion of the manifesting set, we expect that the group of resources still exhibits high degree of similarity. Table 5-4 shows that the lengths of diameter vector of each portion are comparable, which expresses the resources within each portion are similar.

These results indicate that diameter is a reliable measurement to test the semantic centrality of groups of resources in folksonomies. In the following section and in the next chapter we will take advantage of this measurement to explore semantic relations between different tags.

## 5.3 Similarity evaluation between groups of resources

Now we have defined the diameter of a group of resources in folksonomy, we can subsequently define the semantic similarity measure between different groups of resources. For two groups of resources, a seemingly promising approach is to combine the two groups together to form a larger group and apply diameter analysis, as demonstrated already in table 5-4. However, as in this thesis the purpose of clustering is to discover sets of resources that convey homogeneous meanings, the topics that the specific sets of resources convey should not be ignored during distance calculation. Otherwise the similarity measure could bring false results.

Recall that the diameter vector contains common tags shared by a group of resources. These tags are the most shared tags within the group, which denote the main topics of the group. If we have two groups of resources with different main topics, after combining the two groups and producing the diameter vector for the new group, we may find that the new main topics of the group is actually different from both of the initial resources groups. For example, our experiment shows that if the manifesting set of tag “entertainment” is combined with the manifesting set of tag “iraq”, the new group of resources produces a diameter vector with main topics about “news”, “blog”, and so on, which did not related to either “entertainment” or “iraq”. So the newly formed group of resources are somehow similar in the sense that they are related to the

topic “news”, although we expected that the similarity between manifesting set of “entertainment” and “iraq” is low. In this situation the length of diameter of the combined resources group cannot reflect this fact.

Indeed as resources in folksonomies always span several topics, they can exhibit relatedness to several topics. Combining two separated resources sets enables the new set to exhibit topic dimensions that are hidden previously. Thus, we concluded that by combining two different resource groups, the diameter measure of the new resource group cannot faithfully reflect the semantic similarity of the two resources groups. In extreme cases if one group of resources is selected randomly, which may cover a significant number of less related topics, combining another group of resources will not influence the length of diameter at all.

Instead, we employed Jaccard similarity index [54] on the diameter vectors of the two resources group as the semantic similarity measure. Jaccard similarity in our case is defined as the length of the intersection of two diameter vectors divided by the length of the union of the two diameter vectors. The index is between 0 and 1 inclusive, and the closer the index is to 1, the more related the two groups of resources. We carried out some experiments in the next section.

## 5.4 Experiment on assessing similarity between two groups of resources

In this section, we present some experiment results on using Jaccard similarity score of two diameters as the quantitative measure of similarity between two groups of resources.

| Group 1 | Group 2 | Jaccard similarity |
|---|---|---|
| Partial manifesting set of tag “tech”, size 722 | Partial manifesting set of tag “tech”, size 229 | 0.833 |
| Manifesting set of tag “iraq” | Manifesting set of tag “entertainment” | 0.083 |
| Manifesting set of tag “entertainment” | Randomly selected 100 resources, each of which has at least 2 tags | 0.009 |

Table 5-5 Jaccard similarity measures of several test setups

From the above examples we can see the highest Jaccard similarity is obtained from two groups of resources which belong to the same manifesting set of the tag “tech”. It also shows that the tags “tech” and “computer” are more related in the folksonomy than the

tags “entertainment” and “iraq”. Lastly the similarity between manifesting sets and randomly selected groups of resources are very low.

## 5.5 Conclusion

In this chapter, we developed a novel measurement to test the homogeneity of resources groups. Diameter of resources groups can also be employed to reflect semantic relatedness between resources groups. In the next chapter, we utilize this measurement to evaluation to what degree two tags are semantically related.

# Chapter 6. Tag precedence relations

As the semantics of tags are expressed on the resources marked by the tag, similarity scores between manifesting sets of different tags are a reliable measure to decide if the two tags are semantically related. But to calculate the score, one database query needs to be executed to collect tag vectors for each and every resource in a group. For any non-trivial folksonomy dataset, such operations are inefficient when the relations of a large number of tags are expected to be discovered. In this chapter we present an alternative approach for solving this problem.

## 6.1 Using manifesting sets for tag relation discovery

If a tag is a super concept of another tag, we would expect that the super concept tag is associated with a greater number of resources than the sub concept tag. However, because of the basic level variance problem stated in chapter 3, it is not guaranteed that all the resources being associated with the sub concept tag are also associated with the super concept tag. Domain experts tend to tag with more concrete concept tags. For some resources that are of interests of domain experts, they may be out of the interests of users who tag with more abstract concept tags. In turn the resources set of the sub

concept tag will not be a subset of the resources set of the super concept tag. We may still discover super concept tag and sub concept tag relations if the following two conditions are satisfied:

(a) The manifesting set of the super concept tag is larger than the manifesting set of the sub concept tag.
(b) The size of the intersection of two manifesting sets is comparable to the sizes of the two manifesting sets.

According to the two conditions, we developed a subsumption model to discover super concept/sub concept tag relations, or ***tag precedence relations*** where the super concept tag is named ***preceding tag***, and the sub concept tag is named ***preceded tag***.

The subsumption model is stated as:

For two tags $t_1$ and $t_2$, and their manifesting sets $ms_1$ belonging to $t_1$ and $ms_2$ belonging to $t_2$, if:

(1) The size of $ms_1$ is larger than the size of $ms_2$,
(2) And the size of intersection of $ms_1$ and $ms_2$ is larger than $p$ times the size of $ms_2$,

then $t_1$ and $t_2$ forms a tag precedence relation. $t_1$ is the preceding tag and $t_2$ is the preceded tag.

In this subsumption model the parameter $p$ can be adjusted according to the desired semantic similarity between the two tags. If the intersection is only a small portion of the two manifesting sets, then the relatedness of two tags will be low. The relatedness can be quantitatively measured by the Jaccard similarity of the two manifesting sets' diameters.

With a collection of tag precedence relations discovered, we can construct an emergent tag ontology based on these relations, the semantics of which represents the entire folksonomy. In the next section, we present some experiment results on applying the subsumption model with test datasets.

## 6.2 Experiment on discovering tag precedence relations with subsumption model

In our experiments, we applied the subsumption model introduced in the previous section on a collection of tags and their associated resources and users. We randomly selected 100 tags along with their tag assignment triples from the dataset for this experiment. The 100 tags cover 1,055,314 tag assignment triples, which are assigned by

29,123 users to 421,469 resources. Considering that in tagging systems there are highly idiosyncratic tags, we ensured that the 100 tags are used by at least 10 users and being applied to at least 20 resources [39].

First we experimented on $p = 0.5$. So the intersection is more than half the size of the manifesting set of the preceded tag and less than half the size of the manifesting set of the preceding tag. We obtained 32 tag precedence relations. They are illustrated in Figure 6-1. Tags appearing in the tag precedence relations are depicted as vertices, arrows are drawn from the preceding tag to the preceded tag of tag precedence relations.

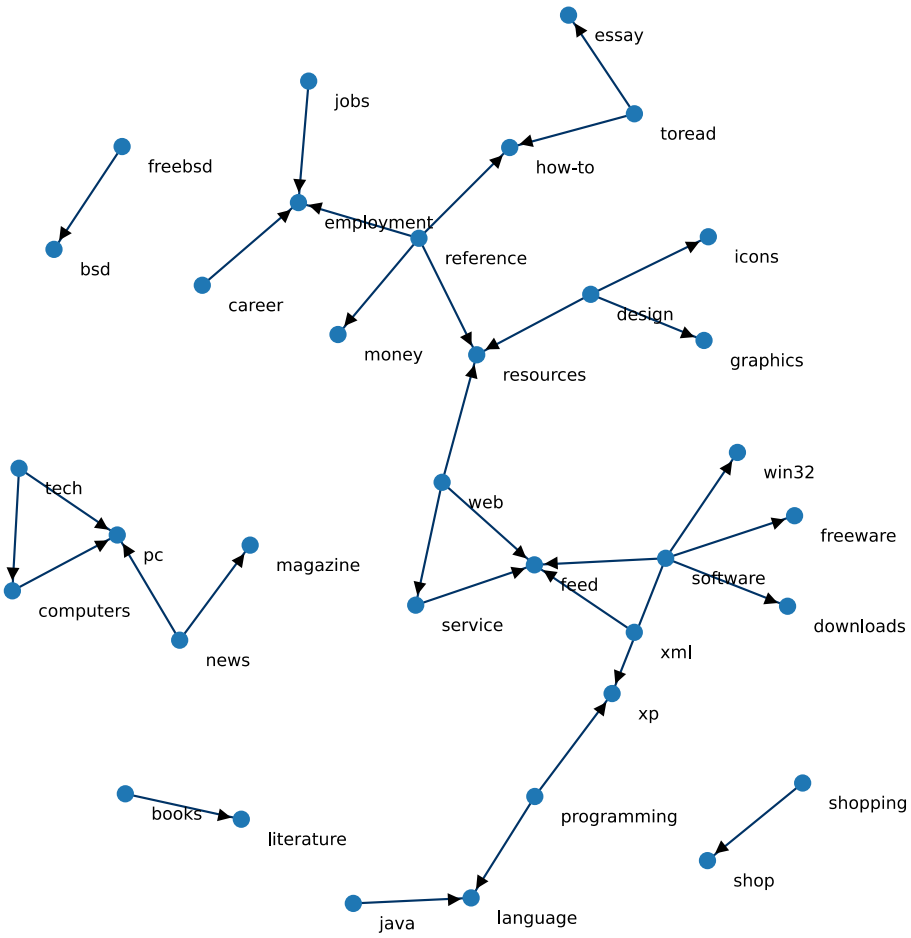


Figure 6-1 Tag precedence relations discovered in the 100-tag dataset

We also examined the similarity between several manifesting sets based on which the tag precedence relations are formed.

| Preceding tag | Preceded tag | Jaccard Similarity |
|---|---|---|
| “tech” | “computer” | 0.5 |
| “books” | “literature” | 0.167 |
| “freebsd” | “bsd” | 0.125 |
| “toread” | “essay” | 0.167 |

**Table 6-1 Jaccard similarity measures on tag precedence relations when $p = 0.5$**

The similarity measures obtained from the tag precedence relations are comparable with results shown in the previous chapter, where two groups of resources from the same manifesting set have the Jaccard similarity of 0.833, and very unrelated tags have similarity smaller than 0.1. To better illustrate the relatedness of tags in a tag precedence relations, we randomly selected 3 pairs of unrelated tags and calculated their Jaccard similarity on their largest manifesting sets.

| Tag 1 | Tag 2 | Jaccard similarity |
|---|---|---|
| Manifesting set of tag “ebooks” | Manifesting set of tag “w3c” | 0.024 |
| Manifesting set of tag “jobs” | Manifesting set of tag “algorithms” | 0.15 |
| Manifesting set of tag “audio” | Manifesting set of tag “searchengines” | 0 |

Table 6-2 Similarities of three randomly selected tag pairs

The average of the similarities on the three randomly selected tag pairs is 0.058, while the sampled tag precedence relations all have higher similarities than this average.

We varied the parameter of $p = 0.3$ and ran the experiment again. This time the number of tag precedence relations discovered is 80, including all tag precedence relations discovered with stricter $p$ value. With $p = 0.1$, the number of tag precedence relations is 244, including all relations discovered before. We then proceeded to sample some newly discovered tag precedence relations and calculate their Jaccard similarity, as shown below.

| Preceding tag | Preceded tag | Jaccard Similarity |
|---|---|---|
| “computers” | “themes” | 0.077 |
| “video” | “dvd” | 0.136 |
| “toread” | “career” | 0 |

Table 6-3 Jaccard similarity of tag precedence relations when $p = 0.1$

Generally speaking, with a lower *p* value, more tag precedence relations can be discovered, but average Jaccard similarity measures between these relations will be lower. However, it is still possible that the meanings of two tags are related even the manifesting sets share a small number of common resources. In these cases the tag precedence relations formed based on a low *p* value still have acceptable Jaccard similarity score. A conservative approach for constructing tag ontology would be using a lower *p* value to form tag precedence relations and then prune off some of them which have similarity scores lower than a preset threshold. Again, in folksonomies the relatedness between resources is a rather subjective matter, so using different *p* values in different domains might be also appropriate.

## 6.3 Conclusion

In this section we combined the tools developed in the two previous chapters to discover super/sub concept relations among tags. A series of experiments were carried out to demonstrate the effectiveness of finding tag precedence relations. Further

analysis on the results showed that with appropriate *p* value, tag precedence relations exhibit above average similarities. Identifying related tags provides foundations for tag ontology construction as well as enables better information retrieval in folksonomies.

# Chapter 7. Applications of tag precedence relations

By identifying all tag precedence relations in a folksonomy dataset, we can construct a tag ontology describing the semantic relations between tags in the folksonomy. Tag ontology can improve the information quality of folksonomy in several ways, both from the perspective of tagging system users and other systems that utilize tags as metadata for the associated resources.

## 7.1 Tag query enrichment

For tagging system users, finding resources related to a tag of interest is one of the major tasks of using the tagging system. Currently tag querying can only return resources that have been assigned with the exact tag. If the tag of interest is the preceding tag in several tag precedence relations, we expect that the preceded tags are the sub concept of the preceding tag. Hence it is reasonable to include resources assigned to the preceded tags when querying the preceding tag. In this way we alleviate the problem of basic level variance where some resources have been tagged with more concrete terms but the more abstract tags which are also appropriate have not yet been assigned to them.

For example, from the 100 tags dataset we have discovered 32 tag precedence relations. If we query the tag "design" in this experimental dataset, we may include the manifesting sets of all its preceded tags which are “icons”, “graphics”, and “resources”. We only include the manifesting sets of these tags that form a subsumption relation with some manifesting sets of “design”, hence unrelated manifesting sets of these tags will not be returned. If a holistic tag ontology is constructed based on the entire folksonomy, more preceded tags and their manifesting sets can be returned with the query upon the preceding tag. This in turn will increase the information recall of the query.

We run the same experiment on 100 tags, 200 tags, and 400 tags datasets respectively, to show how tag precedence relations can enrich query results on a single tag. With the increased number of included tags, more tag precedence relations are discovered. In the meantime, each unique preceding tag is associated with an increased number of tag precedence relations, e.g., several tag precedence relations share the same preceding tag. In other words, each preceding tag has more preceded tags as the size of the dataset increases. If we include the manifesting sets of preceded tags when querying a particular preceding tag, the result of the query will cover more resources that are potentially related to the query. As illustrated in Table 7-1, with more tag precedence

relations discovered, the number of potentially related resources for a given preceding tag increases.

| | 100 Tags | 200 Tags | 400 Tags |
|---|---|---|---|
| Tag precedence relations | 32 | 100 | 258 |
| Unique preceding tags | 17 | 37 | 74 |
| Average tag precedence relations on each preceding tag | 1.88 | 2.7 | 3.49 |
| Average resources marked by each preceding tag | 1928.06 | 1427.92 | 1117.92 |
| Resources included in the manifesting sets of preceded tags, averaged by number of preceding tags | 119.41 | 385.27 | 458.92 |
| Average increase of query result size | 6.19% | 26.98% | 41.05% |

Table 7-1 Tag precedence relations enrich tag queries

Furthermore, we may group the query results according to the sub concepts to which they are related. For instance, we can group query results of the tag “design” into subcategories of “icons”, “graphics”, “resources”, et cetera.

In the meantime, we may also group different manifesting sets of a homonymous tag so that the resources being displayed are related to a single meaning of the tag. As an example, if we query the tag “apache”, we can display resources of the manifesting set related to the web server and the resources of the manifesting set related to the Native American groups separately.

## 7.2 Tag navigation map

Because of the lack of semantic relations between tags, currently users can browse the resources in a tagging system using a list of the most popular tags in the system (tag cloud). With the aid of emergent tag ontology, tagging system users can have another option for browsing the tag space. Initially the user may choose a tag of interest, and the system can display its preceded tags and their preceded tags as a tag network or map.

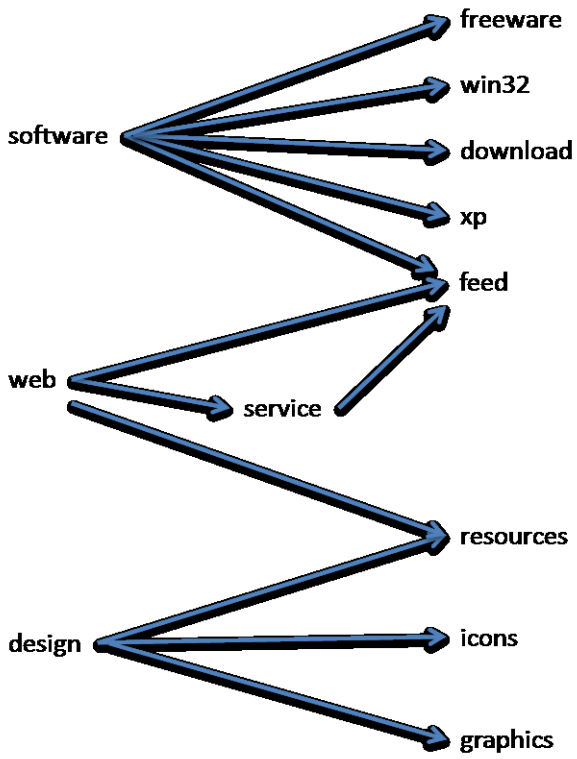


Figure 7-1 Example partial navigation path produced by tag precedence relations

In the above figure, we showed an example partial tag navigation amp constructed from the 100 tag dataset described in the previous chapter. Notably, the navigation map resulted is not a taxonomy/tree structure, but a directed graph where child nodes can have more than one parent nodes.

Tag precedence relation is not transitive. In a tag navigation map we may have tags that are both preceding tags and preceded tags. But this does not imply that their preceding tags are the super concepts of their preceded tags. In figure 7-1, the tag “feed” is the

preceded tag of both tag “web” and “service”. In the meantime the tag “service” is also the preceded tag of tag “web”. Only in this case do the three tags form a hierarchical super/sub concept relation where the tag “feed” is the sub concept of both “web” and “service”.

## 7.3 Tag ontology as another metadata dimension for understanding underlying resources

Tags can be assigned to online resources with a variety of media forms. Hence tags are a preferable format of metadata when the task of parsing the content of the underlying resources is difficult. For example, file sharing or video sharing websites may find tags a desirable addition besides metadata such as resource title and description. Tag ontology on top of these web services can not only improve query recall for a set of given keywords, but also provide another dimension for browsing related resources.

Tag recommendation systems could benefit from emergent tag ontology. If a user is assigning a tag to a resource, based on the tag and the user, the system may cluster the resource into a manifesting set, thus provide related tags based on the specific meaning of the user assigned tag.

Tag ontology could be applied in the field of semantic culturomics [55] as knowledge base for term relations. Combined with statistical analysis on keywords, we may decide some news articles share the same topics if the most used keywords are close positioned in tag ontologies.

## 7.4 Conclusion

Tag ontologies extracted from folksonomies reflect the vocabulary of tagging system users. Hence tag ontologies can be used to assist machines in understanding relations between tag words. As Hendler noted: “A little semantics goes a long way” [56]. With a little more metadata, new ways of utilizing the entire dataset will emerge.

# Chapter 8. Conclusion

In this thesis we have introduced a novel approach for enriching folksonomy data with more semantic metadata. Our work was limited on the scale of dataset and implementation of real use cases that can be presented to the end users for feedbacks, which we deem as interesting future research directions. Despite the quantitative measure we introduced for assessing the similarities for tag precedence relations, ultimately the question whether two tags are closely related should be answered by end users themselves. Hence the overall quality of the emergent ontologies discovered also requires further investigation from the perspective of users. Nonetheless the research described in this thesis sheds light on how some overlooked regularities presented in folksonomies can help us discover useful information that is not available otherwise.

Internet, as the symbol of freedom for this century, enabled that the voice of almost every human being can be heard. But too many voices uttered together without organization become indiscernible and noisy. Tagging systems provide users an easy and unique way to self-organize the content that they created or consumed. In this thesis we introduced a set of tools to augment state of the art tagging systems so that the

meanings of a tag can be identified and the relations between tags can be discovered. With this improvement, a user-defined emergent ontology can be constructed automatically based on the folksonomy. Emergent ontologies can help tagging systems overcome several weaknesses that are often the center of discussion when comparing tagging with other types of metadata such as traditional information expert created taxonomies.

Although the number of websites that are dedicated to the sole purpose of using tagging to organize online resources is declining in recent years, tagging is increasingly becoming an indispensible feature for more and more online services. Popular browsers allow users to organize their bookmarks in folders and store online; online storage services also retain personal directory information on files that appear in more than one user's online space; social media sites often allow users to use hashtags to participate in discussions of certain topics; and even web search services that each search essentially involves a user, some keywords, and a resource that the user considers most relevant. All the above services generate data that is of the structure of folksonomy. With the help of this research, services can provide more interesting metadata that suits the need of end users.

# Bibliography


1. Mathes, A., *Folksonomies-cooperative classification and communication through shared metadata.* Computer Mediated Communication, 2004. **47**(10): p. 1-13.
2. Vander Wal, T., *Folksonomy.* online posting, Feb, 2007. **7**.
3. Reisberg, D., *Cognition: Exploring the science of the mind*1997: WW Norton & Co.
4. Sinha, R., *A cognitive analysis of tagging.* 2005.
5. Guarino, N., D. Oberle, and S. Staab, *What is an Ontology?*, in *Handbook on ontologies*2009, Springer. p. 1-17.
6. Sinclair, J. and M. Cardew-Hall, *The folksonomy tag cloud: when is it useful?* Journal of Information Science, 2008. **34**(1): p. 15-29.
7. Heymann, P. and H. Garcia-Molina, *Collaborative creation of communal hierarchical taxonomies in social tagging systems.* 2006.
8. Shepitsen, A., et al. *Personalized recommendation in social tagging systems using hierarchical clustering*. in *Proceedings of the 2008 ACM conference on Recommender systems*. 2008. ACM.
9. Mika, P., *Ontologies are us: A unified model of social networks and semantics*, in *The Semantic Web–ISWC 2005*2005, Springer. p. 522-536.
10. Markines, B., et al. *Evaluating similarity measures for emergent semantics of social tagging*. in *Proceedings of the 18th international conference on World wide web*. 2009. ACM.
11. Cattuto, C., et al., *Semantic grounding of tag relatedness in social bookmarking systems*, in *The Semantic Web-ISWC 2008*2008, Springer. p. 615-631.
12. Song, Y., B. Qiu, and U. Farooq. *Hierarchical tag visualization and application for tag recommendations*. in *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011. ACM.

13. Basso, C.A.M., J.M. Ferreira, and S.R.P. da Silva. *An unsupervised approach for the emergence of ontologies from personomies in tagging-based systems*. in *Web Congress, 2009. LA-WEB'09. Latin American*. 2009. IEEE.

14. Specia, L. and E. Motta, *Integrating folksonomies with the semantic web*, in *The semantic web: research and applications*2007, Springer. p. 624-639.

15. Angeletou, S., et al., *Bridging the gap between folksonomies and the semantic web: An experience report.* 2007.

16. Laniado, D., D. Eynard, and M. Colombetti. *Using WordNet to turn a folksonomy into a hierarchy of concepts*. in *Semantic web application and perspectives-fourth italian semantic web workshop*. 2007. Citeseer.

17. Van Damme, C., M. Hepp, and K. Siorpaes, *Folksontology: An integrated approach for turning folksonomies into ontologies.* Bridging the Gap between Semantic Web and Web, 2007. **2**(2): p. 57-70.

18. Lohmann, S., P. Díaz, and I. Aedo. *MUTO: the modular unified tagging ontology*. in *Proceedings of the 7th International Conference on Semantic Systems*. 2011. ACM.

19. Kim, H.L., et al. *The state of the art in tag ontologies: a semantic model for tagging and folksonomies*. in *International Conference on Dublin Core and Metadata Applications*. 2008.

20. Gruber, T., *Ontology of folksonomy: A mash-up of apples and oranges.* International Journal on Semantic Web and Information Systems (IJSWIS), 2007. **3**(1): p. 1-11.

21. Newman, R., *Tag ontology writeup.* DOI= http://www. holygoat. co. uk/projects/tags, 2005.

22. Kreiser, A., et al. *A web 3.0 approach for improving tagging systems*. in *Proceedings of the International Workshop on Web*. 2009.

23. Tanasescu, V. and O. Streibel, *Extreme Tagging: Emergent Semantics through the Tagging of Tags.* ESOE, 2007. **292**: p. 84-94.

24. Lachica, R. and D. Karabeg, *Metadata creation in socio-semantic tagging systems: Towards holistic knowledge creation and interchange*, in *Scaling Topic Maps*2008, Springer. p. 160-171.

25. Passant, A. and P. Laublet. *Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data*. in *LDOW*. 2008.

26. Passant, A. *Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs*. in *Proceedings of International Conference on Weblogs and Social Media*. 2007.

27. Torres, D., et al. *Semdrops: A Social Semantic Tagging Approach for Emerging Semantic Data*. in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*. 2011. IEEE.

28. Veres, C. *LexiTags: An Interlingua for the Social Semantic Web*. in *Proceedings of the 4th International Workshop on Social Data on the Web In conjunction with the International Semantic Web Conference (ISWC2011), Bonn*. 2011.

29. Camino, S., et al., *Enabling Semantics-Aware Collaborative Tagging and Social Search in an Open Interoperable Tagosphere.* 2008.

30. Marchetti, A., et al. *Semkey: A semantic collaborative tagging system*. in *Workshop on Tagging and Metadata for Social Information Organization at WWW*. 2007.

31. Lezcano, L., E. García-Barriocanal, and M.-A. Sicilia, *Bridging informal tagging and formal semantics via hybrid navigation.* Journal of Information Science, 2012. **38**(2): p. 140-155.

32. Veres, C., K. Johansen, and A. Opdahl. *SynsetTagger: a tool for generating ontologies from semantic tags*. in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. 2013. ACM.

33. da Silva, S.R., et al., *An approach to enrich users' personomy using the recommendation of semantic tags.* Journal of the Brazilian Computer Society, 2012. **18**(4): p. 283-298.

34. Wetzker, R., et al. *I tag, you tag: translating tags for advanced user models*. in *Proceedings of the third ACM international conference on Web search and data mining*. 2010. ACM.

35. Heckner, M., M. Heilemann, and C. Wolff. *Personal Information Management vs. Resource Sharing: Towards a Model of Information Behavior in Social Tagging Systems*. in *ICWSM*. 2009.

36. Parsons, J. and Y. Wand, *Emancipating instances from the tyranny of classes in information modeling.* ACM Transactions on Database Systems (TODS), 2000. **25**(2): p. 228-268.

37. Parsons, J. and Y. Wand, *Attribute-based semantic reconciliation of multiple data sources*, in *Journal on Data Semantics I*2003, Springer. p. 21-47.

38. Sanderson, M. and B. Croft. *Deriving concept hierarchies from text*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM.

39. Schmitz, P. *Inducing ontology from flickr tags*. in *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*. 2006.

40. Golder, S.A. and B.A. Huberman, *Usage patterns of collaborative tagging systems.* Journal of Information Science, 2006. **32**(2): p. 198-208.

41. Tanaka, J.W. and M. Taylor, *Object categories and expertise: Is the basic level in the eye of the beholder?* Cognitive psychology, 1991. **23**(3): p. 457-482.

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

43. Au Yeung, C.-m., N. Gibbins, and N. Shadbolt. *Contextualising tags in collaborative tagging systems*. in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. 2009. ACM.

44. Gemmell, J., et al. *Personalization in folksonomies based on tag clustering*. in *Proceedings of the 6th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems*. 2008.

45. Görlitz, O., S. Sizov, and S. Staab. *PINTS: peer-to-peer infrastructure for tagging systems*. in *IPTPS*. 2008.

46. Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Vol. 1. 2008: Cambridge university press Cambridge.

47. Ramze Rezaee, M., B.P. Lelieveldt, and J.H. Reiber, *A new cluster validity index for the fuzzy c-mean.* Pattern recognition letters, 1998. **19**(3): p. 237-246.

48. Rendón, E., et al., *Internal versus External cluster validation indexes.* International Journal of computers and communications, 2011. **5**(1): p. 27-34.

49. Ingaramo, D., et al., *Evaluation of internal validity measures in short-text corpora*, in *Computational Linguistics and Intelligent Text Processing*2008, Springer. p. 555-567.

50. Salton, G., A. Wong, and C.-S. Yang, *A vector space model for automatic indexing.* Communications of the ACM, 1975. **18**(11): p. 613-620.

51. Fernandez-Amoros, D., et al., *Automatic word sense disambiguation using cooccurrence and hierarchical information*, in *Natural Language Processing and Information Systems*2010, Springer. p. 60-67.

52. Li, H. and N. Abe. *Word clustering and disambiguation based on co-occurrence data*. in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. 1998. Association for Computational Linguistics.

53. Yarowsky, D. *Unsupervised word sense disambiguation rivaling supervised methods*. in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. 1995. Association for Computational Linguistics.

54. Huang, A. *Similarity measures for text document clustering*. in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008.

55. Suchanek, F.M. and N. Preda, *Semantic Culturomics (Vision paper).*

56. Hendler, J., *The dark side of the semantic web.* IEEE Intelligent Systems, 2007. **22**(1): p. 2-4.