2010 12th International Asia-Pacific Web Conference

An Asymmetric Similarity Measure for Tag Clustering on Flickr

Xiaochen Huang
line 1: School of Information Technologies
The University of Sydney, NSW, Australia
Email: xhua2034@sydney.edu.au

Ying Zhou
School of Information Technologies
The University of Sydney, NSW, Australia
Email: ying.zhou@sydney.edu.au

***Abstract*—Web 2.0 tools and environments have made tagging, the act of assigning keywords to on-line objects, a popular way to annotate shared resources. The success of now-prominent tagging systems makes tagging "the natural way for people to classify objects as well as an attractive way to discover new material". One of the most challenging problems is to harvest the semantics from these systems, which can support a number of applications, including tag clustering and tag recommendation. We conduct detailed studies on different types of tag relations and tag similarity measures, and propose a scalable measure that we name *Reliability Factor Similarity Measure* (RFSM). We compare it with two other measures having similar scalability by integrating them into hierarchical clustering methods and performing tag clustering on a subset of Flickr data. The results suggest that RFSM outperforms those two measures when it is applies for tag clustering purpose. We also present an alternative way of utilizing discovered tag relations to set up tag refining rules in order to deal with some noise in the initial tag sets, which can in turn improve the precision of tag relations.**

***Keywords*-Reliability Factor Similarity Measure, Clustering, Tag, Folksonomy, Web 2.0**

I. Introduction

The social tagging systems provide Internet users new options to organize, categorize, search and explorer resources. Unlike formal classification systems, tagging systems do not have an agreed structure of tags or detailed taxonomy. A flat, non-hierarchical name space contributed by authors and/or consumers from an uncontrolled vocabulary serve as the basis of tagging-based classification. It is commonly referred to as "folksonomy". The overall costs for users of these tagging systems in terms of time, effort and cognition are far lower than the costs of systems that rely on complex hierarchal classification and categorization schemes[8].

While tagging systems have many benefits, there are also limitations and weaknesses stemmed from the uncontrolled vocabulary. Tags can be redundant or ambiguous due to their open nature, which will greatly limit the performance of search and exploration in tag space[2]. Extensive studies have been conducted to improve user experience in the tag spaces, especially in the areas of tag ranking, recommendation, classification, clustering and query expansion. Among these active research areas, automatic tag clustering is widely adopted to overcome the major challenge brought by the ambiguity of user queries. Generating clusters from search results instead of a single result set help users quickly locate the information they are looking for.

One of the key input to a clustering algorithm is the distance measure, which in this case is the measurement of tag relations. Thus tag similarity measures can have a significant impact on the outcome of clustering algorithms. Accurate measuring similarity between tags depend on an in-depth understanding of relations of tags used to annotate a particular resource. In this paper, we examine several useful types of tag relations and present a similarity measure that can best quantify these relations in general. We also present a way of refining and treating a particluar type of tag relations for better clustering results.

In summary, the key contributions in this paper are

- A new similarity measure to better describe tag relations. We identify some problems in previous similarity measures and propose a new measure called Reliability Factor Similarity Measure (RFSM). We compare RFSM with two other similarity measures for tag clustering purpose and the results show that giving the same data sets and queries, RFSM can help find more meaningful clusters than the other two can do most of the time.
- A novel way of utilizing discovered tag relations to improve clustering results. Currently, discovered tag relations are likely to be directly used as inputs for clustering algorithms. We observe that some of the relations can only bring noise to clustering results if they are used that way. We find an alternative way of using these relations to set up tag refining rules, which can raise the precision of extracted tag relations and in turn improve the quality of clustering results.

The paper is organized as follows. In section II we briefly describe different approaches for utilizing tag information. Section III presents a survey on tag relation and the features and problems of similarity measures proposed so far. We also present our own similarity measure trying to mitigate those problems. Section IV describes refinement strategy for a particular type of tag relations. We evaluation out similarity measure and refinement strategy using flickr data on section V. Section VI concludes the paper.

978-0-7695-4012-2/10 $26.00 © 2010 IEEE
DOI 10.1109/APWeb.2010.65

II. Background

The popularity of the tagging systems makes the tagged resources grow exponentially. However, the uncontrolled vocabulary far exceeds the semantics of a hierarchical ontology or taxonomy such as WordNet, which brings significant challenges to organizing and ranking search results for users to better understand and further explore the result set.

Hotho et al.[6] suggest an adapted PageRank-like algorithm FolkRank to improve efficient searching via personalized and topic-specific ranking within the tag space. Bao et al.[1] optimize web search by using tags from two aspects: similarity ranking and static ranking. They propose a new way of utilising tags as a metadata for the similarity measure between a query and a resource. They also argue that the amount of tags assigned to a resource implies its quality in some sense. A novel algorithm SocialPageRank is proposed in the light of the above argument to measure the popularity and quality of web pages.

Meanwhile, a number of studies try to classify tags into semantic categories. Sigurbjornsso and van Zwol[13]map Flickr tags onto WordNet semantic categories using straight forward string matching between Flickr tags and WordNet lemmas. Rattenbury et al.[12] cluster tags from Flickr using temporal and spatial metadata, to assign event and place semantics. Overell et al.[9] present a novel method of categorizing Flickr tags as WordNet semantic categories. They first categorise Wikipedia articles and then map Flickr tags onto those categorised articles.

Compared with the supervised classification approach, clustering is more suitable for the dynamic features of tag usage in tagging systems, and its unsupervised nature is also preferable in large systems where it is not feasible to recognize and predefine categories. Ramage et al.[11] use tags from large-scale social bookmarking websites such as del.icio.us as a complementary data source to page text and anchor text for improving automatic clustering of web pages. Zhou et al.[17] apply a similar idea. In addition, they identify several pitfalls of treating tags the same as terms in document content and considering them as additional terms of the documents. They point out that a tag is generated differently than a document content term and it represents an abstract of the document from single perspective of a single user. They also indicate that the semantics of the tags and the differences in domain expertise of users should be taken into consideration.

III. TAG SIMILARITY MEASURES

Considering the limitations inherited in the uncontrolled vocabulary, much noise is expected in user-defined tags. Therefore, similarity measures should be carefully designed to ensure that it will not be affected by the noise representing meaningless relations. Some insights on tagging motivations and tag relations can greatly help us achieve this goal.

A. Tag Relations

One of the first analyses of tagging systems appears in the work of Golder and Huberman[4]. The authors perform their analysis of the information dynamics in the Del.ic.ious system. They identify seven different types of tags in Del.icio.us: *identifying what (or who) it is about, identifying what it is, identifying who owns it, refining categories, identifying qualities or characteristics, self reference and task organizing*. They also discuss how tags by individual users are used over time and how tag proportions stabilize over time. Halpin et al.[5] take one step further by proving that tagging distributions tend to stabilize into power law distributions.

Marlow et al.[8] offer a comprehensible taxonomy, which allows classifying tagging systems according to user incentives and motivations. They also discuss how the tagging motivations may influence the resultant tags in a tagging system. By conducting an initial study of the tagging dynamics on Flickr, the authors report on different incentives for tagging comparing to those of del.icio.us. Six motivations are identified in Flickr by the authors, including *future retrieval, contribution and sharing, attract attention, play and competition, self presentation and opinion expression*, which can be further categorized into two high-level practices: organizational and social.

Bischoff et al. [3] analyze tag usage in different tagging systems, including Flickr. The authors define tag types as *topic* (describing what a tagged item is about), *time, location, type* (corresponding to file, media or Web page type), emphauthor/owner, opinions/qualities, usage context (suggesting what to use a resource for, or the context/task the resource was collected in and grouped by) and *self reference*. Their study on Flickr tag set shows that the most important category for Flickr is topic, and location also plays an important role. The rest of the tag types only represent a very small part of the Flickr tags. The authors report that most of the tags can be used for search, and in most cases tagging behavior exhibits approximately the same characteristics as searching behavior.

Based on these previous works and the analysis of Flickr's data set, we recognize four types of useful tag relations:

- **Parent-Child Relations**, such as *Sydney - Harbour Bridge* and *Japan - Tokyo*. Parents of a child can be used to construct context of that child which can greatly enhance users' understanding.
- **Hypernym-Hyponym Relations**, such as *dog - Labrador* and *furniture - desk*. The terms form an "is-a" relationship, while the **parent-child** tags form a "has-a" relationship. Both types of relations are very useful in solving basic level variation semantic difficulty when the tags involved in these relations are both of topic or location type.
- **Token-Phrase Relations**. Some phrases entered by

users for tagging purpose are wrongly broken down into separate tokens by the system, which in turn become individual tags. Such separation generate the token-phrase relation. Some examples are *new - york* and *san - francisco*. We name those tags decomposed from phrases as "token tags", such as *new* and *york*, and we name those tags formed by an entire phrase as "phrase tags", such as *newyork* and *sanfrancisco*. Capturing these relations can help us reconstruct users' original intentions.

- **Synonym Relations**. In a broad sence, abbreviations, singular-plural forms and language variations are also included in this type. For example, *manzana* and *apple* may form a sysnonym relation as *manzana* is a Spanish word meaning apple on the web.

B. Current Measures and their problems

An early and simple measure is presented by Begelman et al. [2]. Tag co-occurrence is used as similarity measure to construct an undirected weighted graph. Graph partition is then applied to obtain clusters. Sigurbjornsso and van Zwol[13] point out that using the raw tag co-occurrence for computing the quality of relation between two tags is insufficient, as these values do not take the frequency of the individual tags into account. Therefore, they suggest that co-occurrence count should be normalised by the frequency of one of the tags. The similarity function can be formally written as:

$$sim(t_i, t_j) = \frac{c(t_i, t_j)}{c(t_i)} \tag{1}$$

where t_i and t_j are two tags, $c(t_i, t_j)$ is the tag co-occurrence count of t_i and t_j, and $c(t_i)$ is the tag frequency of t_i. We denote this measure by TCSM (Tag Co-occurrence Similarity Measure). Normalizing tag co-occurence by individual tag occurence creates an asymmetric measure since $sim(t_i, t_j) = sim(t_j, t_i)$ if and only if $c(t_i) = c(t_j)$ or $c(t_i, t_j) = 0$. Therefore, for each pair of tags, two similarity values will be computed: $sim(t_i, t_j)$ and $sim(t_j, t_i)$.

Sigurbjornsso and van Zwol[13] show that symmetric measures such as the Jaccard symmetric coefficient are good at identifying equivalent tags. They could efficiently discover most of the synonym relations and some of the token-phrase relations. Such observation is supported in [7]. However, only a small proportion of tag relations are symmetric. In particular, **parent-child** relations and **hypernym-hyponym** relations are asymmetric. Taking *bird* and *seagull* as an example, *seagull* is closely related to *bird* as it is a type of *bird*. But this by no means implies that *bird* is strongly related to *seagull* for there are many types of bird other than seagull.

Asymmetric measures can be used to explore those asymmetric relations. Moreover, symmetric relations can also be discovered. As two similarity values will be computed for a pair of tags, possible symmetric relations can be identified by checking tag pairs having both similarity values comparatively high. Therefore asymmetric measures will provide a better result in extracting and quantifying useful tag relations.

TCSM can be problematic when some active users tagged much more resources than others. The similarity measurement will be easily biased in favor of those active users. Wang[15] indicates that, in systems like Flickr, it is possible and quite typical for a user to use a similar set of tags over and over again to describe a collection of resources. As a result, there is a high possibility for any tag pair in that set to be identified as highly related tags by TCSM. In order to mitigate the bias towards active users, Wang[15] proposes macro-aggregation similarity measure. Instead of giving each tag occurence same weight, it assigns same weight to each tag user. The similarity measure function is computed as:

$$sim(t_i, t_j) = \frac{u(t_i, t_j)}{u(t_i)} \tag{2}$$

where $u(t_i, t_j)$ is the number of users that assign both t_i and t_j to a same resource , and $u(t_i)$ is the number of users that use t_i. We denote it by UCSM (User Count Similarity Measure). UCSM is also an asymmetric measure.

UCSM has its own limitations. It may over-emphasize the user impact by totally ignoring the tag co-occurrence value. For instance, a user may upload 100 photos of dogs, one of which is of a black Labrador. Suppose all photos have *dog* as a tag, and that particular one also has an aditional tag *labrador*. The UCSM similarity from *dog* to *labrador* is computed as $sim(dog, labrador) = 1/1 = 1$, indicating that *dog* is highly related to *labrador*. This is not the case for the data set as there is only one evidence of it. However, the TCSM similarity has a better measure as $sim(dog, labrador) = 1/100 = 0.01$.

C. Reliability Factor Similarity Measure

Intuitively, tag co-occurrence normalized by the frequency of one of the tags is a good indicator of tag relations, and thus can be used as a start point of similarity measures, called similarity factor (SF). The formula of calculating similarity factor from one tag, t_i, to another, t_j, is defined as:

$$SF(t_i, t_j) = \frac{c(t_i, t_j)}{c(t_i)} \tag{3}$$

Where $c(ti, tj)$ is tag co-occurrence of t_i and t_j, and $c(t_i)$ is the tag frequency of t_i.

The computation of similarity factor is ensentially the same with equation (1). As discussed above, it cannot ensure usefulness of generated tag relations. Specifically, when high frequency tag pairs are used by only a small portion of users, relations between these tags are highly unreliable. Therefore, we add a reliability factor $RF(t_i, t_j)$ to measure the reliability of relations between two tags, t_i and t_j. Our strategy for $RF(t_i, t_j)$ computation is based on the following observations:

- When tag co-occurrence count are the same, relations between tags that have less user count should be considered as weaker relations and thus gain less reliability score;
- When user count is relatively high, fluctuation of tag co-occurrence count should not have a significant impact on the reliability score;
- When the ratio between tag co-occurrence count and user count are stable, relations with higher counts are more reliable than the ones with lower counts, and thus should get a higher reliability score.

The reliability factor $RF(t_i, t_j)$ is computed as:

$$RF(t_i, t_j) = \frac{1}{1 + \frac{\lg c(t_i, t_j)}{\lg^2 u(t_i, t_j)}} \tag{4}$$

This formula requires $u(t_i, t_j)$ to be greater than 1, implying that $c(t_i, t_j)$ has to be greater than 1. As relations between tags that are used by only one user are extremely unreliable, it is reasonable to prune these relations. For simplicity, we rewrite equation (4) as:

$$RF(t_i, t_j) = RF(u, c) = \frac{1}{1 + \frac{\lg c}{\lg^2 u}} \quad (c \geq u > 1) \tag{5}$$

Where u stands for $u(t_i, t_j)$ and c stands for $c(t_i, t_j)$. The range of this function is within (0, 1). The function reaches its limit of 0 when $c \to \infty$ and u is a constant less than c, and it reaches its limit of 1 when $u \to \infty$. At the meantime, when c is fixed, RF reaches its upper boundary $1/(1 + 1/\lg^2 c)$ if $u = c$. Figure1 and 2 demonstrate several notable features of this function:

1) When c is fixed, RF will increase as u increases. This is consistent with our assumption that when tag co-occurrence counts are the same, the more users use a tag pair, the stronger the relation between them.
2) When u is fixed at a relatively high value RF will decrease slightly with the increase of c values. This will compensate the sharp value gains in the SF score calculated by equation (3) to meet the criterion that when user count is high enough, high co-occurrence count should not inflate the reliability score.
3) When $^c/_u$ is fixed, RF will increase as u and c increase. This is in accordance with the intuition that when the ratio between u and c are identical, relations with higher user and co-occurrence count should be more reliable.

With the reliability factor, similarity from t_i to t_j is computed as:

$$sim(t_i, t_j) = SF(t_i, t_j) \times RF(t_i, t_j) \tag{6}$$

Note that the new similarity measure is also asymmetric.

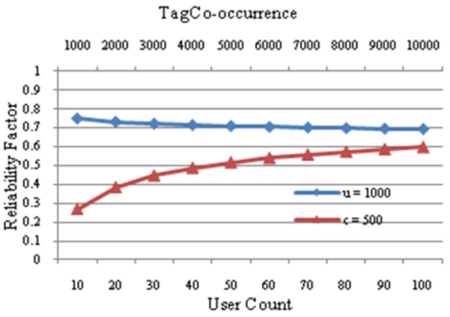

Figure 1. The blue line shows the RF results given different co-occurrence count when u = 1000, and the red line shows the RF results given different user count when c = 500

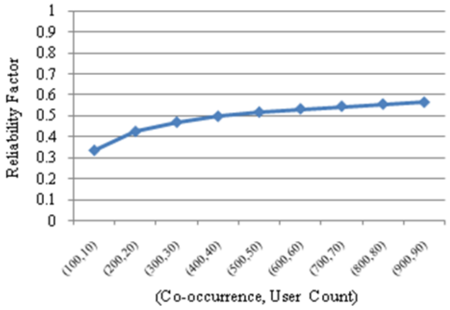

Figure 2. Reliability factor result given different co-occurrence count and user count when c/u = 10

IV. Tag Refinement

In the processing of integrating different similarity measures into tag clustering process, we observe that some tags can be clustered together for wrong reasons and form useless concepts for users. An example is a cluster containing three tags {*new, york, zealand*}. They are clustered together because of two token-phrase relations, *new* - *york* and *new* - *zealand*. The cluster itself does not make much sense. We suggest that although token-phrase relations and synonym relations are useful, certain refinement is ncessary before directly input them for clustering algorithms. Before we present our solution for this problem, let us first explorer how these two types of relations could reduce the quality of generated clusters.

A. Problematic Relations

Problem caused by **Token-phrase** relations occurs frequently. Cluster {*new, york, zealand*} is incorrectly clustered together because *New York* and *New Zealand* are separated into individual words. More examples can be found, such as cluster {*san, diego, francisco*}, which is caused by breaking down *San Francisco* and *San Diego*.

This problem becomes worse when many tags are involved and when only type of tags is assigned a resource. For example, a group of photos in flickr may have *new york* but neither *new* or *york* in their tag lists while another group may have *new* and *york* but not *new york* in the tag lists. As a result, some tags supposed to be closely related to *New York* are now considered to have connections with both *new* and *york*. Tags related to *New York* and tags related to *new* are likely to be grouped into one cluster. It is hard to discover the relations between token tags and its corresponding phrase tags using existing similarity measures if they do not co-occur a lot.

The most serious problem caused by synonym relations is tag ambiguity. Although clustering is widely used to resolve this problem and is believed to have some immunity to it, the reality is that tag ambiguity still has large impact on clustering results. People usually use acronyms to tag photos for convenience. However, most of the acronyms have multiple interpretations. For instance, *AI* can be expanded to *artificial intelligence, art institute, Allen Iverson* and etc. With the sole presence of *AI*, all words associated with these different concepts may be grouped as one cluster. Synonym formed by singular-plural relation will also cause some problems. For example, some tags may only be computed as closely related to *apple* while some others are computed to be closely related to *apples*. If the relation between *apple* and *apples* is not strong enough, two clusters may be generated albeit there should be only one.

B. Tag Refining Rules

Detection of acronyms is quite straightforward as its definition gives us a clear direction. By testing if one tag is formed by the initial components of another tags source phrase, we can estimate whether or not the former one is an acronym of the latter one. To avoid problems brought by acronyms, when a tag and its acronym appear in a photos tag set, the acronym tag will be discarded. However, as some users prefer to use acronyms only, such simple elimination strategy may cut off bonds between tags and their acronyms and cause false cluster decomposition. One possible remedy is to introduce an acronym extension mechanism. Singular-plural relation can be detected using simplified version of the Porter Stemming Algorithm presented by Porter in[10]. The discovered relations will then be used to convert plural words to their singular forms. The accuracy is limited by the intrinsic simplicity of the Porter Stemming Algorithm, but on balance it may improve the performance of clustering algorithms.

As our algorithm only uses relations between a pair of tags, the discovered token-phrase relations are restricted to two-word phrases. If two token tags form a token-phrase relation, they are likely to co-occur a lot. However, token-phrase relations are not necessarily symmetric. This is especially true when one of the tags is also widely used in other context. For example, *francisco* may be strongly related to *san*, whereas *san* may not be perceived as strongly related to *francisco* since it can form token-phrase relations with a number of other tags as well. Therefore, we need to look into those asymmetric relations too. Based on previous observations, we detect token-phrase relations by examining if one tag is strongly related to another while they together can form a phrase tag. Once such a relation is discovered, all the co-appearances of the two token tags involved will be replaced by the phrase tag.

Table I
SUMMARY OF FLICKR DATASET

Query Tag	Number of Photos	Number of Tags	Number of distinct tags	Number of Users	Tags per photo
ai	16071	144953	16681	2379	9.02
apple	29502	302469	33023	9717	10.25
australia	27190	234979	23452	2227	8.64
bird	27579	279825	32640	9122	10.15
bridge	27058	337719	34510	10342	12.48
dog	29227	253948	30231	8837	8.69
jaguar	20265	237121	18171	5580	11.70
japan	28612	252960	21446	2438	8.84
java	22627	215439	16654	2496	9.52
mac	24130	218083	25357	6830	9.04
pluto	12431	145958	10208	2325	11.74
tiger	24313	237796	23121	7676	9.78

V. EVALUATION

A. Data Sets

Our experiments are carried on several data sets obtained from Flickr. Each data set corresponds to a single-word query. The data is gathered via Flickr API functions `flickr.photos.search` and `flickr.photos.getInfo` during the period from February to April 2009. We first generate a list of photos that contain a given query tag using `flickr.photos.search`, all the information of photos in that list is then downloaded to local database for further analysis through `flickr.photos.getInfo`.

Twelve query tags are chosen featuring different types of potential outcomes. For example, clusters generated from *dog* and *bird* are likely to form a taxonomy of the concepts, that are, different breeds of dogs and birds. Meanwhile, clusters generated from *AI* would be a lot of interpretations of this acronym. Table I shows the query tags and the statistics of their corresponding data sets.

B. Clustering Result Comparison

We implemented the three similarity measures mentioned above (TCSM, UCSM, RFSM) and integrated them into clustering algorithms to perform tag clustering on several Flickr data sets. We then compare these measures using the top 20 clusters generated by them to see which one can find more clusters that are potentially useful for users. Judging the usefulness of a cluster is undoubtedly a subjective process. We are expected to see clusters fulfilling some

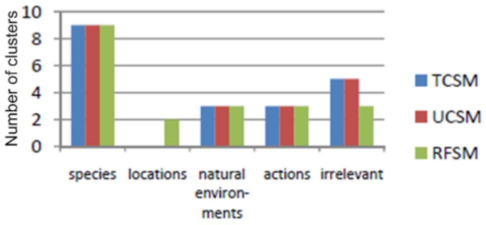

Figure 3. Cluster result comparison of query bird

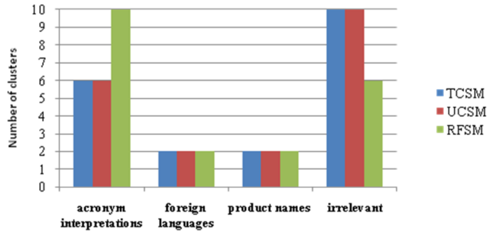

Figure 4. Cluster result comparison of query *AI*

peoples intentions while being considered as useless by others. In case of *bird*, someone may want to see all types of birds within a particular area while some others may prefer only to browse through photos of a single type of bird. There are also people who are curious about both. To minimize human factors in the evaluation process, for each query tag, we first set up several cluster categories which have the potential to meet certain information needs of users. Then, we select the top 20 clusters from cluster results of that tag generated using each similarity measure and assign them into the predefined categories. Finally, we use the number of clusters in each category to compare the effectiveness of the three measures. The ranking of clusters is based on the number of tags contained. Clusters having more tags will be ranked higher.

Figure 3 shows the comparison results using query tag *bird*. We define four categories for the potentially useful clusters - species, locations (consisting of cities, countries and/or continents), natural environments (such as *water, sky*) and actions (such as *flying, swimming* and *taking off*). All the clusters that cannot be classified into these four categories will be deemed as irrelevant, such as cluster consists of keywords {*delete, delete2, delete3*} only. While all three similarity measures find same amount of clusters in species, natural environments and actions categories, our RFSM successfully reveals two clusters under category locations.

Figure 4 shows the comparison results using query tag *AI*. The categories we define for it include acronym interpretations (such as *Air India, Artificial Intelligence, Allen Iverson* and *Adobe Illustrator*), foreign languages (such as Japanese and Chinese) and product names (such as Custom Houses Ai dolls, Nikons AI lenses and Bronica SQ-AI medium format camera). Clearly, our RFSM outperforms the other two by generating four more acronym interpretations clusters.

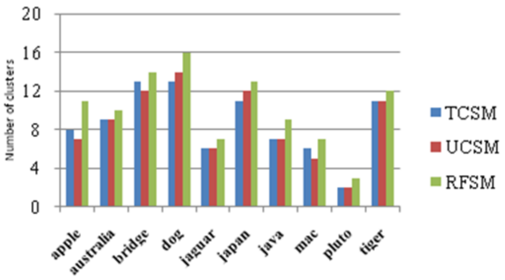

Figure 5. Relevant cluster number comparison of all the other queries. All three similarity measures seem to have an extremely poor performance when the query is pluto. This is due to lack of implications of pluto which leads to only 9, 10, 9 clusters generated by TCSM, UCSM and RFSM respectively.

Table II
ALL CLUSTERS OF QUERY PLUTO GENERATED BY RFSM WITH AND WITHOUT TAG REFINEMENT

Without Tag Refinement	With Tag Refinement
disney, disneyland, goofy, mickey, waltdisneyworld, florida, disneyworld, mickeymouse, minniemouse, characters	disney, goofy, disneyland, mickeymouse, waltdisney-world, florida, donaldduck, character, magickingdom, disneyworld
saturn, mars, jupiter, mercury, venus, earth, uranus, planet, neptune, moon	saturn, mars, jupiter, mercury, venus, earth, uranus, planet, neptune, solarsystem
newyorkcity, newyork, nyc, ny, manhattan	dog, cane, pet
dog, cane, pet	newyorkcity, newyork, manhattan
system, solar	princecharming, cinderella
d80, nikon	flagstaff, arizona
flagstaff, arizona	powershot, canon
powershot, canon	d80, nikon
princecharming, cinderella	

For the other ten query tags, Figure 5 illustrates that RFSM also performs better than TCSM and UCSM by discovering more relevant clusters. In summary, RFSM has better performance in terms of both revealing new types of clusters and finding more useful clusters.

C. Effectiveness of Tag Refining Strategies

To evaluate our tag refining strategy, we first measure the accuracy of the tag refinement rules we've generated, and then compare the cluster results generated with and without tag refinement. In these experiments, RFSM is used as the similarity measure.

In total, there are 250 tag refinement rules extracted from the tag relations we discover using RFSM, in which 182 are derived from singular-plural relations, 32 are derived from acronyms and the other 36 are from token-phrase relations. The number of photos affected by these rules is 47527.

The accuracy of a tag refinement rule is defined as the ratio between numbers of photos that have a descriptive and/or concise tag set after applying the rule and the numbers of photos that have their tag set changed by this rule. For the 182 tag refinement rules coming from singular-plural relations, 181 rules have an accuracy of 100% while the other one, using ai to replace ais, has an accuracy of 0% because the tag relation $AI \rightarrow ais$ caused by Nikon's AI/AIS lenses is misinterpreted as a singular-plural relation. For the 32 rules related to acronyms, all of them achieve a 100% accuracy. And for the 36 rules generated using token-phrase relations, 31 are 100% accurate while the other five rules' accuracy are slightly less than 90%. The overall accuracy for all the changes we've made to the photos' tag sets using these tag refinement rules is 95.3%.

When comparing the cluster results generated with and without tag refinement, we could observe some distinct improvements. In Table II, we can see that less informative *mickey, nyc* and *ny* are excluded from the clustering results. Moreover, we can see that the cluster $\{$*system, solar*$\}$in the left column disappears, and a new tag solarsystem emerges in the second cluster of the right column.

VI. Conclusion

Tag clustering techniques are carried out to identify concepts from search results. Clustering techniques have been studies intensively in the IR field for decades, but when they are applied to solve the tag clustering problem in social tagging systems, people tend to overlook the impact caused by tag relations. Although different tag similarity measures have been proposed to compute tag relations, few works have analyzed the descriptiveness of these tag relations. Meanwhile, discovered tag relations are likely to be directly used as an input for the clustering algorithms.

We proposed a new similarity measure RFSM which can better quantify the tag relations. We also presented an alternative way of utilizing discovered tag relations to set up tag refining rules, which can in turn improve the precision of tag relations. Experiments suggest that our method can significantly improve the tag clustering results.

Although we limit the discussion of RFSM to tag clustering purpose only, it can also be directly applied to other applications or indirectly improve the quality of tag-clustering based applications, such as tag recommendation and query expansion.

References

[1] Bao, S., Xue, G., We, X., Yu, Y., Fei, B., and Su, Z. "Optimizing web search using social annotations". In *WWW'07*, Banff, Alberta, Canada, 2007.

[2] Begelman, G., Keller, P., and Smadja, F. "Automated tag clustering: Improving search and exploration in the tag space". In *Proc. of the Collaborative Web Tagging Workshop,(WWW'06)*

[3] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. "Can all tags be used for search?" In *Proc. of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA, 2008.

[4] Golder, S. A., and Huberman, B. A. "The structure of collaborative tagging systems", HP Labs, 2006.

[5] Halpin, H., Robu, V., and Shepherd, H. "The complex dynamics of collaborative tagging". In *WWW,07*, Banff, Alberta, Canada, 2007.

[6] Hotho, A., Jschke, R., Schmitz, C., and Stumme, G. "Information Retrieval in Folksonomies: Search and Ranking". In *The Semantic Web: Research and Applications*, 2006.

[7] Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., and Stumme, G. "Evaluating similarity measures for emergent semantics of social tagging". In *WWW'09*. Madrid, Spain, 2009

[8] Marlow, C., Naaman, M., Boyd, D., and Davis, M. "HT06, tagging paper, taxonomy, Flickr, academic article, to read". In *Proc. of the seventeenth conference on Hypertext and hypermedia*, Odense, Denmark, 2006.

[9] Overrel, S., Sigurbjornsson, B., and Van Zwol, R. "Classifying tags using open content resources". In *Proc. of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, 2009.

[10] Porter, M. F. "An algorithm for suffix stripping". In *The Porter Stemming Algorithm*. Available online at http://www.tartarus.org/~martin/PorterStemmer.

[11] Ramage, D., Heymann, P., Manning, C. D., and Garcia-Molina, H. "Clustering the tagged web". In *Proc. of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, 2009.

[12] Rattenbury, T., Good, N., and Naaman, M. "Towards automatic extraction of event and place semantics from flickr tags". In *SIGIR'07*, Amsterdam, The Netherlands, 2007.

[13] Sigurbjornsson, B., and Van Zwol, R. "Flickr tag recommendation based on collective knowledge". In *WWW'08*, Beijing, China, 2008.

[14] Van Rijsbergen, C. J. *Information Retrieval*, 2nd edition. Dept. of Computer Science, University of Glasgow, 1979.

[15] Wang, D. "Overcoming the Semantic Problem of Collaborative Tagging Systems with Tag Clustering". Master Thesis, The University of Sydney, Sydney, Australia, 2007.

[16] Wang, S., and Tanaka, Y. "Topic-oriented query expansion for web search". In *WWW'06*, Edinburgh, Scotland, 2006.

[17] Zhou, D., Bian, J., Zheng, S., Zha, H., and Gies, C. L. "Exploring social annotations for information retrieval". In *WWW'08*. Beijing, China, 2008.