Association for Information Systems
AIS Electronic Library (AISeL)

ACIS 2010 Proceedings

Australasian (ACIS)

2010

Adding Context to Social Tagging Systems

JooHee Song
The University of Sydney, {json8472@uni.sydney.edu.au

Ying Zhou
The University of Sydney, y.zhou@uni.sydney.edu.au

Hyungsoo Jung
The University of Sydney, hyungsoo.jung@uni.sydney.edu.au

Joseph Davis
The University of Sydney, j.davis@uni.sydney.edu.au

Adding Context to Social Tagging Systems

JooHee Song
Ying Zhou
Hyungsoo Jung
Joseph Davis
School of Information Technologies
The University of Sydney
NSW 2006, Australia
Email: {json8472, y.zhou, hyungsoo.jung, j.davis}@uni.sydney.edu.au

Abstract

Many of the features of Web 2.0 encourage users to actively interact with each other. Social tagging systems represent one of the good examples that reflect this trend on the Web. The primary purpose of social tagging systems is to facilitate shared access to resources. Our focus in this paper is on the attempts to overcome some of the limitations in social tagging systems such as the flat structure of folksonomies and the absence of semantics in terms of information retrieval. We propose and develop an integrated approach, social tagging systems with directory facility, which can overcome the limitations of both traditional taxonomies and folksonomies. Our preliminary experiments indicate that this approach is promising and that the context provided by the directory facility improves the precision of information retrieval. As well, our synonym detection algorithm is capable of finding synonyms in social tagging systems without any external inputs.

Keywords

Folksonomy, directory, social tagging, Flickr, and search improvement.

INTRODUCTION

In the spirit of Web 2.0, information sharing, user-focused design, and collaboration on the Web have been realized in many web applications. Many of the features of Web 2.0 encourage users to actively interact with each other. Social/collaborative tagging systems are one of the good examples that reflect this trend on the Web. The users annotate their resources such as bookmarks, photos, and videos with *tags or* keywords. The act of annotating resources is called *tagging* in such systems. Similar to the role played by tools such as directories, the primary purpose of social tagging system is to facilitate the access to resources; tagging helps users organize and retrieve resources on the Web.

Traditional information classification methods such as file systems represent a top-down approach; this approach stratifies unstructured information into narrower and more specific sub-categories. A folksonomy or social tagging system is a bottom-up approach because information is organized by users and the structure builds up in a bottom-up fashion. These two approaches are known to be orthogonal to each other although their primary objective of organizing information is the same.

Independent studies have been conducted using both these approaches in order to improve information retrieval and organization. Our focus is on the attempts in social tagging systems. In this paper, we propose a combined strategy that can overcome the limitations of both traditional information organization and social tagging systems, especially the latter. We tackle the existing problems in social tagging systems such as the flat structure and the lack of semantics by exploiting some of the features of both social tagging systems in combination with the more traditional information organization tool, directories.

We make use of the context that is created by grouping related objects into sets or collections in a social tagging system. We show how this context can be used to reduce the ambiguities in the form of synonyms and polysemous words to improve search in such systems. We present our experiments and evaluation, which demonstrate improvements in search based on the context provided directory-type structures extracted from Flickr.

The rest of the paper is organized as follows: Firstly, we briefly introduce the background and the previous works that are related to ours. Secondly, we present social tagging systems with directory facility and how we benefit from such a directory facility. Thirdly, we explain how our system tackles the flatness of structures and how we reduce ambiguity in social tagging systems followed by the description of how to detect synonyms in a social tagging system. Next, we discuss our experiments and evaluation and proceed to present our conclusions and suggested future work in the final section.

BACKGROUD AND RELATED WORK

The activities associated with organizing and categorising resources have been traditionally carried out by experts. Cataloguers and librarians are good examples. Although these professionals create high-quality organization of information, it requires advanced education and training in order to generate high-quality information organization (Mathes 2004). In addition, due to the huge amount of resources created as a result of information explosion on the Web, the professional organization of web resources has become increasingly impractical.

In light of these circumstances, social tagging systems have attracted significant attention. Their popularity arises from the two key strengths; a social tagging system does not require any professional organizers and organizing huge amount of information can be achieved collaboratively by ordinary users in such systems. These advantages make such systems practical in terms of cost-efficiency and feasibility for large data organization.

In social tagging systems, users can share their resources and annotations with other users. A *folksonomy,* a shared and evolving ad hoc classification structure, is created by the users who share resources and annotations. One of the drawbacks of the folksonomies is the flat structure of the information in the absence of semantic links among the tags

Due to the flat structure, the keyword search in such systems fails to find the precise and relevant information that is not properly tagged with the corresponding keyword. The flatness of information structure also interferes with the interaction between social tagging systems and traditional hierarchical taxonomy (Zhou et al. 2007). To mitigate these difficulties, many researchers have tried finding hierarchical relations from such a flat structure of information so as to improve the search as well as to make it easy to interact between different systems or machines.

Uncontrolled vocabulary is another issue in social tagging systems. When describing information with tags in a social tagging system, there is no limitation on the choice of vocabulary. Such systems allow users to use any free-form strings as tags. This uncontrolled vocabulary can lead to problems such as variations in naming tags, synonymy, and polysemy (Mathes 2004). Synonyms are the terms that represent the same or similar concepts or objects. These terms can impede retrieval performance in social tagging systems.

To overcome these limitations of folksonomies from a search perspective, many researchers attempted to construct hierarchical structures of tags from social tagging systems. Machine learning and data mining techniques have been used to induce hierarchical relations, but finding such relations among tags is difficult because of the limited information contained in tag sets (Heymann and Garcia-Molina 2006).

Focused on constructing hierarchical semantics from social tagging systems, Schmitz employed a statistical subsumption model based on the co-occurrence of tags (Schmitz 2006). The idea is as follows:

- Tag T_1 subsumes tag T_2 if $P(T_1| T_2) >= t$ and $P(T_2| T_1) < t$ representing T_1 is more general than T_2, where t is a specified co-occurrence threshold.

The subsumption model suffers from finding context of tags especially for polysemous terms because the model is purely based on tags.

Hierarchical structures can be constructed by means of clustering, for example, by splitting tags into more specific clusters recursively (Zhou et al. 2007). This clustering approach has trouble with identifying the context of tags as well.

Researchers have also tried to solve the problems caused by uncontrolled vocabulary such as detecting synonyms (Cattuto et al. 2008; Clements et al. 2008). Such problems are even harder to solve within social tagging systems by themselves.

Knowing the contexts of tags is useful because it can reduce the ambiguity of tags. For example, one can differentiate two distinctive objects tagged with the same polysemous word given some context. Or an object tagged with two synonymous words can be identified as the same if the two tags exhibit contextual similarities.

To overcome the lack of semantics in tags, Plangprasopchok and Lerman suggest the use of user-specified relations using sets and collections in Flickr instead of tags (Plangprasopchok and Lerman 2009). However, this work deals largely with how to build a hierarchical structure using only sets and collections without explaining how it can contribute to search performance and how it is connected to tags and objects in a social tagging system. Tags are after all the basic component in social tagging systems.

Directories provide context to tags/objects

A file- or directory-hierarchy can be useful when it is well organized (Golder and Huberman 2006). A well-organized hierarchy of directories unambiguously bounds directory's contents; grouping relevant contents into a directory disambiguates the contents within. Tags are often ambiguous. In this section we present how the ambiguities of tags in social tagging systems can be alleviated by integrating directory facility into such systems.

Directories in Social Tagging Systems

Typically, users in a social tagging system organize information using tags. Tags are the main tool to organize and retrieve information in such systems, and all the information is organized into a flat structure. In such a flat structure of information the retrieval of objects solely depends on the tags that are attached to them. This sole dependency limits the retrieval of information in such systems. To overcome such a limitation, social tagging systems need to be improved in terms of information organization as well as information retrieval.

In this situation, some of the social tagging systems now support the directory facility to help users organize their resources hierarchically. Flickr, a photo sharing web site, provides directory-like containers, called *sets* and *collections*, for grouping related photos. By grouping relevant resources in sets and collections in Flickr, directory-like information organization forms. To illustrate this idea, a user of Flickr may group her photos about her trip to Rome into a set titled 'Rome Trip'. She may have other sets about other trips around the world, and then she groups these sets into a broader collection titled 'Travel around the World'. These sets and collections help her organize the photos as if she uses the computer file system. Similarly, Del.icio.us, a social bookmarking site, provides users with *bundles* for grouping related web pages.

Context of Tags

Besides the hierarchical organization of information, the directory facility can also offer contextual information on tags. Figure 1 shows how objects (photos) are organized along with tags and sets in Flickr. Dashed-lines represent sets and two of the objects are annotated by the tag, 'Crane'. In a typical social tagging system, the system is not aware of the difference between these two objects because they are tagged with the same tag and no other information (i.e., sets here) is given. However, the additional information provided by directory facility in social tagging systems can remedy this ambiguity.

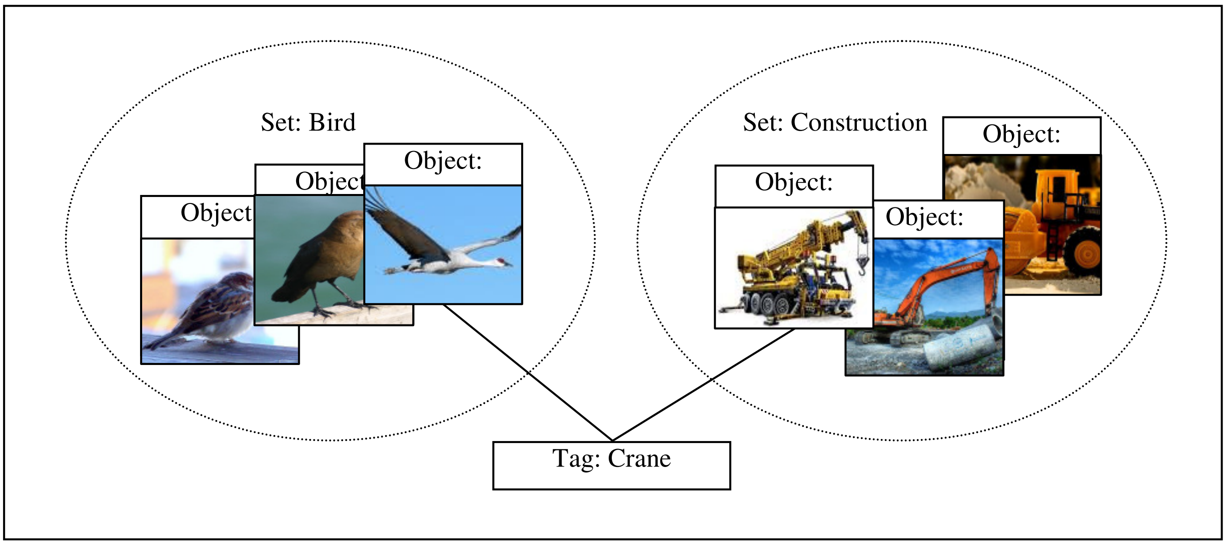

Figure 1: Context creation

As an example, the two photos in Figure 1 annotated with the tag 'Crane' can be distinguished because the photos and the attached tags are given context by grouping related objects into sets; the two photos with the same tag, '*Crane',* in two different sets, *Bird and Construction,* provide context to each object. We can utilize this context to disambiguate objects with the same tag.

The directory facility implemented in a social tagging system can further be utilized to facilitate the retrieval of information. The search in social tagging systems can be improved if we have more explicit information organization. For example, if one makes a search with the term, 'Crane', she can retrieve only the relevant objects (of either bird or construction equipment) because the system is already aware that there are two kinds of 'Crane'.

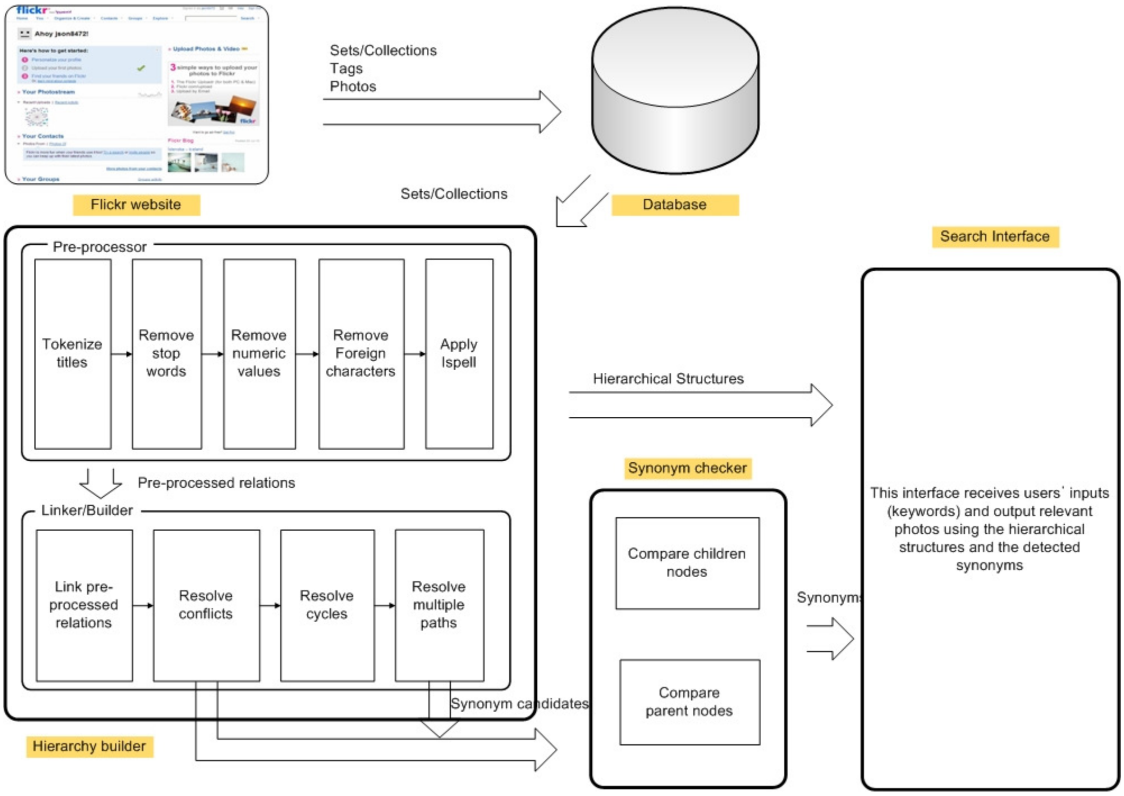

Figure 2: Overall system architecture

Overall system architecture

In most cases, the semantics of tags are not clearly specified in social tagging systems. The lack of such semantics degrades the quality of the information retrieval in such systems. As a result, researches to enrich semantics in social tagging systems have been carried out. The rationale behind these researches is that finding semantic relations among tags and building hierarchical structures based on these relations can improve information retrieval in social tagging systems (Angeletou 2009).

We exploit the sets and collections in Flickr, a tool for grouping photos, to build hierarchical structures instead of tags. We show how such hierarchies can improve the information retrieval in social tagging systems. In the following sections, we first present how the tags and objects are related to such structures and discuss how it can improve social tagging systems. Next, we describe the overall process of building hierarchical structures.

Connection between the structure and tags/objects

Although the hierarchical structure of sets and collections from Flickr can explicitly reveal some information about the system, it does not show how the tags and objects are linked to the sets and collections. There are three components in a typical social tagging system: users, objects, and tags; in our case, directories (sets and collections) are additionally considered. Users, objects and tags are linked to one another in the flat structure; there are no explicit hierarchical relations among these components. Meanwhile, directories create hierarchical relations along with users, objects, and tags in the sense that they can be placed in a hierarchy of directories either directly (i.e., objects) or indirectly (i.e., tags and users via objects). We propose that this information organization with the directory facility improves the information retrieval in social tagging systems where the components are organized in the flat structure.

To illustrate the idea, a keyword search is performed in order to retrieve relevant objects in a tagging system. As discussed in the foregoing, keyword search often misses relevant objects or returns irrelevant objects due to the ambiguities associated with tags such as polysemous words, synonyms, and different levels of specificity (Golder and Huberman 2006). These ambiguities can be reduced if we utilize the directory facility in a social tagging system. For example, if one makes a search in a social tagging system with the directory facility, a keyword-based search will be carried out in the flat information organization first and the objects linked to the keyword will be returned. The returned objects then can be filtered by the context provided by the directories (i.e., sets and collections in Flickr).

Building hierarchical structures using sets and collections

Our model to build the hierarchical structures is based on the model of (Plangprasopchok and Lerman 2009). In this section, we focus more on the difference instead of explaining the details of every step. The process of building hierarchies can be seen in Figure 2.

Collections and sets in Flickr have many features in common with tags; there are no restrictions to vocabulary in naming the titles of collections and sets. In addition, the users often use more than one word, when naming the titles, combined by a special character, preposition, and conjunction. As the result the titles are very noisy. To mitigate such noise, the titles need to be pre-processed (Plangprasopchok and Lerman 2009).

The titles are first tokenized to split the combined words. After tokenizing the titles of combined words, we remove stop words (i.e., 'I', 'you', 'the' and so on), numeric values and foreign characters in the titles. The titles with stop words are removed because stop words can create meaningless relations in the structure. Numeric values and foreign characters are discarded because our study focuses on the general concepts in English words. In the last step, we apply Ispell, a spelling checker, to correct spelling and support the morphological normalization.

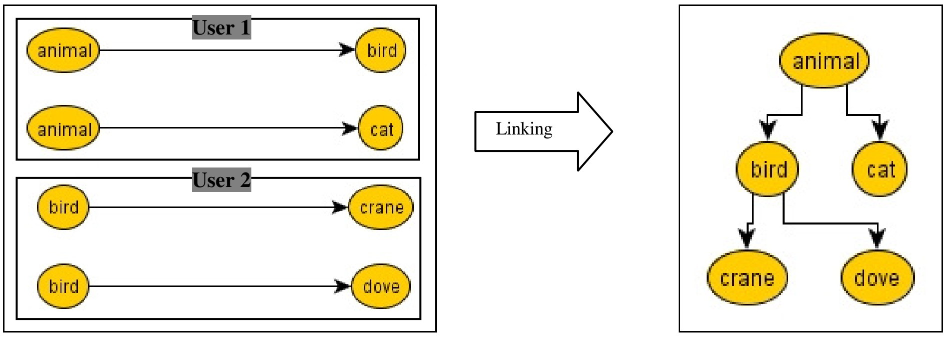

Figure 3: Linking relations

After the titles of sets and collections are pre-processed, relations are extracted from them. As an example, suppose that a user has a normalized collection titled 'animal' and two normalized sets titled 'bird' and 'cat' within the collection. Suppose again that another user has a pre-processed collection named 'bird' and two sets named with the normalized terms 'crane' and 'dove'. Users in Flickr tend to group more specific, related sets into a more general collection, which creates broad-narrow relations between them as shown in Figure 3. The relations from these two users are linked together into a deeper hierarchical structure on the right-hand side of Figure 3.

After linking all the relations from sets and collections, the overall structure becomes quite complex. Thus, we extract sub-hierarchical structures for the research purposes by specifying roots. Starting from a root, we follow the outgoing edges from the root to get the child-nodes and follow the child-nodes again to get their child-nodes, and so on, five levels down (to avoid infinite loops) or until we reach a leaf node.

The sub-hierarchical structures still require further processing; there could be conflicting relations, cycles, and multiple paths in the structure. The conflicting relations are caused by two opposite-directional edges between two nodes. The conflicting relations in the structure can be resolved based on the numbers of the users in both directions; for example, one of the edges is discarded if it is outnumbered by the other. However, there is a situation where both edges should be retained. This will be discussed in more detail in a later section. We employ a graph theory algorithm, Tarjan's algorithm to find the cycles in the structure (Tarjan 1972). The weights, numbers of parents and children, and depths of the nodes in cycles are taken into consideration for the resolution of cycles. Different from (Plangprasopchok and Lerman 2009), the multiple paths between two nodes are processed as follows.

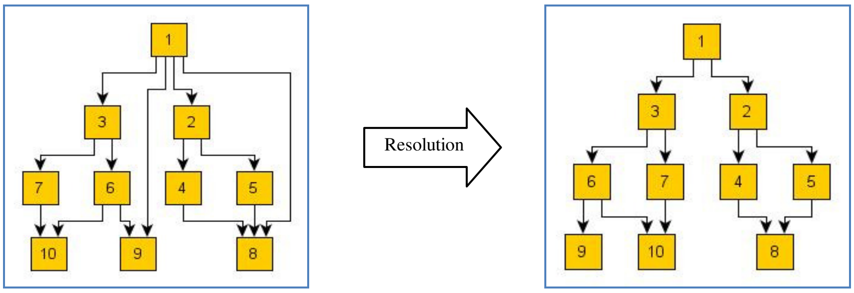

Figure 4(a, b): Multiple paths resolution

Figure 4 shows an example of a sub-hierarchical structure before and after applying the algorithm to resolve multiple paths. Given a specified root, the depth of each node is defined as follows. The depth of a node is the longest distance from the root node. For example, the depth of node 8 in Figure 4 before the resolution is 3; the direct edge from node 1 is not counted for the depth. After all the depths are specified, the algorithm for resolving multiple paths proceeds, as explained in table 1. We start from nodes 8, 9, and 10 at the detph of 3, which is the lowest level. Node 9 has two parent nodes 6 and 1. We examine if there are multiple paths between 9 and 1. If one of the multiple paths goes through node 6, which is one of the parents for node 9, the other path is removed. The longest path is kept here because it is more informational. Next, node 10 has two parents, node 6 and 7. In this case, although there are multiple paths between node 10 and 3, neither is discarded. We will explain the reason later. For node 8, the shortest path between node 8 and 1 is removed but the other two paths stay.

Table 1. Pseudo code of resolving multiple paths

```
Input: an ordered list L of the vertices V in the hierarchical structure S with multiple paths
Output: the hierarchical structure S without multiple paths;
Initialization: curV = S.L.head;

while (cur V != S.L.tail)
    if curV.parents.size > 1
        ResolveMultiplePaths(curV) ;
    curV = S.L.next;
return S;

Procedure: ResolveMultiplePaths (Vertex V)
Initialization: the current parent node  P = V.parents.head;
while (P != V.parents.tail)
    if (P.depth + 1 != V.depth && there is other path between V and P )
        V.removeParent(P);
        P.removeChild(V);
    P = P.next;
if V.parents.size > 1
    this is a clue to possible synonyms;
```

Synonym detection

While constructing the hierarchical structures, we come across conflicting relations and multiple paths. These often are considered unnecessary or redundant in the resulting structure. However, the conflicting relations and multiple paths sometimes give a clue to detect synonyms in certain circumstances. In this section, we describe what the circumstances are and how we detect synonyms.

Finding Synonym candidates from conflicting relations

Conflicting relations appear in the structure because two groups of users may have different understandings of how the resources should be organized; one group may have a set 'america' within a collection 'united states' while the other have a set 'united states' within a collection 'america'.

These two contrary information organization reflect each group's preference to organize the objects (Clements et al. 2008). If one group's preference overwhelms the other with far more number of users, this conflict can be resolved by discarding the preference of the minority. However, if two groups have the equal or similar number of users, this can be a clue to finding synonyms in the structure. To draw synonym candidates from conflicting relations, the following conditions have to be satisfied:

- The conflicting relations in both directions should have significant numbers of users.
- The two conflicting nodes should have the same or similar parent- and child-nodes.

The second condition is necessary because if two words are to be synonyms to each other, they must show much similarity in parent- and child- relationships. To compare the parent- and child-nodes, similarity measurements are needed. The details of the comparison will be discussed later.

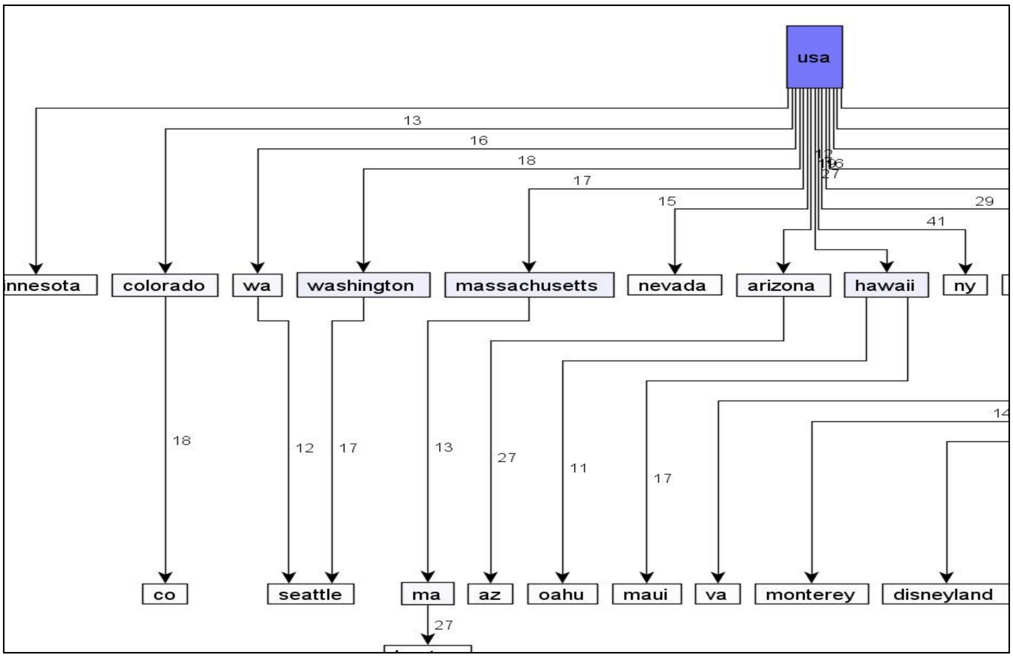

Figure 5: An example of synonyms ('wa' and 'washington') in a hierarchical structure

Finding synonym candidates from multiple paths

As mentioned earlier, not all the multiple paths are discarded by the algorithm. The retained multiple paths are considered significant because they can provide some clue to detecting synonyms, as in the case of conflicting relations.

A hierarchical structure is shown Figure 5 after applying the resolution algorithm for multiple paths. Two paths between the nodes 'usa' and 'seattle' still remain in the structure. This reveals the following facts:

- 16 users express that 'wa' should be a child of 'usa'.
- 18 users express that 'washington' should be a child of 'usa'.
- 12 users express that 'seattle' should be a child of 'wa'.
- 17 users express that 'seattle' should be a child of 'washington'.

The four edges are supported by more than 10 users. None of them is considered trivial.

The following assumption is made to select synonym candidates from the structure:

> When there are two paths between two nodes in the structure, the middle nodes in each path have high probability of representing the same concept if the two middle nodes have similar parents and children nodes in the structure.

Parent- and children- nodes similarities for synonym candidates

Once all the pairs of the nodes from the conflicting relations and the multiple paths are found, the parent- and child-nodes of the synonym candidates are ranked by measuring the similarities. One of the applicable similarity measurements is the cosine similarity (Cattuto et al. 2008). Suppose that nodes 'new york city' and 'nyc' are one of the synonym candidate pairs found during the construction, and that the node 'new york city' has the children nodes of 'central park', 'manhattan', 'brooklyn', and 'building' and the node 'nyc' of 'manhattan', 'brooklyn', 'building', 'metropolitan museum', and 'wall street'. Two row vectors from the nodes A and B are created as follows. Firstly, the domain, the union of all the children from the nodes 'new york city' and 'nyc', is found. Secondly, the first row vector (which denotes the node A's children) is created with the size of the domain. Thirdly, each column is marked with one if the node A has the child in the column or zero otherwise. Finally, the second vector is created in the same way. These two vectors can be compared with the bitwise-AND to measure the similarity. The comparison for the parents is carried out in the same manner. The bitwise-AND of the children vectors is shown in Table 1. In this case, the similarity of the children between the nodes 'new york city' and 'nyc' is 50.0%.

Table 2. Bitwise AND of two vectors for children nodes

Node 'new york city'

'central park'	'manhattan'	'brooklyn'	'building'	'metropolitan museum'	'wall street'
1	1	1	1	0	0

Node 'nyc'

'central park'	'manhattan'	'brooklyn'	'building'	'metropolitan museum'	'wall street'
0	1	1	1	1	1

Bitwise AND of 'new york city' and 'nyc'

'central park'	'manhattan'	'brooklyn'	'building'	'metropolitan museum'	'wall street'
0	1	1	1	0	0

Empirical Results and Evaluation

Two experiments were conducted: one for detecting synonyms during the construction and the other for contextual search. Hierarchical structures using sets and collections in Flick were built in the first experiment, focusing more on the detection of synonyms than on the construction of the structure. In the second experiment, the search with a list of ambiguous keywords was carried out to see if the context provided by the directory facility improves the quality of information retrieval.

Data Set

We gathered data about collections and sets created by a subset of Flickr users. To gather data, we first randomly chose a group of 74238 members. Flickr API methods have been used to collect the collections and sets of these users. The total number of sets and collections are 798849 and 136027 respectively. For the contextual search, the information of the photos and the tags grouped in the sets and collections were collected. 38817700 photos and 5413130 tags have been gathered so far.

After pre-processing the sets and collections, we obtained 1136270 relations. Note that we extracted the relations not only from collection/set relations but also from collection/collection relations; it is possible for a collection in Flickr to contain other collections within it.

Experiment for synonym detection

The first experiment was conducted during the construction of hierarchical structures. While building hierarchical structures with various roots, all the pairs of synonym candidates were found from the conflicting relations and multiple paths. As previously mentioned, the pairs of synonym candidates were compared to measure the similarity of their children and parents.

Several hierarchical structures with various roots (i.e., 'usa', 'nature', 'music', 'architecture', and so on) were tested to get the synonym candidates. Once the synonym candidates were extracted from the hierarchies, the candidates were tested using cosine similarity to make the final decision. The results indicate that our synonym detection algorithm is capable of detecting correct synonyms to some extent; the accuracy varies from 20% to 50% for the top ten pairs in each structure.

Table 3. Synonyms detected from conflicting relations/multiple paths

Set titles	Collection titles	# of shared nodes	# of domain	Cosine similarity (%)
washington	**dc**	**48**	**133**	**36.09022556**
utah	arizona	30	92	32.60869565
florida	**florid**	**36**	**132**	**27.27272727**
chicago	nyc	62	235	26.38297872
usa	ny	42	162	25.92592593
us	**usa**	**46**	**181**	**25.41436464**
wa	seattle	37	147	25.17006803
ma	**massachusetts**	**14**	**56**	**25**
boston	nyc	46	187	24.59893048
california	**ca**	**44**	**188**	**23.40425532**

Table 3 lists top 10 pairs of the detected synonyms in the descending order of the cosine similarity from the hierarchical structure with the specified root, ‘usa’. As indicated in Table 2, five (marked bold) of the pairs are correct synonyms; we consider abbreviation of word as synonyms as well because they represent the same concepts.

Preliminary result for contextual Search

When performing a search with a keyword in a typical social tagging system, the system only returns the objects that are tagged by the keyword. Even when performing a search with more than one keyword, such a system finds objects tagged by each keyword separately and the union of the objects are returned. The keyword-based search is poor in terms of the information retrieval precision and recall.

We suggested that the context provided by the directory facility (sets and collections in Flickr) can improve the search performance in social tagging systems earlier. To test this, a preliminary experiment was conducted as follows.

A list of polysemous keywords, which can ambiguously result in returning two types of objects, was prepared. For each keyword, a search is carried out without using any context at first. Next, a search with each keyword plus a descriptive word is carried out to specify the kind. Each descriptive word is used to exploit the context in our system; each object belongs to a set or collection; if the object is in the set or collection titled with the descriptive word, it remains, otherwise, is filtered. Table 4 shows that the search improves when searching with keywords ‘crane’, ‘windows’, and ‘bank’ with proper descriptive words.

Table 4. The results of the searches with/without the context

Search term	% of precision with context	% of precision without context
crane (Building)	77.9	38.2
windows (Operating system)	74.2	26.8
bank (Financial institution)	58.2	28.4

The above approach for the search improvement is rather naïve; because it improves the precision of information retrieval but not all the relevant objects are returned; the descriptive word does not always guarantee the search improvement if it is chosen improperly (i.e., too broad or narrow word) or simply the relevant objects might be in the set titled with other descriptive words reducing the number of relevant returned objects.

To overcome such limitations, we make use of the hierarchical structures as well as the detected synonyms. For example, a search with the keyword ‘crane’ returned hundreds of photos tagged by the keyword in our dataset. Table 5 lists some of the titles of the sets and collections that are linked to the tag ‘crane’ via photos. If one makes a search with the keyword ‘crane’ plus a descriptive word ‘building’, only the objects that are in the set or collection titled ‘building’ will be returned. However, if we exploit the hierarchical structure, all the objects within some domain (the sets/collections in red circle in Figure 6) can be returned. To illustrate this, 59 photos were initially returned when searching with the keyword ‘crane’ plus the descriptive word ‘building’ in our dataset; 46 out of 59 were relevant photos. If we exploit the hierarchical structure as in Figure 6, a search with the keyword ‘crane’ along with descriptive words such as ‘architecture’, ‘doors’, ‘church’ and so on can be carried out. The actual number of returned photos with these descriptive words is 102 and the precision is 76.4%; in this case, the number of returned objects is increased, maintaining approximately the same precision rate (78 out of 102).

Table 5. A subset of sets/collections that are linked to the tag ‘crane’

Title of set	Title of collection
bridges	structure
building	architecture
church	building
architecture	scenic views
bird	animal
grey crowned crane	bird

In addition, we can utilize the detected synonyms to improve the information retrieval. For example, the above search with the key and the descriptive words can be carried out along with ‘construction’, one of the synonyms to the descriptive word, ‘building’ which was detected during the construction of the hierarchy. This search returned 111 photos in total and the precision is 78.4% in our dataset.

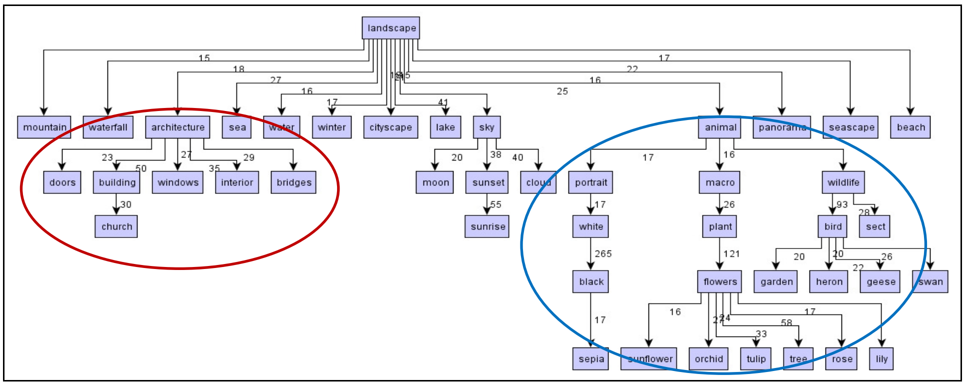

Figure 6: Sub hierarchical structure with the root 'landscape'

CONCLUSION AND FUTURE WORK

In this paper, a combined strategy with a traditional information organization tool, directories, was introduced in order to improve social tagging systems from the perspective of improved information retrieval. We also introduced a synonym detection algorithm and two sources of synonym candidates: conflicting relations and multiple paths. The experimental results indicate that search with the context provided by the directory facility significantly improves the information retrieval in terms of precision and the number of distinct objects (photos) returned. As well, our synonym detection algorithm is capable of finding a good proportion of the synonyms.

The hierarchical structures used can be further exploited to provide more information to improve social tagging systems. For example, integration with existing ontological structures for the domain can further improve the accuracy of the context and this can enhance the search facility in social tagging systems.

Another future work can be the visualization of the overall information. For example, a visual interface for users to browse information can be achieved from the hierarchical structures. This can provide users easier access to the overall information with visual navigation experience.

REFERENCES

Angeletou, S. 2009. "Semantic Enrichment of Folksonomy Tagspaces," *The Semantic Web - ISWC 2008*), pp 889--894.

Cattuto, C., Benz, D., Hotho, A., and Stumme, G. 2008. "Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems." from http://arxiv.org/abs/0805.2045

Clements, M., de Vries, A., and Reinders, M. 2008. "Detecting Synonyms in Social Tagging Systems to Improve Content Retrieval," *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore: ACM, pp. 739-740.

Golder, S., and Huberman, B. 2006. "Usage Patterns of Collaborative Tagging Systems," *Journal of Information Science* (32:2), pp 198-208.

Heymann, P., and Garcia-Molina, H. 2006. "Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems," Stanford University.

Mathes, A. 2004. "Folksonomies - Cooperative Classification and Communication through Shared Metadata."

Plangprasopchok, A., and Lerman, K. 2009. "Constructing Folksonomies from User-Specified Relations on Flickr," in: *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, pp. 781--790.

Schmitz, P. 2006. "Inducing Ontology from Flickr Tags," *WWW 2006*, Edinburgh, UK.

Tarjan, R. 1972. "Depth-First Search and Linear Graph Algorithms," *SIAM Journal on Computing* (1:2), pp 146-160.

Zhou, M., Bao, S., Wu, X., and Yu, Y. 2007. "An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations," in: *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea,* A. Karl, C. Key-Sun, N. Natasha, A. Dean, L. Kyung-Il, J.B.N. Lyndon, G. Jennifer, M. Peter, M. Diana, S. Guus and C.-M. Philippe (eds.). Berlin, Heidelberg: Springer Verlag, pp. 673--686.

COPYRIGHT