A Concept-Relationship Acquisition and Inference Approach for Hierarchical Taxonomy Construction from Tags

Eric Tsui, W.M. Wang, C.F. Cheung, and Adela S. M. Lau*

Knowledge Management Research Centre,

Department of Industrial and Systems Engineering,

The Hong Kong Polytechnic University, Hung Hum, Kowloon, Hong Kong

Abstract

Taxonomy construction is a resource-demanding, top down, and time consuming effort. It does not always cater for the prevailing context of the captured information. This paper proposes a novel approach to automatically convert tags into a hierarchical taxonomy. Folksonomy describes the process by which many users add metadata in the form of keywords or tags to shared content. Using folksonomy as a knowledge source for nominating tags, the proposed method first converts the tags into a hierarchy. This serves to harness a core set of taxonomy terms; the generated hierarchical structure facilitates users' information navigation behaviour and permits personalizations. Newly acquired tags are then progressively integrated into a taxonomy in a largely automated way to complete the taxonomy creation process. Common taxonomy construction techniques are based on 3 main approaches: clustering, lexico-syntactic pattern matching, and automatic acquisition from machine-readable dictionaries. In contrast to these prevailing approaches, this paper proposes a taxonomy construction analysis based on heuristic rules and deep syntactic analysis. The proposed method requires only a relatively small corpus to create a preliminary taxonomy. The approach has been evaluated using an expert-defined taxonomy in the environmental protection domain and encouraging results were yielded.

Keywords: Collaborative Tagging, Folksonomy, Natural Language Processing, Knowledge Capture, Semantic Web

* Corresponding Author Tel: (852) 2766 4026
Fax: (852) 2364 9663
Email: Adela.Lau@inet.polyu.edu.hk

1. Introduction

Folksonomy is the end product of a process by which many users add metadata in the form of keywords or tags to shared content. Folksonomy is also known as collaborative tagging, social classification, social indexing, and social tagging (Golder and Huberman, 2005). Recently, folksonomy has grown in popularity on the web, on sites that allow users to freely tag bookmarks, photographs and other content. Folksonomy is a common way of organizing content for future navigation, filtering, visualization, and search. Terms in a folksonomy can be freely chosen by the user; often there is no restriction or prior assumption imposed on the user to provide a tag. Because of this, folksonomy has certain advantages and disadvantages. Its advantages include easy-to-create and build up (as there is no prior learning needed), free of control and completely user-driven. As the tags are often created by the originator (and sometimes by other users as well), the tags often reflect the context of the prevailing knowledge sources. It is a convenient, low cost and dynamic framework for indexing user-generated content. However, folksonomy also has many disadvantages. Firstly, as it is uncontrolled, redundancies, incompleteness, and possibly inconsistencies commonly found in a folksonomy. Terms (Tags) can be in different word forms e.g. plural, singular, various word tenses and pronouns. Secondly, a folksonomy does not boast any hierarchy nor relationship between/among any of the terms. These shortfalls render reasoning with these terms extremely difficult without resorting to third party background knowledge. Up to now, nearly all the public sites adopting folksonomies are merely supporting keyword search on the tags and navigation via a tag cloud map (where more popular tags are displayed in larger fonts and in bold.) Thirdly, folksonomy tags may reflect user viewpoints and biases, and this may be in contrast with the nature and the context of the information indexed by the tags.

The word of “Folksonomy” is a portmanteau of the words folk and taxonomy. However, folksonomies are excluded from the concept of taxonomy (Mathes, 2004). Taxonomy is the type of controlled vocabulary where all the terms are connected by means of any structural model (hierarchical, tree, faceted, etc.) and specially oriented to browsing, organization systems and search of contents of the web sites (Centelles, 2005). Taxonomy is a representation of a reality in the most appropriate way for the purpose and interests of the entity. From the information architecture perspective, taxonomy is a basic or auxiliary tool for the various browsing, organization and content search, labeling and personalization systems. Taxonomy benefits include search, support, navigation, data control/mining, schema management, and personalization/information delivery (Benbya, et al., 2004). Any organization that needs to make significant volumes of information available in an efficient and consistent way to its customers, partners or employees needs to understand the value of a serious approach to taxonomy design and management. Hence, taxonomy is always considered as the initial step or general investment in managing organizational information.

Taxonomy is also a central part of most semantic models. Properly structured taxonomies help bring substantial order to elements of a model, are particularly useful in presenting limited views of a model for human interpretation, and play a critical role in reuse and integration tasks (Welty and Guarino, 2001). Its simple hierarchical structure makes it as the fundamental component for enriching other types of semantic models which consist of more complex structures. It provides identity, unity, essence, parthood, and dependence for ontology, which can be used as the foundation of a methodology for conceptual modeling (Guarino and Welty, 2000). However, taxonomy creation is a resource-demanding, top down, and time consuming effort. Without ongoing maintenance, any established taxonomy may become obsolete and compromise its effectiveness in reflecting the prevailing context of the captured information (Wood, 2004; Chosky, 2006; Connelly, 2007).

From the above description, taxonomy is a top-down, regulated way of classifying and representing information whereas folksonomy is a bottom-up, uncontrolled way of qualifying user-generated information. Both approaches have their pros and cons. However, in reality, the boundary between work and life interests is increasingly blurring; knowledge worker's learning and behavior are being influenced by, among others, their interactions with enterprise applications and various internet (public) systems. These applications and systems often have taxonomy terms and folksonomy tags embedded in them.

This paper addresses the integration of folksonomy tags into a taxonomy. The ultimate goal is to use the improved taxonomy to enhance knowledge navigation in a corporate environment. In order to solve the existing problems of taxonomy construction, this paper proposed a novel approach to convert tags into a taxonomy. The proposed method treats folksonomy as a rich knowledge source for harnessing tags for taxonomy construction and maintenance with particular reference to preventing taxonomy terms from being outdated easily. Newly acquired tags can be progressively integrated into the taxonomy in a largely automated way to facilitate the taxonomy creation process. The proposed method involves a heuristic rules analysis and a concept-relationship acquisition algorithm. The approach is evaluated with an established taxonomy in the environmental protection domain, an area in which abundant tags have been collected and that several authors have access to subject matter champions in the field. The method provides organizations with a low-cost way to articulate corporate taxonomies with publicly accessible folksonomies for enhanced organizational knowledge navigation.

2. Taxonomy construction

There are at least four different approaches to building a taxonomy, a manual method and a method that customizes one or more off-the-shelf taxonomies, using a taxonomy engine, and a combination of the above methods. This paper in particular deals with the automatic method of constructing a taxonomy.

2.1 Automatic taxonomy construction

In automatic taxonomy construction, many insights can be gained from prior work in ontology building. In this paper, we have adopted the approach of Hakeem and Shah (2004) and Dogac et al. (2002) in which taxonomy and ontology are not considered as interchangeable. According to them, taxonomy is the practice and science of classification. Taxonomy arranges entities in a hierarchical structure. It provides exact names for things within a domain and show which things are parts of other things (sometimes called parent-child or broader-narrower relationships). An ontology is analogous to a taxonomy in that it contains all the entities in the domain, and captures the relationships they have with each other. However, an ontology encapsulates a lot more than a taxonomy: it has strict format and domain theories about those entities and relationships (Dogac et al., 2002). This paper deals with the automatic construction of a taxonomy.

The automatic construction of taxonomy can be divided into 3 main approaches. The first approach is clustering approaches which are based on the distributional hypothesis that assumes terms are similar to the extent of sharing similar linguistic contexts (Harris, 1968). The second approach is based on matching lexico-syntactic patterns which often conveys a certain relation among texts in a corpus (e.g. Hearst, 1992). The third and dominant approach is to use machine-readable dictionaries for searching relations among terms (e.g. Alves et al., 2002; Rajaraman et al., 2002).

Clustering is a technique for partitioning data so that each partition (or cluster) contains only points which are similar according to some predefined metrics. Clustering approaches can be grouped in two classes: similarity-based methods and set-theoretical approaches. Latent Semantic Analysis (LSA) (Deerwester, et. al., 1990) underpins many of the clustering algorithms and it is a way of finding patterns among a collection of text documents. It is an instance for comparing similarities between terms from documents by using Singular Value Decomposition (SVD) and bag-of-words representation of text documents for detecting words with similar meanings (Pereira, et.al., 1993; Hindle, 1990; Faure and Nedellec, 1998; Caraballo, 1999; Bisson, et.al., 2000). An instance of set-theoretical clustering approaches can be found in (Sanderson and Croft, 1999). They proposed that a hierarchy between nouns is derived automatically by considering the document a certain term appears in as context. In particular, they present a document-based definition of subsumption relationship according to which a certain term t_1 is more special than a term t_2 if t_2 also appears in all the documents in which t_1 appears.

For the heuristic lexico-syntactic approach, it automatically discovers lexico-syntactic relations by searching for lexico-syntactic patterns in large text collections. Hearst (1992) identified several lexico-syntactic patterns and aimed at the acquisition of hyponym relations from Grolier's American Academic Encyclopedia. These patterns include:

1. NP_0 "such as" NP_1 ,NP_2 ... , (and | or) NP_n
   - e.g. …natural disasters such as earthquakes and shoreline erosion…
2. "such" NP_0 as NP_1 ,NP_2 ... , (and | or) NP_n
   - e.g. …works by such authors as Herrick, Goldsmith, and Shakespeare…
3. NP_1 ,NP_2 ... , (and | or) other NP_0
   - e.g. …bruises, wounds, broken bones or other injuries…
4. NP_0 (including | especially) NP_1 ,NP_2 ... , (and | or) NP_n
   - e.g. …all common-law countries, including Canada and England…

They imply for all NP_i, i > 1, hyponym(NP_i, NP_0), NP_i stands for a noun phrase. Similar to Hearst's approach, Charniak and Berland (1999) proposed for learning part-of relations. They proposed other 5 patterns:

1. NN_0's NN_1
   - e.g. …building's basement…
2. NN_1 of (a | and | the) NN_0
   - e.g. …basement of a building…
3. NN_1 "in" (a | and | the) NN_0
   - e.g. …basement in a building…
4. NN_1s of NN_0s
   - e.g. …basements of buildings…
5. NN_1s "in" NN_0s
   - e.g. …basements in buildings…

They imply part-of(NN_1, NN_0), where NN stands for a noun.

The other related approach is an automatic acquisition from machine-readable dictionaries, such as WordNet. WordNet (Miller, 1995) provides common knowledge information to match words based on linguistic relations between them (e.g., synonyms, hyponyms). Alves et al. (2002) used WordNet to extract an initial hierarchy of nouns from a document to build an initial list of concepts, followed by several user feedback iterations to deduce relationships between pairs of concepts and hypothesize about their relations. Paik et al. (2001) defined a representation called Concept-Relation-Concept (CRC) triples. Their system analyzes raw text to construct a database of CRC triples based on semantic relations. In contrast, Rajaraman et al. (2002) proposed a concept frame representation which focuses on searching for Noun-Verb-Noun structures in sentences based on syntactic relations. It applies WordNet for generalization of terms through sense disambiguation so that it does not require constructing any such domain dependent databases and it is able to identify concepts directly from the input text documents. There are also some methods in automatic ontology building, such as the idea of emergent relationships between tags (Boast et al., 2006) and the creative use of multidimensional clustering (Henegar et al., 2006). However, the present study is focused on building only a hierarchical taxonomy, so the issue of studying multiple relationships among tags is outside the scope of this paper.

In summary, taxonomy generation from unstructured data is a compelling problem in its own right. Naive clustering approaches do not cater for tag hierarchy generation, and this seems to be due to the structure of the data itself. Since similarity between parents and their children in a hierarchy does not seem to be sufficiently convincing for purely similarity based hierarchical clustering to produce useful results (Heymann and Garcia-Molina, 2006); the possibilities of establishing equivalence and hierarchical relationships between the categories is very limited. The result is usually a flat taxonomy, closer to a clustering of resources than a classification in itself (Centelles, 2005). While lexico-syntactic approaches are characterized by a relatively high precision in the sense that the quality of the learned relations is very high. However, lexico-syntactic approaches suffer from a very low recall rate which is due to the fact that the patterns are very rare. Lastly, the machine-readable dictionaries approach requires a priori conceptual hierarchy database to be constructed, which are often difficult to be extended/augmented. Furthermore, the precision and recall rate are highly dependent on the provided dictionaries.

3. Proposed methodology

In this paper, a new approach is proposed for taxonomy construction. Firstly, the tags are collected from the tag clouds of domain folksonomy websites. The folksonomy tags serve as the terms of the target domain taxonomy. Secondly, the taxonomy is automatically constructed based on heuristic rules and deep syntactic analysis which have been under explored when comparing with the other 3 approaches. The learning of taxonomic relations can be seen as a classification task. For example, given two terms, A and B, they could either be classified into four different ways: parent(A,B), parent(B,A), neighbor(A,B), or they could also be taxonomically unrelated. Heuristic rules approach traditionally has the characteristic of relatively low recall but high precision rate. In contrast, deep syntactic analysis has a higher recall but lower precision rate. Our methodology combines the 2 algorithms by applying a heuristic rules analysis first and then a concept-relationship acquisition algorithm to avoid the low recall due to heuristic patterns are rare to be found in tags. In addition, our experimental results have shown that the heuristic rules analysis has a higher priority due to its higher precision rate. Therefore, heuristic rules analysis is applied prior to the concept-relationship acquisition algorithm.

3.1. Heuristic rules analysis

Heuristics is a technique designed to solve a problem that ignores whether the solution has a correct proof, but which usually produces a good solution or solves a simpler problem that contains or intersects with the solution of the more complex problem (Clancy, 1993). Heuristics is often applied to enhance computational performance or achieve conceptual simplicity. In the present study, the heuristic rules analysis has been applied to detect the simple syntactic patterns among tags in a timely and effective manner. We apply 3 basic heuristic rules as shown below:

a) Rule 1: When one term is same as the other term and additionally modified by certain words or adjectives, the longer term is always categorized as a part of the shorter terms. Thus, a rule to detect vertical relations by a simple is-a rule, which is adapted from the research of Velardi et al. (2001). Basically, given two terms t_1 and t_2, if t_2 matches t_1 and t_1 is additionally modified by certain words, then the relation is-a(t_1, t_2) is derived. For example, t_1 = 'credit card' and t_2 = 'card', the relation is-a('credit card', 'card') is derived.

b) Rule 2: A rule to detect abbreviations. Given two terms t_1 and t_2, if t_2's alphabet letters matches the first letters of words of t_1, then t_2 is the abbreviation of t_1 and t_1 and t_2 are considered to have a neighbor relationship. For example, t_1 = 'natural language processing' and t_2 = 'NLP', the relation neighbor ('natural language processing', 'NLP') is derived. As abbreviations are abundant in folksonomy tags, this rule serves to group the abbreviation under its original (expanded) phrase so as to minimize the complexity of the taxonomic structure.

c) Rule 3: Some folksonomy tags may be made up of two different kinds of things into a single tag by using the word "and" or "or". In order to deal with this situation, a rule to detect vertical relation by simple syntactic patterns is proposed. Considering t_1 has a pattern NP_1, NP_2 ... , (and | or) NP_n, and $t_2 = NP_1$, or $t_2 = NP_2$, ... , or $t_2 = NP_n$, where NP_i stands for a noun phrase, then the relation is-a(t_2, t_1) is derived. For example, t_1 = 'Business Intelligence and Data Warehousing', and t_2 = 'Data Warehousing', the relation is-a('Data Warehousing', 'Business Intelligence and Data Warehousing') is derived. Since Rule 3 is contradicted with the Rule 1, hence Rule 3 is assigned a higher priority than the Rule 1. By combining the all the three rules, some other relationships among tags can be classified. For example, when t_1 = 'Computer Management and IT Management', and t_2 = 'Information Technology Management', t_3 = 'IT Management Strategy', the relation is-a(t_2,t_1), is-a(t_3,t_1), and is-a(t_3,t_2) can be derived.

3.2. Concept-relationship acquisition and inference

The second part of our method is based on a concept-relationship acquisition and inference algorithm which is adapted from an automatic concept mapping algorithm (Wang, et. al., 2008). The algorithm converts raw text documents into a "concept-relationship-concept" format. However, in our research, the relationships among the tags are classified into parent-child or neighbor relationship by mapping the tags with the concepts when there are two tags appear within a same sentence of a document. The tags are inferred and converted into a taxonomy based on the classified relationships. By considering the "concept-relationship-concept" format, the tags are regarded as the concepts and the semantic relationships among tags can be deduced based on their appearances in the raw text documents.

As shown in figure 1, the text document is first preprocessed by an ad hoc sentence boundary detection algorithm based on regular expressions. Then each individual sentence is parsed based on an augmented grammar for syntactic analysis. This step simultaneously tags each word with its part-of-speech (POS) using an in-house developed POS tagger and produces a parse tree for the sentence. The POS tagger is adopted from WordNet (Fellbaum, 1998). The notation of the tagset is shown in the Appendix. A detailed description of the tagset is available in (Satorini, 1990). Simultaneously, interjections, list item markers, and short sentences are filtered out in this process. For instance, in a simple sentence, the subject of a sentence represents a concept and the object of a sentence represents the second concept. The relationship between two concepts is identified by the main verb in the sentence. For a complex sentence, the sentence can be divided into several simple sentences. In order to solve the ambiguities arising in this process, anaphora resolution is applied based on rule based reasoning (RBR) and case based reasoning (CBR) (Kolodner, 1993) techniques. Traditional methods utilize simple rules to solve the anaphora resolution problem. They use more than a hundred rules which are difficult to resolve the conflicts among rules as well as the rules cannot be learnt adaptively. By integrating CBR and RBR into anaphora resolution, a smaller amount of rules needs to be maintained and the self-learning capability can also be achieved by continuously updating the case base.

Take in Figure 1

In RBR, syntactic rules are applied for extracting the noun phrases and verb phrases. Two examples of the rules are given as following:

np: <det><jj><np>+<in><vbg><jj>*<np>**

vp: <rb><vp)+<in><rb><vp>**

where <*det*> is determiner, <*jj*> is an adjective, <*np*> is a noun phrase, <*in*> is a preposition, <*vbg*> is a verb gerund, <*rb*> is an adverb, <*vp*> is a verb phrase, * is an operator that means zero or n occurrences of an item, and + is an operator that means at least one occurrence.

Since there are numerous different writing styles, RBR is not able to solve all the cases. If RBR is unable to solve the problem, it will pass the problem to the CBR module. CBR is a machine learning technique that searches for a similar case that was resolved in the past (Kolodner, 1993). In present study, each case for CBR consists of a problem and a solution. When there is an ambiguous sentence, the model extracts the POS classification of the sentence as a new problem. Then, it searches for similar case(s) in the case base that were successfully resolved in the past. The solution of the most similar case is adapted to this new situation finding the word that has the same syntactic function. The similarity between the new case and old cases is calculated based on the nearest neighbour matching (Aamodt & Plaza, 1994). The similarity is determined as follows:

$$Similarity = \frac{\sum_{j=1}^{m} w_j sim(v_j^o, v_j^r)}{\sum_{j=1}^{m} w_j} \qquad (1)$$

where m is the number of inputs, w_j is the weighting of the jth POS, v_j^o and v_j^r are types of the jth POS of the input case and that of the retrieved cases, $sim(v_j^o, v_j^r)$ is the similarity function for the jth POS as follows:

$$sim(v_j^o, v_j^r) = 1 \text{ if } v_j^o = v_j^r \qquad (2)$$

$$sim(v_j^o, v_j^r) = 0 \text{ if } v_j^o \neq v_j^r \qquad (3)$$

An example about the conjunction problem is depicted in figure 2, in which <*vp*> is verb phrase, <*np*> is noun phrase, and <*cc*> is conjunction. And the case is encoded into the case base based on the format show in figure 3. The problem set is encoded by storing the POS of the sentence, and the solution set is encoded by storing the corresponding positions of the POS, which is separated by commas and semicolons to form the concept-relationship-concept propositions.

Take in Figure 2 and Figure 3

By adopting the anaphora resolution, it not only helps to deal with the ambiguity, but also extracts the concept-relationship-concept format. The relationship between the concepts is classified based on a look-up table by a simple matching approach. In the look-up table, a list of words is classified as parent-child relationship (e.g. compose of, consist of, such as, etc.) and child-parent relationship (e.g. part of, type of, etc.). As shown in figure 4, if the relationship matches any of these words, the corresponding concepts are classified into corresponding relationship. For example, c_1 = 'Content Management' and c_2 = 'Content Protection' and the relationship between c_1 and c_2 is 'consists of' then, c_1 is parent of c_2. Hence, a concept hierarchy can be built, in which the links are representing parent-child relationships or unknown relationships among the concepts.

Take in Figure 4

After that, a mapping between tags and concepts is carried out. If a tag matches with the words of a concept, the tag is mapped with the concept. Therefore, after the mapping process, the original concept hierarchy is transformed into a tag and concept hierarchy which includes mapped tags and unmapped concepts. The relationships among tags are then inferred based on the hierarchy by 2 kinds of rules as defined below:

a) Rule 1: IF parent(t_1, c_1) AND {parent(c_1, c_2) AND parent(c_2, c_3) AND … } AND parent(c_n, t_2) THEN parent(t_1, t_2)
where t_i is a tag and c_i is an unmapped concept. An example is depicted in figure 5. If a tag t_1 is a parent of an unmapped concepts c_1, and c_1 is a parent of tag t_2, then t_1 is a parent of t_2. This rule identifies the non-direct is-parent relationship which is difficult to be found directly from the raw text data.

b) Rule 2: IF r_1(t_1, t_2) AND r_1(c_1, t_1) AND r_1(c_1, t_2) THEN neighbor(t_1, t_2)
where r_1() is an unknown relationship. An example is depicted in figure 6. If a tag t_1 has an unknown relationship with an unmapped concepts c_1, and t_1 has the same unknown relationship with a tag t_2, and c_1 has the same unknown relationship with a tag t_2, then t_1 and t_2 have a neighbor relationship. This rule identifies the non-direct is-neighbor relationship which is also difficult to be found directly from the raw text data.

Take in Figure 5 and Figure 6

In some cases, the relationships parent(t_1, t_2), parent(t_2, t_1) and neighbor(t_1, t_2) may occur at the same time. In this research, parent(t_1, t_2) and parent(t_2, t_1) are set in a higher priority than neighbor(t_1, t_2), since the parent-child relationships are more strictly inferred than that of the neighbor relationship. For parent(t_1, t_2) and parent(t_2, t_1), the relationship that has a higher number of counts (evidences) is always selected.

4. Evaluation

4.1 Evaluation Methodology

The above approach has been evaluated against an expert taxonomy in the environmental protection domain developed by U.S. environmental protection agency (USEPA) (http://www.epa.gov/). USEPA was established to consolidate in one agency a variety of federal research, monitoring, standard-setting and enforcement activities to ensure environmental protection. It employs 17,000 people across the country, including headquarters offices, 10 regional offices, and more than 12 labs. Its partnership programs address a wide variety of environmental issues by working collaboratively with companies, organizations, communities, and individuals. There are more than 13,000 firms and other organizations participating in the programs. The EPA website provides a comprehensive coverage of environmental topics and information. The taxonomy provided in its website is constructed by the domain experts for facilitating users' information navigation. It is a hierarchical structure, it is frequently updated, free of charge and open to public, and the quality is highly assured. Furthermore, there are plenty of tags created for the EPA website in the social bookmarking sites (e.g. http://delicious.com) and several authors have access to the very subject matter champions in this domain through one of their knowledge management consultancy projects. Hence, the EPA taxonomy is considered to be not only served as a representative taxonomy in the environmental domain, but also provides a significant and reliable benchmarking data for evaluation. All in all, 255 unique tags with 293 direct is-parent relations, 246 non-direct is-parent relations, and 1263 is-neighbor relations have been selected for different sets of evaluations. Examples of the relations are provided in figure 7.

To verify the scalability of our algorithm, we conduct tests with 25 tags to 125 tags with a 25 tag increment and a test with 255 tags. As shown in figure 8, tags are extracted from the taxonomy and converted from plural into singular form for analysis. As for the underlying corpus, 172 documents were collected from Wikipedia based on the search results of the extracted tags of the testing taxonomy (i.e. EPA expert taxonomy in this case). These documents represent the set of first 100% matched document for each tag; this ensures all the documents are definitely related to the selected tags. In this case, the number of documents ends up with less than the number of tags.

Take in Figure 7 and Figure 8

In the experimental evaluation, we compare our proposed approach with the 3 well-known approaches in taxonomy construction discussed in Section 2.2, they comprising of a clustering approach – LSA (Deerwester, et. al., 1990), a lexico-syntactic approach – Hearst (Hearst, 1992), and a machine-readable dictionary approach – WordNet (Miller, 1995). The results of the 4 approaches are shown in Tables 1 to 4 and Figures 9 to 12 in respectively in terms of recall and precision on the dataset. The recall and precision rates are measured based on the following equations:

$$recall_{direct-is-parent,a\lg orithm} = \frac{C_{direct-is-parent,a\lg orithm}}{N_{direct-is-parent}} \tag{4}$$

$$recall_{indirect-is-parent,a\lg orithm} = \frac{C_{indirect-is-parent,a\lg orithm}}{N_{indirect-is-parent}} \tag{5}$$

$$recall_{is-neighbor,a\lg orithm} = \frac{C_{is-neighbor,a\lg orithm}}{N_{is-neighbor}} \tag{6}$$

$$recall_{overall,a\lg orithm} = \frac{C_{direct-is-parent,a\lg orithm} + C_{indirect-is-parent,a\lg orithm} + C_{is-neighbor,a\lg orithm}}{N_{direct-is-parent} + N_{indirect-is-parent} + N_{is-neighbor}} \tag{7}$$

$$precision_{a\lg orithm} = \frac{C_{a\lg orithm}}{S_{a\lg orithm}} \tag{8}$$

where $N_{direct-is-parent}$, $N_{indirect-is-parent}$, $N_{is-neighbor}$ are the numbers of actual direct is-parent relations, indirect is-parent relations, and is-neighbor relations, respectively. $C_{relationship,a\lg orithm}$

is the correct number of a particular relation suggested by a particular algorithm. $S_{a\lg orithm}$ is the number of relations suggested by a particular algorithm.

4.2 Result and Discussions

With the LSA method, a tag-document occurrence matrix is formed and decomposed by singular value decomposition (SVD). The similarities among tags are calculated by comparing the vectors of the corresponding tags by cosine similarity. The most similar tag for each tag is classified to have a neighbor relationship. As shown in table 1 and figure 9, LSA has a relatively high is-neighbor recall rate and precision rate compared with the other approaches. However, the is-neighbor recall rate and precision rate drop significantly when the number of tag size increase. On the other hand, LSA is used for classifying neighbor relationship only, the recall for is-parent relation is 0.

Take in Table 1 and Figure 9

Using Hearst method, the following four syntactic patterns are used to detect is-parent relations:

1. NP_0 “such as” NP_1 ,NP_2 ... , (and | or) NP_n
2. “Such” NP_0 “as” NP_1 ,NP_2 ... , (and | or) NP_n
3. NP_1 ,NP_2 ... , (and | or) other NP_0
4. NP_0 (including | especially) NP_1 ,NP_2 ... , (and | or) NP_n

They imply for all NP_i, i > 1, hyponym(NP_i, NP_0), NP_i stands for a noun phrase. Results are shown in table 2 and figure 10. Since the patterns are rarely found in the documents, low recall rate is resulted in classifying direct and in-direct is-parent relations. Similar results are found by other researchers in the literature (Cimiano et. al, 2004b) with around 2% to 4% recall in applying Hearst to classify is-parent relations. Since Hearst approach is used to deduce is-parent relationship only, the recall of is-neighbor is 0. In general, Hearst has a relatively high recall in non-direct is-parent measurement and high precision rate compared with the other approaches.

Take in Table 2 and Figure 10

In WordNet method, the noun dataset of WordNet 2.1 is used in this evaluation -. It contains 117,097 nouns and 81,426 senses. There are 19 pointer symbols for nouns for representing the relationships between 2 nouns (e.g. hypernym, meronym, etc). They are classified into is-parent, is-child and neighbor relations in this research. All senses and non-direct is-a relations of the classification are taken into consideration. From the results as shown in table 3 and figure 11, the WordNet approach has the lowest recall. In the research by Cimiano, et.al (2004b) which for a given domain and evaluate with regard to a given concept hierarchy by using WordNet, the recall is 3.77% which is higher than this evaluation. By investigating the results manually, it is found that the incorrect classified relations are due to the different point of views of the taxonomy. For example, WordNet classifies that agriculture is a parent of aquaculture, however, USEPA classifies agriculture and aquaculture under different branches.

Take in Table 3 and Figure 11

The results of the proposed approach are shown in table 4 and figure 12. Based on the results, the recall of classifying is-parent relations is around 16% to 25% and 2 to 6% recall in classifying is-neighbor relations.

Take in Table 4 and Figure 12

A summary of evaluation results of all approaches with 255 tags is provided in table 5. From the results, we can see that the proposed method has outperformed the other three methodologies in the recall of classifying direct is-parent relations (improved from 0.0307 to 0.1638), the recall of classifying direct is-neighbor relations (improved from 0.0024 to 0.0578) and the overall recall (improved from 0.0161 to 0.0782). Although the precision of WordNet (0.2) is higher than that of proposed method (0.1610), WordNet has a very low recall compared with that of proposed method.

Take in Table 5

A series of student's t-tests was conducted to compare the recall (direct is-parent), recall (non-direct is-parent), recall (is-neighbor), overall recall, and precision of different tag sizes (i.e. 25, 50, 75, 100, 125, and 255) in LSA, HEARST, and WordNet with those in the proposed method, respectively. The results are shown in Table 6. Based on the results, one can see that nearly all the measures of the proposed approach were significantly better than those of the other approaches ($p < 0.05$). The exception is the comparison of recall (is-neighbor) in LSA with the proposed approach. The result shows that the proposed approach is not significantly better than that of LSA ($p = 0.099$). However, the results of mean, variance, standard deviation and standard error of the proposed approach all perform better than that of LSA.

Take in Table 6

To summarize, the proposed method improves the existing algorithms in certain degree. Traditionally, the clustering approach mainly focuses on the statistical analysis of tags among the documents and considering the neighbor relationships among tags. While, the proposed method makes further consideration on the semantic relationship among tags. It identifies not only the neighbor relationships, but also the parent-child relationships, so that it enables the construction of a hierarchical structure among tags. For the heuristic lexico-syntactic approach, it discovers the semantic relations by searching for lexico-syntactic patterns among the given documents. The proposed approach makes further inference among the extracted patterns so as to deduce hidden relationships among tags which are not directly written in the documents. The dictionary approach provides common knowledge information to match the tags based on linguistic relations between them. It is fast, easy to use, and reliable. However, it requires huge amount of effort to construct and maintain the dictionary. The proposed approach can be regarded as an automatic tool for suggesting new terms and relations for the dictionary approach.

5. Conclusion

This paper presented a novel approach of learning tags hierarchies based on a hybrid heuristic rules analysis and a concept-relationship acquisition algorithm. The approach is evaluated with regard to a public taxonomy of the environmental protection domain. A relatively small size of corpus was used for analysis and encouraging results were found by systematically comparing with other common taxonomy construction methods. In particular, the ability in classifying is-parent (i.e. is-a) relation have relatively high recall and precision rate. The proposed method was also compared with the other approaches for testing. The results showed that the combination leads to increased in both recall and precision.

The model serves as the fundamental component of other semantic models by analyzing the basic hierarchical relationships among tags. It provides an extra evidence for semantic relationship analysis. For future work, the model should be enriched for considering more different types of semantic relationships. The method should be also applied into different domains and data set in order to find out the scope, limitations, or optimal point of the method. The primary application of this research is used for constructing taxonomy based on folksonomy tags. These tags can be used to support facet navigation (Cheung et al, 2005). Further analysis will be carried out for filtering and selecting the meaningful tags for taxonomy construction. Future work will also be focused on integrating folksonomy and taxonomy so that their respective advantages can be leveraged.

Acknowledgements

The authors would like to express their sincere thanks to the Research Committee of The Hong Kong Polytechnic University for its financial support of this research (Project code G-YF58). The authors also like to thank Dr CC Kiu for her valuable contributions towards the experimentations and feedback on several versions of this paper.

References

Alves, A., Pereira, F., Cardoso, A. (2002). Automatic Reading and Learning from Text. Proceedings of the International Symposium on Artificial Intelligence.

Benbya, H., Passiante, G. and Belbaly, N.A. (2004), Corporate portal: a tool for knowledge management synchronization, International Journal of Information Management 24, 201–220

Bisson, G.; Nedellec, C.; and Canamero, L. (2000) ‘Designing clustering methods for ontology building - The Mo’K workbench’, Proceedings of the European Conference on Artificial Intelligence (ECAI) Ontology Learning Workshop, pp. 13-19.

Boast, Robin; Bravo, Michael; and Srinivasan, Ramesh (2006) Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge, The Information Society, Volume 23, Issue 5 October 2007, p.395-403

Centelles, Miquel (2005) Taxonomies for categorisation and organisation in Web sites, Hipertext.net, num. 3, <http://www.hipertext.net> ISSN 1695-5498

Charniak, E. and Berland, M. (1999) Finding parts in very large corpora, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 57–64.

Cheung, C.F., Lee, W.B., Wang Y. (2005) A multi-facet taxonomy system with applications in unstructured knowledge management, Journal of Knowledge Management, Volume: 9 Issue: 6, pp. 76-91

Chosky, Carol E. B. (2006). 8 Steps to Develop a Taxonomy, The Information Management Journal, Vol. 40, No.6, November/December, pp. 30-41.

Cimiano, P.; Hotho, A.; and Staab, S. (2004a) Clustering ontologies from text, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pp. 1721-1724.

Clancy, W. J. (1993) Notes on "Heuristic Classification", Artificial Intelligence, 59, p.191-196.

Connelly, J. (2007). Eight Steps to Successful Taxonomy Design, The Information Management Journal, Vol. 41, No. 6, November/December, pp. 41-16.

Deerwester, S., Dumais, S., Furnas, G., Landuer, T., and Harshman, R. (1990) Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, vol. 41, no. 6, 391-407

Dogac, A., Laleci, G., Kabak, Y., Cingil, I. (2002) ExploitingWeb Service Semantics: Taxonomies vs. Ontologies, IEEE Data Engineering Bulletin, Vol. 25, No. 4, December 2002.

Fellbaum, C. (1998), WordNet: An Electronic Lexical Database, MIT Press.

Golder, S.A. and Huberman, B.A. (2005), The Structure of Collaborative Tagging Systems, Information Dynamics Lab: HP Labs, Palo Alto, CA, available at: www.hpl.hp.com/research/idl/papers/tags/tags.pdf.

Guarino, Nicola and Welty, Christopher (2000) Ontological Analysis of Taxonomic Relationships, A.H.F. Laender, S.W. Liddle, V.C. Storey (Eds.): ER2000 Conference, Lecture Notes in Computer Science, vol. 1920, pp. 210-224

Hakeem, A. and Shah, M. (2004) Ontology and taxonomy collaborated framework for meeting classification, Proceedings of the 17th International Conference on Pattern Recognition, Volume 4, 23-26 Aug. 2004 p.219-222

Harris, Z. (1968) Mathematical Structures of Language, Wiley.

Hearst, M.A. (1992) Automatic acquisition of hyponyms from large text corpora, in Proceedings of the 14th International Conference on Computational Linguistics (COLING), pp. 539-545.

Henegar, Corneliu; Clément, Karine; and Zucker, Jean-Daniel (2006) Unsupervised Multiple-Instance Learning for Functional Profiling of Genomic Data, Machine Learning: ECML 2006, Volume 4212/2006, p.186-197

Heymann, P., Garcia-Molina, H., (2006) Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems, Technical Report InfoLab <http://dbpubs.stanford.edu/pub/2006-10>.

Mathes, Adam. (2004). Folksonomies: cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Miller, A. G. (1995) WordNet: A lexical database for English. Communications of the ACM, (38(11)):39–41, 1995.

Paik, W., Liddy, E. D., Liddy, J. H., Niles, I. H., and Allen, E. E. (2001) Information Extraction System and Method Using Concept-Relation-Concept (CRC) Triples. US Patent 6,263,335, Jul, 2001.

Rajaraman, K. & Ah-Hwee Tan. (2002). Knowledge Discovery from Texts: A Concept Frame Graph Approach. The 11th International Conference on Information and Knowledge Management.

Sanderson, M. and Croft, B. (1999) Deriving concept hierarchies from text, Research and Development in Information Retrieval, 206–213.

Santorini, Beatrice (1990) Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Steinbach, M.; Karypis, G. and Kumar, V. (2000) A comparison of document clustering techniques, KDD Workshop on Text Mining.

Velardi, P.; Fabriani, P.; and Missikoff, M. (2001) Using text processing techniques to automatically enrich a domain ontology, Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS), pp. 270-284.

Wang, W. M.; Cheung, C. F.; Lee, W. B.; and Kwok, S. K. (2008) Mining Knowledge from Natural Language Texts through Fuzzy Associated Concept Mapping, Information Processing & Management, Volume 44, Issue 5, September 2008, Pages 1707-1719

Welty, Christopher, and Guarino, Nicola (2001) Supporting ontological analysis of taxonomic relationships, Data & Knowledge Engineering, Volume 39, Issue 1, October 2001, Pages 51-74

Appendix – Tag set

1. cc: Coordinating conjunction
2. cd: Cardinal number
3. det: Determiner
4. ex: Existential there
5. fw: Foreign word
6. in: Preposition or subordinating conjunction
7. jj: Adjective
8. jjr: Adjective, comparative
9. jjs: Adjective, superlative
10. ls: List item marker
11. md: Modal
12. nn: Noun, singular or mass
13. nns: Noun, plural
14. nnp: Proper noun, singular
15. nnps: Proper noun, plural
16. pdt: Predeterminer
17. pos: Possessive ending
18. prp: Personal pronoun
19. prp$: Possessive pronoun
20. rb: Adverb
21. rbr: Adverb, comparative
22. rbs: Adverb, superlative
23. rp: Particle
24. sym: Symbol
25. to: to
26. uh: Interjection
27. vb: Verb, base form
28. vbd: Verb, past tense
29. vbg: Verb, gerund or present participle
30. vbn: Verb, past participle
31. vbp: Verb, non-3rd person singular present
32. vbz: Verb, 3rd person singular present
33. wst: Wh-determiner
34. wp: Wh-pronoun
35. wp$: Possessive wh-pronoun
36. wrb: Wh-adverb