Inducing Ontologies from Folksonomies using Natural Language Understanding

Marta Tatu, Dan Moldovan

Lymba Corporation
Richardson TX 75080 USA
marta@lymba.com, moldovan@lymba.com

Abstract

Folksonomies are unsystematic, unsophisticated collections of keywords associated by social bookmarking users to web content and, despite their inconsistency problems (typographical errors, spelling variations, use of space or punctuation as delimiters, same tag applied in different context, synonymy of concepts, etc.), their popularity is increasing among Web 2.0 application developers. In this paper, in addition to eliminating folksonomic irregularities existing at the lexical, syntactic or semantic understanding levels, we propose an algorithm that automatically builds a semantic representation of the folksonomy by exploiting the tags, their social bookmarking associations (co-occurring tags) and, more importantly, the content of labeled documents. We derive the semantics of each tag, discover semantic links between the folksonomic tags and expose the underlying semantic structure of the folksonomy, thus, enabling a number of information discovery and ontology-based reasoning applications.

1. Introduction

Social bookmarking has rapidly emerged as a tool that allows users to associate subjective descriptions to web pages. It helps its users organize and recall information of interest. Moreover, by sharing their bookmarks, users are able to identify other users with common interests as well as other resources of interest. Examples of popular social bookmarking sites include Delicious (www.delicious.com), Flickr (www.flickr.com), and BibSonomy (www.bibsonomy.com).

The labels used within social bookmarking settings generate a *folksonomy* (*folk* + *taxonomy*). This flat lexicon with all user tags contains inconsistencies: the users' uncontrolled vocabulary includes different types of variations and ambiguity, e.g., case sensitivity of tags, use of space or punctuation as delimiters, both singular and plural forms, same tag applied in different context, and synonymy of concepts (Golder and Huberman, 2005). Adam Mathes[1] notes *The sheer multiplicity of terms and vocabularies may overwhelm the content with noisy metadata that is not useful or relevant to a user*.

Advanced linguistic processing of tags results in a better organization and management of folksonomies as well as improved sharing of resources. By explicitly capturing and representing tag semantics in a more formal taxonomy (an ontology), the information structure of user tags is revealed, thus, facilitating machine understanding of user interests.

In this paper, we describe our efforts to derive ontological structures from folksonomies using natural language processing (NLP) and automatic ontology generation technologies.

1.1. Related Work

Various research groups have proposed algorithms that attempt to structure folksonomies. First, analyzes and comparisons of folksonomy with taxonomies and ontologies have emerged (Li et al., 2009; Peterson, 2006; Quintarelli, 2005). Various models that conceptualize tagging activities were proposed (Echarte et al., 2007; Gruber, 2008). Approaches to tagging in folksonomies have been dominated by a focus on the statistical analysis of tag usage patterns (Golder and Huberman, 2006), and information retrieval and navigation (Halpin et al., 2006), based on tag data. Statistical models for subsumption were used to derive folksonomic ontologies from annotations of image bookmarking (Clough et al., 2005; Sabou et al., 2006). The links identified between tags cover few relations types, most notable are: *type of*, *aspect of*, and *same-as*. Support for spelling corrections as well as integration of morphological tools have not been addressed yet. Methods that exclusively explore the social bookmarking annotations, more specifically, the tag co-occurrences among resources and users were also investigated (Mika, 2007). These algorithms employ graph-clustering procedures to connect tags, which were used by the same users for the same resources. Other studies propose an approach that combines the user-generated tag set with controlled vocabulary in order to develop an ontology (Chen and Qin, 2008). We note that no understanding of tags has been attempted.

Identifying mappings among large ontologies manually is an enormous task. Developing algorithms that automatically find candidate mappings is a very active area of research (Euzenat and Shvaiko, 2007). Existing techniques focus on calculating similarities between entities of two ontologies by utilizing various types of information in ontologies, e.g., entity names, taxonomy structures, constraints, and entities' instances.

1.2. Overview of Technical Approach

Because folksonomies are collections of tags, our initial efforts in designing a formal representation of folksonomies focused on the tags and their representation. For each tag, we derived a rich semantic representation that captures its concepts and any semantic relations that link them. Thus, each tag becomes a rich semantic graph that can be eas-

[1] http://adammathes.com/academic/computer-mediated-communication/folksonomies.html

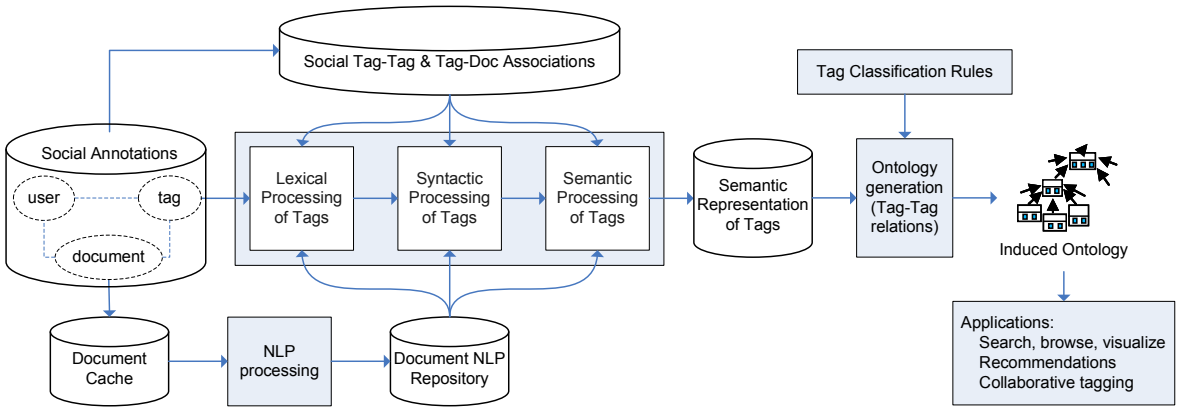

Figure 1: System architecture

ily exploited during the process of organizing the tags. For example, the tags *americanhistory* and *read-now* are represented as $American|JJ|1 \overset{TOPIC}{\longleftarrow} history|NN|2$ and $now|RB|3 \overset{TEMPORAL}{\longleftarrow} read|VB|1$.

We note that each concept part of a tag representation is linked to its corresponding WordNet synset ($American|JJ|1$ is part of synset id 2785615). These links enable the system to identify synonyms ((word, sense) pairs that denote the same WordNet concept). Moreover, metadata, including language and bookmarking information and frequency statistics, accompany each tag.

At the folksonomy level, semantic relations, such as ISA (ISA), PART-WHOLE/MERONYMY (PW), or SYNONYMY (SYN), link the tags, inducing the folksonomy's rich semantic structure. Thus, folksonomies become rich semantic graphs whose links are the semantic relations connecting the folksonomic tags, which constitute the nodes of the representation.

In order to derive rich semantic representations of tags, we developed mechanisms that normalize the lexical, syntactic, and semantic variations present in the folksonomic data. For this purpose, we exploited not only a tag's textual information, but also its associations with other tags and documents as created by the social bookmarking users. Once we captured each tag's meaning in a rich semantic representation, we used a series of classification steps that produce numerous tag-tag relationships, which complete the folksonomic ontology. SYN, ISA, PW, SIMILARITY (SIM), DOMAIN (DOM), ATTRIBUTE (ATR), and other relations between tags expose the folksonomy's ontological organization. Figure 1 displays our system's architecture.

1.3. Experimental Data

For evaluation purposes, we collected real-world social bookmarking data from the Delicious social bookmarking service. In Table 1, we present the statistics of our dataset (all bookmarks were stored publicly within Delicious between May 19^{th} and June 4^{th} of 2009). Using Lymba's suite of NLP tools, we processed cached versions of the bookmarked documents focusing on English textual documents whose content will be used to capture each tag's semantics (Section 2). We note that 10% of the tags (randomly selected) were used to score the performance of each tag understanding processing step. For this purpose, the system's output was compared with the gold annotations.

(user,document,tag)	148,709
(user,document)	113,313
unique number of users	58,198
unique number of documents	83,827
unique number of tags	8,460

Table 1: Statistics of experimental dataset

2. Capturing Tag Semantics

We broke down the tag understanding process into eight different linguistic processing steps. Each stage uses three sources of information that provide complementary information to our system: (1) **the tag space**: the text of each tag is used to derive information about the tag, (2) **the social bookmarking data**: tag associations augment and refine the initial understanding of a given tag, and (3) **the content of textual documents**: situating a tag within the larger semantic context of the documents it was assigned to enhances the existing understanding of a given tag.

2.1. Lexical Understanding of Tags

The lexical understanding of a tag includes the following stages: language identification, spelling corrections, tokenization and capitalization restoration.

During the tag **language identification** step, each tag text was matched against various dictionaries. For this purpose, we made use of Lymba's language identification module, which was expanded to include the 24 most frequent languages that we identified for our social bookmarking data[2], including Arabic, Chinese, Japanese, Russian, and many European languages. For definite match cases, the tag language was identified. If two or more languages had similar matching scores, the decision was made based on the

[2]We collected a document's language information as part of the metadata we downloaded when we created our local cache for each bookmarked URL. 34 languages were used to create the URL contents. British and American English (EN), German (GE), and Spanish (ES) and among the most frequent languages. Tamil (TA), Breton (BR), Glacian (GL), and Serbian (SR) are among the least frequently used languages.

language of the documents labeled with it. We note that "universal" words, such as the numbers and most technical terms and names (e.g., linux, css, google), were tagged as belonging to the English language.

By verifying whether a tag belongs to a certain language vocabulary, we also determine whether it is a single token. The tag **tokenization** step is important because many social bookmarking tools use *space* as a tag separator for user input (*americanhistory* is a valid Delicious tag). If not found as part of a language vocabulary, then (1) the tag contains two or more words "glued" together, which should be tokenized for a correct understanding of the tag, (2) the single token tag was misspelled and should be corrected, or (3) a combination of (1) and (2).

Typographical errors and spelling variations are abundant in folksonomies, requiring a mandatory **spelling correction** step. Thus, for each tag for which the system did not identify a matching vocabulary item, candidate concepts that correctly spell the tag include valid vocabulary concepts that minimize the edit distance to our input tag. Furthermore, depending on the length of the tag, the system will attempt to break the tag into multiple vocabulary items. Generated candidates are scored based on their presence within the content of the documents labeled with this tag. If they are part of the folksonomy and share documents or users with the analyzed tag (co-occurring tags), their scores are further boosted. For all tags split into multiple words, we score the generated phrase using probability values of English bigrams - random word phrases that do not "go together" are scored lower than valid English phrases. For further processing, the system will use the highest scoring variation of the tag text.

In order to **restore the capitalization** of tags that may denote proper names, we compare each tag text with the content of its corresponding documents. We note that the bookmarks include the correct spelling and capitalization information for a tag. Any competing values for a tag's capitalization are scored based on the position of the candidate within a document (English headlines capitalize the initial letter of all their content words; Sentences begin with a capitalized word regardless of the common capitalization of the word). The capitalization of a tag plays an important role in the process of identifying named entities.

All these processing steps transform the folksonomy from an unstructured set of tags into a collection of phrases, which are correctly spelled, capitalized, and tokenized. Links to the original tags exist. Some of the tags remain unchanged during this process. However, most tags are lexically normalized into well-formed phrases, which will be accurately processed by Lymba's suite of NLP tools.

Examples of tags modified by this processing step include: *linux* → *Linux*, *xhtml* → *XHMTL*, *bbq* → *BBQ*, *javascript* → *JavaScript*, *diy* → *DIY*, *christian_fiction* → *Christian fiction*, *amish* → *Amish*, *twitter,* → *Twitter*, *bradley/colin* → *Bradley / Colin*, *latex* → *LaTeX*.

2.2. Syntactic Understanding of Tags

For the **part-of-speech tagging** step, preference is given to the NOUN part-of-speech for single word tags, which cannot be tagged within context. Ambiguities are resolved by selecting the part-of-speech marked within the content of the tag's documents.

For tags with multiple tokens, the **syntactic parsing** step identifies the type of the tag phrase, its syntactic head as well as any grammatical dependencies between its constituents (Glaysher and Moldovan, 2006). This information is needed by Lymba's semantic parser as well as the ontology generation procedure (Section 3).

Several non-trivial examples which demonstrate the system's processing for the syntactic understanding of a tag include:

ushistory → *US history* → *US/NNP history/NN* → *(NP (NNP US) (NN history))*;

10.000+words → *10.000 words* → *10.000/CD words/NNS* → *(NP (CD 10.000) (NNS words))*;

christopher_hitchens → *Christopher Hitchens* → *Christopher/NNP Hitchens/NNP* → *(NP (NNP Christopher) (NNP Hitchens))*; and

toread → *to read* → *to/TO read/VB* → *(VP (TO to) (VB read))*.

2.3. Semantic Understanding of Tags

This processing stage covers the understanding of abbreviations/acronyms, the sense disambiguation of tags and the discovery of semantic relations within multi-word tags. The first two processing steps are the most challenging ones, as they require a broad context for the tag usage. Within folksonomies, social tagging systems, it can be argued that tags are primarily used to help the particular end-user who is submitting them (a tag is a set of words that defines a relationship between the online resource and a concept in a user's mind, freely chosen by the user without any formal guidelines). Thus, every user-selected word actually has a unique meaning. However, the increasing popularity of tagging systems and its social, collaborative effort to label existing content enabled users to browse and search vast bookmark collections, which lead to a natural convergence of tags (and their meaning) with few single-use tags (10-15%). Consequently, we depend on the content of the bookmarked documents to provide the context much needed for the disambiguation of each tag. For tags associated with non-textual documents (images, videos, audio files), we use co-occurring tags existent as part of the social bookmarking data.

First, we **disambiguate abbreviations** that either form or are part of tags using Lymba's abbreviation dictionary (118,055 abbreviations, 25% with multiple definitions). We rely on co-occurring tags and associated document content to determine the correct disambiguation. We build lexical chains of WordNet synset-synset relations between tag-describing concepts (co-occurring tags and associated document content) and candidate definitions (Moldovan and Novischi, 2002; Novischi and Moldovan, 2006). Because short chains indicate strong semantic similarity, we narrow the set of possible interpretations for an abbreviation. Further disambiguation is done using co-occurring tags and their meanings. Also, by aligning the abbreviation text with the document content (more specifically, its list of simple noun phrases), new definitions for abbreviations can be accurately identified and associated with the tags.

For instance, in our dataset, the tag *PR* is used to label 1409 documents. In our dictionary, there are 87 distinct definitions for this abbreviation, including, *Press Release*, *Public Relations*, *Puerto Rico*, *Page Rank*, *Public Radio*, *Permanent Resident/Residency*, etc. The contents of the documents labeled with this tag were vital to the semantic understanding of the abbreviation. For instance, when used to tag http://prsarahevans.com/2009/06/do-you-have-a-strategy-for-online-comments, *PR* denotes *public relations*, a phrase that appears in the content of the document six times. Other tags used to label the same document in our dataset include *public* and *relations*. On the other hand, when used to label http://www.bbc.co.uk/pressoffice/pressreleases/category/new_media_index.shtml, *PR* refers to *press releases* - also a frequent phrase in the document's content. A less frequent interpretation of *PR* is derived when it is used to tag http://escape.topuertorico.com. We note that none of the three documents included the abbreviation *PR*.

The semantic disambiguation process continues with the **sense identification** step which assigns each tag or tag concept its corresponding WordNet sense number. We rely on the content of the documents labeled with the tag. The word sense disambiguation process exploits the linguistic context of the analyzed word, which, in the case of folksonomic tags, is provided by their corresponding documents and the set of co-occurring tags. For tags that appear within their corresponding documents, we use the sense numbers derived by Lymba's word sense disambiguation module (Novischi et al., 2004; Novischi et al., 2007) during the semantic processing of the documents. For instance, tag *sign* used to label http://www.signingsavvy.com (Signing Savvy: Your Sign Language Resource) occurs in the document content and its linguistic context on sign language, American sign language, fingerspell, etc. pinpoint to its WordNet sense number 9 (a gesture that is part of a sign language). This sense value is also assigned to the tag concept. For tags that do not appear in the content of their associated documents, that label non-English or non-textual documents, we use the set of co-occurring tags to determine the correct sense of the tag (senses for the tag constituents). For example, when tag *sign* is attributed to http://www.nikonet.or.jp/spring/sanae/report/suusiki/suusiki.htm (Japanese document), we use the set of tags used to label this document to disambiguate *sign*. One of these tags is *mark*, concept synonymous with *sign|NN|1*. Part of another *sign* example, where this tag labels an image file (http://img179.imageshack.us/img179/6307/2172685295d8860567cbb.jpg), its co-occurring tags (*graffiti*, *pics*) pinpoint to its second sense (a public display of a (usually written) message).

We note that we use WordNet as our sense inventory. For non-WordNet concepts, Lymba's named entity recognizer associates named entity classes to tags. These can be derived from the content of the tag's documents, or based on the grammar rules and lexicons the module uses (no context is needed for certain named entity classes such as date, number, money, etc). For instance, the tag *christopher_hitchens* is used to label URL http://www.salon.com/news/1998/07/13news.html. The content of the document includes two mentions of this tag (in its normalized formed), both marked as human named entities during the document's processing through Lymba NLP pipeline. Tags such as *2009MAY* or *1960s* are easily identified as dates.

For single-word tags, this processing step produces a semantic representation of the tag, the system is now able to use the information extracted to link the tag with other tags as part of the ontology building process.

For multi-word tags, a **semantic parsing** step is required - the discovery of semantic relations that connect the tag concepts, thus completing the semantic understanding of the tag. Lymba's semantic parser (Bixler et al., 2005; Badulescu and Srikanth, 2007) uses a combination of semantic rules and machine learning classifiers to identify 26 semantic relations, including ISA, PW, AT-TIME/TEMPORAL (AT-T), INSTRUMENT (INS), and AGENT (AGT) (Balakrishna et al., 2010). The semantic parser relies of the senses assigned to each word as well as on their syntactic/grammatical dependencies (the syntactic parse of the phrase) to derive the correct semantic relation. Examples of semantic relations identified within tags include TEMPORAL(later, read) for tag *readlater*, PROPERTY-ATTRIBUTE-VALUE(primary, source) and ISA(primary source, source) for tag *primary_sources*, and INSTRUMENT(stick, fight) and PURPOSE(fight, stick) for tag *fightstick*.

2.4. Tag Understanding Evaluation

In Table 2, we summarize the performance of the system's tag understanding stages. Most errors occur when tags cannot be identified within their corresponding documents. Propagation errors from the capitalization restoration step account for future mistakes made during the part-of-speech tagging and named entity recognition stages.

Tag processing stage		*Accuracy*
lex	language understanding	97.87 %
	tokenization / spelling correction	97.16 %
	capitalization restoration	89.00 %
syn	part-of-speech tagging	93.26 %
	syntactic parsing	93.02 %
sem	abbreviation disambiguation	95.05 %
	word sense disambiguation	82.51 %
	named entity recognition	85.81 %
	semantic parsing	94.50 %

Table 2: Accuracy of individual tag processing stages

For our dataset of 8,460 tags, 91.65% of the tags were marked as belonging to the English language. Other most frequent languages include Spanish (1.85%), Portuguese (1.72%), and German (1.47%). We manually verified the correctness of language value attributed to 10% of the tags and the system is 97.87% accurate in assigning a language to a folksonomic tag. Because our dictionaries were built using Wikipedia in various languages, most errors occurred while assigning English as a language to non-English tags, which appear in English Wikipedia articles that contains many more entries when compared with other Wikipedia collections. Fewer errors are caused because the content

of the document labeled with the given tag was not available for language analysis. Examples include *davidleeroth* and *vanhalen*, which are used to tag an all-flash website whose textual content was not derived by our system (http://www.thetyser.com).

For our dataset of 7,754 English tags, the tokenization and spell correction procedure altered 25.03% of the tags. Its accuracy is 97.16% when evaluated on a randomly selected set with 10% of the folksonomic tags. The main source of errors stems from the "richness" of the English vocabulary as derived from the English Wikipedia articles. The unigram language model derived from this document collection includes, as single words, concepts that could be tokenized into multiple words, e.g., *googlemaps*, *blogpost*, *macosx*, *screenprinting*, *todo*, *searchengine*, etc. Because the untokenized version of the tag is found in the dictionary, no changes are attempted by the system.

The accuracy of the capitalization restauration processing step is 89.00% when compared to the manual annotations for a subset of 846 folksonomic tags. We note that errors made by previous tag understanding steps will propagate. Most errors occur because certain tag constituents appear only in document headlines/title and cannot be correctly disambiguated. An additional processing step that may improve these results will identify the correct capitalization of a tag or tag constituent within a much larger set of documents, not only documents labeled with that tag.

For the part-of-speech tagging task, the system's accuracy is 93.26% when the automatically generated output is manually verified for 10% of the folksonomic tags. Most errors are sourced by bad capitalization errors: adjectives whose first letter is capitalized by the previous processing step are wrongly identified as proper nouns (e.g., *international*, *urban*).

Bad part-of-speech tags lead to a bad syntactic parse of the tag. Thus, the system's accuracy for the syntactic parsing of folksonomic tags is 93.02%. We note that the syntactic structure of the tags is not complex, easing the parser's task.

The system's accuracy for the abbreviation disambiguation processing step is 95.05% for a randomly selected set of 10% folksonomic tags. We note that most tags are not abbreviations nor do they contain abbreviations or acronyms. Within the entire dataset, the system modified 3.86% of the folksonomic tags by expanding them or some of their components to the definition it considered appropriate. Most of the errors made at this processing step are due to well-established computer concepts such as *HTML*, *ASCII*, *USB*, *PDA*, which are not defined within the contents of the documents they label despite the fact that they may appear within the document. Many of these concepts are defined in many English dictionaries, such as WordNet. Very few abbreviations were not found in our compiled dictionary of acronyms and abbreviations and were not expanded.

The accuracy of the word sense disambiguation step on a randomly selected set of 10% folksonomic tags is 82.51%. There are several sources of error: (1) inherent word sense disambiguation errors caused by semantically close WordNet senses, (2) word sense disambiguation errors within document contents that propagate to the tags, (3) limited linguistic context for certain tags problem alleviated for analyses of a larger dataset.

For the named entity recognition step, our proposed system achieves an accuracy of 85.81% when evaluated on a randomly selected set of 10% of folksonomic tags. Most errors are due to the fact that named entity tags are not recognized as named entities (e.g., *Twitter*, *Apple*, *Java*, or *BBC*) when the tag is analyzed, but also within the content of the documents they appear in. Fewer errors are caused by the labeling of a tag with the wrong named entity information (e.g., *cycling* as award, *Dewey* as town, *rubik* as town). We note that within the entire dataset, 10.04% of the tags were marked as named entities.

The overall accuracy of the semantic parsing step is 94.50% when measured on a randomly selected set of 10% folksonomic tags. We note that the evaluation was not restricted to multiple word tags. Thus, errors propagated from previous processing steps are accounted for during this evaluation process. Within the set of analyzed tags, 17.6% of the tags were multi-word concepts. Most of the tags marked with an inaccurate semantic relation understanding were missing relations that would complete the tag's meaning. For instance, *cycling_blogs* (normalized to *cycling blogs*) is marked as having an ISA(cycling blogs, blogs) semantic relation without an additional TOPIC(cycling, blogs) relation. Fewer errors were cause by the "abnormality" of the tag. Because the word order is reversed, the semantic parser cannot derive the correct semantic relations that link the tag concepts (e.g., *things_japanese*, *Radio_Online*).

3. Deriving the Folksonomy's Structure from Tag Semantics

Once the process of understanding what each tag represents is complete, the focus of our research is shifted to the derivation of the folksonomy structure from the derived tag semantics. We begin by connecting tags using EQUALITY and SYNONYMY relations.

EQUALITY (EQ) relations link tags semantically normalized to the same form. Thus, EQ relations are created between tags with the same lemma, part-of-speech, and sense number (e.g, EQ(activity, activities), EQ(after-effects, AfterEffects), and EQ(opinion, Opnion)). This relation type links tags that are semantically normalized to the same form.

SYN relations link tags with identical synset ids, e.g. *Archeology* and *Archaeology*, *OS* and *operating.system*. These tags belong to the same synset in WordNet (for single word tags), thus deemed synonyms within WordNet. The synset id is derived based on the lemma, part-of-speech and sense number of the tag. For non-WordNet concepts, we use the named entity and abbreviation information to identify SYN relations between tags that refer to the same concept using different wordings, e.g. SYN(LA, losangeles), SYN (nyt, nytimes). We also create SYN relations between multi-word tags that have synonymous constituents linked by the same semantic relations. All SYN relations connect semantically similar tags. These links are not as strong as the EQ relationships.

Existing WordNet relations that link two folksonomic tags are also added to our ontology, e.g. ISA(vegan, vegetarian), PW(Businesses, markets), ENTAIL(proofreading, +read),

SIM(important, general), and DOM(light, physics). We note that this step is possible because each tag is sense disambiguated and a link to its corresponding WordNet synset is created. This procedure added 23.66% of the total number of relations to the automatically created ontology.

Furthermore, we build lexical chains (Moldovan and Novischi, 2002; Novischi and Moldovan, 2006) of size two between tags ($tag_1 \xrightarrow{rel_1} synset \xrightarrow{rel_2} tag_2$). We use Lymba's semantic calculus rules (Tatu and Moldovan, 2006; Tatu and Moldovan, 2007), which derive new semantic relations by combining two existing relationships, to add new tag-tag relations to our ontology (41.02% of the ontological relations were added using this procedure); e.g., ISA(integration, events,) derived from ISA(integration, group_action/NN/1) and ISA(group_action/NN/1, events,); PW(lobby, hotels) derived from PW(lobby, building/NN/1) and ISA(hotels, building/NN/1). We note that the concept connecting the two tags is not part of the folksonomy. If synset were itself a tag, then the semantic calculus rules would create a redundant relation, which would be removed by further processing.

Additional ISA relations are created between named entity tags and WordNet synsets that describe their corresponding named entity class, e.g., ISA (OracleCorporation, organization), ISA(davidfosterwallace, person). We note that most named entity tags are not defined within WordNet and these ISA relations are vital in describing the hierarchical structure of the folksonomy. These relations denote a directional semantic subordination of their arguments.

For complex tags of the form *mod head*, where *head* $\in$ folksonomy and REL(*mod*, *head*), we add ISA(*mod head*, *head*) relations, e.g. ISA(book-cover, covers), ISA(theoryofmind, theory), and ISA(photoshoptutorials, tutorials,). The relation linking *mod* and *head* can be a ATR, PW and even a TEMPORAL relation. This procedure accounts for 17.12% of the total relations added to the ontology. Examples include ISA(book-cover, covers), ISA(theoryofmind, theory), and ISA(photoshoptutorials, tutorials,).

For complex tags of the form mod_i $head_i$ where there exists a semantic relation REL(mod_i, $head_i$), (i=1,2), we analyze any semantic connections between mod_1 and mod_2 as well as between $head_1$ and $head_2$ in order to derive semantic links between mod_1 $head_1$ and mod_2 $head_2$. Thus, we add a new ISA relation between the tags if (1) ISA(mod_1,mod_2) and ISA($head_1$,$head_2$), (2) ISA(mod_1,mod_2) and SYN($head_1$,$head_2$), or (3) SYN(mod_1,mod_2) and ISA($head_1$,$head_2$). Also, if SYN(mod_1,mod_2) and REL($head_1$,$head_2$), where REL could be any semantic relation, then a new REL relation is added to the ontology. Examples include ISA(build-solar-panel, create-solar-panel), SIM(socialnetworks, socialweb) (based on the SIM(networks,web) which was derived using Lymba's semantic calculus rules - both nouns are derivations of the concept web|VB|1).

Sanity checks that ensure a consistent ontology structure, which can support applications involving the input folksonomy, include (1) identity and resolution of conflicts as well as (2) redundancy checks. We show a small portion from a generated ontology in Figure 2. Its nodes contain richer information not shown in Figure 2 (sample in Figure 3).

3.1. Relation Generation Evaluation

Given perfect input, the classification rules described above derive a highly-accurate set of tag-tag semantic connections. However, given the sense disambiguation errors, tags are placed into incorrect SYN clusters more than 17% of the time, affecting the relation generation process. The accuracy of this processing step is 80.30%, as measured on a randomly selected set with 20% of the total 5,439 relations.

For our social bookmarking dataset, our system created 9,820 EQ clusters for the 8,460 folksonomic tags. Most tag strings that belong to multiple EQ clusters are abbreviations expanded to different definitions for different bookmarks (e.g., *ST*, *OS*, or *AI*). However, there are EQ clusters that combine multiple unique tag strings (e.g., *tutorial*, *tutorials*, and *tutorials,*) - these tags were normalized (lexically, syntactically, and semantically) to the same concept.

Within the same dataset, our system derived 8,801 SYN clusters. Most of the tag strings that find themselves within different EQ clusters belong to different SYN clusters also. The largest SYN clusters groups 133 (user,document,tag) triplets where tag can be *car, automobiles, auto, autos, cars*, or *automobile*. Other large SYN clusters include {*movies, movie, Movies:, film, films*}, {*gadgets, widget, widgets, gadget, appliance*}. The SYN clusters are determined by the semantic understanding of each tag (associated with a certain bookmark). Thus, any errors made by the system when creating the clusters of synonymous tags were caused by mistakes made during earlier processing stages, most notably the word sense disambiguation step.

Among these SYN clusters, our system identified 5,439 ontological relations using the classification procedures described above. This set of relations uses 11 types of semantic relations. The most frequent is ISA with a total of 3,869 instances, followed by SIM (601 instances), PW (429 links) and others such as DERIVATION, DOMAIN, ANTONYMY, etc. The SYN cluster that is linked the highest is {*humans, person, human*} which participates in 89 semantic relations. The most prolific source of semantic relations is WordNet when combined with Lymba's semantic calculus rules. There were 1,778 ontological relations derived using this procedure.

4. Conclusion

In this paper, we described a method that exposes the latent semantic structure of folksonomies, thus, eliminating their inconsistency problems and linking semantically related tags. The resulting structure is a rich graph with nodes that represent clusters of synonymous tags and labeled directed links that denote the semantic relations that connect the folksonomic tags. Projections of this graph, which include only relationships such as ISA and PW, reveal hierarchical organizations of the folksonomy which can be exploited by social web applications.

5. References

Adriana Badulescu and Munirathnam Srikanth. 2007. LCC-SRN: LCCs SRN System for SemEval 2007 Task 4. In *Proceedings of the Fourth International Workshop*

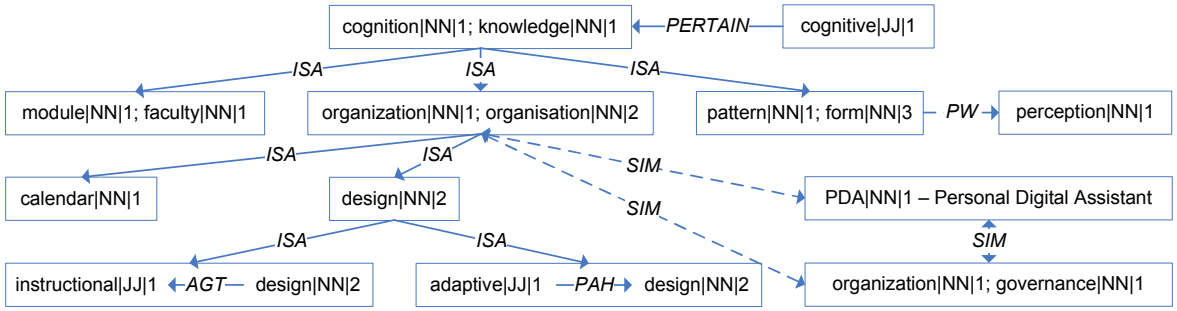

Figure 2: Sample portion of automatically induced folksonomic ontology

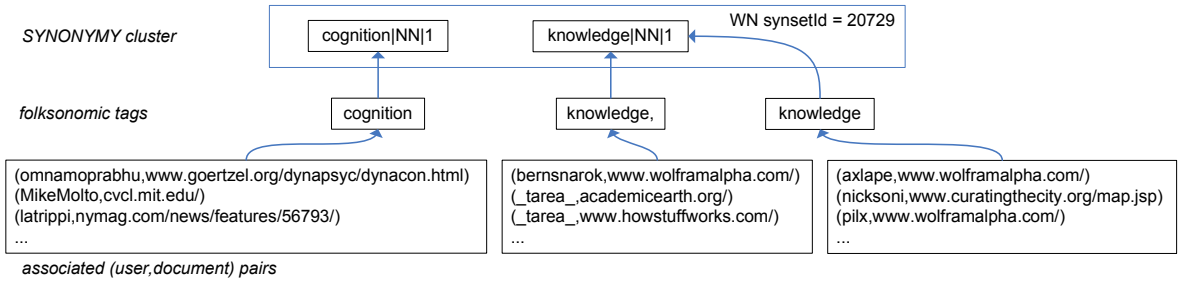

Figure 3: SYN cluster of normalized tags with social meta information

on Semantic Evaluations (SemEval-2007), pages 215–218, Prague, Czech Republic, June.

M. Balakrishna, D. Moldovan, M. Tatu, and M. Olteanu. 2010. Semi-automatic domain ontology creation from text resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May.

David Bixler, Dan I. Moldovan, and Abraham Fowler. 2005. Using Knowledge Extraction and Maintenance Techniques to Enhance Analytical Performance. In *Proceedings of the 2005 International Conference on Intelligence Analysis*, Washington D.C.

M. Chen and J. Qin. 2008. Deriving ontology from folksonomy and controlled vocabulary. In *Proceedings of iConference*, Los Angeles, California, February.

P. Clough, H. Joho, and M. Sanderson. 2005. Automatically organizing images using concept hierarchies. In *Proceedings of the SIGIR Workshop on Multimedia Information Retrieval*, Salvador, Brazil, August.

F. Echarte, J. J. Astrain, A. Cordoba, and J. Villadangos. 2007. Ontology of folksonomy: A new modeling method. In *Proceedings of Semantic Authoring, Annotation and Knowledge Markup (SAAKM), CEUR Workshop*, Whistler, British Columbia, Canada, October 28-31.

J. Euzenat and P. Shvaiko. 2007. *Ontology Matching*. Springer, New York, NY, USA.

Elliot Glaysher and Dan Moldovan. 2006. Speeding Up Full Syntactic Parsing by Leveraging Partial Parsing Decisions. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 295–300, Morristown, NJ, USA. Association for Computational Linguistics.

S. Golder and B.A. Huberman. 2005. The Structure of Collaborative Tagging Systems. Technical report, HP Labs. Available at http://www.hpl.hp.com/research/idl/papers/tags.

S. A. Golder and B. A. Huberman. 2006. The structure of collaborative tagging systems. *Journal of Information Sciences*, 32(2):198–208.

T. Gruber. 2008. Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics*, 6(1):4–13.

H. Halpin, V. Robu, and H. Shepard. 2006. The dynamics and semantics of collaborative tagging. In *Proceedings of the First Semantic Authoring and Annotation Workshop (SAAW)*, Athens, GA, USA, November.

J. Li, J. Tang, Y. Li, and Q. Luo. 2009. Rimom: A dynamic multi-strategy ontology alignment framework. *IEEE Transaction on Knowledge and Data Engineering*, 21(8):1218–1232, August.

P. Mika. 2007. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, 5(1):5–15.

Dan I. Moldovan and Adrian Novischi. 2002. Lexical Chains for Question Answering. In *Proceedings of COLING 2002: The 19th International Conference on Computational Linguistics*, pages 1–7, Taipei, Taiwan, August.

Adrian Novischi and Dan I. Moldovan. 2006. Question Answering with Lexical Chains Propagating Verb Arguments. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 897–904, Sydney, Australia, July.

Adrian Novischi, Dan I. Moldovan, Paul Parker, Adriana Badulescu, and Bob Hauser. 2004. LCC's WSD Systems for Senseval 3. In *Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Sys-*

tems for the Semantic Analysis of Text, Barcelone, Spain, July.

Adrian Novischi, Munirathnam Srikanth, and Andrew Bennett. 2007. LCC-WSD: System Description for English Coarse Grained All Words Task at SemEval 2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 223–227, Prague, Czech Republic, June.

E. Peterson. 2006. Beneath the metadata: Some philosophical problems with folksonomy. *D-Lib Magazine*, 12(11), November. www.dlib.org/dlib/november06/peterson/11peterson.html.

E. Quintarelli. 2005. Folksonomies: Power to the people. ISKO Italy-UniMIB meeting: http://www.iskoi.org/doc/folksonomies.htm, June.

M. Sabou, M. D'Aquin, and E. Motta. 2006. Using the semantic web as background knowledge for ontology mapping. In *Proceedings of the 1st International Workshop on Ontology Matching*, pages 1–12, Athens, Georgia, USA.

M. Tatu and D. Moldovan. 2006. A Logic-based Semantic Approach to Recognizing Textual Entailment. In *Proceedings of COLING/ACL 2006*, Sydney, Australia, July.

M. Tatu and D. Moldovan. 2007. Cogex at rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 22–27, Prague, Czech Republic, June.