Applied Ontology 8 (2013) 95–130
DOI 10.3233/AO-130124
IOS Press

Exposing, formalizing and reasoning over the latent semantics of tags in multimodal data sources

John Tyler [a], Jon Pastor [b], Michael N. Huhns [c,*], Shad Kirmani [c] and Hongying Du [c]

[a] *Altamira Technologies Corporation, McLean, VA, USA*
E-mail: John.tyler@altamiracorp.com
[b] *College of Information Science and Technology, Drexel University, Philadelphia, PA, USA*
E-mail: jon.a.pastor@drexel.edu
[c] *Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA*
E-mails: huhns@sc.edu, shad.kirmani@gmail.com, du5@email.sc.edu

Abstract. This paper describes our solution to the problem of inducing ontological information from metadata provided informally. The metadata is in the form of linguistic tags attached to items in an on-line domain. We formulate four hypotheses about the structure implicit in a set of tags and evaluate them using data from public tag sets. The results confirm three of the four hypotheses and show that it is feasible for ontological information – specifically subclass relationships – to be made explicit and hence available for inferencing.

Keywords: Folksonomy, tagging, inducing an ontology

Accepted by: Guido Boella

1. Introduction

We have been investigating the problem of inducing information from metadata provided informally via a set of tags describing objects – both physical and conceptual – and mapping the induced information to an ontology. An ontology is a computational model of some portion of the world. It is often captured in some form of a semantic network – a graph whose nodes are concepts or individual objects and whose arcs represent relationships or associations among the concepts (Huhns & Singh, 1997). The goal is to support deductive reasoning over the data, which receives much attention since the Semantic Web (Berners-Lee et al., 2001) was proposed. Sheth and Stephens (2007) introduced basic concepts for the Semantic Web and summarized real world applications of Semantic Web technologies. As a result of our investigations, we have established the core requirements for a suite of tools that could enable users, such as intelligence analysts and military strategists, to derive significantly improved utility from unstructured information in a wide range of domains.

[*] Corresponding author: Michael N. Huhns, Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA. Tel.: +1 803 777 5921; Fax: +1 803 777 3767; E-mail: huhns@sc.edu.

We have surveyed current technologies in the areas of tagging, tags, folksonomies, semantic analysis, integration of heterogeneous information systems, and ontology matching and reconciliation. Section 2 presents an overview of the findings from the technologies surveyed, and how this prior work supports our investigations.

In order to identify salient features common to many tagged data repositories and to manage the complexity of this typically socially-driven information domain, we have articulated the concepts for describing and analyzing *Folksonomy Space*: a framework that supports knowledge extraction from a folksonomy. This framework includes the strategy and methods to combine statistical induction (bottom-up) techniques with lexical and semantic mapping (top-down) techniques, to achieve robust extraction of semantics from unstructured information spaces. The dimensions of the Folksonomy Space help delineate salient features of tag sets, tags, taggers, and referenced objects. Section 3 presents this analytic framework.

We have identified and developed several analytic methods and software techniques needed to combine the top-down and bottom-up techniques to maximize the information that can be extracted from unstructured domains and make any implicit structure explicit. Section 4 provides an overview of these methods and how they can be applied to a variety of tagged data sources.

We applied these methods to existing folksonomies, including tagged data from amazon.com and flickr.com. We consider these tagged data sources as representative of free-form, socially-driven, unrestricted, heuristic tagging. Through several experiments we were able to demonstrate both induction of ontological information from these folksonomies as well as mapping these induced structures to existing ontologies. Section 5 presents these results, while Section 6 summarizes this research and development effort and recommends promising directions to pursue.

2. Related work in tagging and folksonomies

The concept of tagging emerged to answer a need for flexible access to objects and information, and the concept of a folksonomy emerged from the growth of collections of tags. We review the origins and evolution of these terms in the following sections.

2.1. Tags and tagging

A *tag* is defined as a non-hierarchical linguistic *keyword* or *term* assigned to a piece of information, such as an Internet bookmark, digital image, or computer file. This kind of *metadata* helps describe an item and allows it to be found again by browsing or searching. Missier et al. (2007) addressed general requirements of metadata management.

The history of tags on the World-Wide Web begins in 2003 when Joshua Schachter, the founder of the first widely used social bookmarking site http://delicious.com, pioneered the use of tags for a user's bookmarks. In 2004, *Flickr* included the same concept and allowed users to tag pictures and videos (numbering more than 6 billion in 2011), making them easier to be searched. With the success of *Flickr* and del.icio.us, collaborative tagging gained popularity and, consequently, many other websites, such as *YouTube*, *Picasa* and *Technocrati*, began implementing tagging.

There are at least two general kinds of tagging, typically with rather different characteristics. The original (conventional) tagging model embedded the tags within a document; such *intrinsic* tags were typically keywords added by persons knowledgeable about the content of the document. Tagging on the Web was originally in this form: tags were added to Web pages by website developers, and generally only visible when viewing the page's HTML source. This sort of *embedded keyword* tagging is still the most widely used approach, and the keywords are used to build search indices for the Web and other mechanisms for users to find information (and for businesses to market website content).

One problem encountered by Web users is that it is often difficult for them to add the metadata accurately, much less precisely and formally, because they are often not sufficiently knowledgeable about the content (or the domain in which the content is relevant) as to be able to assign meaningful or accurate tags.

As a partial solution, some websites in Web 2.0 environments have achieved success by enabling users to add metadata in the form of natural language tags. The result is that increasing amounts of on-line information are being categorized by associating with each piece of information tags created by the users themselves in ways that are comfortable and natural for them. In this *social tagging* model, users classify the content according to their liking or needs, with no restrictions regarding their choice of terms. The users are not given any guidance or restrictions about the form or structure of the tags they may use, keeping the tagging process intuitive, easy, and straightforward for them.

Some examples of websites where users commonly tag content are:

- http://delicious.com/ – for bookmarks,
- http://www.flickr.com/ – for photographs and videos,
- http://www.amazon.com/ – for a variety of products offered for sale,
- http://www.librarything.com/ – for books,
- http://www.gmail.com/ – for e-mails,
- http://www.odeo.com/ – for podcasts.

These sites generally place no limit on the number of tags an item can have. Associating a larger number of tags with an item facilitates the finding of more relevant information (i.e., greater recall). As is known from information retrieval, as recall increases, precision typically decreases, because retrieving additional relevant information also retrieves additional irrelevant information, and it can be difficult to separate the two.

Tags have been used to study the growth of social networks (Wu, 2011). Accelerating growth patterns appear in the virtual world. The phenomenon confirms that assigning user-chosen keywords to a piece of information to facilitate searches does not correlate in a linear way to the number of social media users using Internet tagging. Wu (2011) used the tagging behavior on the *Flickr* and del.icio.us social media sites to study the growing activity of online communities. Although the number of tags and the population fluctuates, communities have heterogeneity in individual tagging activity that remains constant over time, but differs across systems. The average individual activity will grow as the system expands and lead to the accelerating growth of overall activity. Such modeling of online activity growth could be used to predict the server capacity needs of social media sites on the basis of historical data. It is believed that use of the links on the Internet and related tags would greatly help the categorization system (Shirky, 2005). Uren et al. (2006) identified seven requirements of annotation/tagging including ontology support and automation, reviewed existing annotation systems, and concluded that challenges remained after all the requirements were integrated in these annotation systems. Automatic annotation systems have also been studied (e.g., Kiryakov et al., 2004).

Though very useful, tags have quite a few disadvantages. First, since tags are freely chosen, synonyms, homonyms, and polysemy are very likely to arise, thereby degrading the efficiency of searching. For example, a user could tag an item as `Sport` or `Sports`, and searches for items having one of these tags separately would yield different results.

Second, tags are used only as keywords, which do not convey information about their semantics. Consequently, when an item is tagged with a word that can have more than one meaning, the search results are bound to display some results that might be irrelevant to the user. For example, a user can tag an item as `Orange`, which can refer either to the color orange or the fruit orange.

Third, items in an application domain or at a website might be tagged by many different people, and often idiosyncratically. As a result, searches might have to be repeated a number of times with different search terms before appropriate information is returned.

Last, tags are not related by any explicit structure, so any tag-based search returns only the content containing exactly the same keyword. For example, consider three pictures where the first is tagged as `(Sports, Soccer)`, the second is tagged as `Sports`, and the third is tagged as `Soccer`. A search for the tag `Sports` fetches the first and second pictures as its result, but not the third picture, even though *Soccer* is a kind of *Sports*. Without knowing this semantic relationship, we are left with just the first two pictures (i.e., low recall). Therefore, many researchers try to measure the relationship between tags effectively and doing semantic analysis according to these measures. Cattuto et al. (2008a) analyzed three measures of tag relatedness: tag co-occurrence, cosine similarity (Salton, 1989) of co-occurrence distributions, and FolkRank (Hotho et al., 2006) using data from the social bookmarking system *del.icio.us* and based on a semantic grounding derived from WordNet (Fellbaum, 1998). They determined the measures that were most appropriate for a given semantic application. For example, cosine similarity is best for synonym discovery among the three measures. A systematic methodology was developed to characterize these measures by Cattuto et al. (2008b), which explored more measures of tag relatedness. Körner et al. (2010a, 2010b) classified users as Categorizers and Describers and presented findings about evaluations of semantically grounded tag relatedness measures.

Although tags can be assigned easily, because they are unstructured, there might be an implicit structure within a set of tags that emerges bottom-up as tags are added to a collection of items. The term *folksonomy* (Wal, 2007) refers to this implicit structure. In the following section, we trace the evolution of this term and provide a more precise definition that is used in the remainder of this paper.

2.2. Folksonomies

Aggregating the tags of many users creates a *folksonomy*. Folksonomies began gaining popularity in 2004 as a part of social software applications on the Web (Peters & Becker, 2009; Gruber, 2007). The term “folksonomy” was coined by Wal (2007) as a combination of the words “folks” and “taxonomy”, although it is only loosely related to a taxonomy. A *taxonomy* refers to a categorization of data in an explicitly hierarchical structure – a kind of informal and under-specified ontology – whereas a folksonomy categorizes content with tags, which do not have any explicit hierarchy defined and are all treated as being at the same level, i.e., they are theoretically “equal” to each other. Using these tags, a folksonomy is intended to make information retrieval extremely easy and fast. It can also be used, as demonstrated with the tags from http://delicious.com (Yeung et al., 2008), to customize searches. Halevy et al. (2009) provides many challenges and possibilities by learning on unlabeled data.

The Semantic Web offers great potential to facilitate human activities (Berners-Lee et al., 2001) and the task of discovering semantic relations between concepts (e.g., subsumption, disjointness, or named relations) is core to productive use of the Semantic Web. As such, *Scarlet* (Sabou et al., 2008) harvests the Semantic Web by automatically finding and exploring multiple and heterogeneous online knowledge sources. The result is discovered relations, which can be used for tasks such as ontology matching, ontology learning, word sense disambiguation, and ontology enrichment. The sources for this effort are not tag sets, but partial ontologies of a domain.

Because of their demonstrated utility, there have been several attempts to improve upon the semantics of the tags to increase their utility. In one of these, a domain was seeded with a taxonomy and user-generated tags were then added to it, so that the resultant set of tags would have a structure (Hayman, 2007). Jäschke et al. (2007) developed algorithms based on FolkRank and PageRank for tag recommendation and compared these two algorithms. Another research effort tried to disambiguate tags (Yeung et al., 2007) and another utilized it for better information retrieval (Zhou et al., 2008). In contrast, our approach is to induce a structure from an existing unstructured set of tags.

In a similar vein, Angeletou et al. (2007) and Angeletou (2008) developed *FLOR*, a tool that performs semantic enrichment of folksonomy tag collections by exploiting online ontologies, thesauri, and other knowledge sources. The result is improved semantics for tags, but not relationships among tags. Helic et al. (2011) proposed a pragmatic framework that used hierarchical structures learned by different folksonomy algorithms as background knowledge for decentralized search to evaluate the usefulness of folksonomies for navigating social tagging systems. It compared the results of four folksonomy algorithms on five different social tagging datasets and found that folksonomies produced by tag similarity graph algorithms performed better than hierarchical clustering algorithms for navigation.

Several researchers have developed frameworks and algorithms for learning ontologies, taxonomies, or folksonomies from tags. Mika (2005) presented a methodology to generate lightweight ontologies from a tripartite model that integrated a social dimension into a traditional bipartite model of ontologies. Schmitz (2006) used a subsumption-based model to induce an ontology. Heymann and Garcia-Molina (2006) developed an algorithm to acquire a hierarchical taxonomy from tagging systems. The approach aggregated tags into tag vectors, calculated the similarity between tags, and discovered the latent hierarchical taxonomy from the similarity graph. Cattuto et al. (2007) studied network characteristics by considering folksonomies as tri-partite hypergraphs. Solskinnsbakk and Gulla (2010) used tag vectors and built a semantic hierarchical structure based on folksonomies using association rule mining. Other folksonomy learning methodologies from social tagging systems include Benz et al. (2010) and Li et al. (2007).

Learning folksonomies from smaller structures has also been studied. Plangprasopchok and Lerman (2009) introduced statistical frameworks that use user-specified relations and aggregate individual hierarchies. Later they (Plangprasopchok et al., 2011) described an unsupervised probabilistic approach that integrated smaller, noisy structures into a few complex structures. Plangprasopchok et al. (2010) learned folksonomies from social metadata on Flickr by using relational clustering.

Ontology matching or reconciliation, which integrates heterogeneous databases and tagging systems, plays an important role in studying ontologies. Mahalingam and Huhns (1997) developed a GUI tool called *JOE* to create and edit ontologies, which was used later for semantic reconciliation (Mahalingam and Huhns, 2000). Doan et al. (2003) introduced an ontology matching system *GLUE* that uses machine learning strategies, while Dou et al. (2003) developed *OntoMerge* for ontology merging with web languages as its input. Ontologies have been reconciled on an enterprise level (Huang et al., 2006; Huhns & Stephens, 2002; Stephens & Huhns, 2001) also. Other ontology matching and resource integration works include (Castano et al., 2003; Huang & Huhns, 2006; Stephens et al., 2003; Giunchiglia et al., 2005; Doan & Halevy, 2005; Collet et al., 1991; Huang et al., 2005a, 2005b, 2005c). Stephens and Huhns (2001) worked towards reconciling semantics from different sources without a global ontology and determining the efficacy of doing so. Noy and Musen (2000) proposed a semi-automatic approach for ontology merging and alignment.

Schema matching, which differs from ontology matching by usually not providing explicit semantics for the data, has been studied for enterprise applications. Madhavan et al. (2001) proposed a new algorithm called *Cupid*, which integrated the use of linguistic, structural matching, and so on to do generic schema matching. Melnik et al. (2002) presented a schema matching algorithm taking graphs as inputs. Pastor et al. (1992) found that little research has been done on mapping data structures to a database without a semantic schema. They proposed a View-Concept model which used the knowledge representation language *LOOM* (Macgregor & Bates, 1987) to define the semantic schema of a database and implemented a prototype system.

The paper by Van Damme et al. (2007) provides a thorough analysis of tags and the tagging process. It identifies and analyzes all of the information that is relevant to the interpretation of a folksonomy from an ontological viewpoint. The information includes a statistical analysis of tags, lexical resources, existing ontologies, and existing ontology alignments and mappings. We also advocate such a comprehensive approach. The authors describe the approach and its facets, but have not yet applied it to a folksonomy, as we have done and report on in this paper.

In the work most relevant to our own, Eda et al. (2009) produced an organization of the tags in a domain by using probabilistic latent semantic indexing and a measure based on entropy. Each tag was represented by a vector whose elements are the number of times each tag was applied to an item by a group of taggers. The entropy of each tag determined its place in a directed acyclic graph, as well as whether the tag was subjective or objective. It is most appropriate for domains having a relatively small number of items that are tagged by a large number of taggers, which are different than the domains considered herein.

Henceforth in this paper, when we refer to a *folksonomy* we intend it to connote *the implicit hierarchical structure that emerges from a collection of tags*, and the intent of our research has been to make this implicit structure explicit and available to applications interested in the objects that are referenced and described by the tags. As we explain in detail in subsequent sections, we consider all of these to be in scope for our investigation:

- the *folksonomy* itself (i.e., the collection of all tags in a particular repository),
- the *tags* themselves,
- the referenced *objects* (whether tangible or intangible, concrete or abstract) to which they are applied, and
- the population of *taggers* from which a given tag collection arises.

Our hypothesis is that all of these can be used to help elucidate the relationships among tags, and hence to induce the structure implicit in the tags.

We emphasize the breadth of our definition of "object", because we intend it to apply very broadly to anything that can be accumulated and tagged. We also note that tags, as we define them, subsume any sort of metadata that has been applied to objects, whether or not it is structured or constrained in any way: objects tagged with terms from a controlled vocabulary are a specialized form of *ad hoc* tagging, so any results we obtain from analyzing the tags at amazon.com and flickr.com will be equally applicable to more organized domains, such as military situation representations (SITREPs) containing tactical information or scientific papers categorized using the ACM classification hierarchy.

3. Exposing latent information in folksonomies for reasoning

While the integration of structured data sources into reasoning systems has received a great deal of study and many successful systems have been implemented, integration of unstructured information has remained difficult. The principal reason for this is that most practical reasoning systems are *deductive*: reasoning proceeds according to well-defined rules of inference that assume a declarative model representing knowledge in a particular logical form that encodes the semantics of the domain. Deductive logic cannot easily be applied to information that is not encoded in this form.

There are other forms of reasoning, however, that lack the formal rigor of deductive reasoning, but are nonetheless valuable in that they can discover structure where none obviously exists. In particular, *inductive* reasoning looks for patterns in data, and uses statistical and other techniques to discover similarities that are often strongly analogous to the kind encoded in ontologies as classes.

We believe that success in integration of unstructured data and structured data into a unified reasoning system will require both techniques.

3.1. Problem analysis

Whereas structured sources typically have semantics encoded in their structures, specifying the semantics of unstructured data has required humans to perform a semantic analysis and encode the meaning of the data in an ontology. This is labor-intensive and impractical given the massive volumes of unstructured data available.

We presume a collection of objects (entities), as depicted in Fig. 1. The objects might be physical (e.g., products on amazon.com) or informational (e.g., intelligence data). A formal ontology (e.g., Cyc), which at least partially covers the objects of interest, is also assumed to be available. The objects are described informally by a collection of tags produced by a population of individuals (*taggers*). The tagging of objects (see Fig. 2) usually takes place without the taggers being constrained by any guidelines. The taggers are in many cases unorganized and unknown to each other, so the process of tagging is *ad hoc* and the resultant collection of tags typically does not have any formal structure, although it often has an *implicit* structure Lalwani and Huhns (2009), as described in Section 2.2.

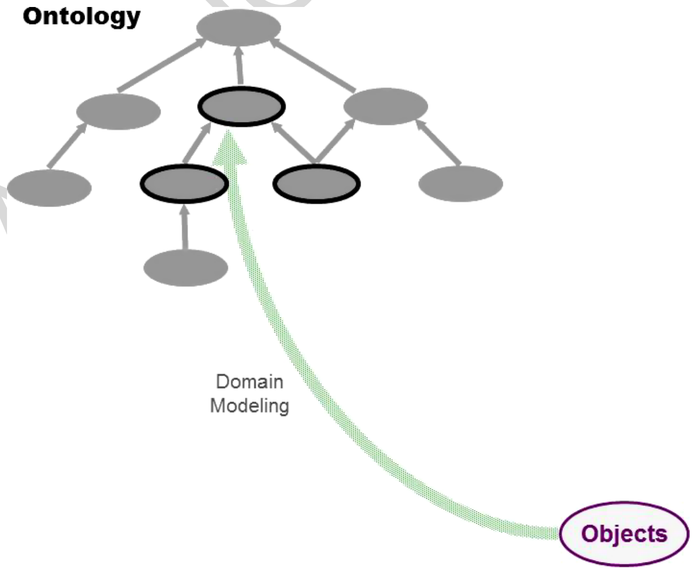


Fig. 1. The ontology is presumed to already exist and to have at least partial coverage of the domain based on models of the objects in the domain. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

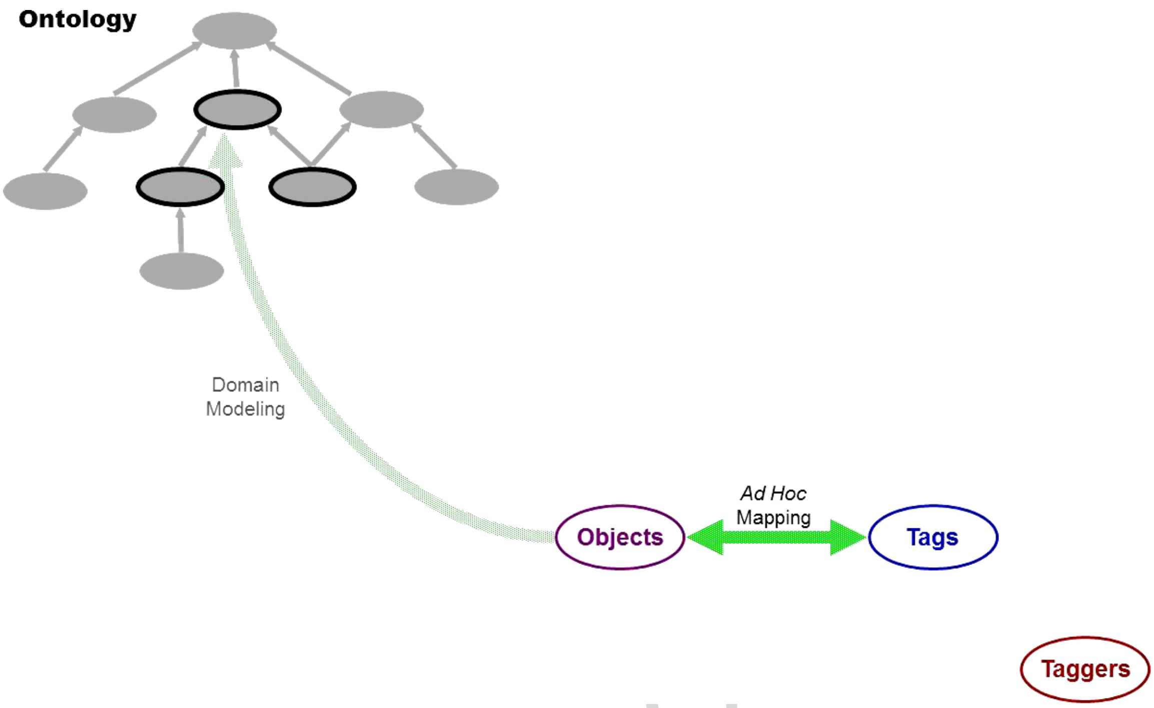


Fig. 2. A population of Taggers has now applied Tags to the Objects, introducing an *ad hoc* mapping from the objects to the tags. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

The purpose of our effort is to develop tools that can explicate any implicit structure from a collection of tags and induce from it an informal ontology. This can be further mapped to a formal ontology, and thus can support deductive reasoning. This is depicted in Fig. 3.

We believe that the process of inducing the implicit structure of the tags and making it explicit can be improved if information about the objects, the domain, the population of taggers, etc., is considered. For example, if the taggers are electrical engineers, then the tag "conductor" is more likely to be about a wire and not the leader of an orchestra.

Although there have been other attempts to reveal the structure explicit in a collection of tags, none has fully related the tags to ontological and lexical knowledge that is available, and none has made use of information about the characteristics of the domain, the folksonomy itself, the tags, the objects, and the population. The basis of our research is our confidence that applying multiple techniques synergistically will permit explication of far more of this implicit knowledge, and permit far more accurate and comprehensive mappings to formal structures – ontologies – so as to permit previously inaccessible data to be used in automated reasoning systems.

Our approach consists of three general categories of techniques for exposing and formalizing the latent semantics in existing (large) data sets of tags (illustrated in Figs 1–3):

(1) Using inductive and machine learning techniques to derive a set of probabilistic clusters from the tags. This is shown in Fig. 3 as the curved arrow on the right leading from the tags to the induced structure.

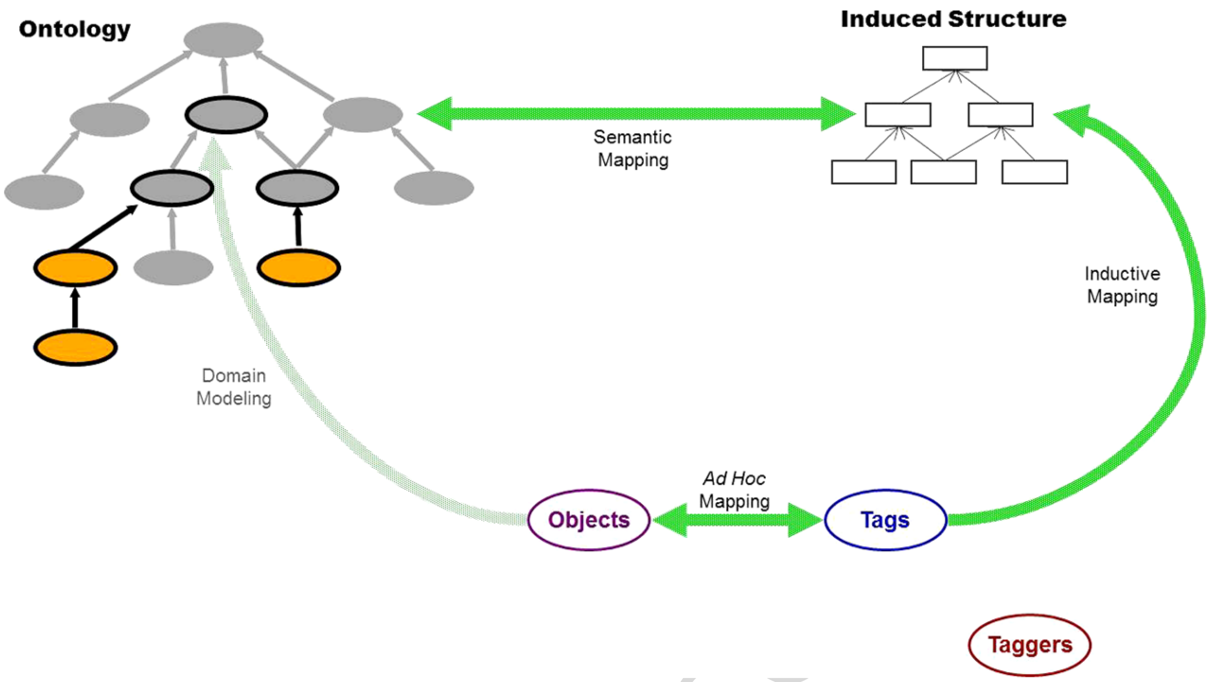


Fig. 3. Using inductive techniques, such as clustering, a structure can be induced over the tags, introducing induced mappings from the tags to a new structure that is similar to an ontology. Using techniques from database matching and ontology alignment, preliminary semantic mappings can be introduced. The nodes shaded in gold correspond to new classes derived from the induced structure. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

(2) Using intermediate structures, such as lexicons, to map – largely via conventional deductive techniques – from both the tags themselves and the informal structures identified by induction, to a formal ontology. This is illustrated in Fig. 4.

(3) Using a conceptual structure that we call Folksonomy Space (cf. Section 3.2) to inform both the inductive and deductive processes by tailoring them to the context. This is shown in Fig. 5.

In the literature, the first two of these have been called, respectively, bottom-up and top-down approaches. We do not believe that either of these general approaches can succeed on its own: bottom-up techniques lack the ability to produce formal structures that can be used in reasoning, and top-down techniques cannot derive novel information except by deduction from existing information.

Inductive techniques generally attempt to discover patterns and structures in data by statistical clustering mechanisms. The clusters might be composed from individual terms (to discover classes), pairs of terms (to discover binary relations), or the co-occurrences of sets of terms (to discover nth-order relations).

Lexicon-based approaches to the general problem of mapping text to ontologies have existed for years, but we believe that they have been under-exploited, because they have typically operated on terms that are either unstructured, or structured as parse trees and similar linguistic artifacts. Our integration of this approach is depicted in Fig. 5.

Extant mappings from intermediate structures, such as lexicons (e.g., WordNet) or thesauri, are used to map from individual tags and the clusters identified by inductive techniques. These intermediate structures are used not only to identify induced categories with classes in the ontology, but also to

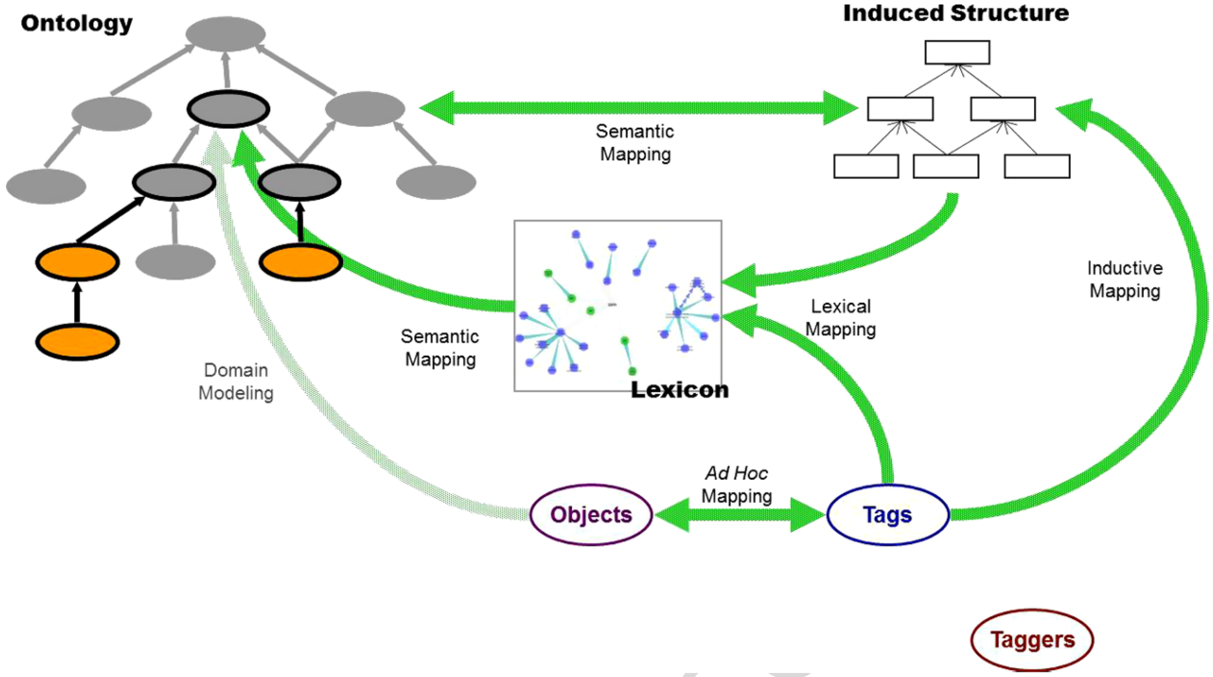


Fig. 4. A lexicon (e.g., WordNet) or other lexical structure is introduced, and both the raw tags and the terms in the induced structure are mapped to the ontology via lexical and semantic mappings. These mappings may augment, confirm, and otherwise improve the semantics of the mappings and any new classes in the ontology that have been derived from the inductive processes. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

provide empirical support for existing class structures in the ontology. In other words, in addition to compensating for weaknesses in the individual top-down and bottom-up approaches, our approach uses the strengths of one technique to augment the other.

We believe that the third element of our approach – Folksonomy Space – is unique to our effort and can be compared to a "tag cloud", that is, a clustering of tags with no inherent structure or formal properties Knautz et al. (2010).

Three of its axes are illustrated in Fig. 5:

- Tagger population (in this case, the level of heterogeneity of the population).
- Folksonomy structure (in this case, the degree of organization among the tags).
- Object type (in this case, the level of abstraction of the objects).

For the examples in this paper,

- The tagger population is situated well toward the "heterogeneous" end of the Tagger axis.
- The tags are situated well toward the "tags" (unstructured) end of the Structure axis.
- The objects are on the concrete end of the Object Type axis.

In the section that follows, we elaborate on the concept of Folksonomy Space.

3.2. *Folksonomy Space*

The concept of an N-dimensional organizing framework, which we call Folksonomy Space, guides our selection of data sets and is useful in understanding their differences. As described below, we have

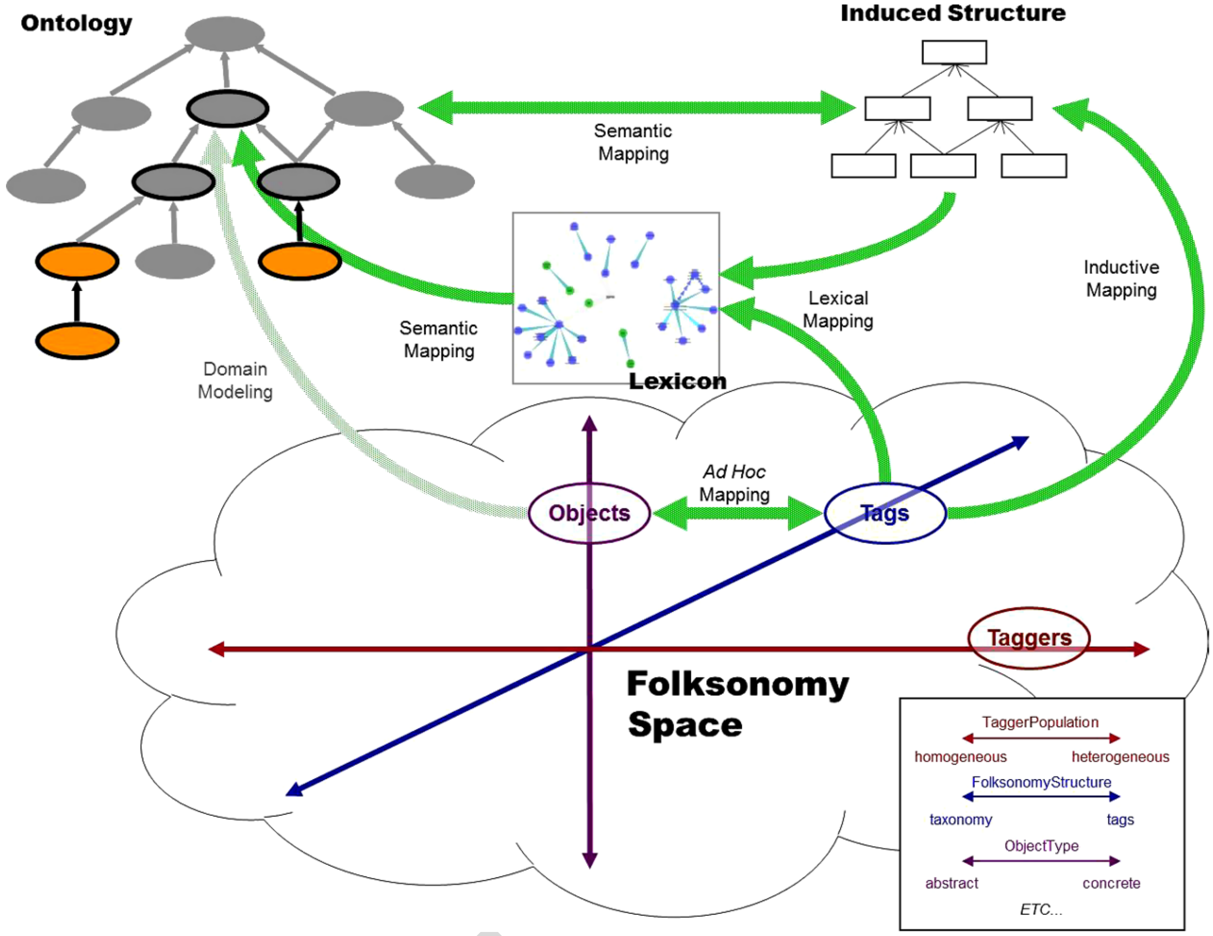


Fig. 5. An N-dimensional space – Folksonomy Space – has been introduced to the system; the Objects, Tags, Taggers, and a number of other features of a particular context are all situated at specific locations on the axes of Folksonomy Space. These locations yield an N-tuple that can be interpreted as a coordinate in Folksonomy Space; we intend to use this coordinate to guide the selection of tools, the strategy adopted, and the techniques applied in our processing. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

identified 18 dimensions of this space, but have not yet established those that are most salient in predicting which techniques and which structural artifacts (e.g., lexicons and ontologies) will be most effective in processing a data set that maps to a particular locus in Folksonomy Space.

We view this 18-dimensional space as comprising four clusters of axes, where the clusters are neither correlated nor orthogonal, but are grouped only according to the entity (Folksonomy, Tag, Tagger, Object) to which they apply.

3.2.1. Folksonomy dimensions

One axis in Folksonomy Space corresponds to characteristics of the folksonomy itself and is depicted in Fig. 6.

(1) *Structure* refers to whether or not there is an explicit structure to the tag data. For example, the *Amazon* tag data has no structure of any kind, while ACM and IEEE tags within computer science

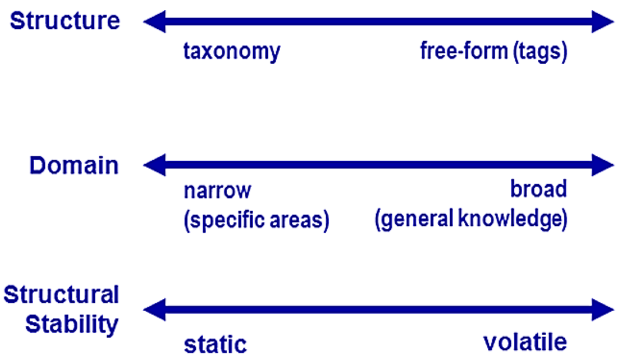


Fig. 6. The dimensions of Folksonomy Space relating to the space as-a-whole. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

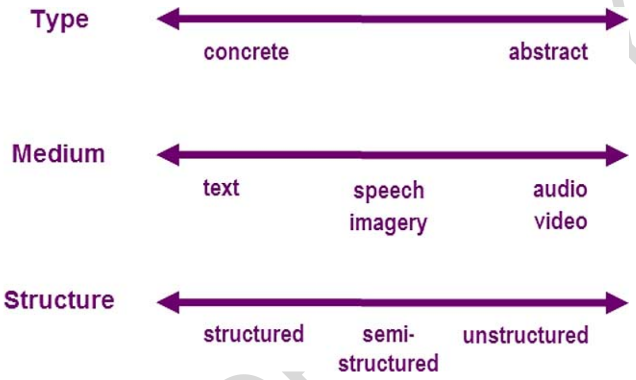


Fig. 7. The dimensions of Folksonomy Space relating to the objects being tagged. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

papers use a vocabulary that is predetermined. We believe that *ad hoc* tag collections will present much greater challenges to our approach.

(2) *Domain* refers to the breadth of coverage of the tags.

(3) *Structural stability* refers to the degree to which whatever structures exist in the data persist over time; for example, in an unconstrained tag set like that at amazon.com, the structure of the tag collection changes constantly, while in a more constrained collection it might persist.

3.2.2. *Folksonomy object dimensions*

Our description of objects is not based on an analysis of the objects themselves: techniques for processing and analyzing objects are moot for folksonomic structures, except for establishing ground truth, and are outside the scope of our inquiry. However, our approach can incorporate components such as a Google-like indexing engine that can preprocess the untagged objects. Our system operates on (meta)data associated with objects, not the objects themselves.

For the objects themselves, as depicted in Fig. 7, the dimensions are

(1) *Type* relates to the entities that are the objects of the categorization: are they concrete objects or abstract entities?

(2) *Medium* is primarily relevant for abstract objects; they may have physical manifestations, but the medium is not the message, and it is the content rather than the container that is typically of interest.

(3) *Structure* is primarily, although not exclusively, related to textual objects, where the meanings of the various points on the scale are fairly well understood.

3.2.3. Tag dimensions

The dimensions in Folksonomy Space related to the tags (data), as shown in Fig. 8, are:

(1) *Range* refers to limits on the number of tags in the space; a finite (and relatively limited) set of tags permits full exploration and analysis of the space. For Amazon.com, the tag space is limited only by the imagination of the tagging population.

(2) *Accessibility* refers to the level of protection placed on the data by its owner. This is the only dimension on which *Amazon* data is not worst case, but inaccessible data (like intelligence data) cannot be analyzed.

(3) *Explicitness* describes how difficult the data is to extract from its milieu; *Amazon* data is, despite being freely available, not structured to facilitate analysis, but rather to support browser functionality.

(4) *Location* describes where the tag data exists – *intrinsic*, or within the tagged object (e.g., keywords in a technical paper), or *extrinsic*, i.e., external to the document. Both aspects are widely used on the Internet. The folksonomy data examined in this study were all comprised of extrinsic tags.

(5) *Tag structure* refers to the actual lexical, typographic, orthographic structure – spelling, punctuation, capitalization, character set, presence or absence of whitespace, etc. The term "tag" suggests a single word, composed of contiguous non-blank characters; this is not the case for *Amazon* data and, we expect, most data sets.

(6) *Vocabulary derivation* refers to the origin of the terms in the vocabulary: Are they imposed externally or evolved internally? Again, the *Amazon* vocabulary is entirely evolved, as would likely not be the case for intelligence analysts.

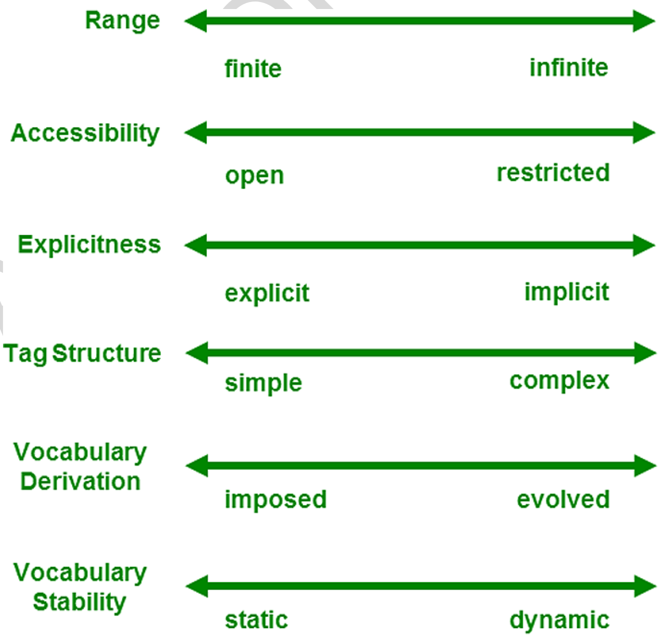


Fig. 8. The dimensions of Folksonomy Space relating to the tags. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

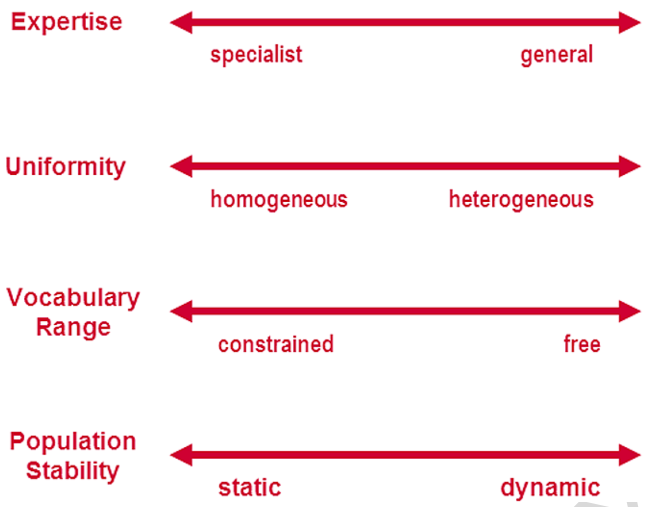


Fig. 9. The dimensions of Folksonomy Space relating to the population of taggers. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

(7) *Vocabulary stability* refers to change in the composition of the tag set over time, which may or may not be directly related to the tagger population; again, *Amazon* vocabulary changes constantly and rapidly as new products and technologies enter the market (e.g., `1080p` would not have been in the *Amazon* tag space until the advent of high-definition televisions).

3.2.4. Tagger dimensions

The dimensions in Folksonomy Space related to the tagging population (taggers), as shown in Fig. 9, are:

(1) *Expertise* is correlated with the scope of the terminology that the tagger population can be expected to use.
(2) *Uniformity* refers to whether the tagger population is similar or dissimilar in terms of relevant background, experience, and knowledge.
(3) *Relation to objects* refers to whether the author of the tags is also the author/creator of the tagged object, or the tagger is an outside agent.
(4) *Vocabulary range* refers to whether or not there are explicit constraints on the vocabulary.
(5) *Population stability* refers to change in the composition of the population over time. For *Amazon* taggers, we presume this to be highly dynamic, as would not be the case with, e.g., intelligence analysts.

4. Methods for extracting knowledge from folksonomies

Our process for extracting knowledge from folksonomies has the following five phases:

Domain identification: the choice of a domain determines virtually every subsequent operation.
Extraction: gaining access to the raw data, and extracting it from the original source.
Preprocessing: filtering and other preprocessing to increase the signal-to-noise ratio and generally improve the quality of the raw data.

Induction: applying inductive and statistical methods to the filtered data set to construct candidate ontology fragments.

Mapping: using lexical, statistical, and other mapping techniques to integrate and align induced ontology fragments with formal ontologies.

We discuss each of these steps in the following sections.

4.1. Identification of the domain

We considered and evaluated many alternative data sets for our evaluation effort, and this process helped to shape our views of a typical space of tag data. Among the early candidates were military and intelligence data sets, but both of these presented problems with accessibility (one of the dimensions of Folksonomy Space related to tags).

Another class of candidates – scientific papers (e.g., in computer science) – were accessible, but we felt that the range of tags within the papers was relatively narrow (the Range dimension for tags), imposed (the Vocabulary Derivation dimension for tags), and intrinsic (the Location dimension for tags). In addition, along the tagger dimensions, all of the points on the Expertise, Uniformity, Relation to Objects, and Vocabulary Range dimensions would be at or near the midpoints of the ranges.

In order for our results to have maximum generality and coverage, we felt that a data set at or near the extreme ends of most of the dimensions of Folksonomy Space would be more suitable, if more challenging. Hence, we chose to adopt the *Flickr* and *Amazon* tag data sets – the former for preliminary experiments using hand-culled data, and the latter for more sophisticated experiments using larger volumes of automatically-culled data. We expect that any results derived from the *Amazon* data set will be applicable to any less *ad hoc* data set.

4.2. Extraction

Locating and extracting the raw data for processing depends to a large degree on the characteristics of the folksonomy, e.g.,

(1) Is it accessible to external processes?
(2) Is it in the form of a taxonomy or free-form tags?
(3) Is it static or volatile?

Each of these can present significant challenges to processing. In the case of the *Amazon* data, the tags are embedded in undocumented, proprietary formats that are intended only for the purpose of supporting the amazon.com website: extraction required analysis of the HTML source code, identifications of "landmarks" in the HTML that indicate the existence of different components of the raw data, and the use of fairly complex pattern-matching and assembly to produce usable data.

In the case of more structured tags, and particularly those intended for external use, the extraction process would not be quite as tedious. Embedded (internal) tag data – such as internal tags in a relatively standardized location and format – will typically be relatively straightforward to extract and interpret, although it may require identification of the type of resource and parsing of the content.

4.3. Preprocessing

The amount of preprocessing required to make the raw data usable is to some extent a function of the nature of the tags and the tagger population. For example, tags from an unconstrained vocabulary, applied by a heterogeneous population of taggers, will tend to show far more variation and less objective accuracy (in terms of fidelity to ground truth) than those from a constrained vocabulary applied by specialists in the domain from which the tagged entities are drawn.

With *ad hoc* tag and tagger populations, accepting every tag proposed for a given entity can lead to lowered precision, due to high proportions of inappropriate or idiosyncratic tags. In such situations, it is necessary to apply some sort of filtering to eliminate as much noise as possible, preferably without eliminating valid data. We have investigated a few statistical techniques to perform this filtering, but we regard this as a fruitful area for further research.

Other sorts of preprocessing may be necessary for both the induction phase and the mapping phase. For example, even among single-term tags, it may be useful to apply stemming and other conventional text-processing techniques; in addition, compound tags may require more sophisticated techniques to disambiguate among different interpretations of the tag (`baby oil` is oil for use on babies; `linseed oil` is oil made from linen seeds). These too are areas in which we recommend additional research.

4.4. Induction

In order to expose the structure implicit in a set of tags, so that we can take advantage of the structure to improve user queries, we employ the following procedure:

(1) We begin with graphs of tags, taggers, and objects being tagged, with our primary data source being graphs of co-occurrences of tags.
(2) Based on what we know of the tag domain (e.g., *Amazon* products, intelligence community evidence, or *Flickr* photos) and the taggers (e.g., intelligence analysts or *Amazon* customers) supplying the tags, we form hypotheses about the structure. Each hypothesis concerns an ontological component of the tag space.
(3) We test the hypotheses using statistical techniques, primarily the cardinalities of various features of the tag set, and ensure that they are statistically significant.
(4) We evaluate the results in terms of the standard metrics for information retrieval: precision and recall.
(5) For validated hypotheses, we apply the conditions of the hypotheses to all of the data and generate ontology fragments. These resulting fragments are then passed to the mapping stage.

4.5. Mapping

The final stage in processing is mapping of induced ontology fragments to ontologies, either directly or via lexicons and other intermediate structures. Direct mapping between ontology fragments and ontologies has been successful, but mapping via lexicons provides additional flexibility, since terms from the ontology fragment can be matched against any member of a synset, and thence from the synset to the ontology via mappings provided by the authors of at least three of the major upper ontologies (Cyc, SUMO/MILO and DOLCE). This means that it is not necessary to match a single term in the ontology fragment to a term in the ontology: if the former is a synonym of the latter, it will map to the ontology as long as the synset to which it belongs is mapped.

There are many techniques that can be used to implement this mapping. During our initial experiments, we have used syntactic and lexical matching – without any conventional text-extraction preprocessing, as described in Section 3.2 – but a better approach would be to use techniques that have been developed for ontology alignment (cf. http://www.ontologymatching.org/publications.html for a representative bibliography), as well as additional information derived from lexicons and ontologies.

Examples of the additional information available to us from lexicons and ontologies include internal relationships among synsets within WordNet (e.g., hypernym and hyponym) and tagging of the SUMO/MILO-WordNet mappings that identifies the relationship (e.g., subclassof, instanceof) between the synset and the corresponding SUMO/MILO term.

One limitation of the system we have developed is that it does not take into account the fact that the mappings we identify are intrinsically probabilistic. As we develop improved mapping techniques, particularly as we begin to apply statistical techniques to our matching processes, we expect to move to a probabilistic model of the mappings that will allow reasoning to proceed without making unwarranted assumptions about the accuracy of the mappings.

4.6. Using tags and folksonomies to enhance search

The tags that are linked to Web content can be used for keyword-based searches. The tags ideally are used to filter out the enormous amount of data present on the Web and display only the data of interest. When a user runs a tag-based search on the Web, only the information that has been tagged by other users using the same tags is displayed in the search results. The user must then select from among these results the most appropriate, desired content.

Superficially, tagging of data and subsequent searching for that and other data by specifying tags of interest seems to be simple and effective. However, tag-based search filters out only some of the irrelevant data, because of the idiosyncratic and uncontrolled nature of tagging. More importantly, it also filters out a large amount of relevant information that is marked with similar, but not identical, tags as used in the search. This is because in tag-based search – absent an organizing ontology – there is no information about the semantics of a tag, and therefore no practical way to find related tags. For example, a search for “utensils” on a cooking website would return only the items that have been tagged with that exact keyword, ignoring all the other items that may be related to it. The user performing this search would almost certainly be interested in seeing all utensils, whether they have been tagged as `utensils`, or as `spoons`, `forks` or `knives`.

The simple process illustrated above – expanding the search to include terms that are related, but not identical, to the provided keyword – is actually a simple form of deductive reasoning, and could be supported by an ontology, if the implicit structure among the tags could be made explicit and related to the ontology. Furthermore, when the implicit structure of the tags in the folksonomy is made explicit, the ontology that emerges can be integrated with an existing ontology, and even used to extend, elaborate, and validate it.

By mapping tags into an ontology – and thus, enabling semantic search – enhanced search results and other advantages accrue over strictly keyword-based search.

- Semantic search can retrieve more of the available relevant information, resulting in better scores for the information retrieval metric ‘recall’ (the total number of items retrieved divided by the number of relevant items in the source data set).

- Semantic search can correctly ignore more of the available irrelevant information, resulting in better scores for the information retrieval metric ‘precision’ (the number of relevant items retrieved divided by the number of relevant items in the source data set).

This brings a range of previously inaccessible data to play in systems using tag data.

5. Deriving ontological structure from a folksonomy

The structure of the Web has changed substantially via transitions from *Akamai* to *BitTorrent*, *Britannica Online* to *Wikipedia*, personal websites to personal blogs, publishing to participation, content management systems to Wikis, stickiness to syndication feeds, and directories to folksonomies are some of the indicators of the changes underway. The resultant Web now consists of Internet communities, social networking sites, data sharing sites, wikis, blogs, and tagging systems. Consequently, huge amounts of information of different types and organizing principles are now accessible to users.

One aspect of the Semantic Web vision is that Web pages will have metadata that helps to specify the semantics of the contents of the pages. The metadata will be machine-understandable and machine-processable, which are needed for computers to be able to assist humans with use and management of the massive amounts of data.

5.1. Induction of ontological information from folksonomies

To make tagging systems more efficient, a methodology needs to be devised so that all the data related to keywords of interest gets displayed, rather than just the data containing those keywords. The objective of our research is to make the structure and semantics of folksonomies explicit, and to integrate it with ontologies so that reasoning can be performed over the entire body of knowledge. Our approach is to formulate plausible hypotheses based on our previous work in the construction of ontologies Huang et al. (2007); Huhns and Stephens (1999); Singh and Huhns (2005); Stephens et al. (2004), and then evaluate the hypotheses using data from existing on-line systems of tagged items. Hence, in this paper, we outline several hypotheses to derive additional utility from the tags that people are associating with items that are on the Web or, more precisely, we are investigating how to derive ontological structure from a folksonomy.

Specifically, we have formulated the following four initial hypotheses that we believe might describe the implicit structure in a folksonomy.

Hypothesis 1. For a group of items, if the number of occurrences of `Tag1` is less than the number of occurrences of `Tag2`; and if there are items where `Tag1` co-occurs with `Tag2`, then `Tag2` is a subclass of `Tag1`. The general heuristic rule we hypothesize is that the more a tag is used, the higher the level (closer to a root) it will be in an ontology, because it will cover more instances.

Hypothesis 2. Two tags can be claimed to be related if and only if the ratio of their co-occurrences to the subclass is greater than some threshold value, with confidence based on cardinality.

Hypothesis 3. For a group of items, if `Tag1` co-occurs with `Tag2` and `Tag1` also co-occurs with `Tag3`, but `Tag2` does not co-occur with `Tag3`, then `Tag2` and `Tag3` are subclasses of `Tag1`, and `Tag2` is disjoint with `Tag3` with a confidence based on the respective cardinalities.

Hypothesis 4. For a group of items, if Tag1 co-occurs with Tag2 more than a large fraction of the average cardinalities of Tag1 and Tag2 (heuristically meaning that Tag1 and Tag2 almost always occur together), then Tag1 and Tag2 are synonyms.

To evaluate the veracity of these hypotheses, we have conducted a number of searches on *Flickr*. Our first set of searches concentrated on evaluating the first hypothesis. Each search contained three parameters: Tag1 – displaying the number of items tagged with Tag1, Tag2 – displaying the number of items tagged with Tag2, and (Tag1, Tag2) – displaying the number of items tagged with both Tag1 and Tag2. The evaluation was based on the number of occurrences.

For the second hypothesis, each search contained two parameters (Tag1, Tag2) and Tag2, and then a mathematical analysis was done to calculate the threshold ratio along with allowable variance.

For the third hypothesis, the numbers of co-occurrences were compared to the threshold value and then analyzed to identify whether some relationship exists or the tags are disjoint. This hypothesis inherently considers Hypothesis 2 to be true.

The fourth hypothesis relies on tags frequently occurring together. The analysis is done two tags at-a-time.

5.2. Statistical induction from flickr.com tags

To understand the ontological knowledge among tags used by people in different domains, we have performed a mathematical analysis based on the results obtained from the experimental searches done on http://www.flickr.com (only on pictures that were available publicly). Specifically, this analysis provides the additional knowledge implicit in the tags that can be used to derive a hierarchical structure (along with a few non-statistically significant exceptions). The obvious cause of these exceptions is the use of tags freely chosen by users.

The Folksonomy Space used for the preliminary bottom-up experiments consisted of a set of tags $T = t_i$, with the nth-order relation R (read as "occurs together"), where

$$R : T_1 \times T_2 \times \cdots \times Tn \rightarrow N.$$

For example, the tags Sports and Basketball occurred together more than $N = 100$ times. N is the set of positive integers (in this case the cardinality of the relation tuple).

For the experiments, we only considered pairs of terms (such as Sports and Basketball), so the relation was

$$R : T_1 \times T_2 \rightarrow N.$$

By sampling R, we were able to induce an ontology.

The dimensions of the Folksonomy Space for our experiments, using the dimensions specified above in Section 3.2, are:

- *Object type*: abstract (photographs in *Flickr*),
- *Object structure*: unstructured,
- *Medium*: image (photographs in *Flickr*),
- *Population*: general,
- *Vocabulary*: free,

- *Folksonomy structure*: free-form (tags),
- *Domain*: broad.

Other features not captured in the space: the identity of the taggers was not known and the taggers generally did not know each other, so the tagging was done independently.

5.2.1. Hypothesis 1

For a group of items, if `Tag1` *occurs less frequently than* `Tag2` *and if we have a reasonable count for items where* `Tag1` *co-occurs with* `Tag2`*; then the class* Tag2 *can be termed a subclass of the class* Tag1.

Note, we consider here the co-occurrences of tags only if their number is at least 5% of the number of less occurring tags, to cover the margin for errors that can occur due to the use of an uncontrolled vocabulary. This assumption is based purely on observation.

This hypothesis was an intuitive guess taken initially while investigating how users have tagged content on Flickr. The first thing we observed was that for any given hierarchy of classes, users tend to use the leaf nodes as tags more often than they use superclass terms for tags. For example, the tag for the class *Spoon* is used more frequently than the tag for its superclass, *Utensil*. The results returned from the experimental searches were evaluated by counting the occurrences of `Tag1` and `Tag2` individually and then counting the occurrences of both tags together. The cardinality ratio of `Tag2`/`Tag1` for all the individual observations was calculated and then its arithmetic mean was taken. The average of the cardinality ratio was calculated to be 3.86, which might seem to conclude that most of these experimental searches abide by this hypothesis. But the reality is just the opposite. The average came out to be on the higher end, because a few search results gave a high ratio of up to 32. A large number of these searches defied the hypothesis, which is clearly shown in the histogram in Fig. 10.

According to the hypothesis, the frequency for *subclass*/*superclass* ratio $>$ 1 should have been more than the frequency for *subclass*/*superclass* ratio $<$ 1, but the histogram reveals the opposite. The main cause of this inconsistency is the use of an uncontrolled vocabulary by the users, which inhibits determining the actual count of the tag and consequently the failure of this hypothesis. For instance, consider a picture of a "human being". Different users can tag this picture with `Human, Humans, Homo sapiens, Man, Woman`, etc., which, although they mean the same thing, are different when counting tags.

Another reason observed here is that users tag content differently for different domains. For some categories, users tend to use the superclass rather than the leaf nodes, but vice versa for other categories.

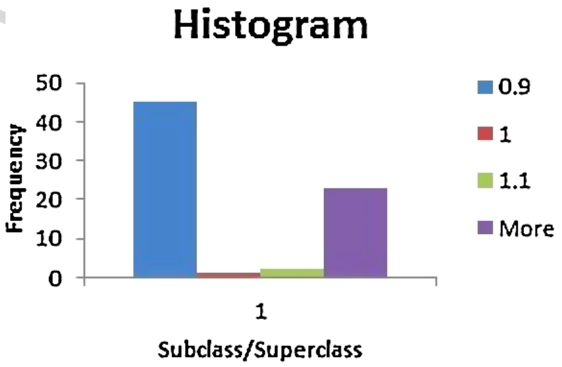


Fig. 10. A histogram of the experimentally found rations for (subclass cardinality)/(superclass cardinality). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

For example, the cardinality ratio for pictures tagged as `Fork` vs. the pictures tagged as `Utensil` is 32, causing the hypothesis to appear true, whereas the cardinality ratio for pictures tagged as `Eagle` vs. the pictures tagged as `Bird` is 0.1, causing the hypothesis to be considered false.

There are many similar examples and hence no certain conclusions can be drawn about the subclass–superclass relationships of tags based on this hypothesis. Hence, it is not verified.

5.2.2. *Hypothesis* 2

Two tags can be declared to be co-related if and only if the ratio of their co-occurrences to the subclass is greater than some threshold value, with confidence based on cardinality.

In the first hypothesis, we made a vague assumption about tags being co-related. So, in this hypothesis, we claim that for two tags to be co-related, the ratio of the cardinality of their co-occurrence to the cardinality of the individual tags must be greater than a plausible value, determined experimentally. In order to calculate this value, three consecutive searches were done – first having both the tags and then a search for obtaining each individual tag. The result of the first search is then divided by the result of both the other searches separately to calculate two ratios for cardinality. Here, two ratios of cardinality have been calculated separately because, currently, we do not know whether any relationship between the tags exists and, therefore, we need to consider both cases. Similar calculations have been made for approximately two-hundred search results on *Flickr*. The mean of the ratios calculated, thus, gives us a good estimate of the final cardinality ratio to decide whether a *subclass*/*superclass* relationship exists or not; or more precisely, it gives us the threshold value.

For example, the cardinality of items with tag `(Animal, Dog)` is divided both by the cardinality of items with tag `Animal` and the tag `Dog`. The cardinality ratios are calculated to be 0.12 and 0.06. So, the final cardinality is the mean of these two, which is 0.09.

The histogram in Fig. 11 shows the results of our experiment. The tags for this histogram are deliberately chosen to have an inherent subclass–superclass relationship, which will then help us in determining the threshold value.

The mean of the cardinality ratio of the tags used for the above histogram is 0.09 and the variance is 0.005. But as can be seen from the histogram, the frequency is highest between cardinality ratios of

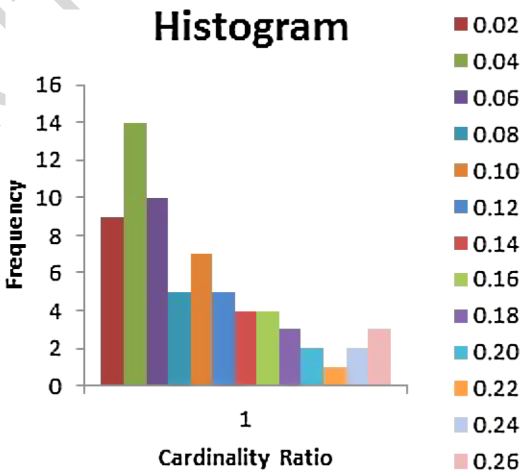


Fig. 11. A histogram of the cardinality ratios used to determine a threshold for deciding whether two tags are related or not. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

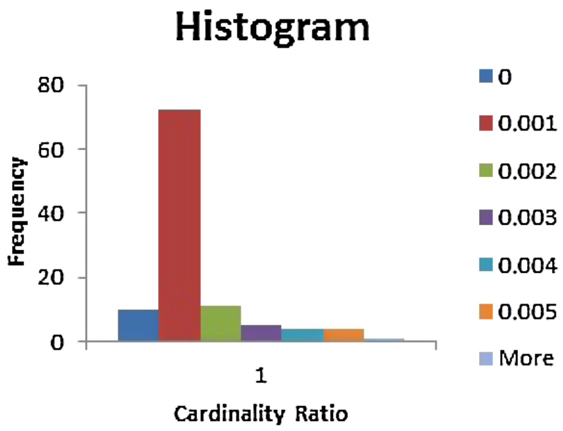


Fig. 12. A histogram of the cardinality ratio among tags that are unrelated to each other, to provide a contrast with the histogram in Fig. 11. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

0.04 and 0.08, so it can be concluded that the threshold value for two tags to be in a subclass–superclass relationship is 0.06, based on our statistical analysis.

To verify the above finding, we constructed another histogram (shown in Fig. 12) for tags that are not related to each other in any way. The mean of the cardinality ratios for these tags was calculated to be 0.0008 and the variance was 1.5E−06. Hence, our experiment clearly shows that the threshold value of the cardinality ratio works well indeed and Hypothesis 2 holds.

There are certain cases where this hypothesis does not hold. For example, consider the tags Grass, Green. Obviously, Green is just the color of the grass and does not have any superclass–subclass relationship with Grass, but if this hypothesis were true, then it would result in *Green* being declared to be a superclass of *Grass*. (However, if *Green* is interpreted as the set of all things that have color green, then the superclass relationship is reasonable.) This again shows that inconsistency in tagging content from the user's end can introduce inconsistencies in the behavior of any such ontological structure-finding technique, and so errors are expected to occur. For the number of searches made, this hypothesis gave a success rate of more than 85%. Hence, this hypothesis is considered to be verified.

5.2.3. Hypothesis 3

For a group of items, if Tag1 co-occurs with Tag2 and Tag1 also co-occurs with Tag3, but Tag2 does not co-occur with Tag3, then Tag2 *and* Tag3 *are subclasses of* Tag1*, and* Tag2 *is disjoint with* Tag3 *with confidence based on the measured cardinalities.*

This hypothesis is a direct consequence of Hypothesis 2, where we were able to derive a conclusion that some relation exists between two tags, but could not define the exact relation. So, in this Hypothesis, we add one more tag to the observations and record the results as explained: the co-occurrence of (Tag1, Tag2) is recorded, then of (Tag1, Tag3) and finally of (Tag2, Tag3). All these co-occurrences are then analyzed separately and cardinality ratios for each of them are calculated. The cardinality ratios for the first two searches are greater than the threshold and, hence, from Hypothesis 2 it can be concluded that there exists a relationship between them. No such relationship exists between Tag2 and Tag3, as their cardinality ratio is far below the threshold. So, it can again be concluded that *Tag2* and *Tag3* are disjoint classes. Now, since Tag1 co-occurs both with Tag2 and Tag3 with a high cardinality ratio, it becomes obvious that it is one level higher than the other two tags and hence becomes the superclass for the other two tags. The combined results, shown in the histograms of Figs 11 and 12, describe the statistical analysis of Hypothesis 3. The following example provides intuitive justification for the hypothesis.

Example. For pictures having tags `(Soccer, Sports)`, the cardinality ratio was calculated to be 0.065 (greater than the threshold value, 0.06). For `(Basketball, Sports)` the cardinality ratio was 0.10 ($>$0.06) and, finally, for `(Soccer, Basketball)`, the cardinality ratio was 0.005 ($<$0.06). Indeed the hypothesis holds true, as *Basketball* and *Soccer* are subclasses of *Sports*.

Again, there are certain exceptions where this hypothesis fails. But considering the behavior of folksonomies and the reasons specified, exceptions of about 15% are assumed to be tolerable. Since the analysis for this hypothesis is dependent on Hypothesis 2, it primarily fails in cases when Hypothesis 2 fails. Hence, this hypothesis is considered to be verified.

5.3. *Statistical induction from amazon.com tags*

The goals of the statistical induction effort are to (1) determine the correct classification of an item, so that it can be found precisely by a user's query, and (2) find the classification of all items, so that a user's query can be answered completely. The Folksonomy Space being used for the preliminary bottom-up experiments is assumed to consist of a set of tags $T = T_i$, having three nth-order relations. The first is R_1 (read as "occurs together"), where R_1 is:

$$R_1 : T_1 \times T_2 \times \cdots \times T_n \to N.$$

N is the set of positive integers (in this case the cardinality of the relation tuple). For example, the tags `Sports` and `Basketball` occur together in *Flickr* more than $N = 100$ times.

For the first experiments, we only considered pairs of terms (such as `Sports` and `Basketball`), so the relation simplifies to:

$$R_1 : T_1 \times T_2 \to N.$$

The second relation R_2 (read as "has same tag") is:

$$R_2 : I_1 \times I_2 \times \cdots \times I_n \to M,$$

where I_i is an item and M is a positive integer.

The third relation R_3 (read as "tagged the same item") is:

$$R_3 : P_1 \times P_2 \times \cdots \times P_n \to L,$$

where P_i is a tagger and L is a positive integer.

By sampling R_1, R_2 and R_3 we were able to induce fragments of an ontology. For example, we were able to induce that *Basketball* is a subclass of *Sports*.

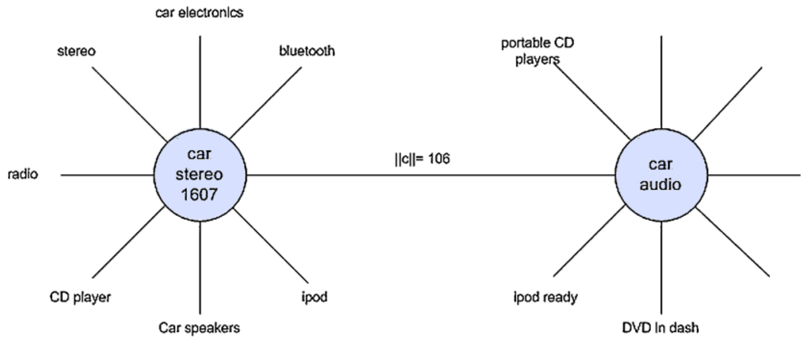


Fig. 13. Example of Graph #1 of tags and their co-occurrences. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

5.3.1. *Graphs constructed from Amazon product tags*

Our basic objective is to find one or more relationships among the tags in a tag space for some domain, and then use those to find relationships among the domain objects that the tag space describes. The latter relationships can then be used to assist users in navigating the domain.

To investigate this, we derived three graph structures from the website amazon.com, as follows:

(1) Graph (non-directed) of *Amazon* tags (see Fig. 13):

- Each node is a tag.
- Each node has a cardinality indicating the number of items (products) for which the tag was used. For example, the tag `Apache` was associated with 180 items at amazon.com.
- Each link between two tags represents a co-occurrence of those tags for an item.
- Each link is nondirectional.
- Each link has a cardinality indicating how many instances of the co-occurrence there are.
- The graph can be captured by choosing one or more "seed" tags as starting points for product searches and then performing a search on the tags for each of the returned products recursively in a breadth-first or depth-first order.

(2) Graph (non-directed) of *Amazon* items:

- Each node is an item type (class).
- Each node has a National Stock Number (NSN), UNSPSC, UPC, or equivalent product code in some namespace/registry.
- There is a link between two items if they share at least one common tag.
- Each link has a cardinality indicating the number of tags shared by the two nodes.

(3) Graph (non-directed) of taggers, i.e., people who provide tags for items:

- Each node is a tagger.
- Each node has a unique identifier (the person's *Amazon* name).
- Each link between two taggers represents that the taggers have tagged the same item.

For Graph #1, the resultant space of tags might be very large (on the order of ten times the number of products at *Amazon*), so we reduce the space by merging the nodes whose tags are first from the same stem, second that have been deemed to be synonyms based on Hypothesis 4 (described below), and third from the same synset in WordNet. Alternatively, we split a node if its elements are linked to disjoint sets of words in WordNet.

From the resultant co-occurrence graph, we can induce *disjoint subclass* and *disjoint partOf* relationships (with an associated likelihood or probability). One procedure for this is to apply and then test the following hypothesis.

Hypothesis 3a. For a group of items, if `Tag1` co-occurs with `Tag2` and `Tag1` also co-occurs with `Tag3`, but `Tag2` does not co-occur with `Tag3`, then *Tag2* and *Tag3* are subclasses of *Tag1*, and *Tag2* is disjoint with *Tag3* with a confidence based on the respective cardinalities.

A problem with this hypothesis is that it can reveal not only *disjoint subclasses* in a taxonomy, but also *disjoint superPart* classes in a meronymy, and these cannot be distinguished except by appealing to another source of knowledge, such as a thesaurus (e.g., WordNet) or an ontology (e.g., Cyc). For discovering meronymic information, we formulate the following hypothesis:

Hypothesis 3b. For a group of items, if `Tag1` co-occurs with `Tag2` and `Tag1` also co-occurs with `Tag3`, but `Tag2` does not co-occur with `Tag3`, then *Tag2* and *Tag3* are superparts of *Tag1*, and *Tag2* is disjoint with *Tag3* with a confidence based on the respective cardinalities.

For example, if `Tag1` is `Bolt`, `Tag2` is `Engine`, and `Tag3` is `File Cabinet`, then these tags would satisfy both Hypotheses 3a and 3b and there would be no way to distinguish *subpart* from *subclass*. However, if the tags could be mapped to equivalent terms in WordNet or concepts in Cyc, then the relationships might be distinguished.

Similarly, we can form hypotheses involving the relationships *owns* and *causedBy*, as indicated in Fig. 14, and these are similarly conflated.

5.3.2. *Precision and recall*

Out of 25,000 tags extracted from the *amazon.com* website, the total number of triples that have the structure `(tag1, tag3)`, `(tag1, tag2)` and `(tag2, tag3)` is 173,097. Out of these, the total number of triples that are declared to satisfy Hypothesis 3 is = 9064. We analyzed manually 100 of these and 100 of the (173,097 − 9064) = 164,033 that do not satisfy Hypothesis 3. We determined that the number of triples that satisfy Hypothesis 3 and have the *disjoint subclass* relationship is 26 out of 100. So, precision is 0.26.

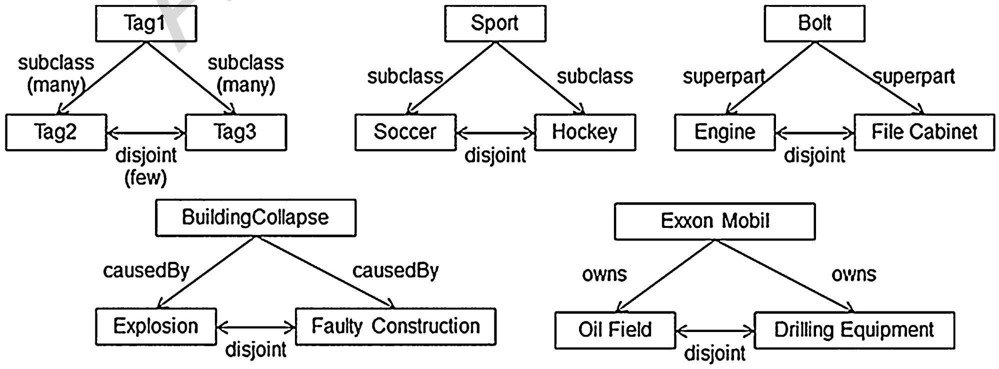


Fig. 14. Examples of the relationships derivable, but not separable, via our statistical induction techniques.

Of the triples that do NOT satisfy Hypothesis 3, but are determined to have the disjoint subclass relationship, there are 6 out of 100. So, recall is $26/(26+6)=0.81$. Changing the thresholds changes the balance between precision and recall, and this can be adjusted in practice based on a particular application of folksonomy-aided information retrieval.

We have done some initial data analysis using F-measure, which is the harmonic mean of precision and recall, i.e.,

$$F_1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}).$$

The thresholds we use for deciding whether or not a particular triple of tags satisfies Hypothesis 3 can be chosen to maximize this F-measure.

Examples:

- Tag pairs that are declared to be *disjoint subclasses* according to Hypothesis 3 and actually are:
  - `nfl` and `college football` are *disjoint subclasses* of *football*,
  - `cambodia` and `vietnam` are *disjoint subclasses* of *southeast asia*,
  - `bustiers` and `shapewear` are *disjoint subclasses* of *slenderizers*.
- Tag pairs that are declared to be subclasses according to Hypothesis 3 and actually are NOT:
  - `confederacy` and `abraham lincoln` are not *subclasses* of *civil war*.

5.3.3. *Hypotheses* 1, 2 *and* 4

For a tag space, such as exists at *Amazon* or *Flickr*, we evaluate Hypotheses 1, 2 and 4 as follows. For each pair of distinct tags, t_i and t_j, where $i \neq j$, and c_{ij} represents the co-occurrence between t_i and t_j.

If $\|c_{ij}\| < \alpha(\|t_i\| + \|t_j\|)/2$, then there is no relationship between t_i and t_j. The constant α is determined experimentally with data from *Flickr* and *Amazon* to be 0.06.

If $\|c_{ij}\| > \beta(\|t_i\| + \|t_j\|)/2$, then there is a synonym relationship between t_i and t_j. The constant β is determined experimentally to be 0.75.

Otherwise, if $\alpha(\|t_i\|+\|t_j\|)/2 \leqslant \|c_{ij}\| \leqslant \beta(\|t_i\|+\|t_j\|)/2$, then there is a disjoint *subclass–superclass* relationship between t_i and t_j, where $t_i \subset t_j$, if $\|t_i\| < \|t_j\|$ and $t_i \supseteq t_j$, if $\|t_i\| \geqslant \|t_j\|$.

We can also link each node to a synset in WordNet and from there to a concept in an ontology (in the current system, Cyc or SUMO/MILO). A user query based on a given tag can be broadened or narrowed, as needed, by using either more general or more specific concepts from Cyc or WordNet.

The construction of Graph #3 is difficult, because the needed data is not publicly available from *Amazon*. It is not yet clear what folksonomic structure, and thus utility, can be derived from Graphs #2 and #3. We expect that continuing research will reveal this.

The possibilities for graphs over tags, taggers, and items that can be analyzed to reveal implicit structure are:

- Co-occurrences of tags based on items (Graph #1).
- Relationship on tags based on taggers.
- Relationship on items based on tags (Graph #2).
- Relationship on items based on taggers.
- Relationship on taggers based on items.
- Relationship on taggers based on tags (Graph #3).

Besides the inability to distinguish among several different kinds of semantic relationships, as described above, our statistical induction techniques cannot distinguish between classes and objects (class instances). For example, it cannot distinguish between the class *OilCompany* and the instance `ExxonMobil`. We anticipate that mappings to ontologies will facilitate making distinctions like this, and we suggest investigating this in future research.

5.3.4. *Additional results from amazon.com*

Our Hypothesis 4 concerns the detection of synonymy. We present here the results for the amazon.com data. Each pair of tags has three numbers associated with it: two counts for the individual tags and the count of co-occurrences. We first compute the means of the three numbers. If all three lie within 0.75 (more about this value below) times the standard deviation (for these three numbers), then the two tags are said to be synonyms.

If we choose the threshold value to be 0.75 times the standard deviation, then about 3212 pairs are said to be related and 22,493 are not said to be related of the total 25,000 data items extracted from the amazon.com website (i.e., pairs that have cardinalities). If the threshold is 0.70 times the standard deviation, then NONE of the tag pairs is deemed to be related. At the other extreme, if the threshold is 0.85 times the standard deviation, then ALL of the tag pairs are found to be related. Just as for Hypothesis 3, we plan to choose the threshold based on maximizing the F-measure.

5.3.5. *Precision and recall*

According to Hypothesis 4, of the 100 tag pairs that are said to be synonymous, 35 are actually synonyms. So, precision is 0.35.

According to Hypothesis 4, of the 100 tag pairs that are said to be NOT synonymous, 3 are actually synonyms. So, recall is $35/38 = 0.92$.

Examples:

- Tag pairs that are declared to be synonymous according to Hypothesis 4 and actually are:

```
christian music, christian rock
slimmer, shapewear
forum, forums
james bond, 007
```

- Tag pairs that are declared to be synonymous according to Hypothesis 4 and actually are NOT:

```
homesensors, smokedetectors
door chimes, bells
civil war, abraham lincoln
```

- Tag pairs that are NOT declared to be synonymous according to Hypothesis 4 and actually are NOT:

```
super reader, horror
scented pitcher, glassware
```

- Tag pairs that are NOT declared to be synonymous according to Hypothesis 4 and actually are:

```
home safety, home security
seatpost, bicycle seat post
lead test, lead check
```

5.4. Discussion of induction results

In this section, we have examined tagging and folksonomies: their emergence, importance, and future use. Tagging began with one website pioneering this idea and now many developers are trying to use some variation of a tagging system for their websites. Considering the future of tagging systems and the way users are tagging items on the Web, it is going to become more difficult to find things using tags. Therefore, our research has proposed and investigated a few hypotheses that can facilitate tag-based search. The idea is to identify the network of related tags for a given tag and, for searches, retrieve all items within a short "tag distance". Based on the results of the experiments already completed, we can conclude that it is feasible to derive an ontological structure from a given folksonomy and use it to retrieve additional relevant information.

Although our first hypothesis was not verified, a number of important inferences can be drawn from the results that we obtained. That is, users tend to tag data with different keywords that may or may not be the leaf nodes in a hierarchy depending on the domain to which the data belongs. For some domains, this hypothesis yields good results, whereas for other domains it fails. Consequently, work can be done to categorize data in different domains and hence a more domain-specific hypothesis can be made that might yield more accurate results.

Our second hypothesis was a direct consequence of the first. The second hypothesis does not tell us clearly what relationship exists between two tags, but it does reveal whether a relationship exists or not. This hypothesis extracts a vague relationship, which is then used along with our third hypothesis to determine a more accurate relationship. Our third hypothesis is an extension of the second one.

Our continuing work will be based on expanding this research to either more specific domains or trying the same approach with different on-line tagging systems. Experiments on *Flickr* and *Amazon* have produced surprisingly consistent results, but they are insufficient to be generalized for other tagging systems. The primary reason is that the method and hypothesis given for tags used on *Flickr* and *Amazon* may or may not be applicable to other systems, such as del.icio.us, since these are two very different tagging systems. For instance, a *superclass–subclass* relationship derived in *Flickr* may turn out to be a *subclass–superclass* relationship in del.icio.us or the tags may not be related at all. The results of our experiments convey the same message.

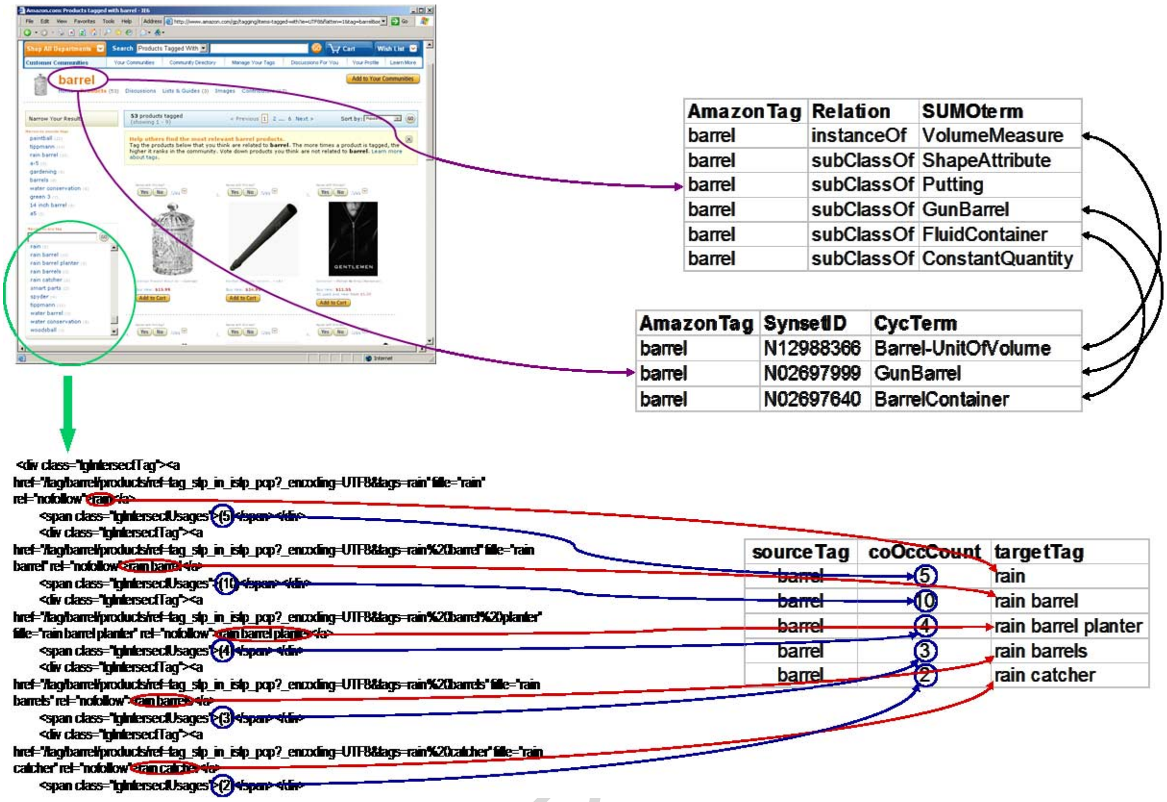
5.5. Semantic/lexical mapping

A general discussion of the process of mapping tags to structured artifacts, such as lexicons and taxonomies, was presented in Section 3.1. In the following sections, we review the processing we performed on the *Amazon* tag dataset.

5.5.1. Extraction

As noted previously, the Amazon tag dataset was deliberately chosen as a "worst-case" example. It is relatively inaccessible, buried in voluminous raw HTML in undocumented formats. Our extraction technique utilized regular-expression patterns to find the tag co-occurrence data.

The tag data was obtained by issuing an initial HTTP request to amazon.com with a "seed" tag; seed tags were deliberately chosen to be ambiguous – to have multiple senses and meanings. Once the co-occurrence data was extracted from the initial page, the co-occurring tags were queued and submitted via HTTP request to amazon.com one at a time, and the page returned for each tag was subjected to the same processing as the seed tag's page. The extraction process is illustrated in Fig. 15.



AmazonTag	Relation	SUMOterm
barrel	instanceOf	VolumeMeasure
barrel	subClassOf	ShapeAttribute
barrel	subClassOf	Putting
barrel	subClassOf	GunBarrel
barrel	subClassOf	FluidContainer
barrel	subClassOf	ConstantQuantity

AmazonTag	SynsetID	CycTerm
barrel	N12988366	Barrel-UnitOfVolume
barrel	N02697999	GunBarrel
barrel	N02697640	BarrelContainer

```
<div class="tgIntersectTag"><a
href="/tag/barrel/products/ref=tag_stp_in_istp_pop?_encoding=UTF8&tags=rain" title="rain"
rel="nofollow">rain</a>
     <span class="tgIntersectUsages">(5)</span></div>
     <div class="tgIntersectTag"><a
href="/tag/barrel/products/ref=tag_stp_in_istp_pop?_encoding=UTF8&tags=rain%20barrel" title="rain
barrel" rel="nofollow">rain barrel</a>
     <span class="tgIntersectUsages">(10)</span></div>
     <div class="tgIntersectTag"><a
href="/tag/barrel/products/ref=tag_stp_in_istp_pop?_encoding=UTF8&tags=rain%20barrel%20planter"
title="rain barrel planter" rel="nofollow">rain barrel planter</a>
     <span class="tgIntersectUsages">(4)</span></div>
     <div class="tgIntersectTag"><a
href="/tag/barrel/products/ref=tag_stp_in_istp_pop?_encoding=UTF8&tags=rain%20barrels" title="rain
barrels" rel="nofollow">rain barrels</a>
     <span class="tgIntersectUsages">(3)</span></div>
     <div class="tgIntersectTag"><a
href="/tag/barrel/products/ref=tag_stp_in_istp_pop?_encoding=UTF8&tags=rain%20catcher" title="rain
catcher" rel="nofollow">rain catcher</a>
     <span class="tgIntersectUsages">(2)</span></div>
```

sourceTag	coOccCount	targetTag
barrel	5	rain
barrel	10	rain barrel
barrel	4	rain barrel planter
barrel	3	rain barrels
barrel	2	rain catcher

Fig. 15. Tag extraction process for amazon.com. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

The order in which tags were processed was based on normalized tag co-occurrence values with the tag for which they first co-occurred; this was a heuristic, and we are interested in examining other such heuristics to determine whether the order of processing affects the results in any significant way.

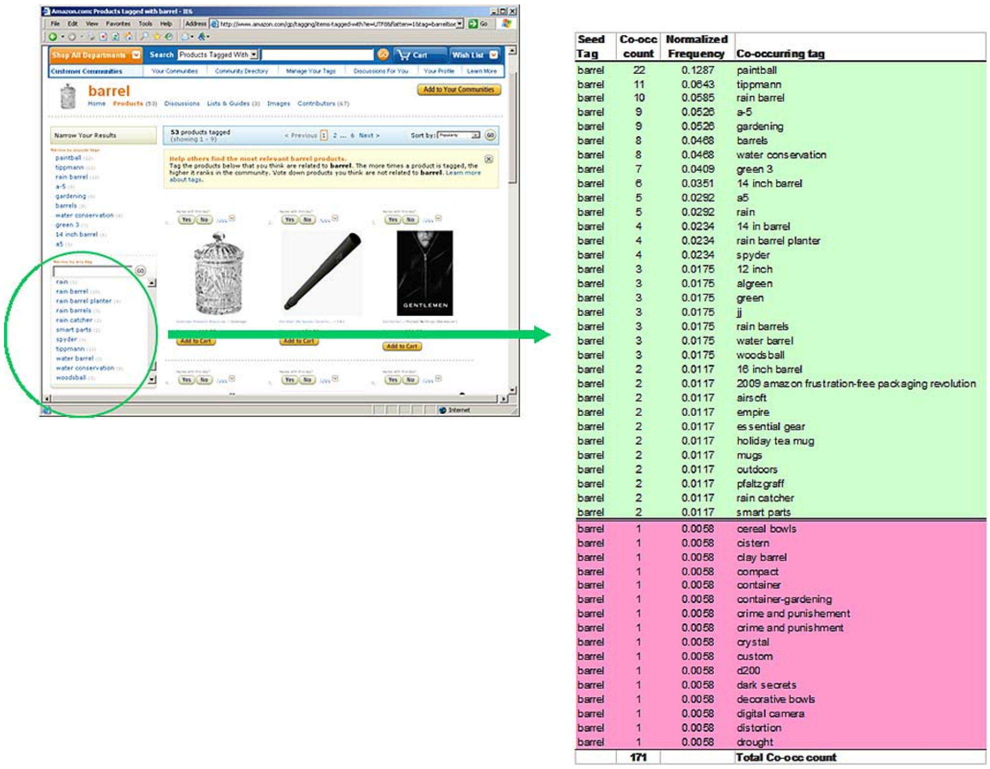
Tag co-occurrences were stored as a graph with tags as nodes and co-occurrence frequency as weighted links, while the tag extraction subsystem used a graph database.

Since *Amazon* tag data can be presumed to be variable and we wished to perform experiments on a stable data set, we experimented on snapshots of the tag set at particular points in time; we note that an operational system would need to allow for growth and change of the tag set and this is a features that warrants further experimentation.

In our experiments, 101,372 distinct tags were extracted from *Amazon's* total tag set. Without direct access to the full tag set, we cannot determine what percentage of the total this extracted subset represents, nor do we know of any way of determining this other than by exhaustive enumeration.

5.5.2. *Filtering*

Not all co-occurring tags were added to the queue: only those exceeding a threshold value were added. This filtering was inserted into the preprocessing based on initial examination of the *Amazon* tag data, which exhibited a large amount of noise in the form of highly idiosyncratic, subjective, and even inscrutable tags. This was not unexpected, given that *Amazon* tag data is at the extreme end of many axes



Amazon.com: Products tagged with barrel - IE6

File Edit View Favorites Tools Help Address http://www.amazon.com/gp/tagging/items-tagged-with?ie=UTF8&flatten=1&tag=barrel&se Go

Shop All Departments Search Products Tagged With GO Cart Wish List

Customer Communities | Your Communities | Community Directory | Manage Your Tags | Discussions For You | Your Profile | Learn More

barrel

Home Products (53) Discussions Lists & Guides (3) Images Contributors (67)

Add to Your Communities

Narrow Your Results

paintball
tippmann
rain barrel
a-5
gardening
barrels
water conservation
green 3
14 inch barrel
a5

rain
rain barrel
rain barrel planter
rain barrels
rain catcher
smart parts
spyder
tippmann
water barrel
water conservation
woodsball

53 products tagged (showing 1 - 9) < Previous 1 2 ... 6 Next > Sort by: Popularity GO

Help others find the most relevant barrel products.

Tag the products below that you think are related to barrel. The more times a product is tagged, the higher it ranks in the community. Vote down products you think are not related to barrel. Learn more about tags.

Seed Tag	Co-occ count	Normalized Frequency	Co-occuring tag
barrel	22	0.1287	paintball
barrel	11	0.0643	tippmann
barrel	10	0.0585	rain barrel
barrel	9	0.0526	a-5
barrel	9	0.0526	gardening
barrel	8	0.0468	barrels
barrel	8	0.0468	water conservation
barrel	7	0.0409	green 3
barrel	6	0.0351	14 inch barrel
barrel	5	0.0292	a5
barrel	5	0.0292	rain
barrel	4	0.0234	14 in barrel
barrel	4	0.0234	rain barrel planter
barrel	4	0.0234	spyder
barrel	3	0.0175	12 inch
barrel	3	0.0175	algreen
barrel	3	0.0175	green
barrel	3	0.0175	jj
barrel	3	0.0175	rain barrels
barrel	3	0.0175	water barrel
barrel	3	0.0175	woodsball
barrel	2	0.0117	16 inch barrel
barrel	2	0.0117	2009 amazon frustration-free packaging revolution
barrel	2	0.0117	airsoft
barrel	2	0.0117	empire
barrel	2	0.0117	essential gear
barrel	2	0.0117	holiday tea mug
barrel	2	0.0117	mugs
barrel	2	0.0117	outdoors
barrel	2	0.0117	pfaltzgraff
barrel	2	0.0117	rain catcher
barrel	2	0.0117	smart parts
barrel	1	0.0058	cereal bowls
barrel	1	0.0058	cistern
barrel	1	0.0058	clay barrel
barrel	1	0.0058	compact
barrel	1	0.0058	container
barrel	1	0.0058	container-gardening
barrel	1	0.0058	crime and punishement
barrel	1	0.0058	crime and punishment
barrel	1	0.0058	crystal
barrel	1	0.0058	custom
barrel	1	0.0058	d200
barrel	1	0.0058	dark secrets
barrel	1	0.0058	decorative bowls
barrel	1	0.0058	digital camera
barrel	1	0.0058	distortion
barrel	1	0.0058	drought
	171		Total Co-occ count

Fig. 16. The process for filtering the extracted tags. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-130124.)

in Folksonomy Space: totally unconstrained vocabulary, maximally heterogeneous tagger population, etc.

Figure 16 illustrates the filtering process. For the tag `Barrel` we find more than 50 co-occurring tags, many (if not most) of which do not appear to be relevant to any of the senses of `Barrel` (or are perhaps meaningful only to the tagger – an example of highly idiosyncratic tagging). Determining which of these are truly relevant is not straightforward, because setting too high a threshold may eliminate viable tags (low precision), while setting one too low helps very little (too high recall). After some experimentation, we determined that by normalizing the co-occurrence count by dividing each count by the total co-occurrence count for the "source" tag, and eliminating any tag with a normalized value below 1%, we achieved a reasonable tradeoff between effective filtering and eliminating viable tags: as can be seen above, had we chosen 2%, several viable tags (e.g., `Rain Barrels` and `Water Barrel` would have been eliminated).

We conjecture that the threshold value may be strongly related to the values of dimensions measuring the *coherence* of the tag set, e.g., homogeneity and level of expertise of taggers, scope of domain, etc. This is another area in which we recommend pursuing additional research.

5.5.3. Mapping

A complication involves the format and consistency of the available mappings between WordNet and the two upper ontologies: SUMO/MILO, in particular, presented a significant challenge given the resources available, since the mappings are implemented by extending the native, idiosyncratic, WordNet data tables with additional fields and notations. While there is software to support manipulating these data tables, expediency dictated a simple text conversion to a form that was more manageable.

A final complication relates to the internal formats of the various lexical and ontology resources with regard to capitalization, spelling, use of delimiters and connectors, e.g., '−' and '_' vs. embedded spaces. Some normalization was done via simple text manipulation, but this – while not a significant challenge – will nonetheless have to be addressed in subsequent research.

All lexicon and ontology data was therefore normalized and formatted as text and imported into a DBMS as separate tables, and the matching of terms was via SQL queries. The initial mappings were therefore based on simple case-insensitive string matching without the use of stemming or other text extraction techniques, and no attempt was made to deconstruct compound tags, which occur frequently in the *Amazon* data.

We also note that we are aware of mappings between WordNet and DOLCE, a third upper ontology. We omitted DOLCE from our experiments because the mappings are to a previous version of WordNet, thereby requiring a translation between WordNet versions.

5.5.4. Experimental results: WordNet ⇔ Cyc

As noted above, experiments were conducted with 101,372 distinct tags extracted from *Amazon's* tag data. Tags were matched (case-insensitive string match) against WordNet terms in Cyc ⇔ WordNet 2.0 mappings; no stemming or other typical text preprocessing was performed. Multiword (compound) tags were not deconstructed; none of the WordNet terms in the Cyc ⇔ WordNet 2.0 mappings contains any white space, so compound tags would not be matched.

Despite the lack of sophistication in preprocessing and matching, 6863 matches were identified – a 6.77% hit rate. Multiple word senses were identified (e.g., conductor ⇒ `ConductingMedium` and conductor ⇒ `MusicalConductor`), as were multiple parts of speech (e.g., combat ⇒ `Battle(v)` and combat ⇒ `Fight-Physical(n)`).

Given the relatively large amount of noise in the data, even after filtering, a success rate of 6.77% for simple string matching exceeded our expectations. Prior experience with information extraction and text processing leads us to believe that increases are very likely when both conventional text-processing techniques and more sophisticated techniques based on statistical matching and ontology alignment are employed, and when compound terms that consist of, e.g., modifier/noun pairs are deconstructed.

5.5.5. Experimental results: WordNet ⇔ SUMO/MILO

Experiments with SUMO/MILO were performed with the same set of 101,372 *Amazon* tags as for Cyc. Tags were matched (case-insensitive string match) against WordNet terms in SUMO/MILO ⇔ WordNet 2.0 mappings. As with Cyc, no stemming or other typical text preprocessing was performed, and multiword (compound) tags were not deconstructed; none of the WordNet terms in the SUMO/MILO ⇔ WordNet 2.0 mappings contains any white space, so compound tags would not be matched.

Despite the lack of sophistication in preprocessing and matching, 18,208 matches were identified – a 17.96% hit rate and, as with Cyc, multiple word senses and multiple parts of speech were identified.

We note here that there is a significant difference between the Cyc and SUMO/MILO mappings to WordNet, and that once the formatting issues have been resolved there is additional information in the SUMO/MILO mappings that will almost certainly provide additional leverage to our matching. Whereas Cyc-to-WordNet mappings are implicitly all equivalences, the SUMO/MILO mappings contain one of a number of relationship tags, including equivalence, subsumption, and type. This information can be used to further refine the mappings.

5.5.6. *Significance of results*

With relatively simple and straightforward data extraction and preprocessing techniques, we achieved reasonable success with tag ⇒ lexicon ⇒ ontology mappings, and we are confident that application of conventional text-extraction techniques, as well as deconstruction of compound tags, will dramatically improve the rate of success.

We note that our mapping experiments were performed independent of the induction experiments, but since *by definition* induced classes will be drawn from the same tag population, mapping from induced class structures should be at least as good as from raw tags, and statistical and ontology alignment techniques should improve mappings even further.

6. Summary and suggested research

Although our first hypothesis was not verified, a number of important inferences can be drawn from the results that we obtained. That is, users tend to tag data with different keywords that may or may not be the leaf nodes in a hierarchy depending on the domain to which the data belongs. For some domains, this hypothesis yields good results, whereas for other domains it fails. Consequently, work can be done to categorize data in different domains and hence a more domain specific hypothesis can be made that will yield more accurate results.

Our second hypothesis was a direct consequence of the first hypothesis. This hypothesis does not tell us clearly what relationship exists between two tags, but it does reveal whether a relationship exists or not. This hypothesis extracts a vague relationship, which is then used along with our third hypothesis to get the exact relationship. Our third hypothesis is an extension to the second one.

A problem with Hypothesis 3 is that it can reveal not only disjoint *subclasses*, but also disjoint *superPart* classes in a meronymy, and these cannot be distinguished except by appealing to another source of knowledge, such as a thesaurus (e.g., WordNet) or an ontology (e.g., Cyc). For example, if Tag1 is `Bolt`, Tag2 is `Engine`, and Tag3 is `File Cabinet`, then these tags would satisfy Hypothesis 3, but *subpart* would be the preferred relation, which might be discernable with the use of WordNet or Cyc. Similarly, we can form hypotheses involving the relationships *owns* and *causedBy*, and these are also conflated with *subclass*.

Experiments on *Flickr* and *Amazon* data have produced surprisingly consistent results, but they are insufficient to be generalized for other tagging systems. The primary reason is that the method and hypothesis given for tags used on *Flickr* and *Amazon* may or may not be applicable to other systems, such as *del.icio.us*, which cover very different domains.

Besides the inability to distinguish among several different kinds of semantic relationships, as described above, our statistical induction techniques cannot distinguish between classes and objects (class instances). For example, they cannot distinguish between the class *OilCompany* and the instance `ExxonMobil`. We anticipate that mappings to ontologies will facilitate making such distinctions.

One limitation of our approach is that it does not take into account the fact the mappings we identify are intrinsically probabilistic. As we develop more sophisticated mapping techniques, particularly as we begin to apply statistical techniques to our matching processes, we expect to move to a probabilistic model of the mappings that will allow reasoning to proceed without making unwarranted assumptions about the accuracy of the mappings.

Finally, to exploit the hypotheses we have formulated and verified, the discovered explicit structure can be captured in a formalism such as OWL. It can then be mapped to an existing ontology expressed in the same formalism. A reasoner for that formalism (such as Pellet for OWL) could then reason over the combination.

We see ample opportunity to improve all aspects of our tag mining strategies and preprocessing techniques. A significant extension of our data extraction and processing system will be to expand the "entity" population to include taggers and the objects being tagged, in addition to the tags themselves. This was extremely difficult to do with the *Amazon* data.

Because tags are usually assigned informally and heuristically, tag collections are *ad hoc* and "noisy" in an information theoretic sense. There is implicit structure in a collection, but it is obscured by the noise and tagging imprecision, as well as by the differing semantics of the taggers. To minimize the noise and derive the maximum amount of structure from a tag collection and make it explicit, all possible knowledge must be utilized. The available knowledge includes information about the tagging population, expressed statistically, as well as cross-correlations among the tags, objects being tagged, and taggers. The data analysis can be guided by relevant high-level ontologies (Cyc and SUMO/MILO) or domain-specific ontologies.

We recommend that our inductive techniques be extended by incorporating the cross-correlations above, as well as by a higher-dimensional analysis involving multiple degrees of co-occurrence. We also recommend that the techniques be enabled to work incrementally, so that they can be applied in an active domain. That is, if new objects (such as items of evidence in an intelligence domain) are added along with new tags, it would be necessary to calculate new and better ontology fragments as the additions are made, rather than completely recalculating ontology fragments.

Acknowledgements

The authors wish to thank the team at Global Infotek, Inc. for their invaluable expertise and guidance on the research reported herein.

References

Angeletou, S. (2008). Semantic enrichment of folksonomy tagspaces. In *Proceedings of the 7th International Conference on The Semantic Web, ISWC'08* (pp. 889–894). Berlin, Heidelberg: Springer-Verlag.

Angeletou, S., Sabou, M., Specia, L. & Motta, E. (2007). Bridging the gap between folksonomies and the semantic web: An experience report. In *Workshop: Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference* (pp. 30–43).

Benz, D., Hotho, A., Stützer, S. & Stumme, G. (2010). Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, Raleigh, NC, USA. Available at: http://www.kde.cs.uni-kassel.de/pub/pdf/benz2010semantics.pdf.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web: Scientific American. *Scientific American*, *284*(5), 34–46.

Castano, S., Ferrara, A. & Montanelli, S. (2003). H-match: an algorithm for dynamically matching ontologies in peer-based systems. In *SWDB* (pp. 231–250).

Cattuto, C., Benz, D., Hotho, A. & Stumme, G. (2008a). Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)* (pp. 39–43). Patras, Greece.

Cattuto, C., Benz, D., Hotho, A. & Stumme, G. (2008b). Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Conference on The Semantic Web, ISWC'08* (pp. 615–631). Berlin, Heidelberg: Springer-Verlag.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M. & Stumme, G. (2007). Network properties of folksonomies. *AI Communications, 20*(4), 245–262.

Collet, C., Huhns, M. & Shen, W.-M. (1991). Resource integration using a large knowledge base in carnot. *IEEE Computer, 24*(12), 55–62.

Doan, A. and Halevy, A.Y. (2005). Semantic integration research in the database community: A brief survey. *AI Magazine, 26*, 83–94.

Doan, A., Madhavan, J., Dhamankar, R., Domingos, P. & Halevy, A. (2003). Learning to match ontologies on the semantic web. *The VLDB Journal, 12*(4), 303–319.

Dou, D., McDermott, D.V. & Qi, P. (2003). Ontology translation on the semantic web. In *Proceedings of the International Conference on Ontologies, Databases and Application of Semantics (ODBASE 2003)* (pp. 952–969).

Eda, T., Yoshikawa, M., Uchiyama, T. & Uchiyama, T. (2009). The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. *World Wide Web, 12*(4), 421–440.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database. Language, Speech, and Communication*. MIT Press.

Giunchiglia, F., Shvaiko, P. & Yatskevich, M. (2005). Semantic schema matching. In *Proceedings of the Thirteenth International Conference on Cooperative Information Systems (CoopIS 2005)* (pp. 347–365).

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

Halevy, A., Norvig, P. & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.

Hayman, S. (2007). Folksonomies and tagging: New developments in social bookmarking. In *Proceedings of the Ark Group Conference: Developing and Improving Classification Schemes*. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.8884.

Helic, D., Strohmaier, M., Trattner, C., Muhr, M. & Lerman, K. (2011). Pragmatic evaluation of folksonomies. In *Proceedings of the 20th International Conference on World Wide Web, WWW'11* (pp. 417–426). New York, NY, USA: ACM.

Heymann, P. & Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Standford University.

Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. (2006). Information retrieval in folksonomies: search and ranking. In *Proceedings of the 3rd European Conference on The Semantic Web: Research and Applications, ESWC'06* (pp. 411–426). Berlin, Heidelberg: Springer-Verlag.

Huang, J., Dang, J. & Huhns, M.N. (2006). Reconciling ontologies for coordination among e-business agents. In *Proc. AAMAS Workshop on Business Agents and the Semantic Web*. Available at: http://www.cse.sc.edu/~huhns/confpapers/BASeWEB_Huang_Dang_Huhns_condensed.pdf.

Huang, J., Dang, J., Huhns, M.N. & Shao, Y. (2007). Ontology alignment as a basis for mobile service integration and invocation. *Journal of Pervasive Computing and Communications, 3*(2), 138–158.

Huang, J., Gutierrez, R.L.Z., Garcia, B.M. & Huhns, M.N. (2005a). Reconciling agent ontologies for web service applications. In Eymann, T., Klugl, F., Lamersdorf, W., Klusch, M. & Huhns, M.N. (Eds.), *Multiagent System Technologies: Third German Conference, MATES 2005*. Lecture Notes in Computer Science (Vol. 3550, pp. 106–117). Springer.

Huang, J. & Huhns, M.N. (2006). Superconcept formation system – an ontology matching algorithm for service discovery. In *Proceedings of Service Discovery on the WWW Workshop (SDISCO 2006)*. Available at: http://www.cse.sc.edu/~huhns/confpapers/sd06-huang-v2.pdf.

Huang, J., Zavala, L., Mendoza, B. & Huhns, M.N. (2005b). A schema-based approach combined with inter-ontology reasoning to construct consensus ontologies. In *Proceedings of the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications* (pp. 80–87). AAAI Press.

Huang, J., Zavala, R.L., Mendoza, B. & Huhns, M.N. (2005c). Sharing ontology schema information for web service integration. In *Fifth International Conference on Computer and Information Technology (CIT 2005)* (pp. 1056–1062). IEEE Computer Society.

Huhns, M.N. & Singh, M.P. (1997). Agents on the web: Ontologies for agents. *IEEE Internet Computing, 1*(6), 81–83.

Huhns, M.N. & Stephens, L.M. (1999). Agents on the web: Personal ontologies. *IEEE Internet Computing, 3*(5), 85–87.

Huhns, M.N. & Stephens, L.M. (2002). Semantic bridging of independent enterprise ontologies. In Kosanke, K., Jochem, R., Nell, J.G. & Bas, A.O. (Eds.), *Enterprise Inter- and Intra-Organizational Integration: Building International Consensus, IFIP TC5/WG5.12 International Conference on Enterprise Integration and Modeling Technique, (ICEIMT 2002)* (Vol. 236, pp. 83–90). Kluwer Academic Publishers.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L. & Stumme, G. (2007). Tag recommendations in folksonomies. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2007* (pp. 506–514). Berlin, Heidelberg: Springer-Verlag.

Kiryakov, A., Popov, B., Terziev, I., Manov, D. & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web, 2*(1), 49–79.

Knautz, K., Soubusta, S. & Stock, W.G. (2010). Tag clusters as information retrieval interfaces. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences, HICSS'10* (pp. 1–10), Washington, DC, USA: IEEE Computer Society.

Körner, C., Benz, D., Hotho, A., Strohmaier, M. & Stumme, G. (2010a). Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 521–530).

Körner, C., Kern, R., Grahsl, H.-P. & Strohmaier, M. (2010b). Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, HT'10* (pp. 157–166). New York, NY, USA: ACM.

Lalwani, S. & Huhns, M.N. (2009). Deriving ontological structure from a folksonomy. In *Proceedings of the 47th Annual Southeast Regional Conference, ACM-SE 47* (pp. 49:1–49:2). New York, NY, USA: ACM.

Li, R., Bao, S., Yu, Y., Fei, B. & Su, Z. (2007). Towards effective browsing of large scale social annotations. In *Proceedings of the 16th International Conference on World Wide Web, WWW'07* (pp. 943–952). New York, NY, USA: ACM.

Macgregor, R. & Bates, R. (1987). The loom knowledge representation language. Technical report, University of Southern California.

Madhavan, J., Bernstein, P.A. & Rahm, E. (2001). Generic schema matching with cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases, VLDB'01* (pp. 49–58). San Francisco, CA, USA: Morgan Kaufmann.

Mahalingam, K. & Huhns, M.N. (1997). An ontology tool for query formulation in an agent-based context. In *CoopIS* (pp. 170–178). IEEE Computer Society.

Mahalingam, K. & Huhns, M.N. (2000). Ontology tools for semantic reconciliation in distributed heterogeneous information environments. In *Intelligent Automation and Soft Computing* (Vol. 6, pp. 1–8). (Special issue on Distributed Intelligent Systems.)

Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In *Proceedings 18th International Conference on Data Engineering* (pp. 117–128).

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V.R. & Musen, M.A. (Eds.), *International Semantic Web Conference*, Lecture Notes in Computer Science (Vol. 3729, pp. 522–536). Springer.

Missier, P., Alper, P., Corcho, O., Dunlop, I. & Goble, C. (2007). Requirements and services for metadata management. *IEEE Internet Computing, 11*(5), 17–25.

Noy, N.F. & Musen, M.A. (2000). Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 450–455). AAAI Press.

Pastor, J., McKay, D. & Finin, T. (1992). View-concepts: Knowledge-based access to databases. In *Proceedings of the First International Conference on Information and Knowledge Management* (pp. 84–91).

Peters, I. & Becker, P. (2009). *Folksonomies: Indexing and Retrieval in Web 2.0*. Knowledge & Information: Studies in Information Science. De Gruyter/Saur.

Plangprasopchok, A., Lerman, K. & Getoor, L. (2010). Growing a tree in the forest: constructing folksonomies by integrating structured metadata. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'10* (pp. 949–958). New York, NY, USA: ACM.

Plangprasopchok, A., Lerman, K. & Getoor, L. (2011). A probabilistic approach for learning folksonomies from structured data. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM'11* (pp. 555–564). New York, NY, USA: ACM.

Plangprasopchok, A. & Lerman, K. (2009). Constructing folksonomies from user-specified relations on flickr. In *Proceedings of the 18th International Conference on World Wide Web, WWW'09* (pp. 781–790). New York, NY, USA: ACM.

Sabou, M., d'Aquin, M. & Motta, E. (2008). Scarlet: Semantic relation discovery by harvesting online ontologies. In *Proceedings of the 5th Annual European Semantic Web Conference (ESWC 2008)* (pp. 854–858).

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Computer Science Series. Addison-Wesley.

Schmitz, P. (2006). Inducing ontology from flickr tags. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*. Edinburgh, UK. Available at: http://www.ibiblio.org/www_tagging/2006/22.pdf.

Sheth, A. & Stephens, S. (2007). Semantic web: Technologies and applications for the real world. World-Wide Web Conference Tutorial. Available at: http://www2007.org/tutorial-T11.php.

Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags*. Available at: http://www.shirky.com/writings/ontology_overrated.html.

Singh, M. & Huhns, M. (2005). *Service-Oriented Computing: Semantics, Processes, Agents*. Wiley.

Solskinnsbakk, G. & Gulla, J.A. (2010). A hybrid approach to constructing tag hierarchies. In *Proceedings of the 2010 International Conference on the Move to Meaningful Internet Systems: Part II, OTM'10* (pp. 975–982). Berlin, Heidelberg: Springer-Verlag.

Stephens, L.M., Gangam, A.K. & Huhns, M.N. (2003). Developing consensus ontologies for the semantic web. In *Proceedings of the Workshop on Semantic Integration* (pp. 86–92).

Stephens, L.M., Gangam, A.K. & Huhns, M.N. (2004). Constructing consensus ontologies for the semantic web: A conceptual approach. *World Wide Web, 7*(4), 421–442.

Stephens, L.M. & Huhns, M.N. (2001). Consensus ontologies: Reconciling the semantics of web pages and agents. *IEEE Internet Computing, 5*(5), 92–95.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics, 4*(1), 14–28.

Van Damme, C., Hepp, M. & Siorpaes, K. (2007). Folksontology: An integrated approach for turning folksonomies into ontologies. In *Proceedings of the ESWC Workshop Bridging the Gap between Semantic Web and Web 2.0*. Springer. Available at: http://www.heppnetz.de/files/vandammeheppsiorpaes-folksontology-semnet2007-crc.pdf.

Wal, T.V. (2007). *Folksonomy Coinage and Definition*. Available at: http://vanderwal.net/folksonomy.html.

Wu, L.F. (2011). The accelerating growth of online tagging systems. *The European Physical Journal B – Condensed Matter and Complex Systems*, *83*(2), 283–287.

Yeung, C.A., Gibbins, N. & Shadbolt, N. (2007). Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, WI-IATW'07* (pp. 3–6). Washington, DC, USA: IEEE Computer Society.

Yeung, C.A., Gibbins, N. & Shadbolt, N. (2008). A study of user profile generation from folksonomies. In *Proceedings of the Social Web and Knowledge Management Workshop, SWKM'08*. Available at: http://eprints.soton.ac.uk/265222/1/swkm2008_paper.pdf.

Zhou, D., Bian, J., Zheng, S., Zha, H. & Giles, C.L. (2008). Exploring social annotations for information retrieval. In *Proceedings of the 17th International Conference on World Wide Web, WWW'08* (pp. 715–724). New York, NY, USA: ACM.