Abstract Framework for Social Ontologies and Folksonomized Ontologies

Hugo Alves
Institute of Computing - Unicamp
Avenida Albert Einstein, 1251
Cidade Universitária, Campinas, Brazil
hugo.alves@students.ic.unicamp.br

André Santanchè
Institute of Computing - Unicamp
Avenida Albert Einstein, 1251
Cidade Universitária, Campinas, Brazil
santanche@ic.unicamp.br

ABSTRACT

Many web-based public repositories are widely adopting tag-based metadata approaches as their main classification mechanism. This phenomenon is fostering initiatives to improve the semantic interpretation of tags, usually involving two main entities: "social ontologies", which emerges from the collaborative tagging (folksonomies), and formal ontologies. In order to analyze and compare these initiatives we present here an abstract framework. It supports our argument that existing approaches do not explore the full potential of the combination between folksonomies and ontologies due to their unidirectionality. The framework is also the basis to evidence that a fusion approach – as proposed in our "folksonomized ontology" – enables to better explore the organic semantics of folksonomies combined to the engineered semantics of ontologies.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

folksonomy, ontology, abstract model, semantic similarity, information content

1. INTRODUCTION

A growing number of web systems offer services for content storage, indexing, and sharing. Most of these systems use tag-based social networks to organize and index the stored content. Their users associate free-form tags with each resource, without a central vocabulary. The term folksonomy – combining the words "folk" and "taxonomy" [13] – has been used to characterize the product which emerges from this tagging in a social environment.

In order to analyze, index, and classify their content, web systems compare tags attached to resources. Instead of considering the semantics of each tag in the comparison, tag-based systems usually rely on string matching approaches. While ontologies are increasingly adopted to enrich tag semantics, one common problem with the proposals to associate tags with formal ontologies concerns their unidirectionality, i.e., ontologies improve tag semantics, or the implicit/potential semantics of folksonomies is extracted to produce ontologies.

In a previous paper, we proposed a fusion approach, called *folksonomized ontology* (FO), which goes beyond this unidirectional perspective [1]. In one direction, the ontologies are "folksonomized", i.e., the latent semantics from the folksonomic tissue is extracted and fused with them. On the other direction, the knowledge systematically organized and formalized in ontologies gives structure to folksonomic semantics, enhancing operations involving tags, e.g., content indexing and discovery. The folksonomic data fused with an ontology will tune it up to contextualize inferences over the repository.

The scenario involving ontologies, folksonomies + derived ontologies (which we call social ontologies) and their relations lacks a preciser characterization, in order to answer open questions as: (i) What is the abstract model behind social ontologies derived from folksonomies? (ii) How this model is related to ontologies? (iii) From the model point of view, how the initiatives explore the relationship between ontologies and folksonomies?

This paper contributes by defining an abstract framework to describe ontologies, folksonomies, social ontologies, and approaches to combine them, including our folksonomized ontology. Through this framework, this paper aims to explicit the limitations of unidirectional relations between folksonomies and ontologies, as well as the demand for our fusion approach and its strengths in:

Tag disambiguation: by finding groups of related tags and mapping them to ontology concepts, the FO can be applied to disambiguate tags and find the ones that are more related, going beyond statistical analyses by using semantic similarity metrics.

Tag suggestion: the current folksonomy systems consider only co-occurrence information to suggest related tags to users; a FO has a richer set of semantic relations among concepts, supporting suggestion of tags that were not used together before – folksonomies cannot do that.

Semantic similarity: a FO can support the computation of semantic similarity between concepts and, by extension, between tags; so, they can expand the usual techniques that focus only at syntactical similarity and co-occurrence of tags, achieving better results in discovery operations.

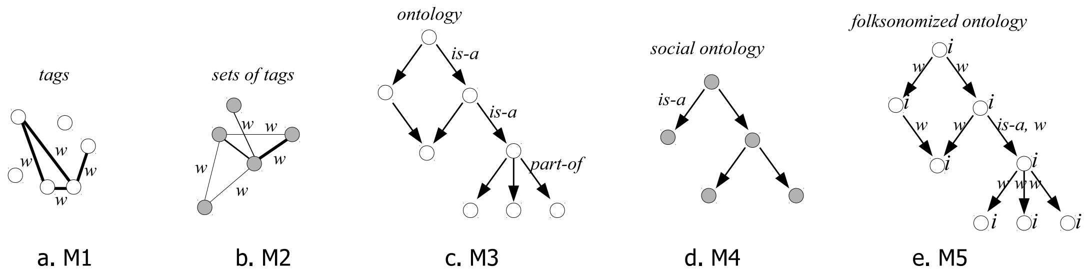

Figure 1: Models of our abstract framework.

Ontology evolvement: a FO can be used to find missing relations in ontologies; the high co-occurrence between two groups of tags, and their corresponding concepts, can indicate a necessary relation in the ontology.

The paper is organized as follows. In section 2 we introduce our abstract framework. In section 3 we compare related work through our abstract framework. In section 4 the folksonomized ontologies are defined through our abstract framework. In section 5 we show the conclusions.

2. ABSTRACT FRAMEWORK

In order to substantiate our analysis of related work and to explicit the main characteristics and the differential of our approach, we have defined an abstract framework composed by a set of models. Although there is related work which present individual abstract models presented here – e.g., folksonomies and ontologies – as far as we know, this paper contributes as the first initiative to embrace a wider scenario, including approaches aimed to relate folksonomies and ontologies.

In our framework we defined five abstract models $M1$ to $M5$ – illustrated in Figure 1 – aiming to model aspects of folksonomies and their relations with ontologies. There are three different classes of nodes in the models: N_t, N_{ts}, and N_c. Each node of the N_t class is a single tag. A node of the N_{ts} class, on the other hand, represents a set of tags grouped together because they share the same meaning – a *tagset*. The N_{ts} nodes are depicted in Figure 1 of the models in gray. Finally, each node of the N_c class is a concept of an ontology. The focus here is in the semantics assigned to each node instead of the label. The models are further detailed:

$M1$ (Figure 1.a) models co-occurrences in a folksonomy – i.e., the relations between tags that were used together – as a tuple (G_T, W_C), where $G_T = \langle T, E_T \rangle$ is an undirected graph with vertex set T formed by tags (members of the N_t class) and edge set E_T representing co-occurrences of tags. W_C is a weighting function $W_C : E_T \to \mathbb{N}$, producing a weight related to each edge, corresponding to the number of co-occurrences of the respective tags annotating the same object. $M1$ is a graph that represents the relatedness among tags. It is the raw material used by many proposals to synthesize ontologies. There are several approaches to define the relatedness [5]. They are mostly variations of co-occurrences of tags annotating resources.

$M2$ (Figure 1.b) models tagsets and their co-occurrences as a tuple (G_S, W_O), where $G_S = \langle S, E_S \rangle$ is an undirected graph with vertex set S formed by tagsets (members of the N_ts class) and edge set E_S representing co-occurrences of tagsets. W_O is a weighting function $W_O : E_S \to \mathbb{N}$, producing a weight related to each edge, corresponding to the number of co-occurrences of the respective tagsets annotating the same object. Since tags with the same meaning are grouped together in $M2$, it is nearer to the way ontologies organize concepts. Many proposals use this model to relate tagsets with concepts from ontologies [9, 11].

$M3$ (Figure 1.c) is a simplified model of an ontology, represented as a tuple (G_O, RT, F_{RT}), where $G_O = \langle C, E_R \rangle$ is a directed graph with vertex set C formed by concepts (members of the N_c class) and arc set E_R representing relations between concepts. RT is a set of relation types between concepts. F_{RT} is a function $F_{RT} : E_R \to RT$, which associates a type with each relation (arc).

$M4$ (Figure 1.d) models tagsets and their typed relations as a tuple (G_S, RT, F_{RT}), where $G_S = \langle S, E_R \rangle$ is a directed graph with vertex set S formed by tagsets (members of the N_ts class) and arc set E_R representing relations between tagsets. RT is a set of relation types between tagsets. F_{RT} is a function $F_{RT} : E_R \to RT$, which associates a type with each relation (arc). $M4$ models "social ontologies", similar to the Kotis et al. proposal [6] in the sense of the social aspect built-in. Compared to $M3$, $M4$ has tagsets (members of the N_{ts} class) in each vertex, rather than concepts. The term "social ontology" will be adopted here, contrasting with folksonomy, to emphasize this structure – mostly derived from a folksonomy – which makes explicit many relations among tags and whose structure resembles ontologies.

$M5$ (Figure 1.e) models our proposed folksonomized ontology; it is derived from $M3$, and incorporates semantic data from folksonomies. It is further detailed in Section 4.

3. RELATED WORK

In this section we discuss the related work, from the perspective of our abstract framework, using the models introduced in Section 2. The main purpose of this section is to show the way related work explore the relation between folksonomies and ontologies. The summary of our analysis is presented in Table 1. Column 1 relates the authors analyzed in this section; column 2 sumarizes the path of the models followed by each approach; column 3 indicates auxiliary resources and models used in the process; column 4 defines the role of the auxiliary resources and models mentioned in column 3. In the last row, we present our fusion approach of folksonomized ontologies. Our main argument here concerns the unidirectionality of the initiatives, i.e., they use folksonomies as the main raw material and ontologies as auxiliary.

All analyzed proposals depart from the model $M1$, since the co-occurrences of tags is a metric to express the latent

Table 1: Work Comparison.

Authors	Path	Auxiliary	Auxiliary Role
Cattuto et al. [4]	tags (M1) → ontology	M3 - Wordnet	Measure tag relatedness (M1)
Specia et al. [9]	tags (M1) → tagsets (M2) → ontology (M4)	M3, Google, Wikipedia	Explicit semantics (M1)
Tesconi et al. [11]	tags (M1) → tagsets (M2) → ontology (M4)	M3 - WordNet, YAGO, Wikipedia	Disambiguate tags (M1)
Damme et al. [12]	tags (M1) → tagsets (M2) → ontology (M4)	M3 - WordNet, Google, Wikipedia	Derive ontologies (M4)
Cantador et al. [3]	tags (M1) → tagsets (M2) → ontology (M4)	M3 - WordNet, Wikipedia	Explicit semantics (M1)
Bang et al. [2]	tags (M1) → tagsets (M2) / ontology (M4)		
Heymann et al. [5]	tags (M1') → ontology (M4)		
Limpens et al. [7]	tags (M1) → ontology (M4)		Propose and review tags (M1)
Alves et al. [1]	(tags/tagsets (M1/M2) ↔ ontology (M3)) → FO (M5)		

semantics of folksonomies. From these co-occurrences, Cattuto et al. [4] calculated several measures of tag relatedness by using an auxiliary ontology, the WordNet ($M3$). They do not group related tags in tagsets; each individual tag of $M1$ is associated with a synset in the WordNet ontology. Synsets are sets of synonyms that play an equivalent role of concepts in ontologies. The similarity of the related synsets are then transferred to the respective tags.

Specia et al. [9] proposed a technique to map clusters of tags to ontology concepts, in order to make explicit the semantics of the tag space. They departed from $M1$ creating clusters of high-related tags (tagsets) and relating them to produce $M2$, using co-occurrence information. The relations between these clusters were aligned with auxiliary external resources like Wikipedia, Google, and ontology bases – following the semantic standards ($M3$) – to produce $M4$. Those resources were used to improve the folksonomic data, mainly making explicit the semantics of the tags in the model $M1$.

In a similar way Tesconi et al. [11] used external resources, namely Wikipedia, and ontologies ($M3$) like WordNet and YAGO [10]. Their objective was disambiguate tags, "semantifying" them. They developed an algorithm to disambiguate tags, grouping them by sense, whose output is an entity like the model $M2$. Its tagsets are finally linked to Wikipedia categories and ontology concepts, producing $M4$.

Damme et al. [12] aimed to use folksonomy data ($M1$) to build and to maintain ontologies. They employ lexical resources, like Leo Dictionary, WordNet, Google, and Wikipedia, to enrich the results of a preprocessing step, in which the tagsets are prepared and cleaned, resulting in $M2$. Then they map tagsets of $M2$ to concepts of $M3$ (ontologies). The relations of $M3$ are mapped back to $M2$, in order to produce $M4$. Finally, the folksonomy's community validates the resulting $M4$.

Cantador et al. [3] proposed a mechanism to filter and classify tags, producing $M2$. Then, they mapped these tagsets of $M2$ to knowledge bases like WordNet and Wikipedia, to discover the corresponding semantic entities. Different from previous approaches, in order to map $M2$ to $M4$ they predefined a set of possible categories and relation types among tagsets. In order to do so, they used direct association or natural language processing heuristics.

Bang et al. [2] proposed the concept of "structurable tags", in which tags can be linked through relations, allowing basic inference operations. They expanded the model $M1$, allowing users to create two types of relations between tags: inclusion and synonymy. These types of relations support the transformation of folksonomic data into more semantic models. Thanks to the synonymy relation, the system transforms the data into the model $M2$, grouping the tags with the same meaning. On the other hand, the inclusion relation led to an hierarchical organization, as a simplified $M4$.

Heymann et al. [5] proposed an algorithm to build a graph $M4$ departing from a variation of $M1$, in which the edges are unweighted. It first aggregates tags in *tag vectors*, in which the $v_{t_l}[o_m]$ corresponds to the number of times that the tag t_l annotates the object o_m. In the resulting unweighted $M1$, the vertexes will be the tags, and there will be an edge for pair of tags whose relatedness is above a threshold. The resulting graph, without weights and maintaining just the relevant edges, contains a "latent hierarchical taxonomy". It is captured by an algorithm that builds a subsumption hierarchy, derived from the centrality of each node in the graph.

As can be observed in our synthesis of related work, all approaches follow almost the same path, producing social ontologies ($M4$) from data extracted from folksonomies. Ontologies appear as adjuncts, making the semantics of tags explicit and helping operations of tag disambiguation and similarity evaluation. Nevertheless, the rich structure of the ontology is not appropriated and the produced $M4$ social ontology is limited to those simple relations – usually subsumption relations – which can be inferred from tags. Our proposal, described in the next section, overcomes this limitation.

4. FOLKSONOMIZED ONTOLOGIES

In this section we describe our *folksonomized ontology* (FO). This section summarizes the main characteristics previously presented in [1] from a new point of view. It describes our FO from the perspective of the abstract framework, introduced in Section 2, and confronts it with related work following the same perspective.

A FO is defined as a tuple (G, RT, F), where $G = \langle V, E\rangle$ is a directed graph with vertex set V formed by ontology concepts (members of the N_c class) and arc set E representing relations between these concepts, and RT is a set of relation types between concepts. F is a set of functions, they are: $\mathcal{F}_1$ is a weighting function $\mathcal{F}_1 : E \to \mathbb{N}$ where the weight of the relation is derived from the total of co-occurrences between tags represented by the respective concepts, the function $\mathcal{F}_2 : E \to RT$ defines the type of the relation as in ontologies (see $M3$) – in its first version, presented here, all relations are subsumptions, but the model is extensible to other types of relations –, and the function $\mathcal{F}_3 : V \to \mathbb{R}$ associates the *information content* (ic) related to each concept, calculated by $ic(c) = -\log p(c)$, where $p(c)$ is the probability of a given concept c. This ic value also derives from a statistical analysis of the folksonomy and will substantiate computations of semantic similarity between the concepts using, for example, Resnik similarity [8].

Existing approaches to integrate folksonomies and ontologies are based on mapping tags or tagsets to ontology concepts. The relations among tags mapped to concepts can be derived from co-occurrence analysis of the tags, or from the relations that already exist in the ontology. As observed in the previous section, the final product of existing approaches is a "social ontology" $M4$. The concepts of this model are limited to those extracted from tags aligned to ontologies. Preexisting concepts from ontologies, not present in the folksonomies, will not be present in $M4$. The semantics of ontologies enriches the analysis of tags in a unidirectional way, i.e., statistical data from folksonomies are not used to improve the similarity analysis in the ontologies. The FO, on the other hand, preserves preexisting ontology nodes that cannot be mapped to tagsets. They are explored to do inferences which are not possible in related work. Moreover, the FO model $M5$ represents more than $M4$ relations among concepts, capturing weights of relations and probabilities, to support better inferences.

A FO is built on the assumption that the semantics of folksonomies can be also applied to refine the ontology itself. For this reason, in one direction FOs support suggestion of related tags that do not appear together in the folksonomy annotations; in inverse direction, other relevant aspect that emerges from the folksonomized approach is the possibility of verifying a relation that does not exist in ontology, but is strong in the folksonomy. This information can be used to evolve and improve ontologies.

5. CONCLUSION

This paper contributed in three important issues concerning FOs, firstly presented in [1]. It introduces an abstract framework to support the presentation of models addressing folksonomies, ontologies, social ontologies and their relationship. This framework substantiated a characterization and comparison of related work, including the FO. As far as we know, this is the first initiative to produce such a framework to compare the models adopted by the related work.

We also presented from the framework perspective the advantages of using the folksonomized ontologies compared to related work, due to its hybrid approach fusing folksonomies and ontologies. It is a symbiotic combination, taking advantage of both semantic organizations. Ontologies provide a formal semantic basis, which is contextualized by folksonomic data, improving operations over tags based on ontologies. Conversely, the folksonomized ontologies can also be used as tools to analyze the ontology quality and to help the process of ontology evolution, showing the discrepancies between the emergent knowledge of a community and the formal representation of this knowledge in the ontology.

We have implemented a practical experiment with 1,049,422 extracted from Delicious. Our prototype can build a FO, restricted to generalization relationships. Future work include: (i) to expand the folksonomized model to include other relations (besides the generalization); (ii) to run tests in specialized contexts applying domain ontologies; (iii) to improve our tool for ontology evaluation and review.

Acknowledgments

This work was partially financed by CNPq, FAPESP, INCT in Web Science (CNPq 557.128/2009-9) and Unicamp/PAPDIC.

6. REFERENCES

[1] H. Alves and A. Santanchè. Folksonomized ontologies - from social to formal. *XVII Brazilian Symposium on Multimedia and the Web*, 2011.

[2] B. H. K. Bang, E. Dané, and M. Grandbastien. Merging semantic and participative approaches for organising teachers' documents. In J. Luca and E. R. Weippl, editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008*, pages 4959–4966, Vienna, Austria, June 2008. AACE.

[3] I. Cantador, I. Konstas, and J. M. Jose. Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9:1–15, March 2011.

[4] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer.

[5] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Standford University, April 2006.

[6] K. Kotis, P. Alexopoulos, and A. Papasalouros. Towards a framework for trusting the automated learning of social ontologies. In Y. Bi and M.-A. Williams, editors, *Knowledge Science, Engineering and Management*, volume 6291 of *Lecture Notes in Computer Science*, pages 388–399. Springer Berlin / Heidelberg, 2010.

[7] F. Limpens, F. Gandon, and M. Buffa. Helping online communities to semantically enrich folksonomies. *Web Science Conference 2010*, 2010.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In C. R. Perrault, editor, *Proceedings of the 14th IJCAI*, pages 448–453, Montréal (Canada), 1995.

[9] L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *Proceedings of the European Semantic Web Conference (ESWC2007)*, volume 4519 of *LNCS*, pages 624–639, Berlin Heidelberg, Germany, July 2007. Springer-Verlag.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[11] M. Tesconi, F. Ronzano, A. Marchetti, and S. Minutoli. Semantify del.icio.us: Automatically turn your tags into senses. In *First Workshop on Social Data on the Web (SDoW2008)*, 2008.

[12] C. Van Damme, M. Hepp, and K. Siorpaes. 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, 2007.

[13] T. Vander Wal. Folksonomy. `http://vanderwal.net/folksonomy.html`, 2007. Retrieved on April, 2011.