Semantic Enrichment of Folksonomy Tagspaces

Sofia Angeletou

Knowledge Media Institute (KMi)
The Open University, Milton Keynes, United Kingdom
S.Angeletou@open.ac.uk

Abstract. The usability and the strong social dimension of the Web2.0 applications has encouraged users to create, annotate and share their content thus leading to a rich and content-intensive Web. Despite that, the Web2.0 content lacks the explicit semantics that would allow it to be used in large-scale intelligent applications. At the same time the advances in Semantic Web technologies imply a promising potential for intelligent applications capable to integrate distributed content and knowledge from various heterogeneous resources. We present FLOR a tool that performs semantic enrichment of folksonomy tagspaces by exploiting online ontologies, thesauri and other knowledge sources.

1 Background and Research Problem

The large-scale content annotation and metadata generation has been realised as Web2.0 applications have become very popular. Despite that, **Web2.0 lacks the explicit semantics** that would allow the content to be used in large-scale intelligent applications. At the same time the advances in Semantic Web technologies imply a promising potential for intelligent applications capable to integrate distributed content and knowledge from various heterogeneous resources. There is significant discussion that the combination of Semantic Web and Web2.0 will lead to an interoperable, intelligent Web ([4,8,10]). The goal of this work is to identify methods for the **automatic semantic enrichment** of Web2.0 generated content with a focus on folksonomies.

Folksonomies are Web2.0 systems whose basic elements are **users**, **resources** and **tags**. A resource is a content object depending on the folksonomy (a photo in Flickr[1], a bookmark in Del.icio.us[2], a video in YouTube[3] and so on). A tag can be any sequence of characters a user can attach to a resource. However the semantics of tags, and as a result the semantics of the resources, are not known and are not explicitly stated. This often hampers the resource retrieval within the individual system as well as the integration of resources in cross platform applications. The goal of this work is **to identify the meaning of the tags** attached to a

---

[1] http://www.Flickr.com
[2] http://del.icio.us
[3] http://www.youtube.com

resource, **to obtain the formal semantics that correspond to these tags** and **to attach the formal semantics to the resource**, automatically creating in that way a semantic layer on top of folksonomy tagspaces. The realisation of the above raises the following research questions.

– **How can folksonomies' tagspaces be semantically enriched automatically?** This research question can be further analysed into the following questions. How to discover automatically the meaning of tags based on their context? How can the Semantic Web be exploited for the semantic enrichment of the tags and what other resources are required in case the Semantic Web falls short of that task?

– **How can the enriched tagspaces be evaluated in terms of content retrieval** against the non enriched tagspaces? What performance measures should be established to measure content retrieval in folksonomies before and after the semantic enrichment?

In the following we describe the existing research on folksonomies and present our work.

2   Related Work

Folksonomy research has focused on comprehending the inherent characteristics of tagging and exploring the semantics that emerge from it. The early works on folksonomies explore the structure, the types of tags and the user incentives of folksonomies ([7] and [12]). There are also works (see [14] for a detailed analysis of the specific methods) based on the assumption that frequent co-occurrence of tags translates to a semantic association among them. These works use various statistical methods to identify clusters of related tags without defining the exact relations among them. An exception is the work detailed in [14], where, in addition to clustering the tags, the semantic relations among them are identified.

More recent research on folksonomies aligns them with knowledge resources such as WordNet and ontologies. For example, in [9] the authors describe a method that presents tag clusters as navigable hierarchical structures derived from WordNet. Using a combination of WordNet based metrics they identify the possible WordNet sense for each tag. They extract the path of this tag from the WordNet hierarchy and they integrate it into the hierarchical structure of the cluster. The TagPlus system [11] uses WordNet to disambiguate the senses of Flickr tags by performing a two step query. The system returns all the possible WordNet senses that define a tag and the user selects (disambiguates) which sense he wishes. Another work aligning folksonomies to a user selected ontology is described in [1]. The system queries the Web with a variety of linguistic patterns between the ontological concepts and the tags. Each tag is categorised under the concept to which it was more related by the Web Search results.

The existing works present methods for tag disambiguation and tag cluster enrichment. Our work aims to address the following additional issues. First, the existing works require some initialising from the user's side (e.g., a priori selecting ontology or knowledge resources for the relevant categories of tags) or they require user contribution to perform the disambiguation of the tags. Our goal is to perform semantic enrichment of folksonomies entirely *automatically* (i.e., without user contribution). Second, we aim to investigate how by using other knowledge resources (e.g., thesauri) and the Semantic Web we can achieve more precise and more complete enrichment of tags compared to the enrichment from single resources (i.e., one ontology, WordNet and so on).

3   FLOR

The goal is to transform a flat folksonomy tagspace into a rich semantic representation. We aim to annotate folksonomy resources with **Semantic Entities** (SEs) rather than raw text tags. However, since tags are the basic description of resources the connection of tags to SEs is the first step prior to connecting the resources to SEs. A SE can ideally be a Semantic Web Entity, SWE (class, relation, instance) defined in an online ontology. Our goal is not just to connect tags to SWEs but also to bring in other knowledge related to these SWEs. In case no SWE exists for a tag the goal is to query other knowledge resources for Semantic Entities.

We present **FLOR**[4], a **FoLksonomy Ontology enRichment** tool, which takes as input a set of tags (either the tagsets of individual resources or clusters derived by the statistical analysis of folksonomies) and automatically relates them to relevant semantic entities (classes, relations, instances) defined in online ontologies. The output of FLOR is a semantically enriched tagset. FLOR performs three basic steps as described in the following.
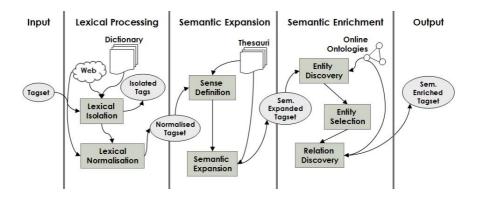

Fig. 1. FLOR Phases

---

[4] `http://flor.kmi.open.ac.uk/`

3.1   PHASE 1: Lexical Processing

Due to the freedom of tagging as a basic rule of folksonomies, a wide variety of different tag types are in use. Understanding the types of tags is the first step in deciding which of them are meaningful and should be taken into account as a basis of the semantic enrichment. Previous work ([7]) has identified different conceptual as well as syntactic categories of tags. For example, there are tags containing special characters, numbers, concatenated tags or tags with spaces and a big number of non-English[5] tags. The Lexical Processing phase is executed in two steps. The **Lexical Isolation** step uses a set of heuristics to identify which tags will not be further processed by FLOR. The **Lexical Normalisation** step aims to bridge the naming conventions used in folksonomies, ontologies and other knowledge resources by producing a list of possible lexical representations for each tag that will be enriched.

   Running Example: Consider the tagset {`buildings, corporation, bw, england, road, neil101`}. Phase 1 isolates the tags {`bw, neil101`} which can't be further processed by FLOR. The Lexical Normalisation step generates the following lexical representations for the tag: `buildings` : {building, buildings}. The rest of the tags are already in a normalised form.

3.2   PHASE 2: Sense Definition and Semantic Expansion

Due to polysemy, the same tag can have different meanings in different contexts. For example, the tag `jaguar` can describe either a car or an animal or an operating system depending on the context in which it appears. The first step of this phase **Sense Definition and Disambiguation** performs sense disambiguation for the tags. The technique described in [2] has been implemented using Word-Net based similarity metrics. Alternative strategies such as [15] and [5] are also considered. Another issue that this phase addresses is the Semantic Web sparseness. While online ontologies might not contain concepts that are syntactically equivalent to a given tag, they might contain concepts that are labeled with one of its synonyms. To overcome this limitation, we perform **Semantic Expansion** for each tag as described in [2]. A combination of thesauri and other knowledge sources is considered in order to achieve an optimal semantic expansion.

Running Example: The Sense Definition step maps a WordNet sense to each of the tags returned from the previous phase. The result in this case is: [`building`: *a structure that has a roof and walls and stands more or less permanently in one place*], [`corporation`: *a business firm whose articles of incorporation have been approved in some state*], [`road`: *an open way (generally public) for travel or transportation*], [`england`: *a division of the United Kingdom*]. Next the Semantic Expansion returns the synonyms and the hypernyms for each tag, i.e.: [`building`: SYNONYMS (edifice) - HYPERNYMS (structure, construction, artefact)], [`corporation`: SYNONYMS (corp) - HYPERNYMS (firm, business, concern)], [`road`: SYNONYMS (route), HYPERNYMS (way, artefact,

---

[5] FLOR deals only with English tags.

object)], [`england`: SYNONYMS (-), HYPERNYMS (European_Country, European_Nation, land)].

3.3 PHASE 3: Semantic Enrichment

The last phase of FLOR identifies the Semantic Entities that are relevant for each tag by leveraging the results of Phases 1 and 2. The relevant Semantic Web Entities are selected during the **Entity Discovery** step by querying the WATSON Semantic Web Gateway [6], which gives access to all online ontologies. Then in the **Entity Selection** step we filter the SWEs in order to identify the ones that correctly correspond to the tags. Finally the **Relation Discovery** step identifies relations between the SWEs using the SCARLET, semantic relation discovery algorithm [13].

Running Example: The Semantic Enrichment phase links the tags to ontological entities. For example the following semantic information has been attached to the tag: [`building`: subClassOf (*Infrastructure, Manmade_Structure, HumanShelterConstruction, SpaceInAHOC*) - superClassOf (*Restaurant, Rail-roadStation*)]. The same happens for the rest of tags.

The output of FLOR for a resource is a semantic layer, containing the definitions of the concepts described in the resource and their relations.

4 Current Status and Outlook

FLOR was designed on the basis of the results presented in [3] where the characteristics of folksonomies versus the characteristics of ontologies were identified. The first functional version of FLOR has been implemented and the results have been manually evaluated. Applying FLOR on a dataset of 250 photos from Flickr with a total of 2819 tags we obtained the results reported in [2]. FLOR enriched approximately the 49% of the tags with enrichment precision of 93%. The main conclusion from this experiment was that folksonomy tagspaces can be automatically enriched with formal semantics extracted from online ontologies and thesauri. Yet, the Relation Discovery (Step 3) of Phase 3 still needs to be implemented. Also the evaluation of FLOR in a large-scale experiment is part of the ongoing work. Additionally the manual evaluation of FLOR revealed the shortcomings of the FLOR algorithm and provided the basis for future work. This is broken down to the following tasks:

– Testing and evaluation of FLOR in a large-scale retrieval task (*M24*)
– Relation Discovery of Phase 3 (*M24*)
– Enhancement of Sense Discovery and Semantic Expansion of Phase 2 (*M30*)
– Testing and evaluation of the improved version of FLOR (*M32*)

Finally the expected contribution of this work is:

– Methodology for semantic enrichment of a set of keywords
– Concept based annotation and retrieval
– Folksonomy content retrieval evaluation strategy.

Acknowledgements

References

1. Abbasi, R., Staab, S., Cimiano, P.: Organizing resources on tagging systems using T-ORG. In: Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0, Innsbruck, Austria, pp. 97–110 (2007)
2. Angeletou, S., Sabou, M., Motta, E.: Semantically enriching folksonomies with FLOR. In: 5th European Semantic Web Conference, Tenerife, Spain (accepted in the Workshop of Collective Semantics, 2008)
3. Angeletou, S., Sabou, M., Specia, L., Motta, E.: Bridging the gap between folksonomies and the semantic web: An experience report. In: Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0, Innsbruck, Austria, pp. 30–43 (2007)
4. Benjamins, R., Davies, J., Baeza-Yates, R., Mika, P., Zaragoza, H., Greaves, M., Gomez-Perez, J., Contreras, J., Domingue, J., Fensel, D.: Near-term prospects for semantic technologies. Intelligent Systems 23, 76–88 (2008)
5. Cilibrasi, R., Vitanyi, P.: The google similarity distance. Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
6. d'Aquin, M., Sabou, M., Dzbor, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Motta, E.: Watson: A gateway for the semantic web. In: Poster Session of the 4th ESWC, Innsbruck, Austria (2007)
7. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
8. Greaves, M.: Semantic Web 2.0. Intelligent Systems 22(2), 94–96 (2007)
9. Laniado, D., Eynard, D., Colombetti, M.: Using WordNet to turn a folksonomy into a hierarchy of concepts. In: Proc.of 4th Italian Semantic Web Workshop, Bari, Italy, pp. 192–201 (2007)
10. Lassila, O., Hendler, J.: Embracing "Web 3.0". Internet Computing 11(3), 90–93 (2007)
11. Lee, S., Yong, H.: TagPlus: A retrieval system using synonym tag in folksonomy. In: Proc. of the Int. Conference on Multimedia and Ubiquitous Engineering, Seoul, Korea, pp. 294–298 (2007)
12. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Position paper, tagging, taxonomy, flickr, article, toread. In: Proc. of the 15th Int. World Wide Web Conference, Edinburgh, Scotland (2006)
13. Sabou, M., d'Aquin, M., Motta, E.: Exploring the semantic web as background knowledge for ontology matching. In: Journal of Data Semantics (accepted for publication, 2008)
14. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Proc. of the 4th ESWC, Innsbruck, Austria, pp. 624–639 (2007)
15. Trillo, R., Gracia, J., Espinoza, M., Mena, E.: Discovering the semantics of user keywords. Journal of Universal Computer Science 13(12), 1908–1935 (2007)