Repurposing Social Tagging Data for Extraction of Domain-level Concepts

Sandeep Purao[1], Veda C. Storey[2], Vijayan Sugumaran[3], Jordi Conesa[4],
Julià Minguillón[4], Joan Casas[4]

[1] College of Information Sciences & Technology, The Pennsylvania State University,
University Park State College, PA 16802
spurao@ist.psu.edu
[2] Department of Computer Information Systems, J. Mack Robinson College of Business
Georgia State University, Box 4015 Atlanta, GA 30302
vstorey@cis.gsu.edu
[3] School of Business Administration, Oakland University
Rochester, MI 48309
sugumara@oakland.edu
[4] Estudis d'Informatica i Multimedia, Universitat Oberta de Catalunya,
Rambla del Poblenou, 156, E-08018, Barcelona, Spain
jconesac@uoc.edu, jminguillona@uoc.edu, jcasasrom@uoc.edu

Abstract. The World Wide Web, the world's largest resource for information, has evolved from organizing information using controlled, top-down taxonomies to a bottom up approach that emphasizes assigning meaning to data via mechanisms such as the Social Web (Web 2.0). Tagging adds meta-data, (weak semantics) to the content available on the web. This research investigates the potential for repurposing this layer of meta-data. We propose a multi-phase approach that exploits user-defined tags to identify and extract domain-level concepts. We operationalize this approach and assess its feasibility by application to a publicly available tag repository. The paper describes insights gained from implementing and applying the heuristics contained in the approach, as well as challenges and implications of repurposing tags for extraction of domain-level concepts.

1 Introduction

The World Wide Web, the world's most valuable information resource, has continued to evolve over time. It is no longer just a place to store information. Increasingly, it presents a platform for web citizens to communicate, collaborate and share content. This combination of information storing and collaboration is changing how we work, as well as how we carry out specific tasks such as information seeking [1]. A significant contributor to this change is the contribution of tags, or weak meta-data, by users. The result is a layer of social tagging data that contains significant potential.

This research explores how to take advantage of this potential by repurposing social tagging data for extraction of domain-level concepts.

The objective of this research is to develop, implement and evaluate an approach to extract concepts that are implicitly represented by tags contributed by web citizens. The weak semantics in this meta-data layer, must be cleaned, structured, and aggregated before extracting and identifying domain-level concepts. This paper develops a set of heuristics and procedures, applies them to tags extracted from a publicly available tagging site, and analyzes the results. The contribution of this research is to demonstrate the feasibility of extracting domain-level concepts from tags. It also shows that it is possible to repurpose content generated by the social web to contribute semantics to information contained in the World Wide Web.

2 Related Work

Tagging is the act of attaching a label that captures an interpretation of the underlying content. It can be a label that evokes what the content is about (e.g. green products), a label that captures an impression or action from the user (e.g. important, to read), or a combination of the two (e.g. important green products). Most tags fall into the first category, and evoke the underlying content [2]. As a result, they provide a short-hand for identifying key aspects of, or elements contained in, the domain. Each tag captures weak semantics about the underlying resources and contains the potential (with aggregation), to identify and extract important concepts in a domain [3].

The collection of tags may be viewed as ‘wisdom of the crowd.’ It reflects the independent and diverse opinions of a group of individuals [4]. The tags in this folksonomy [5] are often related to each other because they may be contributed by a user to tag related resources, or because they are contributed by different users to tag the same resource, or because different users may tag the same resource many times [6, 7]. The patterns that emerge from these tagging practices provide, not only a way to organize information for the users [8], but also a layer of weak semantics [9] that can be aggregated and linked to identify and extract domain-level concepts.

3 Identifying Domain-level Constructs

The approach consists of multiple phases, outlined in Figure 1. This paper focuses on the first two phases: data cleansing and identification of concepts and connectedness. We describe the operationalization of heuristics and the results from investigating the potential feasibility of the approach.

The present work extends our prior work [10] by adding data cleansing and improving the identification of concepts connectedness. Further information about the heuristics that underlie these phases is available in [10].

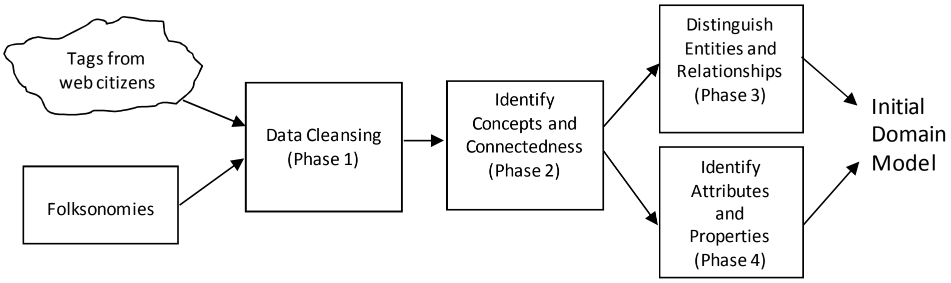
Tags from web citizens
Folksonomies
Data Cleansing (Phase 1)
Identify Concepts and Connectedness (Phase 2)
Distinguish Entities and Relationships (Phase 3)
Identify Attributes and Properties (Phase 4)
Initial Domain Model

Fig. 1. A Multi-Phase Approach to Identify Domain-level Constructs

3.1 Phase 1: Data Cleansing

Before invoking any heuristics, the approach requires that the data obtained from a tagging web site be cleansed. This first phase, data cleansing, involves detecting and correcting corrupt or inaccurate records [11, 12] obtained from the social tagging web site. Often used within the context of data mining, the term 'data cleansing' refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data in order to correct these errors by replacing, modifying or deleting these pieces of data.

The data cleansing phase is important because most tagging websites do not constrain nor validate the tags contributed by users. The result is a considerable amount of incorrect, inaccurate and sometimes, duplicated tags. For example, the use of punctuation in tags is common and often employed to compensate for the inability of delicious.com to deal with multi-word keywords. Tags such as *coolwebsites*, *cool_web*, *cool-site*, *coolsite* are, therefore, used. In addition, several tags represent variations of the same word (e.g., *bank* and *banking*) or different tags for the same concept, but in different languages (e.g. the concept *environment* represented by the tags *meio_ambiente, meioambiente, medio_ambiente, medi_ambient*).

The data cleansing phase, therefore, uses multiple techniques. Some of the operations performed in this phase for cleansing include:

- combining singular and plural forms of the tags (e.g. *auctions* and *auction*).
- converting all tags to lowercase (e.g. *static* and *Static*).
- replacing special characters, such as [, ; , .] (e.g. *web2.0* and *web2_0*).
- converting different conjugations of verbs to their infinitive (e.g. *awaiting* or *awaits* to *await*).
- expanding abbreviations (e.g. *sw* to *SouthWest*).
- dealing with synonyms (*achieved* to *attain*).
- deleting bookmarks that do not contribute any tags.
- deleting tags from users who only tagged few resources: e.g. discarding tags from users who tagged less than *x* resources, with *x* as a configurable variable.
- applying transformations manually created for each domain, which may include common mis-spellings and abbreviations (e.g. expand *diy* for "*do it yourself*").

Examples of these transformations include e-commerce -> Ecommerce (removing the hyphen), Auktionen -> auction (mapping words in different languages), Banking -

> bank (removing the gerund), Financial -> finance (removing ‘ial,’ an adjective-forming suffix), supervise -> oversee (mapping words with similar meaning), fig->figure (by expanding abbreviations), ocw->open course ware (by expanding manually detected abbreviations) among others. Although the actual process is automated and can succeed in reducing the number of tags, much work from natural language processing can be applied, such as the use of *soundex* or similar algorithms to detect tags wrongly written (e.g. *auction* vs *aucktion*) or using bigrams to identify automatically the language of tags. Figure 2 outlines the outcomes that originated with a query on corporate sustainability.

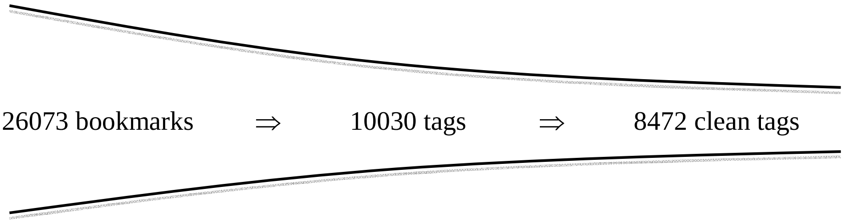

Fig. 2. Results based on retrieval and cleansing of tags for Corporate Sustainability

The figure shows that the query retrieved 26,073 bookmarks. For each bookmark we collected the bookmarked resource, the user who made the bookmark, and all the tags the user used to tag the bookmark. This produced 10,030 tags. The data cleansing trimmed these to 8,472 tags, reducing the number by over 15%. This is relevant because the cleansing allows higher quality results.

3.2 Phase 2: Identifying Concepts and Connectedness

This phase detects the most relevant concepts of the domain and their connectedness. It is supported by five heuristics to assess the probability that a Tag is a *Candidate* domain-level construct and to identify pairwise *Candidate* connections. The heuristics progressively increase the probability that a tag represents a domain-level construct.

Heuristic 1: Number of Users using a tag. This heuristic leverages the number of users who use a tag. It follows the rationale that the larger the number of users who use a tag, the greater the agreement among them that the tag represents a meta-level concept.

Heuristic 2: Number of Resources tagged with a Tag. This heuristic suggests that the number of resources tagged with a given tag indicates the importance of that tag for the domain. It increases the probability that the candidate tag is a domain-level construct.

Heuristic 3: Frequency of Tag Use. This heuristic examines the frequency of tags and uses the rationale that the greater the frequency, the more important the tag. This heuristic examines only resources generated from the previous heuristic, effectively bounding the search, and results in a scoping decision. The heuristics, in effect, move from users to tags to resources, in each step, increasing the confidence in the assessment that a Tag is a domain-level construct.

Heuristic 4: Centrality of Tag. This heuristic leverages the centrality of tags interpreting their size and position in the Tag Cloud. The centrality is an indicator of the importance of the tag to further augment the probability that a candidate tag is a domain-level construct. This paper does not operationalize this heuristic.

Heuristic 5: Connectedness of Tags. The connectedness of the tags is assessed by measuring co-occurrence frequencies. Co-occurrence may indicate that tags are synonyms, or they represent related constructs that should eventually be bridged with a relationship, or they represent related constructs (e.g. one is an entity, the other is an attribute). This heuristic only considers candidate tags; that is, tags identified by previous heuristics as candidates for being domain-level constructs. The connectedness of tags has two facets:

- 5A: *Occurrence*: two tags are connected when they are used to tag the same resource. For example, the tags *Travel* and *Flights*, identified after executing the query “Air Travel”, co-occur 31 times in the first 100 results returned.
- 5B: *Tendency*: two tags are connected when a significant amount of users tend to use the two tags to tag the same resource. In order to calculate this heuristic we used Principal Component Analysis (PCA) [13]. The PCA is conducted with the following parameters: using maximum likelihood as the method for extraction of the components, a Varimax rotation and only fields with weight >= 0.3 are taken into consideration.

The PCA analysis identifies a set of factors that conceptually can be seen as clusters whose tags represent a point of view from which the domain of discourse can be seen. The tags of each cluster may be semantically connected because they deal with the same semantic concept or sub-domain. Therefore, the tags of a given cluster may be used to infer tags potentially relevant for the user. (If a user decides that one of the tags of a cluster is relevant then all its tags are potentially relevant). For example, after applying the PCA analysis to the results obtained from the query “online auction” some of the obtained factors (or clusters) are: 1) craft, handmade, art, design, gifts and diy (related to crafting stuff), 2) daily, technology, gadgets, woot, electronics (related to technology stuff) and 3) paypal, money, finance and bank (related to the payment).

The result of these heuristics is a set of tags that are likely candidates for domain-level constructs, and potential connections among these constructs.

4 An Application

The approach has been applied to tags obtained from del.icio.us. A user can input a term or a phrase for a domain to retrieve documents and tags. The prototype carries out the procedures and applies the heuristics in a semi-automated manner. The result, concepts and connectedness among concepts, is visually represented as local graphs, as shown in Figure 4. At this time, a significant component of the experiments is the PCA analysis that emphasizes the connectedness of tags. For the experiment reported, the PCA has been conducted with the following parameters: using maximum likelihood as the method for extraction of the components; a Varimax rotation and only fields with weight >= 0.3 are considered. Because the PCA is conducted at the last step, the number of tags is relatively low. For example, in the cases described in this paper, the PCA analysis involved approximately 100 tags each.

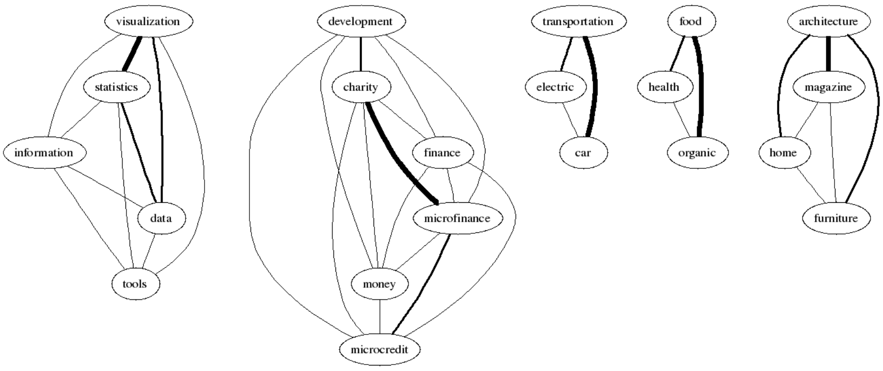

Fig. 3. Local graphs for the term Corporate Sustainability (Partial) obtained from PCA analysis

From the PCA analysis factors (aka components) are obtained. Each is composed of a set of tags. Although PCA is not a clustering technique, the set of tags in each component can be considered as a cluster, whose tags represent the annotation tendencies. For example, in the case of corporate sustainability, we can interpret one of the factors as an indication that a significant number of people tend to use the tags *movie*, *film* and *documentary* together. The tags within each cluster may, therefore, be considered semantically connected (see Figure 3). Here, the thickness of the line indicates the correlation between the connected tags. Table 1 shows the first factors (clusters) obtained from the PCA analysis for the “Corporate Sustainability” query.

An examination of the example of “Corporate Sustainability” shows that each component describes a different domain. It also shows that the first component is not related to the domain. It comes from the bookmarks of one website that is tangential to the domain. The analysis shows that a first-occurrence procedure can bias the results. The fifth component shows a similar bias. It would be useful to detect such tangential sites during the data cleansing process. In that case, the system could ask the users whether to take into account such sites.

Table 1. Tags contained in the main factors /clusters for “corporate sustainability”.

Corporate Sustainability	
1st	visualization + statistics + information + data + tools
2nd	Development + charity + finance + microfinance + money + microcredit
3rd	Transportation + electric + car
4th	Food + health + organic
5th	Architecture + magazine + home +furniture
6th	trends + future + strategy + consulting
7th	Inspiration + flash + portfolio + webdesign
8th	architecture + collaboration + opensource
9th	Movie + film + documentary

Regardless of the insights gained from this experimentation, the local graphs generated for each component were then merged. The merged graph shows several interesting sub-domains that would be difficult to predict based on a layman’s understanding of the terms ‘corporate sustainability.’ Even expert understanding of the term ‘corporate sustainability’ is unlikely to discover the varied sub-domains that this overall graph represents. Consider, for example, the sub-domains that are anchored towards micro-finance (second local graph of the figure) and the architecture (towards the right of the figure). The approach was also applied to other domains with the following outcomes: Online Auctions (102 domain-level concepts from 4,635 resources); Air Travel (150 concepts from 11,760 resources); and Emergency response (225 concepts from 1,664 resources). The implications are discussed further in the next section.

5 Conclusions

This research describes and validates an approach to repurpose tags from social tagging sites to extract domain-level concepts. The process and experiments reported demonstrate the feasibility of the approach. The experiments also point to interesting insights in two areas: domain restrictions as a way to advance the realization of results; and the possibility of additional heuristics. The key contribution of this research is the operationalization of the approach that may be broadly described as “designing with a crowd” to identify domain-level concepts for conceptual modeling and their connectedness. In particular, the PCA analysis provided useful insights to help in the identification of concepts in a domain. The results show that the selection (or rejection) of tags cannot be automatic; rather, it requires some feedback from a user. The results also highlight challenges such as the generic nature and ambiguity of tags that necessitate cleansing, and the need to aggregate tags that represent the same underlying concept. In addition, the components generated from the PCA analysis (such as *microfinance* and *architecture*) point to the possibility that the proposed approach has the potential to greatly exceed the limitations of naïve or even expert perspectives on domain models by accounting for a much wider array of concepts that can be part of the representation of a domain.

Future research will focus on additional experiments, improving the data cleansing process to increase the efficacy of discarding irrelevant tags and merging similar ones, developing techniques to extract more insights from the tags, and using fuzzy variables. Research is also needed to refine the heuristics, implement the remaining phases, and conduct empirical analyses.

Acknowledgments

This work has been supported by Sogang Business School's World Class University Program (R31-20002) funded by Korea Research Foundation, the Georgia State University, the IN3 institute and the Spanish Ministry MICINN (TIN2008-00444).

References

1. J. Hendler and J. Golbeck, "Metcalfe's Law Applies to Web 2.0 and the Semantic Web," *Journal of Web Semantics*, vol. 6, 2008, pp. 14-20.
2. G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, 1987, pp. 964-971.
3. G. Weikum, G. Kasneci, M. Ramanath, and F. Suchanek, "Database and information-retrieval methods for knowledge discovery," *Communications of the ACM*, vol. 52, 2009, pp. 56-64.
4. C. Sunstein, *Infotopia: How Many Minds Produce Knowledge*, Oxford University Press, 2006.
5. S. Angeletou, M. Sabou, L. Specia, and E. Motta, "Bridging the gap between folksonomies and the semantic web: An experience report," *Workshop: Bridging the Gap between Semantic Web and Web 2.0 at 4th ESWC.* Innsbruck, 2007.
6. X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 417-426.
7. D.G. Campbell, "A phenomenological framework for the relationship between the semantic web and user-centered tagging systems," *Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop, Austin, TX, USA*, 2006.
8. M.E.I. Kipp and D.G. Campbell, "Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices," *Proceedings of the ASIST*, Austin, TX, USA, 2006, pp. 1-18.
9. J. Dye, "Folksonomy: A game of high-tech (and high-stakes) tag," *EContent*, vol. 29, 2006, pp. 38-43.
10. V. Sugumaran, S. Purao, V. Storey, and J. Conesa, "On-Demand Extraction of Domain Concepts and Relationships from Social Tagging Websites," *Proceedings of the NLDB*, 2010, pp. 224-232.
11. J.I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis," *Proceedings of the Conference on Information Quality*, 2000, pp. 200-209.
12. K.W. Tan, H. Han, and R. Elmasri, "Web data cleansing and preparation for ontology extraction using WordNet," *Proceedings of the WISE*, IEEE, 2000, pp. 11-18.
13. C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings of the twenty-first international conference on Machine learning*, 2004.