FaceTag: Integrating Bottom-up and Top-down Classification in a Social Tagging System

Quintarelli, E. - Resmini, A. - Rosati, L.

***Abstract* – Facetag is a working prototype of a semantic collaborative tagging tool conceived for bookmarking information architecture resources. It aims to show how the widespread homogeneous and flat keywords' space of tags can be effectively mixed with a richer faceted classification scheme to improve the "information scent" and "berrypicking" capabilities of the system. The additional semantic structure is aggregated both implicitly observing user behaviour and explicitly introducing a compelling user experience to facilitate the creation of relationships between tags directly by end-users. Facetag current implementation is written in PHP / SQL and includes an open API which allows querying and integration from other applications.**

***Index Terms* – Social classification, folksonomy, tagging, faceted classification, information architecture.**

I. INTRODUCTION [*]

Collaborative tagging systems have been largely adopted by end-users as useful and powerful tools to organize, browse and publicly share personal collections of resources on the World Wide Web through the introduction of simple metadata.

The aggregation of user metadata is often referred to as a folksonomy, a user-generated classification, emerging through bottom-up consensus while users assign free form keywords to online resources for personal or social benefit. Del.icio.us <http://del.icio.us/>, Flickr <http://www.flickr.com/>, 43things <http://www.43things.com/>, Furl <http://www.furl.net/> and Technorati <http://www.technorati.com/> are web-based collaborative systems for building shared databases of items, enriched by a flat metadata vocabulary that can be used to perform metadata-driven queries, to monitor change in areas of interest or to discover emergences or trends, such as the hottest / most popular topics in the system [Quintarelli 2005].

In the past, folksonomies have often been seen as orthogonal to taxonomies and controlled vocabularies: the latter rigid, hierarchical and organically hand-crafted by professionals a priori; the former flat, inclusive and emerging from bottom-up users' consensus [Quintarelli 2005]. In a flat tagging system each document can be retrieved through a simple set of keywords, collaboratively introduced by users to describe and categorize the document, very much like in a keyword-based search process in which descriptive terms can be used to get a set of applicable items.

Despite their low cognitive cost, their capability of matching users' real needs and language and their great value in a serendipity research task, folksonomies imply however a lack of precision, a very low findability quotient (especially in a known-item approach) and a limited scalability for the intrinsic variability of language [Quintarelli 2005].

As a result of the inherently inconsistent, evolving and much variable process of associating words and meanings, tagging systems are also implicitly plagued by a number of issues which include polysemy, homonymy, plurals, synonymy, problems of ego-oriented nature and basic level variation which do not appear easy to solve [Golder & Huberman 2005]. Any of these problems can dramatically reduce the effectiveness of the application, mining the benefits brought on by the use of tagging systems.

In addition, tags have recently started to be used by bloggers as reading-aids to help users identify articles and posts of interest, providing as such a complimentary structure over a purely chronological list of text pieces. This approach marks a major shift, in that tagging also becomes a tool to maximize findability and browsability without limiting the reader to only access the most popular or recent tags as in common tag clouds [Feinstein & Smadja 2006].

Tag clouds are widely used visual interfaces for information retrieval that provide a global contextual view of tags assigned to resources in the system. In such a structure, the most popular tags are usually displayed through an alphabetically ordered list with the font size increasing with the tag's relevance. Users browse the cloud, scanning hyperlinks to recognize information of interest [Hassan-Montero & Herrero-Solana 2006].

Flat tag clouds are anyway not sufficient to provide a semantic, rich and multidimensional browsing experience over large tagging spaces:

- Choosing tags by frequency of use inevitably causes a high semantic density with very few well-known and stable topics dominating the scene (as seen on RawSugar, <http://www.rawsugar.com/>);
- Providing only an alphabetical criterion to sort tags heavily limits the ability to quickly navigate, scan and extract, and hence build a coherent mental model out of tags;
- A flat tag cloud cannot visually support semantic relationships between tags. We suggest that these

[*] This paper is the result of a collaborative effort. Nonetheless, Emanuele Quintarelli specifically wrote paragraphs *I-II,* Andrea Resmini wrote paragraphs *V-VI* and Luca Rosati paragraphs *III-IV*.

relationships are needed to improve the user experience and general usefulness of the system;
- Current tag clouds often miss to provide complex logical operation over tags. Simply clicking on a tag is not enough to enable a smooth and powerful exploration or refinement.

Even if Facetag doesn't promise to address all of these issues, we believe our approach can limit the impact of polysemy, homonymy and basic level variation while introducing an innovative, multidimensional and more semantic paradigm for organizing, navigating and searching large information spaces through tags.

To reach this goal, FaceTag mixes three contributions to social tagging systems:
- The use of (optional) tag hierarchies. Users have the possibility to organize their resources by means of father-son relationships**;**
- Tag hierarchies are semantically assigned to editorially established facets that can be later leveraged on to flexibly navigate the resource domain;
- Tagging and searching can be mixed to maximize findability, browsability and user-discovery.

II. OVERVIEW OF FACETAG

Until today, one of the main limitations of hierarchical faceted categories was the lack of a good automated process for both creating the categories and associating items to the hierarchy of labels under each facet [Hearst 2006a].

We decided to avoid the issue entirely and use no algorithmic round-ups: Facetag is built around the notion that the users provide the structure and especially aims to investigate how a hierarchical and faceted metadata structure can be added to user generated content making use of tags provided by end users in collaborative systems, limiting the amount of effort and toil required through a careful user interface design.

III. FACETED ANALYSIS: THE FACETED SCHEME CONSTRUCTION

Although facet*,* faceted have become very common terms in the information architecture field, their application falls often far from its original meaning. The attribute *faceted*, indeed, is used in a large variety of meanings, and is often referred loosely to the availability of means to search by different keys [La Barre 2004]. The full theory of faceted classification, as it has been developed by Ranganathan and the Classification Research Group (CRG) and which includes rules for citation order and notation, is less widespread as a backend for website organization; remarkable exceptions are offered by projects staffing librarians, such as FATKS [Slavic 2002].

So, we thought to apply faceted classification to the IA field itself respecting in full the original library theory, in order to leverage on its potentialities and obtain maximum benefits. In such perspective, our design was inspired by these projects: Flamenco project <http://flamenco.berkeley.edu/>; Facetious <http://demo.siderean.com/facetious/facetious.jsp>; Etsy <http://www.etsy.com> [1].

The choice of facets is based on the CRG theory [Vickery 1960]. Indeed, an aspect often underestimated on the World Wide Web is that both Ranganathan and the CRG described a generic schema for faceted classification, which every actual schema can refer to. Thus, in a faceted classification project one does not have to rebuild the schema from scratch every time, but may follow a constant guideline while building one's main categories (i.e. facets). CRG postulates 11-13 general categories. In the table below we show the matching between CRG standard categories and IA-related categories that were used to define our facets.

TABLE 1: FACETAG FACETS DEFINITION BY CRG STANDARD CATEGORIES.

CRG	**FaceTag**
Thing	[Documents, resources]
Type	Resource Types (e.g. online report, case study...)
Part	--
Property	Language
Material	[Format]
Process	--
Operation	Activities/Subjects (e.g. competitive analysis, faceted classification ...)
Product	[Deliverables]
Byproduct	--
Patient	Usage (e.g. Industry, Health ...)
Agent	People
Space	[Country]
Time	Date

A preliminary analysis of a corpus of IA resources from the Information Architecture Institute Library <http://iainstitute.org/library/> allowed us to define six facets which appeared to be suitable for the classification of IA resources.

1 Both Facetious and Etsy mix proper facets and metadata (formal proprieties of an item).

TABLE 2: FACETAG FACETS AND EXAMPLES OF FOCI

Facet	Examples
Resource Types	white paper, case study etc.
Language	*predefined values (based on ISO Standard ISO 639-2)*
Activities/Subjects	discovery>competitive analysis, classification>facets
Usage	industry, public administration, health etc.
People	dion hinchcliffe, morville
Date	*automatically added by the software*

The foci listed near some of the facets serve the only purpose of making the facets self-explanatory. In the actual implementation, since tags are our foci, foci will be user-generated, with the only exception of the language facet, which will use a predefined list of languages in the ISO 639-2 notation, and the date facet, which will receive a software-generated timestamp upon resource creation.

IV. BERRYPICKING, INFORMATION SCENT AND THE TWO AXIS OF INFORMATION ARCHITECTURE

As a matter of fact, facets constitute an adaptive classification system capable, in force of its own nature, to represent:

- in movement knowledge, like that observable in a social collaborative context;
- several mental models at the same time, such as those playing their role in this context.

Furthermore, facets are particularly suitable to classify a homogeneous collection of items – i.e. a set of resources belonging to a specific disciplinary area.

Besides enforcing order on the flat space of keywords, the blend of tags and facets is able to empower the "information scent" [Chi et al. 2001] and the "berrypicking" [Bates 1989] capabilities of the system. Every information architecture project refers to two different information axes:

- a vertical (or paradigmatic) axis, i.e. the hierarchical relationship that each item of a system engages with the others;
- a horizontal (or syntagmatic) axis, i.e. the semantic, contiguity relationship that each item engages with the others.

In our case, the combination of tags and facets allows for better management of both these axes:

- from the vertical or paradigmatic point of view, when a user is going to associate a keyword to a facet (in order to tag a resource), the system suggests similar tags or hierarchy of tags pertaining to the same facet;
- from the horizontal or syntagmatic point of view, at the same time, the system will allow the user to see all the other tags belonging to the same facet(s).

V. FACETED HIERARCHICAL TAGGING

Facetag deals with users, resources, tags and facets in two quite distinct ways: since it's a social tagging application, it offers both a browsing/searching mode and an administrative/editing mode. These are two different activities, to which the user interface adapts providing different aiding tools (navigation, resource management) and different behaviours (zooming, tag suggestions) respectively.

When a user accesses the application first, Facetag replies in browsing mode and she is presented a page which lists the most recent additions to the system in the main body. Other relevant parts of the user interface are a search box and a sidebar. The sidebar lists facets and pertaining first-level tags with query previews, i.e the number of resourced associated to each tag automatically generated from the schema and data stored in the database.

Inside Facetag, a user can decide to look for content a) by entering keywords b) by choosing first-level tags from a specific facet list.

If the user enters a keyword, Facetag returns the paginated results set of all the resources which either contain that keyword in their tags or in their title, description or notes. The sidebar facet display is adjusted to show only those facets and pertaining first-level tags which are related to the results set.

In case the keyword happens to be an nth-level tag, the corresponding facet will show all nth+1 tags and add any broader tag in the hierarchy up to the nth-1 tag to the facet title as clickable items which allow zooming out. If there is no nth+1 tag, the facet is not displayed.

If the user clicks on a tag from the facet sidebar, Facetag returns the paginated results set of all the resources which have been tagged with that tag. A breadcrumb path is displayed which lists the active facet (the one the tag is a focus for) and the position of the tag in any tag hierarchy it may belong to.

The sidebar facet display is adjusted consequently. The active facet shows all broader tags from the hierarchy the selected tag may be part of alongside the facet title, and all pertaining narrower tags. Inactive facets show first-level tags which relate to the resources pertaining to the results set.

Upon subsequent zooming in and refining the query, when there are no narrower tags, the breadcrumb display is maintained to allow zooming out or what we call *disengaging*, resetting the search, while the active facet display is effectively removed from the sidebar.

Obviously, a user may start searching for a keyword and then adjust her results set using facets, combining the two approaches in any way she prefers until she reaches a satisfactory answer, or proceed viceversa and zoom in and out by using tags. Similarly, tags pertaining to different facets can be used together during a single search to narrow down a results set quickly and efficiently. If there is no disengagement, all subsequent operations are performed on the intermediate results set.

If a user logs in, access to the administrative interface is granted and adding, editing and deleting resources and tags becomes possible.

Upon entering new resources, a user is provided with a

simple form with entry fields for every facet. These tag fields are optional, and can be left empty at will: there is no mandatory facet. But if a user start to enter a tag, the completion tool suggests similar tags from the pertaining facet only. Moreover, since users can optionally identify two or more tags as a hierarchy through a simple syntax (using the '>' character), the completion tool can suggest, again facet per facet, not just similar tags, but similar tags as parts of a hierarchy [2] of tags, hence effectively suggesting an entire hierarchy.

Gradually, with use, these hierarchies acquire complexity and become globally significant in the system.

Editing or modifying can be done seamlessly from the browsing interface, by clicking icons which appear next to one's own resources. Noticeably, the same happens if a user tries to add a resource she already added (based on URI identification): Facetag simply supplies the editing interface preloading the original data.

VI. CONCLUSION

By providing the user with facets to which hierarchical sets of tags relate and pertain and a usable interface which adapts to the ongoing query, Facetag may solve, through contextualization and user-added semantic value, most of the basic issues connected with polysemy, homonymy and base level variations.

While further testing and usability studies are needed to verify to which extent users are motivated to use our prototype and to introduce structure in addition to flat tags, preliminary user evaluations show how the addition of hierarchies and facets can improve and disambiguate the meaning of tags giving them a stronger context and a more coherent organization. For example, by navigating a hierarchy users can make better sense of the meaning of a tag, discover related tags at different levels of specificity and exclude homonimies or find out a large number of other tags that can be of interest. This approach also tends to augment the scalability of the system when addressing the enormous domains presented today by the most appreciated social applications.

Improving on current features, Facetag aims to provide an advanced tagging experience through other innovative tools or widgets, like a Firefox plugin to seamlessly add new bookmarks while browsing, a WYSIWYG editor to offer drag and drop inclusion of texts and pictures from the web page the user is bookmarking, and a history of all the times a bookmark has been tagged.

Future works include testing the application on a real user base and verifying the outcomes, both in terms of internal logic and usability tests to widely prove the benefits of a semantic tagging application.

2 Note that hierarchies are not taxonomies but simply forests of shallow trees.

VII. REFERENCES

Bar-Ilan J., Shoham S., Idan A., Miller Y., Shachak A., (2006) *Structured vs. unstructured tagging – A case study*, WWW20006, Edimburg <http://www.rawsugar.com/www2006/12.pdf>.

Broughton, V. (2001) *Klasifikacija za 21. stoljece: nacela i struktura Blissove bibliografske klasifikacije [= A classification for the 21st century: principles and structure of the Bliss bibliographic classification]*, Vjesnik bibliotekara Hrvatske, 44, 1-4, p. 38-51; trad. it. Una classificazione per il 21' secolo: principî e struttura della Classificazione bibliografica Bliss, AIB-WEB. Contributi, <http://www.aib.it/aib/contr/broughton1.htm>.

Campbell, G.D., Fast, K.V., (2006) *From Pace Layering to Resilience Theory: The Complex Implications of Tagging from Information Architecture*, Proceedings of IA Summit 2006 (Vancouver, March 23-27, 2006), ASIS&T <http://www.iasummit.org/2006/files/164_Presentation_Desc.pdf>.

Chi, E.H. - Pirolli, P. , Chen, K. – Pitkow, J. (2001) *Using Information Scent to Model User Information Needs and Actions on the Web*, Proceedings of the SIGCHI conference on Human factors in computing systems (Seattle, Washington, 2001), ACM Press <http://www2.parc.com/istl/projects/uir/publications/items/UIR-2001-07-Chi-CHI2001-InfoScentModel.pdf>.

English, J., Hearst, M., Sinha, R., Swearingen K., and Yee, P., (2002a) *Hierarchical Faceted Metadata in Site Search Interfaces,* CHI 2002 Conference Companion <http://flamenco.berkeley.edu/papers/chi02_short_paper.pdf>.

-- (2002b) *Flexible search and browsing using faceted metadata*, Unpublished Manuscript <http://flamenco.berkeley.edu/papers/flamenco02.pdf>.

Feinstein, D., Smadja F., (2006) *Hierarchical Tags and Faceted Search. The RawSugar Approach,* Proceedings of SIGIR 2006 (August 6-11, 2006, Seattle, Washington).

Flamenco Group (2002) *How to Build a Flamenco instance* <http://bailando.sims.berkeley.edu/flamenco/howtobuild/howtobuild.html>.

Gnoli, C., Marino, V., Rosati, L., (2006) *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il Web [= Organizing Knowledge. From Libraries to Information Architecture for the Web]*, Tecniche Nuove.

Golder, A.S., Huberman, B.A., (2005) *The Structure of Collaborative Tagging Systems*, Information Dynamics Lab <http://arxiv.org/pdf/cs.DL/0508082>.

Hassan-Montero, Y., and Herrero-Solana, V., (2006) *Improving Tag-Clouds as Visual Information Retrieval Interfaces*, International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006 <http://www.nosolousabilidad.com/hassan/improving_tagclouds.pdf>.

Hearst, M.A. (2006a) *Clustering versus faceted categories for information exploration*. Communication of the ACM April Vol 49, No.4 <http://flamenco.berkeley.edu/papers/cacm06.pdf>.

-- (2006b) *Design Recommendations for Hierarchical Faceted Search Interfaces*, ACM SIGIR Workshop on Faceted Search <http://flamenco.berkeley.edu/papers/faceted-workshop06.pdf>.

-- *The Flamenco Search Interface Project* <http://flamenco.berkeley.edu/pubs.html>.

Heymann, P., Garcia-Molina, H., (2006) *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*, Technical Report InfoLab <http://dbpubs.stanford.edu/pub/2006-10>.

Kome, S H., (2006) *Hierarchical Subject Relationships in Folksonomies*

La Barre, K. (2006) *The Use of Faceted Analytico-Synthetic Theory as Revealed in the Practice of Website Construction and Design*, <http://leep.lis.uiuc.edu/publish/klabarre/facetstudy.html>.

Morville, P., (2005) *Ambient Findability*, O'Reilly.

Quintarelli, E., (2005) *Folksonomies: Power to the People,* Proceedings of 1' ISKO Italy-UniMIB meeting (Milano, 24 giugno 2005) <http://www.iskoi.org/doc/folksonomies.htm>.

Slavic, A., (2002) *FATKS: Facet Analytical Theory in managing Knowledge Structures for humanities,* <http://www.ucl.ac.uk/fatks>.

Travis, W., (2006) *The strict faceted classification model* <http://facetmap.com/pub/strict_faceted_classification.pdf>.

Yee, K.P., Swearingen, K., Li, K., and Hearst, M., (2003) *Faceted Metadata for image searching and browsing,* Proceeding of CHI 2003 <http://flamenco.berkeley.edu/papers/flamenco-chi03.pdf> .

VIII. SCREENSHOT

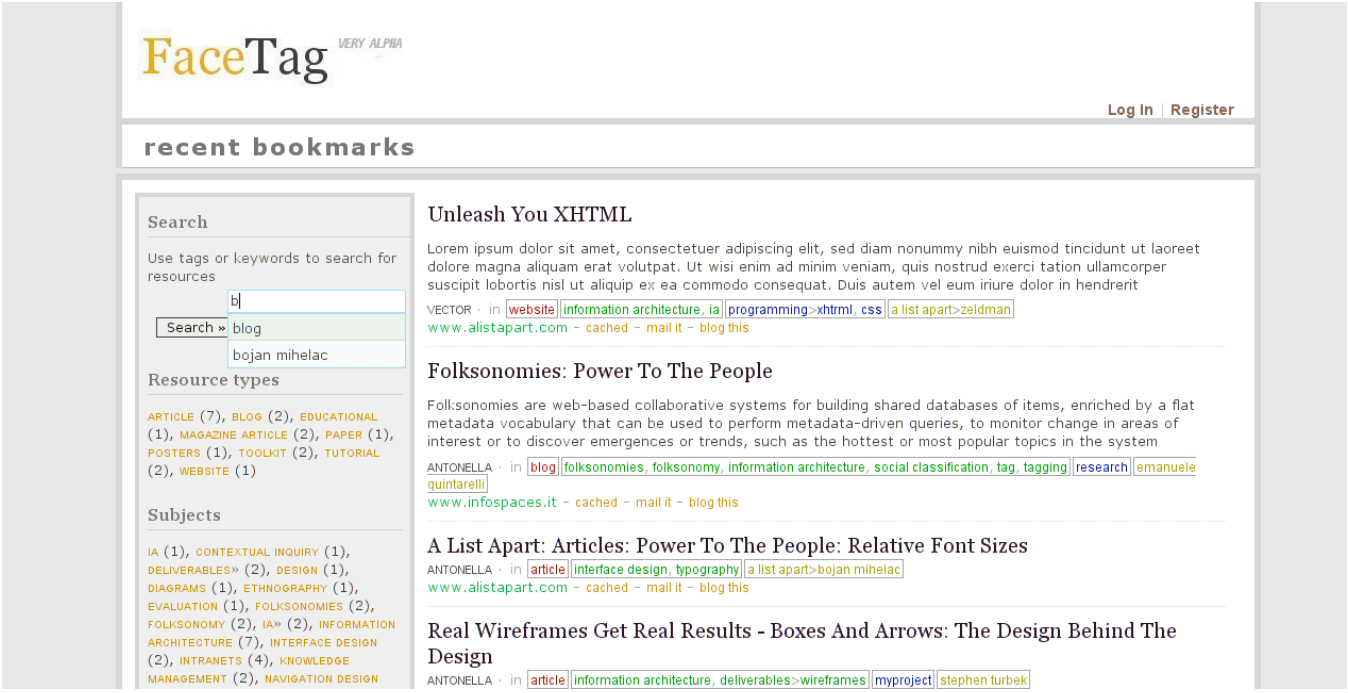

Figure 1: The system interface.

Figure 2: A zooming sample, choosing Resource type > blog + Subjects > Information architecture.