

Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System (*)

Emanuele Quintarelli

Reed Business
Rome, Italy
info@infospaces.it

Luca Rosati

University for Foreigners
Perugia, Italy
luca@lucarosati.it

Andrea Resmini

University of Bologna
Bologna, Italy
root@resmini.net

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 created by users while tagging 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 that facilitates the end-user creation of relationships between tags.

FaceTag current implementation is written in PHP / SQL and includes an open API which allows querying and integration from other applications.

Author Keywords

Information architecture, classification, faceted classification, social classification, folksonomy, tagging.

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' input and 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 linguistic issues which include polysemy, homonymy, plurals, synonymy 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 as currently implemented are not sufficient to provide a semantic, rich and multidimensional browsing experience over large tagging spaces. There are several reasons for this:

1. 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/>);
2. 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;
3. A flat tag cloud cannot visually support semantic relationships between tags. We suggest that these relationships are needed to improve the user experience and general usefulness of the system;
4. 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 linguistic complications such as 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 contributes to social tagging systems in three ways:

1. The use of (optional) tag hierarchies. Users have the possibility to organize their resources by means of father-son relationships;
2. Tag hierarchies are semantically assigned to editorially established facets that can be later leveraged on to flexibly navigate the resource domain;
3. Tagging and searching can be mixed to maximize findability, browsability and user-discovery.

Related work

The widespread adoption of tagging systems by end-users has greatly stimulated discussion about their long-term implications inside the information studies community and the so-called blogosphere.

In *Ambient Findability*, Morville [Morville 2005] states that tagging has its own proper place inside information architecture theory and practice, suggesting that these systems can be productively considered a complementary fast moving layer over slower layers represented by more traditional information architecture practices.

Karl Fast reinforced this position showing how pace layering theory can be combined with the complexity and resilience theories to provide a working model of interaction between folksonomies and conventional information architecture [Campbell & Fast 2006].

Golder and Huberman [Golder & Huberman 2005] analysed the structure and dynamics of collaborative tagging systems trying to extract stable patterns and recurrent tag types in del.icio.us <http://del.icio.us/>.

With Mefedia <http://mefedia.com/>, Dutch information architect Peter Van Dijck was among the first to propose a mixing of facets and tags (even if tags are statically assigned to a fixed number of editorially designed facets).

A discussion of faceted classification models in current web sites and the relationship between facets and tagging has been presented by Travis Wilson at IA

Summit 2006 [Wilson 2006].

Marti Hearst and The Flamenco project investigated for thirteen years how faceted interfaces can help users flexibly navigate and search through large information spaces [Hearst, The Flamenco Search Interface Project].

Hierarchical relationships can be implicitly found in tagging systems as showed by Sam H. Kome [Kome 2006], while Heymann and Garcia-Molina presented a simple algorithm to automatically convert tags associated to objects into a hierarchical taxonomy [Heymann & Garcia-Molina 2006].

The social bookmarking site RawSugar <<http://www.rawsugar.com/>> implements a similar hierarchical approach, mixing it up with clustering techniques.

Finally, Bar-Ilan [Bar-Ilan et al. 2006] compares unstructured (freely assigned tags) and structured tagging (tags assigned to predefined metadata elements), suggesting that structured tagging may be able to produce stronger user guidance, hence possibly resulting in higher quality descriptions.

Overview of semantical structures in tagging systems

Usability studies show how information seekers in large domains of objects prefer meaningful groupings of related items, in order to quickly understand relationships and so decide how to proceed [Hearst 2006a]. In other words, it seems quite clear that without any means to explore and make sense of large quantities of similar items, users feel lost and fail to complete their information seeking tasks.

How to generate and navigate such groups from a flat set of objects is anyway a different matter. Both clustering and faceted classification have been proposed in the past as useful techniques which allow searchers to easily browse and navigate information spaces.

Clusters

Document clustering refers to the act of grouping of items according to some measure of similarity, typically searching for identifiable repetitive patterns of words and phrases.

Some advanced tagging systems like Rawsugar and

Flickr are already using clusters to address the issues that plagued the first generation of folksonomy-based applications: clusters help reduce the semantic density and improve the visual consistency of tag clouds. Moreover, clustering is automatable, can be used to refine vague queries and to disambiguate ambiguous search keywords.

Nonetheless, clustering techniques and algorithms are not perfect and often generate messy groups which are generally hard to predict. These groups also tend to conflate many different dimensions becoming also hard to label in ways that are meaningful for users. Moreover, clustering does not generally allow issuing refinement and follow-up queries, thus heavily limiting the explorative capabilities of the system.

For these reasons, usability results show that users prefer clear hierarchies with categories at uniform levels of granularity over the messy, unpredictable and unlabeled groupings typical of clustering techniques [Hearst 2006a].

Hierarchical Facets

At the other end of the classification line, traditional hierarchical categories are coherent and complete systems of meaningful labels which systematically organize a domain. The main drawback of this approach is that a single a priori and monolithic hierarchical organization rarely has the capability to match the varied ways of thinking and organizing the world of different users.

Hierarchical faceted metadata has shown to be a promising middle ground, able to satisfy the needs of a wide range of users with different mental models and vocabularies [Yee et al. 2003].

Facets are orthogonal descriptors (i.e. categories) within a metadata system. Each facet has a name and it addresses a different conceptual dimension or feature type relevant to the collection. Facets can be flat or hierarchical and they can be assigned single or multiple values. Thus a faceted search interface requires that each object in the collection is classified through labels from different facets.

In a hierarchical faceted navigation tool, choosing a label from one of the facets is equivalent to performing a disjunction over all the labels beneath the selected one, while choosing labels from different hierarchies builds a query that is a conjunction of

disjunctions over the selected labels and their sublabels. In this kind of interfaces, users can navigate multiple faceted hierarchies at the same time [English et al. 2002b]. Usability studies show how this approach is preferred over single hierarchies because users feel in control without getting lost [English et al. 2002b, Yee et al. 2003].

Additional features exposed by faceted based interfaces are the suggestion of logical alternatives at each navigation step and avoidance of dead ends.

For these reasons, faceted metadata can be used to support navigation along several dimensions simultaneously, allowing seamless integration between browsing and free text searching and an easy alternation between refining (zooming in) and broadening (zooming out) the query while retaining a feeling of control and understanding [English et al. 2002b]. The major benefits resulting from this approach are a strong reduction of the mental work, favouring recognition over recall and better support for exploration, discovery and iterative query refinement [Hearst 2006a].

Again, usability studies attest how hierarchical faceted interfaces are preferred over simpler keyword based search interfaces and how they can be easily understood by the average user [Yee et al. 2003] if iteratively designed and tested to address usability issues [English et al. 2002a].

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.

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:

1. the Flamenco project
<<http://flamenco.berkeley.edu/>>
2. Facetious
<<http://demo.siderean.com/facetious/facetious.jsp>>
3. 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]

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

CRG	FaceTag
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.

Table 2: FaceTag facets and examples of foci

Facet	Examples
Resource Types	white paper, case study, blog>enterprise web,
Language	<i>predefined values (based on ISO Standard ISO 639-2)</i>

Facet	Examples
Activities / Subjects	discovery>competitive analysis, classification>facets, web 2.0>folksonomies, information design>navigation design>breadcrumbs>
Usage	industry, public administration, health, software>companies>google, education>conferences>www2006
People	dion hinchcliffe, morville
Date	<i>automatically added by the software</i>

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.

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.

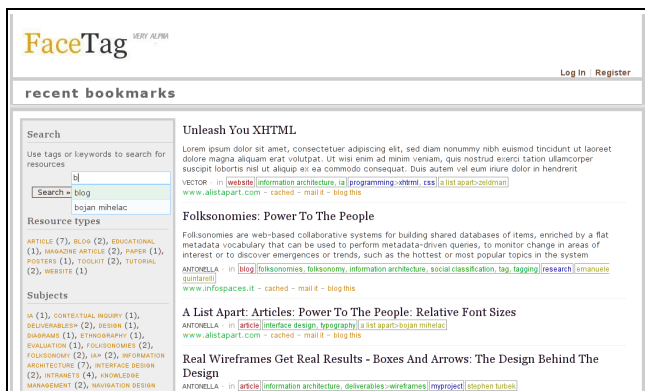
- ◆ In our case, the combination of tags and facets allows for better management of both these axes:
- ◆ from the vertical or paradigmatic point of view, when a user is going to associate a keyword to a facet (in order to tag a resource), the system suggests similar tags or hierarchy of tags pertaining 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).

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.

Image 1: FaceTag main page. Facets and tags are visible in the left sidebar, resources in the main body



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.

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.

References

- Bar-Ilan J., Shoham S., Idan A., Miller Y., Shachak A., (2006) *Structured vs. unstructured tagging – A case study*, WWW2006, 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: principi 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>>.
- Scott, J., (2000) *Social Network Analysis*, 2nd edition,

Sage Publications, London.

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>>

(*) This paper is the result of a collaborative effort. Nonetheless, Emanuele Quintarelli specifically wrote paragraphs “*Related work*”, “*Overview of semantical structures in tagging systems*”, “*Clusters*” and “*Hierarchical Facets*”, Andrea Resmini wrote paragraph “*Faceted Hierarchical Tagging*” and Luca Rosati wrote paragraphs “*Faceted analysis: the faceted scheme construction*” and “*Berrypicking, Information Scent and the Two Axis of Information Architecture*”.

Emanuele Quintarelli is an IT consultant, customer experience expert and information architect at Reed Business Information. After graduating in Computer Sciences, he completed a master in Multichannel User Experience and led a number of projects for content management, document management and portal systems with a specific focus on user centered processes and methodologies. Since 2005 he is working on the evolution of collaborative tagging platforms through his blog, papers and talks. In 2006 Emanuele organized the first Italian IA Summit <<http://www.iasummit.it>> and his website is Infospaces <<http://www.infospaces.it>>.

Andrea Resmini is an information architect. An IT professional since 1989, Andrea holds a master degree in Architecture and has been a teaching assistant of generative design at the Politecnico di Milano, Faculty of Architecture. He is now a partner at exEA, a small IT consulting firm in Italy, and a Ph.D. candidate in History and Informatics at the Department of History, University of Bologna and a visiting researcher at JIBS in Jönköping, Sweden. Andrea specializes in Free and Open Source Content Management Systems (CMS) and his website is [resmini.net](http://www.resmini.net) <<http://www.resmini.net>>.

Luca Rosati is a freelance information architect and assistant professor in Informatics for Humanistic Science (i.e. Information Architecture and Human Computer Interaction) at University for Foreigners of Perugia (Università per Stranieri di Perugia), in Italy. In 2003 he founded Architecta, the first Italian mailing list in IA. Luca runs Trovabile <<http://www.trovabile.org>>, an IA magazine (Trovabile is a neologism and linguistic calque on the English Findable). Luca is co-author of the book "Organizing Knowledge. From Libraries to Information Architecture for the Web" (Milan, 2006). His website is lucarosati.it <<http://lucarosati.it>>.

This paper was originally presented at the EuroIA Conference 2006, Berlin (DE).