BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **6**, No 3

Sofia • 2006

Practical Semantic Web – Tagging and Tag Clouds[1]

Ivo Marinchev

Institute of Information Technologies, 1113 Sofia

Abstract: *In this paper we research the incentives, state-of-the-art, and trends in a recently appeared social activity of tagging and creating tag clouds. We explain how this process works in practice and how it implicitly builds mutual understandable semi-formal semantics of many terms, artifacts and human activities. Although at the moment the results of the tagging are only intended to be used by humans its widespread adoption make it potential building block that may fill the gap between formal semantic web research, with its formalized ontologies and the practical semantic needs of the majority of lay-minded people.*

Keywords**:** *Semantic Web, Tagging, Tag Clouds, Folksonomies*

Introduction

In 1876 Melvil Dewey created his famous system of ordering information, the Dewey Decimal Classification System [2], and since then it was greatly modified and expanded in the course of the twenty-two major revisions. But at his time there weren't web sites, video clips or blogs. Today millions of items are generated on an hourly basis, and even the most sophisticated search engines can't find and, like alone, organize all the useful stuff amongst it. Recently a new wave of approaches under the general name *semantic web* emerged to help solving the problem. They all rely on the idea that the solution is to let the people do the categorization. There were/are a lot of scientific projects in this area but most of them gain little or no widespread influence or applicability. On the practical side of the problem a new practice called tagging came into fashion. It is neither formally defined process nor relies on some set of predefined ontologies (at least at the moment) but leveraging the combined knowledge and experience of internet community the people can collectively label everything from great literature to pictures and create some sort of do-it-yourself classification.

[1] The research has been partially supported by "Technologies of the Information Society for Knowledge Processing and Management" – IIT – BAS Research Project No 010061.

What is tagging?

As the name implies, tagging something means putting a virtual label on some piece of information. What the tag says is totally up to the person that applies it. The important thing is that later one can find things simply by the tag name. Think of tagging as the opposite of search. By leaving linguistic breadcrumbs behind on one's wanderings through cyberspace, one can easily relocate the sights (and sites) he saw along the way (i.e. the general idea is keeping found things found). There's no limit to the number of tags people can slap on an item.

The big question about tagging is whether the lack of rules will lead to chaos. Early results from the Web sites that exploit tagging show quite the opposite: order seems to emerge from the chaos of freestyle labeling. A site called **del.icio.us** [3], participants put tags on their favorite Web sites, making it not only easy to find information on specific topics, but allowing visitors to view the most popular sites of the whole community. The photo-sharing site Flickr [6], which classifies images by user-selected labels, has generated a sometimes quirky but totally coherent form of organization, simply because people can check out which tag words get the best responses from the community, and do their own tagging accordingly. The whole process seems similar to that of forming natural languages (with a difference that with tagging we usually get domain specific languages), which is also a self-organized process. That process is also at work on a Web site called 43 Things, where people express their goals, tag them and comment on the goals of others. It turns out that a lot of people on the site read a book called "Getting Things Done." When someone came up with the idea of making a tag called "GTD," others recognized that the abbreviation was an ideal label, and thereafter anyone who posted a goal inspired by that book stuck a GTD tag on it. That's a classic example of how the group effort of tagging can discover its own kind of compelling logic. Tagging enthusiasts call such systems "folksonomies."

Another idea is that tagging is not designed to share. It's designed to create *walled gardens* [4]. Here is one informal definition of this term: A *walled garden* is a large set of pages that bring in their own organizational or conceptual baggage, and hence integrate poorly with the rest of "the world". The content is appropriate, but the terminology prevents integration. This means that if you bring your own organizational structure to something, it won't fit with the rest, walling it away. One has to use terms that are known only to experts so that he can found information in any tagging system.

How does this "organization" help a newcomer? It doesn't. You have to stumble onto a phrase that returns some relevant hits, and then try to understand why a link has these other terms to decide if they help or hinder your search. The examples above only emphasize this. It speaks of the tag "GTD" as short for relevant to a book, "Getting Things Done". But if one doesn't know that, it's a useless tag for sharing. Oh, it shares with those who already know, but it turns social activities online to setting up cliques and domain specific languages.

One feature del.icio.us has to circumvent this problem is the related tags area. So if you begin your search by looking up a broad tag like just data, then you will see

a list of related tags along with the results. These tags can often lead you down an interesting path, which often does lead to a useful resource.

More general solution is probably in setting two fields of tags – “Standard Tags” (standardized set of tags) and “Free Tags”. Standard tags are keys to the concept of an article or link or page or functionality or whatever. It can be a big list, and it will change, and it would override the old terms where necessary but it will have predefined semantics (not needed formal definition, just really broadly accepted terms). The Free Tags are all additional tags that the user thinks are appropriate. So using “Standard Tags” and “Free Tags” together will allow folks to find useful things and expand their knowledge at the same time, instead of being forced to either find stuff they already know about.

In general, folksnomies lend themselves to the “pivot search” which works like this:

1) search for the most specific tag you can think of;

2) if that yields too many or too few results, keep trying other words or combination of words;

3) once you find a relevant result, look at how other people tagged it (you now know the “right” words);

4) search for those tags;

5) repeat.

For example, I may not know the “right”-word for the nifty auto-suggest feature of Google Suggest [5] but if I look for only the tag “suggest” on del.icio.us, I learn pretty quickly that “ajax” might be a significant term. It is not just finding resources related to a concept but finding labels important to the community. Moreover those labels aren’t just incidental/accidental associations with the topic (as found by search engines) but intentional, conscious labels given to it by people.

Finally, you might fear that tags create walled gardens which hide rather than share information but, in practice, that doesn’t happen because:

1) As more people tag a resource, the “gene pool” of associated tags becomes more diverse. So, if you say “potayto” and I say “potawto”, the community will link those resources eventually.

2) Folksonomy systems are already getting smarter about capturing fuzzy associations. You can already browse related tag clouds. Future improvements would use simple linguistic analysis to group related tags (i.e. “blog, blogs, blogging”) and allow more consistent “phrase tags” so that “social_software”, “social.software” and “social-software” would be equivalent.

Tag clouds

From the user point of view *tag clouds* commonly consist of two elements:

1) a collection of linked tags shown in varying fonts and colors to indicate frequency of use or importance;

2) a title to indicate the context of the collection of tags;

Flickr's tags [6] page is the iconic example of the tag cloud. Fig 1 shows screenshots of one of the first well known tag cloud tagging implementations del.icio.us [7].

Fig. 1. Tag cloud of **del.icio.us**

Tag clouds: as visualizations of semantic fields

The simple structure of tag clouds allows them to perform a very valuable function without undue complexity. That function is to visualize semantic fields or landscapes that are themselves part of a *process of mutual understanding* (a semi-formal process of transformations and steps that identify and understand all the different kinds of people, taggings, and meanings, Fig. 2) linking taggers (tag creators) and tag consumers. The tags in tag clouds originate directly from the perspective and understanding of the people tagging. Tag clouds accrete over time when one person or a group of people associate a set of terms with a focus of some sort – a photo on flickr, a URL / link in the case of **del.icio.us**, an album or song for **last.fm**. A focus can be anything that can carry meaning or understanding. The terms or tags serve as carriers and references for the concepts each tagger associates with the focus. Concepts can include ideas of aboutness, origin, or purpose, descriptive labels, etc. While the concepts may change, the focus remains stable. What's key is that the tag is a reference and connection to the concept the tagger had in mind. This connection requires an initial understanding of the focus itself (perhaps incorrect, but still some sort of understanding), and the concepts that the tagger may associate with the focus. And this is the first step in the process of mutual understanding behind tag clouds, as shown in Step 1 on Fig. 2.

The tag is a sort of label that references a concept or set of concepts. A cloud of tags is then a collection of labels referring to a cluster of aggregated concepts. The combination of tags that refer to concepts, with the original focus, creates a "semantic field". A semantic field is the set of concepts connected to a focus, but in a form that is now independent of the originating taggers, and available to other people for understanding. In this sense, a semantic field serves as a form of reified understanding

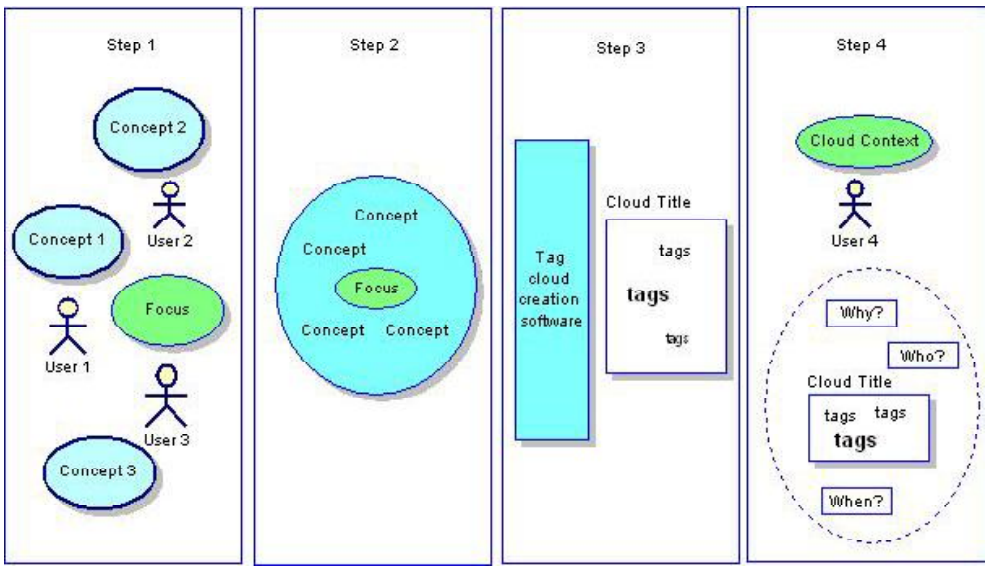

Fig. 2. Process of mutual understanding for tag clouds: Step 1 – taggers choose focus and concepts; Step 2 – taggers apply concepts creating semantic field; Step 3 – software creates tag cloud that visualizes semantic field; Step 4 – consumers interpret cloud context to understand tags and concepts

that the taggers themselves – as well as others outside the group that created the semantic field – can now understand, act on, etc. Step 2 on Fig. 2 shows this second step in the process of mutual understanding; without it, there is no semantic field, and no tag cloud can form. The most important thing to understand is that *tag clouds are visualizations of a semantic field*.

Tag clouds are revolutionary in their ability to translate the concepts associated with nearly anything you can think of into a collectively visible and actionable information environment. In practice, tag clouds can make metadata visible in an easily understood fashion.

To summarize, creating a tag cloud requires completion of the first three steps of the process of mutual understanding that supports social metadata. Those steps are:

1) understanding a focus and the concepts that could apply that focus;
2) accumulating and capturing a semantic field around the focus;
3) visualizing the semantic field as a tag cloud via transformation;

The fourth step of this process involves users' attempts to understand the tag cloud.

How tag clouds are understood by consumers

We must introduce the idea of context, which addresses the questions of which original perspectives underlie the semantic field visualized in a tag cloud, and how those concepts have changed before or during presentation.

Users need to put a given tag cloud in proper context in order to understand the cloud effectively. Their goals may be finding related items, surveying the thinking

within a knowledge domain, identifying and contacting collaborators, or some other purpose, but it's essential for them to understand the tags in the cloud to achieve those goals. Thus whenever a user encounters a tag cloud, they ask and answer a series of questions intended to establish the cloud's context and further their understanding. Context related questions often include:

- Where did these tags come from?
- Who applied them?
- Why did they choose these tags, and not others?
- What time span does this tag cloud cover?
- Etc …

Context in this case means knowing enough about the conditions and environment from which the cloud was created, and the decisions made about what tags to present and how to present them. Once the user or consumer places the tag cloud in context they can use or work with the tag cloud (Step 4 on Fig. 2).

Conclusion and future trends

I use **del.icio.us** to discover smart people on various subjects. I do this by exploring the tags others have used for sites I find useful in a given area. I track from tags back to users. Once I hit upon someone that's smart in a given area (defined as lots more engaging material in a given area than I've been able to collect), I plug his tag of interest into my RSS reader. At the present time I'm "intellectually drafting" off the discoveries of several very smart folks in several different areas. But I tag only things that are relevant for me. And that is always different than what is important for anyone else. Now sometimes, my interests are more closely aligned with someone else's than theirs would be to my expert. In a sense, I'm a human filter for them. I've also discovered that a number of people are drafting off of me in the same fashion – sometimes on the same tags.

To date, tag clouds have been applied to just a few kinds of focuses (links, photos, albums, blog posts are the more recognizable). In the future, we expect to see specialized tag cloud implementations emerge for a tremendous variety of semantic fields and focuses: celebrities, cars, properties or homes for sale, hotels and travel destinations, products, sports teams, media of all types, political campaigns, financial markets, brands, etc.

From a business viewpoint, these tag cloud implementations will aim to advance business ventures exploring the potential value of aggregating and exposing semantic fields for a variety of strategic purposes:

1) creating new markets;
2) understanding or changing existing markets;
3) providing value-added services;
4) establishing communities of interest / need / activity;
5) aiding oversight and regulatory imperatives for transparency and accountability;
6) etc …

Cloud consumers' needs for better context, will drive the addition of features and functionality that identify the context of a tag cloud explicitly and in details. For

example, clouds created by a defined audience will identify that audience, whether it is system administrators, freelance web designers, DJ's, or pastry chefs rating recipes and cooking equipment and provide indication of the scope and time periods that bound the set of tags presented in the cloud.

Diversifying consumer needs and goals for way finding, orientation, information retrieval, task support, product promotion, etc., will bring about inverted tag clouds. Inverted tag clouds will center on a tag and depict all focuses carrying that tag.

Along the same lines, tag clouds will demonstrate more complex interactions, such as spawning other tag clouds that act like magnifying lenses. These overlapping tag clouds may offer: multiple levels of granularity (a general view and zoom view) of a semantic field; thesaurus style views of related concepts; parameter driven term expansion;

Finally, we will show one possible application in the search field. Now Google only sees web pages. Right now that's an advantage since the ratio of web pages to shared bookmarks is high. What will happen when social bookmarking goes mainstream? Then, there will be more social bookmarks than there are "interesting" web pages. It will be the result of the fact that people can (and do) tag far more pages than they write. When social bookmarking reach this turning-point (or even close), tags will be the definitive insight into how people think about and classify information.

References

1. Levy, S. In the New Game of Tag, All of Us Are It. **http://www.msnbc.msn.com/id/7445653/site/newsweek/**
2. **http://en.wikipedia.org/wiki/Dewey_Decimal_Classification**
3. **http://del.icio.us/**
4. **http://c2.com/cgi/wiki?WalledGarden**
5. **http://www.google.com/webhp?complete=1&hl=en**
6. **http://www.flickr.com/photos/tags/**
7. **http://del.icio.us/tag/**