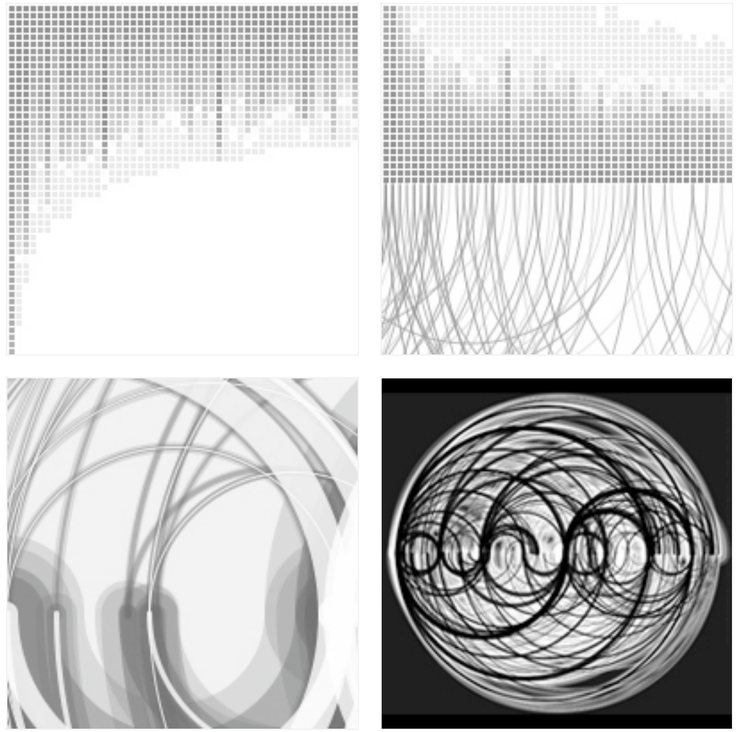
» mandalabrot.net



del.icio.us.discover is a set of visualizations + explorations into "link-spaces" in the del.icio.us collaborative book-marking network. It was created with Processing + Perl, using data gathered early March, 2006.

In this project I've explored and generated:

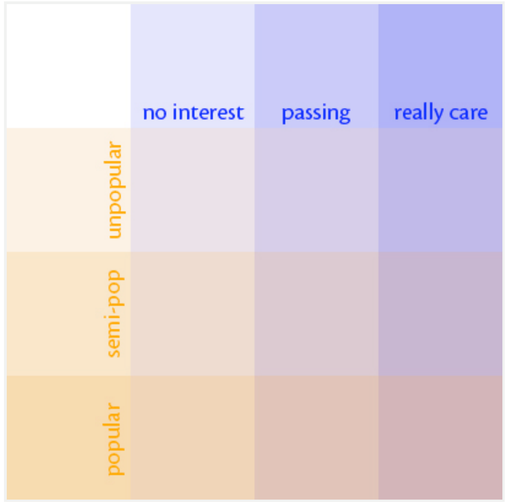
1) direct and abstract visualizations of intra- and inter-user relationships

2) simple, implementable algorithms to recommend "links-of-interest" that might otherwise be missed by a user currently utilizing only popularity-based link aggregators, and ideas for their further usages.

Created by kiddphunk / view code / begin!

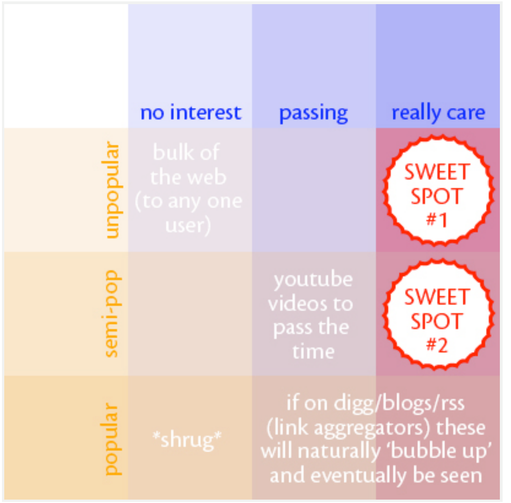
del.icio.us.discover kiddphunk creative commons Processing

» mandalabrot.net



interest level vs. link popularity - 01.01

01.02



one user's interest level vs. link popularity - 01.03

As the amount of information available to everyone increases, it similarly becomes increasingly necessary to seek out "nodes" that gather and filter a subset of all of the available information. We see these nodes take the form of Google News, RSS/feed aggregators, "social bookmarking" sites, and blogs that gather/collate in specialized niches.

For example, I used to keep a running collection of sites that featured visualizations of complex networks and interesting interfaces for browsing large sets of data, until I discovered visualcomplexity.com, a labor of love by Manuel Lima that *"intends to be a unified resource space for anyone interested in the visualization of complex networks."* Because of the narrow focus of the site, for me it has an extremely high signal-to-noise ratio, and I can read this "aggregate blog" through my RSS news-reader alongside other blogs, Basecamp project updates, del.icio.us inbox subscriptions, etc. and know that I have a wide coverage of visualization systems.

It is useful now to consider the empty chart in Figure 01.01 which plots personal interest in a link against its general popularity (on del.icio.us). Consider how many of the links that you receive/view in the day are popular/random, and/or useful/forgettable.

And when considering how we browse the web, there should also be a distinction drawn between various modalities. There are many, but three modes in particular are relevant here: *"information retrieval"* where one has a query or roughly knows what they are looking for, *"show me whats new"* where the aim is to get a feel for the 'news' (or a pulse of the current world vibe), and the *"show me what I need to see right now"* mode where one has a fixed amount of time allotted to bit-viewing yet is flexible as to the content. (Note of course that any one of these modes can cause a user to branch or split off into submodes and sub-queries, a complex, intertwined dance of information gathering.)

One drawback with most aggregation nodes is that they mostly aggregate on popularity or relevance to the general population, acting as band-pass filters attuned to a specific nodal resonance. The tradeoffs are between precision in delivery, signal-to-noise ratios, and volume of information presented. As we wish to get more a wider spectrum of information, these parameters get jostled. I started adding my links and tags to del.icio.us when I realized what an immense predictive tool it was, and that by adding my data to the collective whole I would then be able to use the system as a feedback device for my own exploration.

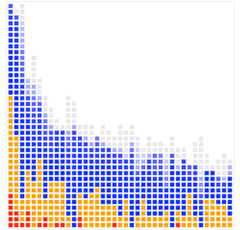
Generally speaking, I think the key shift is to begin to see, understand and utilize *each other* as these nodes/sources of aggregation, and continue to extend our tools and interfaces to facilitate our group-mind interaction.

In the interest/popularity graphs, perhaps tautologically, I am assuming that the probability that a user will see a given link is proportional to its popularity. This means that there are links (that fall into the areas above marked "SWEET SPOT" in Figure 01.03) that are currently of medium- to low-popularity and yet are of high interest. These are the areas that I feel are worth exploring and facilitating access to, and that nodes such as del.icio.us can be utilized in this manner in a much greater degree than at present.

del.icio.us allows a user to subscribe to another user's book-mark stream. This is a great feature, but out of the thousands of users, how do we pick out the users with the highest (for us) signal-to-noise ratio? And how do we best utilize the disparity in link popularities, to find hidden nuggets of personally-relevant information that might otherwise be missed?

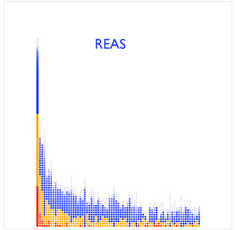
Continue on to part 02 | *Graphs*

« mandalabrot.net

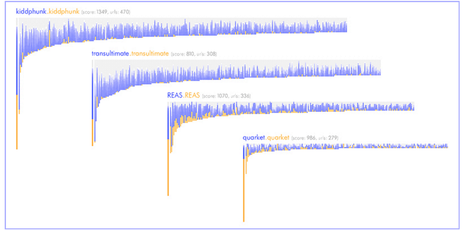


02.01

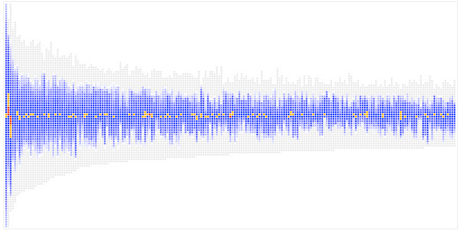
02.02



02.03



02.04



02.05

At the heart of the del.icio.us.discover visualizations is an individual user's graph, a representation of their link-space.

The graphs in this set of experiments/visualizations are tag-agnostic; tags on any link are not taken into account, although one could easily imagine various ways to integrate tag pruning/filtering/matching to assist in searching, or setting the seed directions for an 'information stroll'.

One other obvious element at work is time. These are merely snapshots for a given time period (in this case, the beginning weeks of March, 2006) and one cannot step in the same collaborative link-space river twice (or even once, really).

Every hugely popular link started with one posting and a new link that has 1 person considering it important may have 1000 people within a week. Due to the nature of the link-space and the curves that fall out of it, the temporal aspect of these graphs is not of great concern, although there are many interesting predictive ideas to explore.

The graph in Figure 02.02 is a representation of my (`kiddphunk`) del.icio.us link-space.

Each column of colored squares represents a different del.icio.us user. The 10 pixel x 10 pixel squares that form the body of the columns each represent a particular link from that user's link-space that had intersected my link-space. The leftmost column with the solid bar is a stack of my links, with the height of each square compressed down to 1 pixel to save space.

I can now sort by total number of links matched and graph in descending order. This creates a characteristic power-law distribution curve that will be discussed in greater depth shortly.

The internal algorithm works simply by looking at every link that the main user has book-marked, and hashing all of the users who have also book-marked this same link. However, more interesting than the degree of overlap between del.icio.us users' link-space is the degree of popularity for a given link, especially when considering the "sweet spots". The colors in this particular graph (02.02) plot links of a popularity number 'p' (the number of other people linking to the same link) using the following formula:

```
common = grey = (P > 1000)
popular = blue = (20 < P <= 1000)
semi-popular = orange = (3 < P <= 20)
random = red = (P <= 3)
```

(I'm not implying anything about the actual popularities by these arbitrary lines in the sand; they are however, easy to remember general-level names.)

The link squares are now colored to match popularity and within each column are further sorted with the most popular links at the top in grey, followed by a gradienting blue representing the "middle ground", then the more random (less-popular) links at the bottom in orange and finally red.

A simple weighting algorithm was utilized to 'bubble up' users who matched more of the less-popular links (those in red/orange). The first weighted sorting variant I utilized gave a score of 1 was given for grey links, 2 for blue, 3 for orange and 4 for red. The results of this sorting method is shown for user REAS in Figure 02.03 above. Another weighting method experimented with only assigned scores to links in the red/orange set. Additionally tweaking the thresholds for grey/orange/red in conjunction with various sorting methods gives finer grained control for different visualizations.

One optimization that I did not have time to implement were sorting modes that found users with higher internal red/orange matching percentages by considering the total number of links in an individual's link-space.

Figure 02.04 above shows four user graphs, all recognizably power law distributions. While at first this surprised me, after researching more about power laws and scale-free networks I find this now intuitively sound.

"What matters is this: Diversity plus freedom of choice creates inequality, and the greater the diversity, the more extreme the inequality... The very act of choosing, spread widely enough and freely enough, creates a power law distribution."

(Excerpt from *Power Laws, Weblogs, and Inequality*; see also Wikipedia, *Scale-free networks* for more background information on social networks and the power law distribution.)

This is an important feature to note because it means that the top N user's intersections cover a much wider span of links (which may in turn imply a good future predictive capacity for that user, especially when coupled with an appropriate sorting algorithm) than the next N users.

So a rather small sampling of users taken in the form of monitoring their 'link stream' or as an on-demand summary could give a fairly decent sampling of interesting links based on past link history, assuming an adequate amount of information has been accumulated.

In addition, while Figure 02.04 lacks a value for the red parameter in this particular rendering, a quick comparison of curves still affords a few general observations:

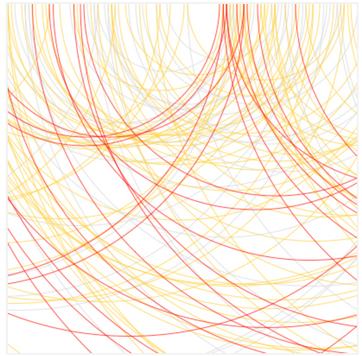
* `quarket` has the highest proportion of less-popular links (the orange) than REAS, who in turn has a higher orange proportion than `kiddphunk`.

* while `quarket` and REAS have roughly the same number of book-marks in their respective link-spaces and a very roughly similar amount of less-popular links, `quarket` overall has much less area under the curve, implying that many of the less-popular links were not shared with anyone else or are very widely distributed across a wider range of users.

Above, Figure 02.05 shows a section of a graph of `transultimate`'s link-space, with a red/orange weighted sort on the top and a url-count weighted sort on the reflected bottom.

Continue on to part *03 | Connections*

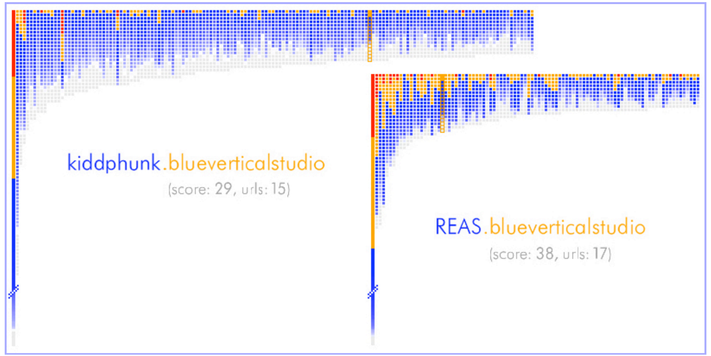
» mandalabrot.net



03.01

Now that a user's graph of their link-space is established, the next step involves placing several graphs side-by-side and connecting them where there is interesting overlap.

Overlap in the context of these experiments consists of sharing a user between the graphs (the column count/number of users is configurable), although one could also map the connections between the actual individual links themselves.

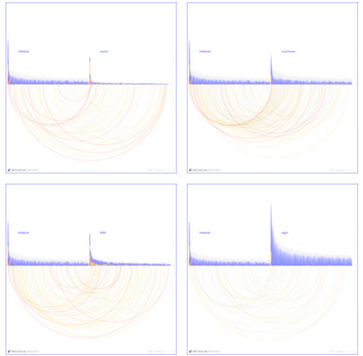


03.02

Originally del.icio.us.discover was an interactive application that displayed several graphs and highlighted users (columns of the graph) that were shared. As you moused-over a column the name of that particular user was shown and was highlighted in orange in all graphs containing that user.

Figure 03.02 above shows this behavior, highlighting the user `blueverticalstudio` who is common to both `REAS` and `kiddphunk`. 'Score' refers to the 1-4 point weighing score discussed in the previous section, and 'urls' to the total intersecting links.

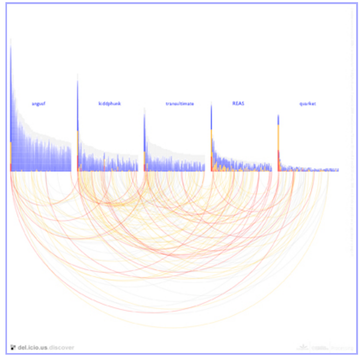
I decided to produce a set of images rather than a true interactive application due to the tradeoff of memory and speed/optimization in the app. The amount of data and processing involved also prohibited doing this with real-time data from del.icio.us, and so I decided on using a sample set of users and pre-gathering the data required to generate their graphs.



03.03

Two users in the examples, `quarket` and `transultimate`, are close friends of mine with whom I love to bounce ideas and information off of. As we have many similar interests and due to exchanging of links via {IM, irc, email, mailing lists}, my graph (`kiddphunk`) and their graphs display an unusually high correlation between our respective link-spaces. Another user `REAS` is a friend-of-(multiple)-friends and the remaining users (to my knowledge) I did not know a priori, and selected them basically at random from the set of my intersecting users that looked 'interesting'.

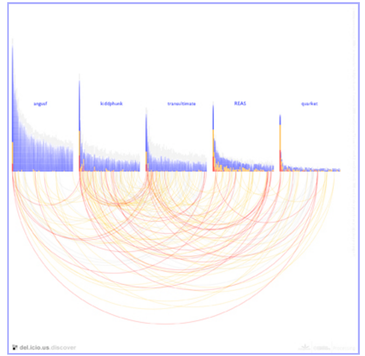
The four individual panels of Figure 03.03 show pairs of graphs connected, between {`kiddphunk`} and {`quarket`, `transultimate`, `angusf`, `REAS`}.



03.04

Two columns are linked by an arc if they represent the same "sub-user" in both graphs. The color chosen for these renderings was red, orange and grey depending on the presence of them in the leftmost (of the pair) graph's user column. That is the colors sort of show a 'leftward' flow or a little bit of information about how the subuser existed in the left graph. I could easily have flipped the polarity or taken a max representation.

The forms of the next section, which I had already started to envision, negated the real need to play around with these representations much further.

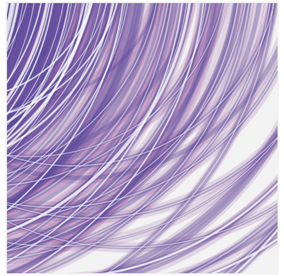


03.05

A few additional graphs in this series (as well as many others from all of the sections) is available in highest resolutions at this *flickr visualization gallery*.

Continue on to part 04 | *Abstractions*

» mandalabrot.net

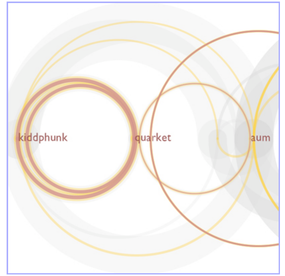


04.01

To correct the asymmetry present in the color coding of the arcs (left->right relationships), I decided to remove the top graph altogether and replace it with the inverse linkages of the bottom (right->left relationships). Having a balanced top-bottom allows for yin-yang symmetries to form in the natural interlock of sub-users-in-common and their respective link-space influences.

By varying the arc widths and color the same data can be contained as in the stacked-pixel graph (although abstractly).

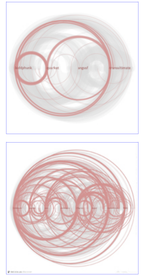
The basic method I used for these particular renderings in this section often ignores the blue realm of the previous graphs, the "middle-ground", in order to more clearly view the opposite extreme ends of the link-spaces.



04.02

Figure 04.02 shows the intersection of several users; this view shows just where the primary users themselves appear on each others' graphs. The radius is 1 pixel/link and the red/orange/grey values should be viewed in an additive sense (that is, the 'width' of a colored band should be read from the end of the previous color band); no data is contained in the alphas. The coloration formula follows that of the graphs from the previous section.

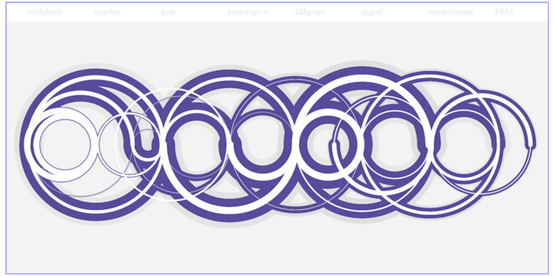
In this example, the arcs can be read quickly to ascertain that `kiddphunk` and `quarket` share a large amount of red-links, but not many greys. Quite a different pattern is seen between `kiddphunk` and `angust` (off-screen), who overlap in a large amount of greys, a few oranges and no reds.



04.03

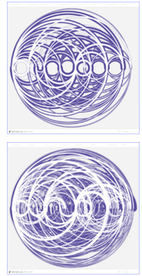
These two renderings show 4 and 8 users, each with a 40 column graph. The arcs contain all links, as the coloration rule used covers all popularity values:

```
grey = (P > 10)
red = (P <= 10)
```



04.04

Figure 04.04 illustrates where the 8 primary users connect to one another; only connections between adjacent user graphs are shown.

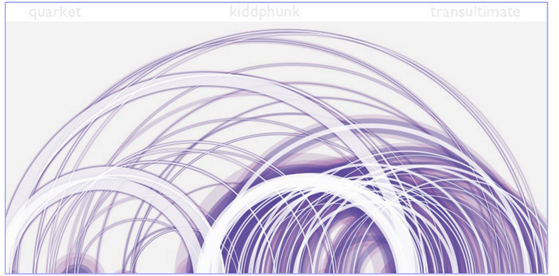


04.05

Figure 04.05 goes further and graphs the full set of primary connections (top) and the set of all connections (bottom).

The coloration rule utilized by these three graphs is:

```
grey = (P > 5000)
blue = (50 < P <= 500)
[white] = (P <= 50)
```



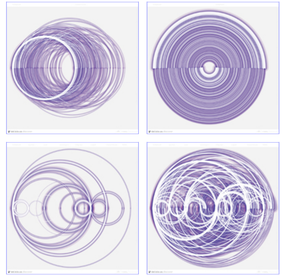
04.06

The top hemisphere of a tri-graph visualization (of `quarket`, `kiddphunk`, `transultimate`) is seen in Figure 04.06, above. The coloration rule utilized by these final graphs remains the same as the previous set; the coloration has however shifted to a purple/blue/white palette with highly transparent arc coloring.

Further explorations may include larger dataset renderings (16-user graphs), more precisely-focused ones (8-user graphs between a primary user and their "top 7") or investigating ways to show the "overlap" more richly (how every column intersects with another users' link-space).

This simple exploration has looked into the connections that arise between a small number of people on del.icio.us, providing a rich set of behaviors, patterns and inter-linkings that are as beautiful to behold in the eye as they are when absorbing what these links, flows and relations represent in non-visual terms.

Footprints and reflections from a global mind learning to work with new forms of input/output, shaping the freedom and flow of information in real-time, utilizing the ever-expanding wave of knowledge individually and collectively.



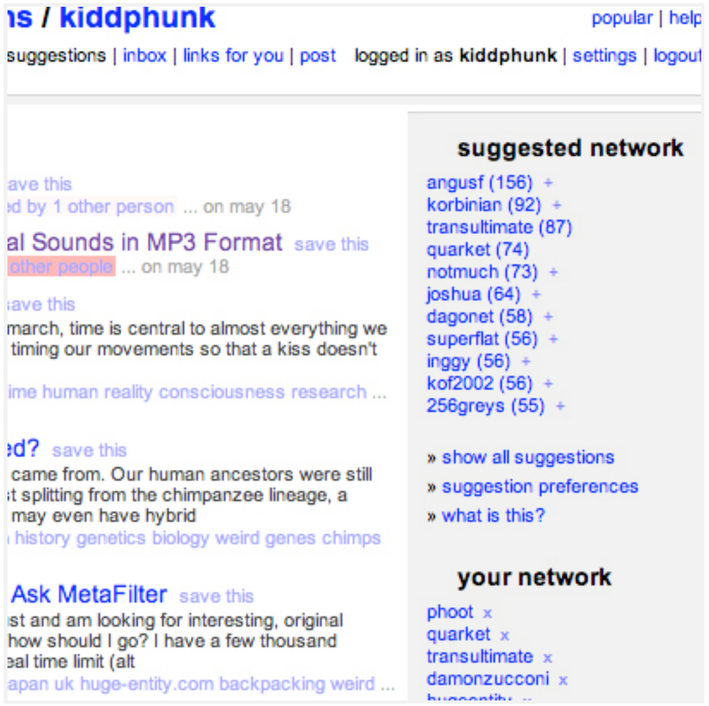
04.07

Above (clockwise from bottom-left):

- connections between 8 primary users
- a detailed pair of graphs (`kiddphunk` and `transultimate`)
- record/grooves effect created by using two mirrored `kiddphunk` graphs
- full purple/translucent version of the 8-user link-space

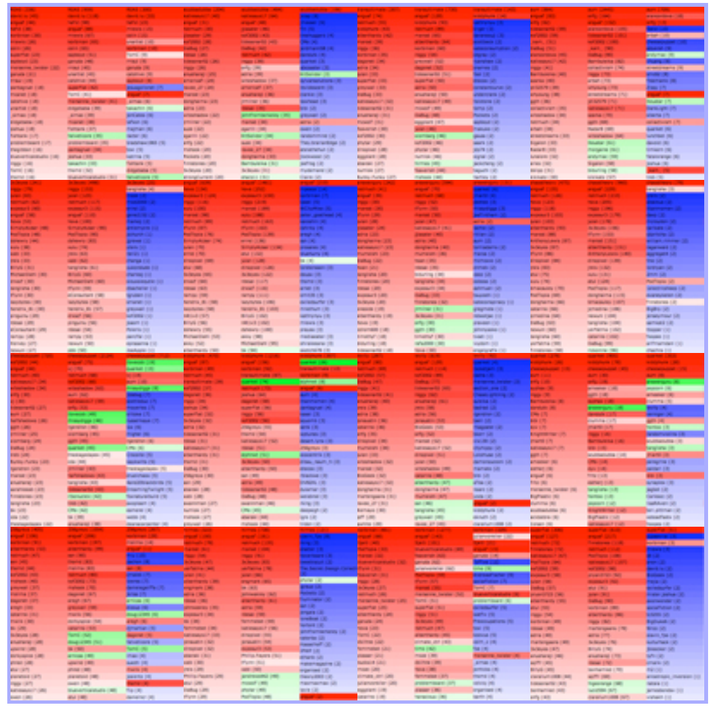
Continue on to part 05 | *Recommendations*

» mandalabrot.net



click for sample
user recommendations:

» 256greys «
» 3x3eyes «
» angusf «
» answerguru «
» aum «
» chaostheory «
» cheesepuppet «
» hirmes «
» kiddphunk «
» korbinian «
» levity «
» quarket «
» REAS «
» scubbadubba «
» superflat «
» transultimate «



It would be easy to add this to the del.icio.us interface (artist's sketch above), and the computations could be run on an on-demand and once-per-N-week-limit to keep loads down.

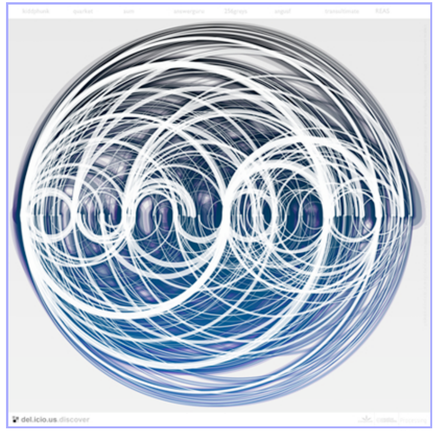
I believe it essential to integrate something like this into del.icio.us; at the low-end to allow users to find others with similar book-marking habits, and at the high-end to have the ability (with sliders!!!) to truly play with the parameters and delve into the richness of the data present.

Click on a grid thumbnail at right or on a name above to view one of sixteen sample recommendations.

Each features a three-column recommendation grid, and are 100% organic and tasty!

Each features a three-column recommendation grid, and are 100% organic and tasty!

» mandalabrot.net



06.01

06.03

Hopefully del.icio.us.discover has provided food-for-thought, and stirred up or augmented related ideas around utilizing each other as data filtering nodes in our own right.

Many sites of course already utilize collaborative filtering for personalization/recommendations, but there are many other popular nodes where the application of these simple algorithms (or more complex collaborative filters would, I think, generate a very useful source of new, more precisely targeted suggestions. Here are just a few examples of other implementations:

del.icio.us

In addition to the suggestions already made, more temporal aspects of the collective book-marking can be exploited. That is, it would be nice to know what other people have book-marked the same links recently as I, as I add a new link, or even to make predictions as to what I might like to see next based on what I've recently added.

toptaggers offers a list of helpful del.icio.us users, which ranks users based on their book-marking links with meaningful tags (as evidenced by others subsequently doing the same). It also offers a mechanism for suggesting users that, for a given user and tag, have the highest overlap. This is a great mechanism for finding the popular links and sources for new links, although it is tag-based and scores heavily on popularity.

flickr

Using essentially the same algorithm, one could find interesting photographer/photostreams or suggested photos. Every photo you add to your favorites is essentially the same as marking a del.icio.us book-mark. By considering the popularity of the photos that are in one's collection (or even the popularity of other photographer/photostreams in your network) the suggestion engine would work in the same manner.

blogger, technorati

Similarly, by considering one's blog to be a repository for 'favorite links' akin to a del.icio.us user's link-space, one could easily discover other blogs and bloggers that are on a similar wavelength. Again, by considering the popularity of the links that are shared, one can avoid just finding popular blogs and turn up ones that are relevant.

One could then imagine a simpler, yet more useful feature, where only the last N-days worth of links that you have blogged are considered, so suggestions are more aligned with your current subjects of interest.

Regardless of the complexity of the algorithms used to do the matching, the key issue is to begin to have the recommendations.

everything, more connected

As nodes grow and become mature, the data that is collected as a function of the node's operation and user connections is usually extremely rich and deserves to be better utilized. "Social" bookmarking sites are hardly social in the conventional sense of the term, they are more often "collective" or "public" bookmarking sites; the challenge is to add more of the social aspects to them, for example by exploiting the power-law shape of the user-link spaces to make predictive suggestions.

I look forward to greater interaction between us all.