Self-organising map based tag clouds

Creating spatially meaningful representations of tagging data

[OPAALS Conference 2007]

Jaakko Salonen
Hypermedia Laboratory
Tampere University of Technology
Tampere, Finland
jaakko.t.salonen@tut.fi

Abstract*—In this article, we present a method and a tool for visualising tagging data with self-organising map (SOM). Tagging as a knowledge management method is described in relation to existing tools and methods. Current approaches of tagging data visualisation are also presented, especially popular "tag cloud" method and its different variations. Finally, our SOM based visualisation method, SOM Cloud, and a proof-of-concept implementation of it is put to test with data from del.icio.us social bookmarking service. As a result of the study, we found out that while the applicability of SOM to "tag cloud" metaphor is limited, we could use it successfully to add spatial encoding to tagging data.

tagging; folksonomies; information retrieval; information visualisation; self-organising maps; knowledge visualisation

I. Introduction

From the new, "web 2.0" breed of online services, a new approach for creating metadta, called tagging, has emerged. Especially in social bookmarking services, users may add tags – i.e. arbitrary keywords associated with resources – to describe and organise resources. Tagging features have also spread to other types of applications as well. Even offline applications such as music players and photo organisers have started supporting them.

Tagging is technically a trivial feature, but it is one of those rare features that give users the ability to create their own structures for knowledge organisation. Tagging data that creates associations between tags and different resources, can be used to various personal navigation structures, including navigating resources by tags, navigating tags by related tags and navigating related resources by their tags.

Motivation to our study of tagging systems lie on research objectives of the European Network of Excellence Project OPAALS. In OPAALS, an overall objective is to build sustainable interdisciplinary research community and develop integrated theoretical foundation for digital ecosystems' research. This objective is approached by building an Open Knowledge Space (OKS) for knowledge creation and sharing.

One of the challenges in building the OKS, is the integration of knowledge from different practise communities. While various scientific domains may have very well defined, formal domain ontologies that accurately specify conceptualisations, they offer little help for leveraging cross-domain understanding of concepts and their relationships, as these ontologies do not capture the entire process, but mainly the results of the vocabulary construction work.

As a side effect of a well-planned tagging system, tagging contributes to development of a folksonomy. A folksonomy is a vocabulary of tags emerging from community of users, not a vocabulary defined by a single user nor by an outside party. We see that in this area, tagging and the resulting folksonomies could be used to support this process of leveraging cross-domain knowledge. As users are given the freedom to create arbitrary tags, it is suspected that more of the otherwise hidden aspects of the vocabulary construction could be tracked.

Tagging, however, has several challenges that need to be addressed. Vocabularies based on tagging data 1) may evolve rapidly, 2) may be incomplete or inconsistent, 3) may not have clear, unambiguous interpretations for specific tags and 4) may not be readily organised into tight hierarchical or taxonomic structures. As such, it is not all obvious, what is the most effective way to display tagging based data. Fixed presentations typical to most knowledge organisation systems (like hierarchies and taxonomies) may not be readily used. Associative navigations based on binary relationships between tags and resources, on the other hand, can be used, but do not scale well to large numbers of tags and resources.

In our approach, we use a combination of data mining and information visualisation to generate more flexible presentations of the tagging data. We see that the freedom in visualisation for choosing the visual properties (shape, colour and position) of the object, offer an opportunity for recreating new and more efficient abstractions. Ideally, this would mean that not only current relationships could be better understood, but also discovery of hidden properties of the data could be enabled.

Our study to tagging data visualisation is organised to in this paper as follows: section 2 defines basic concepts used throughout the rest of the work. In section 3, we explore current approaches used in tagging systems for information retrieval and visualisation. In section 4 we present our approach to tagging data visualisation, based on an implementation of self-organising maps algorithm and a visualisation client in Java. Section 5 concludes and discusses our work.

II. Tagging and folksonomies

Before beginning, let us clarify some of the more generic terminology we use. By *objects* we refer to any individual abstract information entities. *Resources* are digital information objects such as web pages, photos or video clips. In our definition, that has an identity can be tagged is called a resource. Terms resource and object are used throughout this paper interchangeably as differences are somewhat subtle.

Tagging is defined as the process of attaching tags to resources. In the process of tagging, user selects one or several tags. The user performing the tagging is *called the tagger*. *A tag* is a user-defined string, usually a single keyword that is associated to resources in the act of tagging. Note that, in comparison to traditional keyword metadata, tags are not chosen from controlled, third party defined. While old tags may be re-used, tagger may create new ones on the fly. Recommendations based on both user's own and other community members' tagging habits should be made available To support the reuse of existing tags.

We see that the usual motivation of tagging is personal information retrieval. Examples include tagging posts in weblogs for categorisation, tagging songs or videos for playlist generation and tagging pictures for tag-based navigation. This kind of tagging practise much resembles the use of directories for organising file-based resources, with the distinction that a resource may have several all no tags at all, regardless of physical location.

It would make sense that tagging schemes of personal information retrieval would consist of tens, rarely hundreds of unique tags. However, as social bookmarking tools have demonstrated (fc. [1]), is higher number of unique tags surprisingly common. As such, there is a clear difference between the ways how people use tags in the contexts.

The difference on the way of using labels in social software tools from traditional software tools has spurred a lot of discussion. To distinct the traditional use of labelling from socially influenced labelling, Thomas Vander Wal has coined the concept of folksonomy, defined as ”the result of personal free tagging of information and objects (anything with a URL[1]) for one's own retrieval” [2]. The concept is underlain by the idea that tagging does not always lead to creation of a folksonomy. For folksonomy creation, it is required that tagging is 1) personal, motivated for one's own information retrieval 2) done in a social environment and 3) done by the person consuming the information [2]. As according to Vader Wal's point of view, we see that tags in a folksonomy should meet with the following three criteria:

1) *Tags should be personal*. Users may or may not share same keywords for the same resources. Folksonomic tags are – in fact - cumulative, resulting in social indexing of resources in which everyone gets a vote.

2) *Tagging habits should have an influence on the outcome.* The tagging habits of the tagger himself and other community members should have an influence on the outcome tagging.

3) *Tags should not be added automatically.* An implicit assumption in Vander Wal's definition is that the person consuming the information knowingly adds it to his or her personal collection of resources for later reuse.

To distinct different ways of applying tags in social environment, Vander Wal has made a distinction between broad and narrow folksonomies. *Narrow folksonomy* ”provides benefit in tagging objects that are not easily searchable or have no other means of using text to describe or find the object”, very much resembling the way how media organising software tools employ tagging. *Broad folksonomy*, on the other hand, incorporates many people tagging same objects while everyone may choose to apply their own tags. This is also the situation in social bookmarking services, where most of the tagged bookmarks are public and shared. [3].

Let us next consider the formalisation of tagging. While the intuitive interpretation of a tag is a two-place relation between resources[2] and tags, tags are better understood of votes from corresponding taggers. Tom Gruber has formalised tagging as a three-place relation [4]:

```
Tagging(object, tag, tagger)
```

While this relation separates use of tags by different taggers, it still does not commit on the *in situ* nature of tagging. Gruber sees that by adding the source of the tagging data as a fourth property to the relation, may help to understand in which social context the tagging was done [4]. To put it formally, we can define that:

```
Tagging(object, tag, tagger, source)
```

Gruber remarks that while source is easily interpreted as the community in which the act of tagging was done, it may be understood more generically as an explicit notion of source in scope of namespaces or “universe of quantification” for these objects. Whatever the interpretation, it is important that some formalisation of the tagging context is available: tagging objects in a photography service has distinct implications from tagging objects in a scholar reference service [5].

III. Visualising tagging data and folksonomies

A. Current approaches

Perhaps the most common way for creating visually oriented representations of tagging data is the use of tags as search facet. In this way, users may browse resources according to their associated tags. We split these representations into two types: personal and shared tag indices.

Personal tag index consists of only tagging data entered by one person. These indices are seen to contain usually tens, rarely more than hundreds of different tags. For this reason, it is easy to represent personal tag index as an alphabetically ordered list or as another ordered structure.

Shared tag index consists of an aggregation of tagging

1 Universal Resource Locator

2 Or objects, as according to Gruber

data, entered by several users. Aggregations of tags may be used to explore any resources that anyone has tagged with corresponding keywords. In broad folksonomies, indices may be contextual: by defining context as set of users, a contextual set of folksonomic tags may be displayed.

An index of tags may, however, grow larger than what can be feasibly displayed with simple lists. A commonly used approach for dealing with the increasing amount of information, is the use of different kinds of tag clouds. *A tag cloud* is a two dimensional presentation of tags that makes it easier to perceive a large number of tags. Usually in tag clouds, tags are visually weighted according to their use frequency.

To succeed in creation of visualisations, we see that it is especially important to create abstractions that possess meaningful visual interpretation for users. This is not always the case in tag clouds, as figure 1 demonstrates: tags 'java' and 'howto' appear spatial nearby, but the close positioning is more likely based on alphabetical than on semantic proximity.

The lack of meaningful spatial interpretations in tag clouds has already been address for instance by Hassan-Montero and Herrero-Solana. Their solution is based on use of algorithm that organises similar tags close to each other [5]. An alternative solution, as proposed by Bassett is to implement interactivity to tag clouds. In such tag clouds (or "focus clouds"), focusing over a tag will highlight similar tags from the rest of the cloud to support the understanding of associations between different tags [6].

B. *Self-organising maps in tagging data visualisation*

Our approach to visualising tagging data is based on use of self-organising maps (SOMs). SOM (also known as Kohonen networks) is discussed in literature very extensively, and therefore interested readers are encouraged to take a look at the available literature (see especially [8], [9], [10]).

A self-organising map consists of an arbitrary number of neurons. *A neuron* is m dimensional real number vector. These neurons are associated to higher or equal dimensional model vectors, resulting in bijective, one-to-one mapping from neurons to models. In practise this means that map nodes have both [geometric] target space positions and model vectors for source space alignment.

The bearing idea behind SOM is that it can be used to create easily perceivable visual presentations of high-dimensional data sets. Thus, self-organising maps can be used to convert the high dimensions of statistical relationships between tags and resources (or tags and tags) into simple geometric relationships that can be represented efficiently as low, usually two, dimensional maps. (cf. [8])

As an algorithm, SOM is a form of unsupervised learning, based on the neural networks. The algorithm can be roughly split into two phases: 1) Training of the neural network in which initial input data is used to span the map and 2) The use of the trained map either for drawing or query of the map.

SOM is commonly applied for creation of similarity maps from given multidimensional input data [8]. Neurons are associated with selected points of the target space, so that the region is most sensitive to input vectors near that area. (see figure 1). In this way, while the mapping itself is discrete, any input vector can be mapped with SOM.

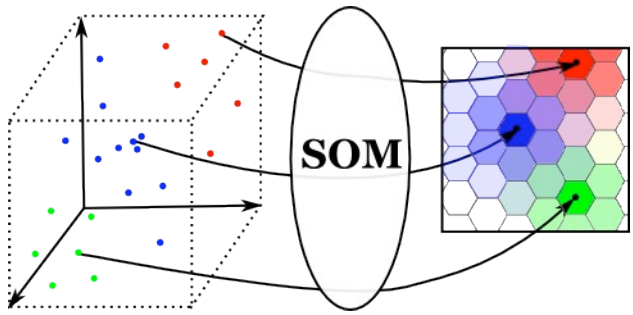

Figure 1: A schematic view of SOM algorithm: neural network is used to assign input vectors to locations in target space

Before the actual learning algorithm takes space, the neurons and their model vectors must be formatted with initial values. For model vectors, these initial values can be randomly selected or algorithmically derived from the input data values.

IV. Experimenting with del.icio.us

A. *Method and data description*

The idea behind our del.icio.us experiment was to compare our SOM based visualisation method with the del.icio.us's integrated tag clouds. Our SOM visualisations were generated using SOM Cloud application: a proof-of-concept Java application created especially for tagging data visualisation. Taking input from an XML[3] file, the application can be used to create self-organising map from virtually any set of tagged resources.

In order to span the self-organising map, a set of n dimensional input vectors was required. In our approach, we simply reserved one input vector component per unique tag. Accordingly, introducing new tagged resource add one new input vector, while every new keyword will add one component to each input vector. Thus, the input data for SOM algorithm consists of data in nxm matrix.

For the study, we collected tagging data from del.icio.us social bookmarking service. The data collections extracted from del.icio.us represent a set of tagging data from a small research community.

The data from del.icio.us included descriptions of both resources and tags associated with them. We experimented with total of eight datasets, from which statistics are compiled into table 1. Seven out of eight datasets were collected from distinct users. The last one, however, was generated from dataset A by removing all tags from first 100 resources. This was done in order to inspect differences in resulting visualisations between tagged and untagged data.

3 eXtensible Markup Language

dataset	distinct resources	distinct tags	tags after dimensionality reduction
A	327	512	188
B	423	582	220
C	402	616	264
D	35	118	21
E	94	84	47
F	49	97	46
G	142	308	52
A'	327	512	188

Table 1: Description of data in del.icio.us experiment

While SOM can be effectively used to reduce dimensions, linear growth in number of dimensions makes map creation exponentially slower. While testing the SOM Cloud application, we found out that when the number of dimensions was around 1000 tags, the visualisation was still usable. Greater number of tags resulted in fast degration of performance. On basis of this experience, it made sense to limit maximum number of tags sent to SOM client. We restricted number of unique tags by excluding all tags that occurred less than thrice. While this approach would not scale to all possible datasets, it was sufficiently effective for our immediate need.

A. Results and discussion

Visualisations of all datasets were successfully generated using both del.icio.us's own user interface and our SOM Cloud application. The tag visualisation readily available in del.icio.us was, however, found out not to scale well to these datasets (see figure 2). When number of distinct tags grew to hundreds, the tag cloud did not any more fit to single page. For this reason, we also ran focus cloud to provide one page visualisations[4].

Figure 2: Del.icio.us's tag cloud with dataset A

4 http://foobr.co.uk/focus/

Figure 3 illustrates how focus clouds rendered datasets. In the visualised focus cloud using dataset A, total of 89 tags were displayed with size of the tags reflecting the number of occurrences. User has moved mouse over the keyword "python" which has resulted in highlighting of 14 associated keywords with yellow background.

Figure 3: Focus cloud with dataset A

The illustrated focus cloud present a usual problem in tag clouds: as layout of tags is not based on tagging data semantics, there is no clear spatial interpretation. Spatially nearby tags may or may not be related. For instance in figure 3 tags 'framework' and 'software+development' are strongly related, while tags 'USB' and 'python' have very little in common. Such organisation of tags is in conflict with well-established laws of perceptual organisation and therefore may give rise to false interpretations (fc. [11]).

In figure 4, the same dataset (A) is visualised by using SOM Cloud. Tagged resources are organised to neurons, visualised as grey balls. Diameter of the circles are relative to number of resources matching the neurons. While neurons are uniformly distributed over the surface, only the ones with associated resources are displayed. As all tags represent one input dimension in the SOM, tags may be assigned to a resource located anywhere in the map. Thus, tags are rarely assigned only to resources associated to a single neuron. Therefore, tags used as labels, may represent only a fraction of tags applied.

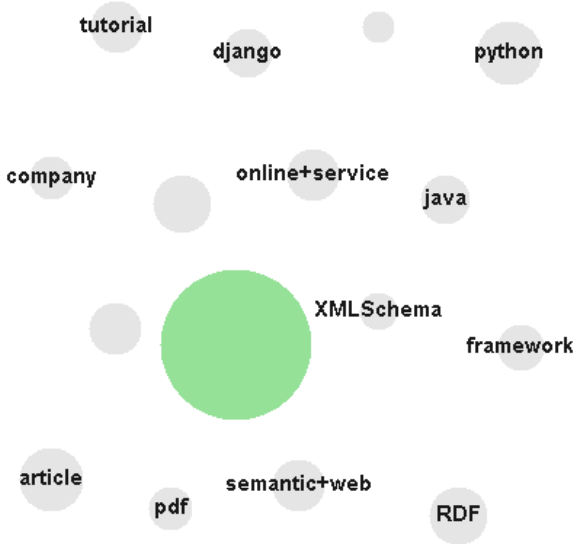

Figure 4: SOM cloud with dataset A

We also implemented a focus feature to SOM cloud. Focus is activated by clicking a neuron, after which focused region is zoomed to fit the screen. As illustrated in figure 5, more details of the neuron are displayed focus mode: tags next in relevance after labelling tag are shown around the circle, along with all resources assigned to the neuron.

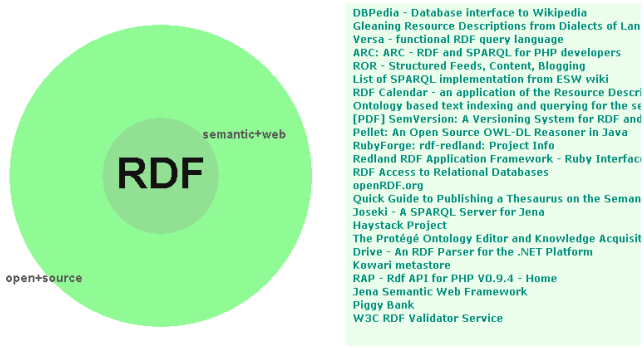

Figure 5: SOM Cloud in focus mode

Our experimenting also pointed out several problematic features in the SOM Cloud. First of all, density of tags in SOM cloud is not always uniform. Labels tend to overlap, making the interpretation of the map more difficult. While most of the overlapping was eliminated by careful text layout, there is no guarantee for overlap not to occur.

Another problematic features that emerged, was the appearance of a large unlabelled neuron, usually located in the centre of the map (also seen in figure 4). The natural interpretation for this feature would be that resources from these region do not contain any tags. This was, however, proven to be false by comparing datasets A and A'.

More likely reason for this feature is that due to the dimensionality reduction, the maps may have less neurons than there are frequently used tags: if two or more sets of completely differently tagged resources fall into same neuron, the neuron may not have a tag that is clearly prevailing. The same applies to the sets of resources that have been assigned multiple, overlapping tags: in such cases, no tag alone best describes these resource sets.

The empty region also reminds us of the fact that there is a fundamental mismatch between SOM and tag cloud topologies. A single neuron in SOM has fuzzy relationships with all tags, while in tag clouds the relationships are exactly one to one. As such, the "SOM cloud" visualisation would better fit the algorithm, if neuron labels could consist of several tags instead of a single one.

V. Conclusions and Future Directions

In this article, we presented idea behind resource tagging and folksonomies, described current approaches to tagging data and folksonomy use and visualisation and described and experimented with a SOM-based tag cloud visualisation method.

In our study, SOM Cloud visualisations converted surprisingly well the high-dimensional tagging data into easily perceivable tag cloud like visualisations. While we were generally satisfied with the overall quality of the resulting visualisation, the SOM Cloud has features that need to be taken into account. Especially the mismatch between SOM structure and visualised one-to-one associations between tags and neurons should be understood as a potential source of false interpretations.

It should be stressed that the power of SOM does not lie on the algorithm, but rather on how it is used: the selection of correct input components is for crucial in creation of meaningful SOM visualisations.

As for future research, we suggest that usability of the SOM cloud visualisation should be further investigated. Especially, efficiency of information retrieval using SOM Cloud, could further be studied.

Acknowledgment

This work has been supported by the European Comission (IST network of excellence project OPAALS of the sixth framework program, contract number: FP6-034824).

References

1] B. Lund, T. Hammond, M. Flack, T. Hannay. "Social Bookmarking Tools (II) – A Case Study - Connotea", D-Lib Magazine, April 2005, Volume II Number 4. ISSN 1082-9873.

2] T. Vander Wal. "Folksonomy Definition and Wikipedia", vanderwal.net, November 2, 2005, Available at http://www.vanderwal.net/random/entrysel.php?blog=1750 [accessed 13.9.2007].

3] T. Vander Wal. "Explaining and Showing Broad and Narrow Folksonomies", personalinfocloud.com, February 21, 2005, Available at http://www.personalinfocloud.com/2005/02/explaining_and_.html [accessed 14.9.2007].

4] T. Gruber. "Ontology of Folksonomy: A Mash-up of Apples and Oranges", 2005, Available at http://tomgruber.org/writing/ontology-of-folksonomy.htm [accessed 5.9.2006].

5] Y. Hassan-Montero, V. Herrero-Solana. "Improving Tag-Clouds as Visual Information Retrieval Interfaces". I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, 2006. To appear.

6] A. Bassett. "Focus Cloud [concept] – What's my current interest", Foobr weblog, June 23, 2007, Available at http://foobr.co.uk/2007/06/focus_cloud_concept/ [accessed 13.9.2007].

7] C. Ware, "Information Visualization, Second Edition: Perception for Design", Elsevier, 2004.

8] T. Kohonen. "Self-organizing maps", third edition, Springer, 2001. ISBN 3-540-67921-9.

9] J. Feldman. "Neural Networks – A Systematic Introduction", Springer-Verlag, 1996. Berlin, New-York. ISBN 3-540-60505-8.

10] E. Oja, S. Saski. "Kohonen Maps", Elsevier Science, 1999. ISBN 0-444-50270-X.

11] M. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, T. Carey. "Human-Computer Interaction", Addison-Wesley, 1994. ISBN 0-201-62769-8.