Visualizing Irish Banter

Tristan Webb, Ian Dickinson

HP Laboratories
HPL-2009-83

Keyword(s):
visualization, semantic web, RDF, self organizing map

Abstract:
Discussion forums could benefit from an exploratory search capability. The large quantity of information on discussion forums can present an opportunity for helping a user through visualization. Self organizing maps have been successfully used as a basis for visualizing high-dimensional data. Our Forum Map prototype, created as an entry into the 2008 SIOC data competition, uses a self organizing map to lay out the "member-space" of a discussion forum. A self organizing map can be used to create an overview of the forum, and an innovative interface can connect this visualization to the content of forum postings through labels and user interaction. Note: This research was conducted over the course of the first author's internship in the Enterprise Informatics Lab at HP Labs, Bristol, during 2008.

External Posting Date: April 21, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: April 21, 2009 [Fulltext]

Visualizing Irish Banter

Tristan Webb[1,2]
tristan.webb@warwick.ac.uk
Ian Dickinson[2]
ian.dickinson@hp.com

[1] University of Warwick
[2] HP Laboratories Bristol

Abstract. Discussion forums could benefit from an exploratory search capability. The large quantity of information on discussion forums can present an opportunity for helping a user through visualization. Self organizing maps have been successfully used as a basis for visualizing high-dimensional data. Our Forum Map prototype, created as an entry into the 2008 SIOC data competition, uses a self organizing map to lay out the "member-space" of a discussion forum. A self organizing map can be used to create an overview of the forum, and an innovative interface can connect this visualization to the content of forum postings through labels and user interaction.

1 Introduction

Contained in Internet discussion forums are answers to questions, opinions on topics, and a social network of members. However, points of interest to a particular user may be hidden by vast quantities of irrelevant messages. The interface that was designed to engage a user in conversation may lack the features to make searching for posts easier. We have identified some key problems with the interface to be that:

- the conventional presentation suffers from a lack of overview
- the basic keyword search query is often an inadequate filtering mechanism
- the navigational layout makes it difficult to quickly find specific content.

In this paper we describe a software prototype called Forum Map. Our Forum Map was created with the goal of displaying an overview of the individuals involved in electronic communication. Its layout attempts to group forum members who converse with one another in similar locations. It also draws labels on the visualization at coordinates according to this layout. The labels displayed are keywords extracted from forum members postings. The viewer gets a sense of who is talking with who, and the nature of these conversations. The goal of the Forum Map is to provide an easy to understand interface for a user to investigate topics of conversations discussed between different members of the forum.

In September 2008, the popular Irish Internet discussion forum *boards.ie* released the entirety of its data to commemorate the site's tenth anniversary. Associated with this release was a data competition run by the SIOC project. The SIOC project has

developed an ontology and a set of RDF vocabulary that describe discussion methods on the web. The data set that was released used this meta-data to describe the content of the forum.

The SIOC data competition was set up in order to encourage research in social ontologies and the semantic web. The organizers specified only that contestants "do something interesting with the boards.ie SIOC Data Set"[3]. Our Forum Map was envisioned based on ideas found from visual analytics, and an implementation was submitted as an entry into the competition.

Visual analytics is the study and creation of interfaces to large knowledge bases in order to provide insight to a viewer. The goal of visual analytics is about changing information overload from a liability into a resource[11]. In order to not miss the opportunity to communicate data to a viewer the visualization must be clear in its presentation and convey useful information. In a board with over 10 million posts, exploring the site at the post level is not practical. Instead, to help users we must follow the information seeking mantra as stated by Schneiderman: "Overview first, zoom and filter, then details on demand."[21]

Our Forum Map is innovative in combining the text of the underlying content with an overview of the member-space. The main visualization is created with a self organizing map (SOM), a form of neural network first described by Teuvo Kohonen[13]. The construction of the self organizing map allows for members to be placed onto a two dimensional grid meaningfully. The data explorer can display the content of the posts by selecting different regions of the SOM. The resulting overview is designed to help in *exploratory search*.

In exploratory search the domain or the goals of the search may be unclear. For instance, users may only be able to submit a tentative query to begin their search[1]. Marchionini explains that this type of search is distinct from traditional "lookup" queries[17]. Recently, researchers have designed interfaces to explore electronic communication and social networks using drawings of networks[9,7]. The Forum Map uses a meaningful layout instead of a network view to help guide users to their search goals.

The next section will detail the structure of the boards.ie forum data, and describe the problem of abstracting the data and transforming it. This paper will then explain the fundamentals of our approach, and an evaluation of the results.

2 Problem Structure

The amount of data released for the SIOC[5] data competition is 2.1GB of plain XML/RDF text. From its beginnings as a forum to discuss the video game Quake, boards.ie has grown to encompass over 700 sub-forums. As of November, 2008 boards.ie is the site with the most traffic in Ireland, with over one million unique visitors per month[19]. The structure of competition data is shown in Figure 1.

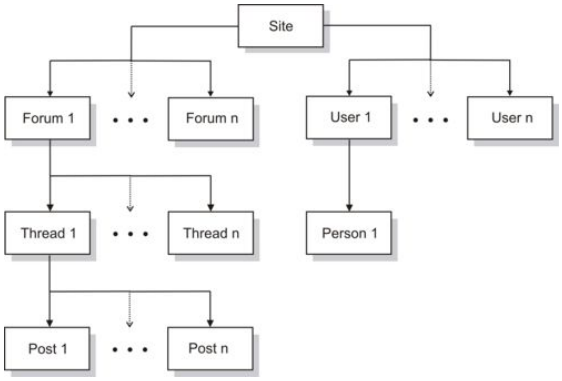

Fig. 1. The structure of the boards.ie forum. Figure provided by the SIOC Project.

The threaded nature of a discussion forum is conducive to conversing, but not to showing an overview. A user may have difficulty when presented with the challenge of keeping up with or exploring a topic. She/he may have difficulty because a single topic may be split across multiple posts and threads. Threads can be pushed off the front page of a forum by newer threads in a matter of hours.

Abstraction can assist in providing a human viewer with an overview of a large dataset. Too much information presented in text format will overtax the viewer. The first step in the abstraction process according to Schneiderman is *overview*. A visualization can present an overview a large dataset by only displaying the most significant features. The human brain is fast at analysing visualizations. Using them opens the door to exploring large knowledge bases. Both the choice of visualization and the method by which one chooses to abstract details are interesting problems.

To many researchers in the field of visual analytics, the transformation of the dataset, and the design of the interface are the key points of interest[24]. The field has also identified that users of exploratory search may be interested in combining their search with some query to guide their search[17]. On the other hand, in exploratory search we can not assume that a user is sure of the goals of thier search. The choice of interface will depend on the use cases we are trying to solve. The Forum Map interface attempts to answer the two use case questions:1) who is talking to whom, and 2) what are they talking about? These two use cases are addressed by the layout and labelling system.

We based our interface on the hypothesis that a user would be interested in looking at a visualization where different regions correspond to different users. Similar users will be aligned close to each other in the visualization, and will have a personal region associated with them. Displayed on the visualization will be labels which describe the topics that this user has been talking about. There can be regions that overlap, and here the labels would describe conversations that users have had with each other.

The Forum Map is a transformation of the original threaded dataset to a visualization that groups similar users into regions that are in close proximity. This results in an overview that shows relationships between users that are not apparent in a conventional view. The result is a clearer presentation of the data for exploratory search.

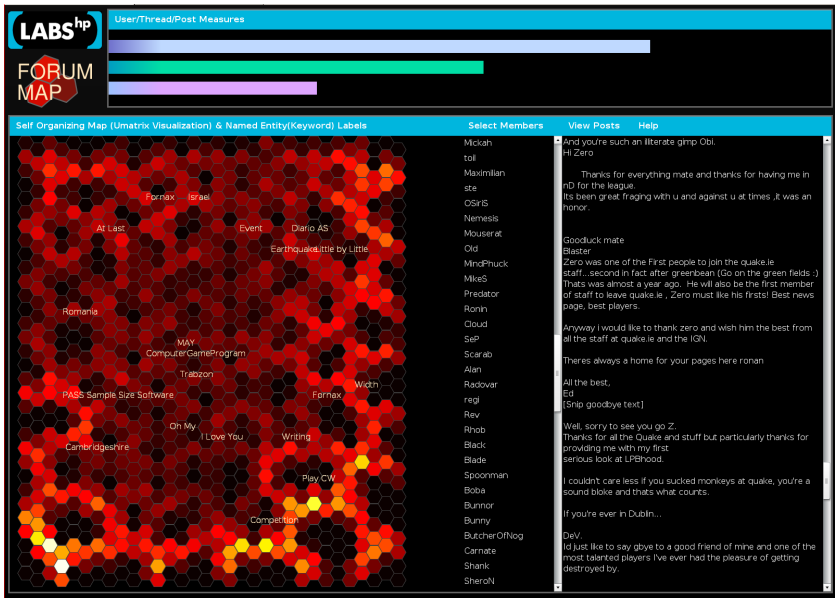

Fig. 2. Illustration of the forum map.

3 Methods and Approach

The power of modern personal computers has made it possible to visualize high dimensional data more quickly than ever before. Processing high dimensional data has brought with it many challenges of how to represent this data. One of these challenges is projecting a high dimensional space onto a 2-D surface, without sacrificing too much information in the process. Many different techniques have been developed in this field of dimensionalilty reduction.

We have chosen to use SOMs here, but there are other techniques that could have been used to visualize the same data. SOMs are, in effect, a non-linear form of principal component analysis (PCA)[10], and share similar goals to multidimensional scaling (MDS)[14]. PCA is much faster to compute, but it has disadvantage of not retaining the topology of the higher dimensional space. MDS is a technique that will preserve distances in the higher dimensional space along with topology. The disadvantage of using MDS is much greater computational complexity compared to SOMs. SOMs strike a good balance between the two. Hybrid approaches different methods have been developed as well[27].

Self organizing maps are a form of neural network with a competitive learning rule which is sensitive to the past history and spatial features of the network[15]. Since the introduction of SOMs in 1985, where they were designed to explain the ordering that is present in the neurological functioning of the brain[18], they have been used in various applications from visualizing gene expression[23], to approximating continuous

functions[20]. SOMs have also been used extensively to success in the areas of data mining and visualization[26].

SOMs are created with a simple and elegant algorithm that preserves the local topology of a high dimensional space. In the world of high dimensional visualization, preservation of topology has become known to mean "that similar input patterns from an input space are projected into nodes that are close to each other in the output space."[22] The SOM algorithm can be thought of as a mesh that has first been cast over a high dimensional data cloud. This mesh is gradually tightened until it has captured the local topology of the data.

3.1 SOM Definition and Application

SOMs use an input pattern comprised of different data vectors, denoted by $\boldsymbol{x}$, in an N dimensional space. This is a data set which we wish to visualize in a lower dimensional space. A self organizing map will project the input pattern onto an output space, usually a regular two dimensional array of nodes as shown in Figure 3(a). Every output node has a fixed position on the grid and has a reference vector in $\mathbb{R}^N$ associated with it. Nodes are denoted by $\boldsymbol{w}_{i,j}$ where (i, j) specifies the output node's position on the two dimensional grid, i.e. nodes are assigned to a coordinate system. Each data vector in the input pattern will be associated with exactly one node in the output space, known as its *best matching unit*(BMU). An input vector's best matching unit is defined to be the node whose reference vector is closest in Euclidean distance. Many different vectors from the input pattern may share the same BMU.

A self organizing map is created through an *unsupervised learning algorithm*[16]. Before training, the map must be initialized to specify its size and the initial values of its node's reference vectors. A commonly used method is to initialize every node's reference vector to random values that are evenly distributed in the area of the corresponding input data vector components. During the training phase, nodes that are in close proximity to each other will learn to be activated by similar inputs [12]. The neighborhood is defined by a *neighborhood function*, defined on the grid, such as the simple bubble function shown in Figure 3(b). Training examines one vector in the input pattern at a time. As each vector is considered, the reference vector of its BMU and the reference vectors the BMU's neighbors on the grid will be updated to this rule:

SOM Learning Rule: If we let i^* denote the index of an input vector's BMU, and let I^* be the neighborhood of nodes defined by a function of proximity to i^* on the two dimensional output grid,

$$\boldsymbol{w}_j(t+1) = \boldsymbol{w}_j(t) + \alpha(t)\Lambda(j, i^*)\boldsymbol{x}(t)$$

for all $j \in I^*$, where t is the time coordinate, $\alpha(t)$ is a learning rate parameter, and $\Lambda(j, i*)$ is the neighborhood function. $\alpha(t)$ can be varied from large values if we want to make big changes at each iteration, or to small values for fine tuning as displayed in Figures 3(b) and 3(c). The learning rate and neighborhood function should be both monotonically decreasing in time[12].

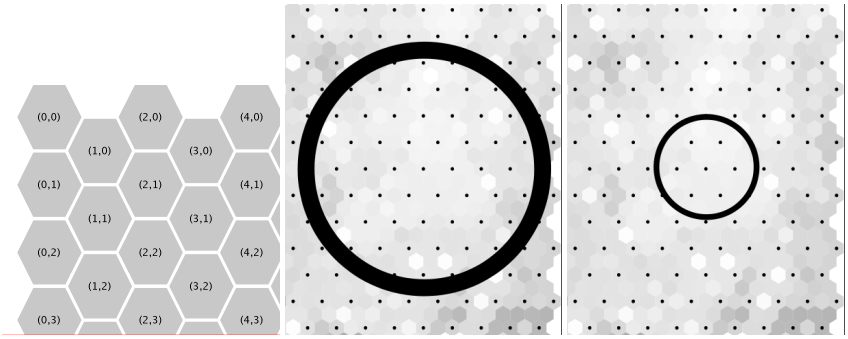

Fig. 3. (a) A visually appealing regular hexagonal lattice. (b) A bubble neighborhood function. (c) The neighborhood function will shrink as training progresses.

To illustrate the SOM learning rule consider the simple example where input pattern is drawn from a uniform distribution in $\mathbb{R}^2$. Training will evenly spread the maps reference vectors over the range of the input space. The final map will also have reference vectors which gradually change in value from neighbor to neighbor.

3.2 Forum Map Implementation

The Forum Map gathers user's posting habits as an input pattern to train the SOM. Posting habits are derived from the number of times each user posts into different threads, which are the lowest level "container" which holds posts from different users. Forum members who post into the same threads a similar number of times will be grouped by the SOM.

The data released for the competition was loaded into Jena to create the input to the SOM. Jena is a framework for the semantic web, which can iterate through the triple statements inside a model or dataset[4]. The availability of the data set in RDF and the Jena framework made the process analysing the data much more high level than if these tools were not available.

The basic construction of each input pattern follows. The input pattern consists of user vectors, which total number every member of the forum. For each post in a thread, the user vector associated with that post's creator is increased by one. In this way, the user vectors describe how many times a user has posted into each of the forums threads. The number of dimensions for each user vector is the number of threads on the website. The ith component in a user vector X corresponds to the number of times user X posted into the ith thread.

These user vectors as input to train a randomly initialized SOM. Users that had less than $U_{\min}$ posts, and threads that contained less than $T_{\min}$ posts were not considered. Our implementation took $U_{\min} = 10$ and $T_{\min} = 3$. Many different map sizes were tested, and a 32x32 array of nodes was chosen because it struck a balance between speed of computation, and quality of the final result. The training was broken down into different phases, with different parameter values for the SOM learning rules.

The way a SOM is visualized is very dependant on the goal in mind. The u-matrix approach has been developed to show basic clusters in the map[25]. In this visualization, nodes are coloured dark colors if the average Euclidean distance to their neighbors is small, and light colors if they are farther away. One can easily spot valleys where the map shows nodes that are similar to each other, and hills that correspond to a contrast in the values of nodes. This visualization technique gives an intuitive explanation to the features of the forum. For example, the first year of the boards.ie data shows a clear body of many members which have their nodes separated by dark colors. One could interpret this to be the "masses" on the forum. The forum administrator on the other hand is positioned near the "masses", suggesting that he interacts with them, yet at the same time has light colors around his region, suggesting that his posting habits are quite different.

Phase	Iterations	Neighborhood Radius	α
1	10'000	23	0.7
2	40'000	17	0.5
3	60'000	9	0.3
4	100'000	5	0.15
5	150'000	2	0.1

Fig. 4. Parameter values used to train a SOM of the first year of the boards.ie data.

The interface to the SOM was programmed in Adobe Flex[2], which is a framework for creating rich Internet applications. The main display is a statically generated SOM, which can be repositioned to be centered around any one user. In the prototype, initially labels are drawn on the map either to indicate a user's BMU.

Clicking on a hex can be used to access the threads associated with that hex. In the current implementation clicking on the hex will display the content of the threads which have the greatest values in that node's reference vector.

The map can be centered around a user by selecting them from a list. This will change the map to display named entities extracted from the conversations of other users in the forum that have posted into the same threads. The map is first redrawn to be centered on the selected user. Each user that has posted into the same thread as the selected user will have their post content from those threads compared to named entities in the UMBEL Ontology[6]. This process produces many different named entity matches, of which a few randomly selected choices are chosen and displayed as conversation keywords.

3.3 Future Work and Evaluation

The forum map represents early work in designing an interface for analysing electronic communication. User trials have not been conducted at this time. These observations are the authors' alone. Evaluations focus on the effectiveness of the visualization.

The choice of a SOM raises some issues regarding the meaning of the finished visualization. Convergence of SOM has only been analyzed in depth for the one-dimensional case[8]. In most cases sufficient numerics have been observed to be sufficient to converge the weight vectors in the map to stable values. Still, the final layout is very dependant on the random initialization of the reference vectors.

The SOM is pre-generated because of the computational cost of training it. The map can be repositioned, but not dynamically generated by the interface. While this

may suffice for many cases, a static image is not as powerful an exploratory tool as a dynamic one.

While on the subject of improving the descriptive capabilities of the SOM, we should mention that choosing a three dimensional space of nodes would capture considerably more of the features of the forum than a two dimensional grid. The extra degree of freedom in the three dimensional space would also allow for more of member's regions to overlap.

A missed objective of the forum map is with the conversation labels extracted from the content of the posts. The named entity extractor's random selection of labels does not capture more abstract details about the conversations. The named entity extractor does not always display most relevant labels to the post content, or the label may reflect the post content relating to one specific word instead of encompassing the entire post. Along with improving the named entity extractor, another avenue that could be explored is the position of the labels. Currently the labels are drawn on the BMUs of users. Our suggestion for improvement is creating shared labels based on topics discussed by many users. Drawing these shared labels on the map to display the weight that topic is associated with different users could bring out more informative detail.

4 Conclusion

The visual display of data has been made more necessary than ever before by the capacity of large databases. Many different visualizations exists, and the choice of which one to use must be tailored to the type of information we wish to extract. A self organizing map can show the major features of the data in an overview, as well helping the user of the interface discover information that was not obvious before. As content extractors improve, an data explorer can quickly see the topics of discussion.

References

1. Supporting exploratory search. *Introduction to Special Section of Communications of the ACM*, pages 36–39, 2006.
2. Adobe - Flex 3, Website accessed December 2008. http://www.adobe.com/products/flex.
3. Home of the boards.ie SIOC data competition, Website accessed November 2008. http://data.sioc-project.org.
4. Jena - a semantic web framework for Java, Website accessed November 2008. http://jena.sourceforge.net.
5. sioc-project.org — semantically-interlinked online communities, Website accessed December 2008. http://sioc-project.org.
6. UMBEL: Upper mapping and binding exchange layer, Web accessed November 2008. http://www.umbel.org.
7. Justin Donaldson, Micheal Conover, Benjamin Markiness, Heather Roinestad, and Flippo Menzcer. Visualizing social links in exploratory search. In *Hypertext and Hypermedia*, pages 213–218. ACM, 2008.
8. J.C. Fort. SOM's mathematics. *Neural Networks*, 2006.
9. J. Heer. Exploring Enron: Visualizing ANLP results. *Applied Natural Language Processing*, 2004.

10. I.T. Jolliffe. *Principal component analysis*. Springer New York, 2002.
11. Daniel A. Keim, Florian Mansmann, Jorn Schneidwind, and Harmut Ziegler. Challenges in visual data analysis. In *Information Visualization*, pages 9–16. IEEE, 2006.
12. Teuvo Kohonoen, Jussi Hynnien, Jari Kangas, and Jorma Laaksonen. The self-organizing map package. Technical report, Helsinki University of Technology, 1995.
13. Teuvo Konohen. *Self Organizing Maps*. Springer Series in Information Science. Springer, 2001.
14. J.B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.
15. S. Y. Kung. *Digital Neural Networks*, pages 85 – 87. Prentice Hall, 1993.
16. David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter 5. Cambridge University Press, 2003.
17. Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 2006.
18. James Matthews. Interview with Teuvo Konohen. *International Journal of Advanced Robotic Systems*, 2005.
19. Aliien O'Toole. State of the net: essential ebusiness intelligence for Irish managers. (11), Winter 1998.
20. Ral Rojas and Jerome Feldman. *Neural Networks: A Systematic Introduction*, chapter 15. Springer, 1996.
21. Ben Schniederman. The eyes have it: A task by data type taxonomy for information visualization. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.
22. J. Si, S. Lin, and MA. Vuong. Dynamic topology representing networks. *Neural Networks*, 2000.
23. Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Suitisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 1999.
24. James J. Thomas and Kristin A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, pages 10–13, 2006.
25. A. Ultsch and C. Vetter. Self-organizing-feature-maps versus statistical clustering methods: A benchmark. Technical report, University of Marburg, 1994.
26. Juha Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 1999.
27. Sitao Wu and Tommy W.S. Chow. PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks*, 2002.