Kohonen Self-Organizing Maps in Clustering of Blog Authors

Eddie Yee Tak Ma

***Abstract*—In this work, Kohonen's Self Organizing Map (SOM) is used as a method to cluster blog posts together. The blog posts are transformed into bitvectors according to the selection of words present in each post. Each index of all bitvectors correspond to a unique word in the pooled corpus of all words seen in the study. We determined that words which occurred within 1% to 7% of the posts allowed the system to perform well. The SOM was successful in grouping authors together and arranging them in a gradient based on similarity. The system did not produce many discrete clusters however. Blogs with the greatest number of posts contributed to the study have the greatest homogeneity while blogs with the least number of posts have the highest separation.**

I. Introduction

In this study, samples of writing are clustered using Kohonen's Self Organizing Maps (SOM)s[4]. This clustering is performed using blog posts as exemplars. Authors contribute their blogs to the study and all unique words from all blogs are pooled together. These words are then sorted based on frequency of use and subsorted in alphabetical order. This unique sorting provides the indexing for a bitvector scheme used to represent each blog post. Since each word occupies a unique index, the presence of that word in a particular post corresponds to a $\{1\}$ in its bitvector at that index. The absence of that word corresponds to a $\{0\}$ in the post's bitvector at that word's index.

SOMs have been used in the past for the characterization of writing. In Section I-A, we overview notable related work.

A. Related Work

The bitvector encoding that has been chosen in this study has appeared previously in the literature. It has been seen for example in the analysis of Spanish language e-mails to find the topics of frequently asked questions [5].

More sophisticated representations include Caseframe Features and Lexical Features used to categorize political blogs as left-leaning or right-leaning in political alignment [3]. These representations take into account the occurrence of a word in the context of a *verb-phrase*, *noun-phrase* pair or a *noun phrase* used as a seed to capture neighbouring words respectively.

Statistical methods also exist that incorporates not only word use frequency but also regularity of use as contextual information into the feature vectors [6].

Furthermore, the abstract notion of style indicated by syntax choice as well as (or without) lexical choice has also received fair attention. These methods may divide recognition at different hierarchical levels of syntax (token-, phrase- and analysis- levels) [8].

Although these methods are interesting, this study opts for a simpler design in order to evaluate the efficacy of word frequency alone for this particular corpus leaving such enhancements for future works.

The new dataset that is gathered for this study is described next in Section I-B.

B. Dataset

The data used in this study is summarized in Table I and Table II. The former lists the blog authors along with the total number of posts contributed to the study and the number of unique words used by that author alone. We of course expect that there is an intersection of words between various authors in order for clustering to occur. The latter table lists the authors and the subject matter of their blogs.

TABLE I
Blog authors by ascending given name. |Posts| indicates number of posts contributed and |Lexicon| is the total number of unique words used.

Author	|Posts|	|Lexicon|
Andre Masella	198	7953
Andrew Berry	46	2630
Arianne Villa	41	1217
Cara Ma	12	854
Daniela Mihalciuc	211	4454
Eddie Ma	161	5960
Jason Ernst	61	3445
John Heil	4	712
Lauren Stein	91	4784
Lauren Stein (cooking)	7	593
Liv Monck-Whipp	30	398
Matthew Gingerich	98	395
Richard Schwarting	238	7538
Tony Thompson	51	2346

The blogs were crawled and posts compiled on 2011, February 28. Blogs were chosen with overlap in subject matter but also some diversity in topic. Authors chosen for this study are either enrolled in, or are graduated from the *University of Guelph* or the *University of Waterloo*. Since these individuals are academically inclined, we expect the word choice to be biased.

The notion of a word should be precisely defined. In this study, we treat substrings composed only of plain Latin characters (without accents) and apostrophes to be words. All other characters are delimiters (including whitespaces, hyphens and other punctuation). No corrections are done to spelling and word stems do not matter ($cat \neq cats$). Apostrophes rendered with the HTML entity $\{\&\#8217\}$ are replaced with the apostrophe available in ASCII.

Eddie Ma is with the School of Computer Science, University of Guelph, Guelph, Ontario, Canada (email: ema@uoguelph.ca).

This work is submitted in partial fulfillment of course requirements for CIS 6050 – Artificial Neural Networks.

TABLE II
AUTHORS AND THE SUBJECT MATTER OF THEIR BLOGS.

Author	Blog Subject Matter
Andre Masella	Synthetic Biology, Cooking, Engineering
Andrew Berry	Drupal, Web Development, Gadgets
Arianne Villa	Internet Culture, Life
Cara Ma	Life, Pets, Health
Daniela Mihalciuc	Travel, Life, Photographs
Eddie Ma	Computing, Academic, Science
Jason Ernst	Computing, Academic
John Heil	Science, Music, Photography
Lauren Stein	Improv, Happiness, Event Announcements
Lauren Stein (cooking)	Cooking, Humour
Liv Monck-Whipp	Academic, Biology, Science
Matthew Gingerich	Academic, Synthetic Biology, Engineering
Richard Schwarting	Academic, Computing, Linux
Tony Thompson	Circuitry, Electronic Craft, Academic

A total of 1250 blog posts are included in this study over 13 authors contributing 14 blogs total (Lauren Stein contributes 2 blogs). After all processing and aggregation of words, a corpus of 21418 words results.

Let us summarize the corpus of words using two visualizations. We introduce the notion of a *rank class* of words. A *rank class* is a set of words which appear in the same number of posts. In the first impulse graph (Figure 1), the frequency of use for the words in the corpus are sorted by rank class. All of the words that have the same frequency are thus shown by the same impulse.

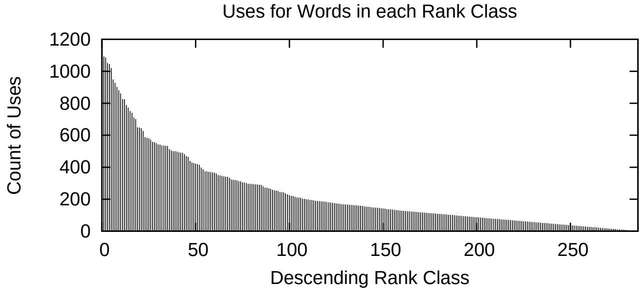

Fig. 1. Impulse graph of words based on descending frequency of use. Each impulse is shared amongst all of the words with the same frequency of use.

Note that the frequency of use descends to zero. This curve is not precisely logarithmic as it drops sharply toward the tail.

The second visualization (Figure 2) shows the number of words in each rank class. It is as expected that there are more rare words than there are common words. These two visualizations complement each other in that a pair of corresponding impulses (one from each graph) fully describes the number of words their occurrences of each rank class.

The ascent in the number of words per impulse experiences exponential growth. The final rank class associated with single-use words is particularly large and contains 51.5% of the words (11020 of 21417). This may however be an outlier considering the number of misspelled words, invented words and specialized jargon that appear in this rank class.

After some preliminary trials, it is determined that using all 21417 elements per vector is not feasible. We therefore use the first $[1,7]\%$ of the corpus. This selection results in 996 words used to construct the bitvectors. The target number was 1000 words, but using 996 allows us to retain full rank classes. The indices of words selected are in $[356, 1351]$.

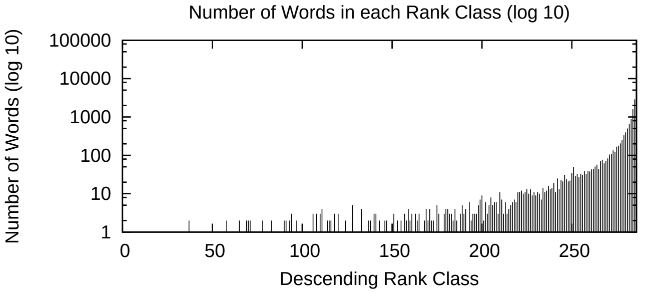

Fig. 2. Impulse graph of words based on ascending rarity. Each impulse is a count of the words which share the same frequency of use. *Note that there are many impulses of size-one along the horizontal axis.*

The remainder of the paper is organized as follows. The methods are next, (Section II) which includes a discussion of the parameters used (Section II-A) and is followed with experimental design (Section II-B). Results are then shown in Section III and is divided apart into graphical results based on the visualization of the SOM map (Section III-A) and a numerical analysis (Section III-B). Finally conclusions are drawn in Section IV.

II. METHODS

We discuss the parameters used to train the SOM and the experimental design in this section.

A. System Parameters

We use the following parameters in the SOM. Since we are using a square configuration of output units, a length of the square is equal to the floored-square-root of the total number of exemplars ($\lfloor\sqrt{x}\rfloor = 35$).

We randomly present exemplars one at a time for ($N = 12500$) training cycles (which is ten times the total number of exemplars). Since this presentation is random, it is not guaranteed that each exemplar is presented.

We adopt the adjustments as seen in Buckland [1] in order to compute the remaining parameters used for the SOM: the initial neighbourhood ($\sigma_0 = 17$) (1); The initial training rate ($\alpha_0 = 0.7$ given preliminary trials); the neighbourhood decay ($\Gamma_\sigma = 4411$) (2); and the training decay ($\Gamma_\alpha = 12500$) (3).

$$\sigma_0 = \frac{\lfloor\sqrt{x}\rfloor}{2} \tag{1}$$

$$\Gamma_\sigma = \frac{N}{\ln(\sigma_0)} \tag{2}$$

$$\Gamma_\alpha = N \tag{3}$$

These equations make use of parameters previously set in order to reduce the number of decisions that must be made. In preliminary trials, these calculated parameters produced reliable results and are thus retained. Note that the decay rates behave as we would like by starting high and approaching zero just as the SOM completes training.

B. Experimental Design

Twelve trials of this experiment are run for the same number of training cycles each. Since a C (c99) implementation of the SOM is used on a UNIX system, the srandom() and random() functions are used to initialize the weights of the SOM. For the sake of reproducibility, each of the twelve trials uses one of the integers in Table III as a random seed.

TABLE III
THE TWELVE RANDOM SEEDS FOR SRANDOM(), ONE FOR EACH TRIAL.

101297	102917	112970	119702	210197	210917
219017	721091	791019	917012	917210	972101

We have covered the experimental design and can now reveal the results.

III. Results

We first visit the graphical results and pair it with the semantic or qualitative analysis of the various densities of the map. We then describe the numerical results after having explained the quantitative values chosen for this task.

A. Graphical Results

Figure 3 shows a representative SOM from one of the twelve trials after the completion of training, and exemplars are clustered.

On inspection, we can see that there is a prominent density in the centre of the map. This density is visible in the SOMs of all other trials, but is not always located at the centre. The posts that fall into this region are either short posts or posts that are so unique that they do not contribute to the definition of the bitvector.

Table IV shows the regions of the SOM corresponding to the example trial. These semantic regions are not located in the same place in each trial, but posts may still be placed together based on topic.

Here are a few observations about the arrangement of posts. Posts by one or more authors that are semantically similar are placed closer together. These regions are not perfect in that other authors' posts can also be found within and some posts are not of the same topic. Authors which write on particularly different topics than the remaining authors are more likely to have unique, well-defined and isolated clusters where the vast majority of posts belong to that single author (e.g. Lauren Stein's improv posts and Dana Mihalciuc's travel posts).

B. Numerical Scoring and Quantitative Results

Three values are used in this study to describe how well the posts of different authors clustered. This evaluation is done after training is complete and all exemplars have been placed onto the SOM. To evaluate how well each exemplar clustered together, the value *homogeneity*[2] (h) is used; this value is given by the average distance between members (u_i, u_j) of a given class (U) (equation 4). A smaller homogeneity is a better score as it indicates that members of the same class are packed closer together.

TABLE IV
THE SEMANTIC REGIONS OF THE SOM WITH RANDOM SEED 721091. PROBING EACH OF THE DENSE REGIONS OF EACH AUTHOR REVEALED THAT POSTS BELONGING TO A SINGLE AUTHOR ARE DISPERSED INTO SEPARATE AREAS WITH DIFFERENT TOPICS. REGIONS MAY ALSO CONTAIN WORK FROM MULTIPLE AUTHORS.

Region	Authors	Topics
Top Left	Liv	Academic Journals
	Eddie	Software Projects
	Jason	Academic
Top Border	Lauren	Human Idiosyncrasies
	Richard	Linux
Top Right	Lauren	Improv
Up & Left of Centre	Daniela	Travel
Centre	(*all*)	(*short posts*)
Right Border	Andre	Cooking
Just Below Centre	Matthew	Software Projects
Bottom Left	Andre	Language Theory
	Andrew	Software Products
	Jason	Software Products
Bottom Border	Richard	Academic
Bottom Right	Eddie	Web Development
	Jason	Software Tutorials

$$h = \frac{2}{|U|^2} \sum_{i=0}^{|U|-1} \sum_{j=i+1}^{|U|-1} \text{dist}(u_i, u_j) \tag{4}$$

The notation ($|U|$) returns the total number of exemplars in class (U). The function dist() returns the physical distance across the SOM for the two exemplars indicated.

The value *separation*[7] (s) is used to describe how unique each cluster is; this value is given by the average distance between members of one class ($u_i \in U$) against members of all other classes ($v_j \in V$) (equation 5). A larger separation is better as it indicates better contrast between classes.

$$s = \frac{1}{|U||V|} \sum_{i=0}^{|U|-1} \sum_{j=0}^{|V|-1} \text{dist}(u_i, v_j) \tag{5}$$

A *quotient* (q) is proposed here which is simply the quotient of the two reversed; a higher combined score indicates overall better clustering (equation 6).

$$q = \frac{s}{h} \tag{6}$$

The numerical results are displayed in Table V.

The lowest scores for homogeneity (< 3.0) come from Arianne and Lauren's Cookbook. These blogs contain posts that are shorter on average. Lauren's Cookbook is both thematically unique and contains only four posts so a strong tendency toward compactness is to be expected. In the high end for homogeneity (> 9.0), we find Andre, Andrew, Jason, Lauren and Liv. These five blogs contain posts which are varied, and thus as we have seen – contain posts which span several different clusters. The variance is lower for blogs that have more posts and higher for blogs that have fewer posts. This suggests that the clustering is more stable when

TABLE V

SUMMARY OF THE THREE VALUES HOMOGENEITY, SEPARATION AND QUOTIENT FOR EACH BLOG (*average*±*standard deviation*) OVER ALL TWELVE TRIALS.

Author	Homogeneity	Separation	Quotient
Andre Masella	9.46 ± 0.72	8.98 ± 0.66	0.94 ± 0.03
Andrew Berry	9.74 ± 2.05	9.35 ± 1.38	0.97 ± 0.07
Arianne Villa	1.48 ± 0.37	5.75 ± 0.53	4.18 ± 1.42
Cara Ma	3.83 ± 1.58	6.48 ± 0.51	2.13 ± 1.41
Daniela Mihalciuc	5.24 ± 1.16	7.37 ± 0.80	1.44 ± 0.22
Eddie Ma	7.77 ± 0.94	8.26 ± 0.65	1.06 ± 0.07
Jason Ernst	11.7 ± 1.61	10.8 ± 1.41	0.92 ± 0.05
John Heil	8.54 ± 5.36	8.69 ± 2.94	1.39 ± 0.96
Lauren Stein	14.9 ± 4.07	14.8 ± 3.55	1.01 ± 0.09
Lauren (cooking)	2.54 ± 1.55	5.88 ± 0.68	3.32 ± 2.26
Liv Monck-Whipp	11.2 ± 1.73	10.1 ± 0.80	0.91 ± 0.09
Matthew Gingerich	5.06 ± 0.68	7.06 ± 0.63	1.40 ± 0.12
Richard Schwarting	6.51 ± 0.86	7.82 ± 0.68	1.20 ± 0.10
Tony Thompson	6.94 ± 1.12	7.85 ± 0.74	1.14 ± 0.12

there are more exemplars to train on. Blogs with a greater number of unique clusters experienced higher separation (> 10.0) including Jason, Lauren and Liv. These additional clusters correspond to those that tend away from the massive cluster of short posts (in the centre for the SOM that we've visualized). By contrast, those blogs with low separation (< 7.0) likely occur in the massive central cluster (Arianne, Cara and Lauren's Cookbook). There does not seem to be a trend with respect to the stability of separation given the behaviour of its standard deviation. Since the quotient is the ratio of separation to homogeneity, we expect that higher quotients belong to authors whose posts are more likely to cluster together than with other blogs. The highest quotients (> 3.0) belong to Arianne and Lauren's Cookbook. Indeed, these blogs have entries that tended to fall into the centre of the map. Low quotients would thus belong to blogs that tend to cluster well with foreign posts. Those blogs that spread out across the map have lower quotients such as Andre, Andrew and Liv's (< 1.0). The stability of the quotient given its standard deviation does not show a trend.

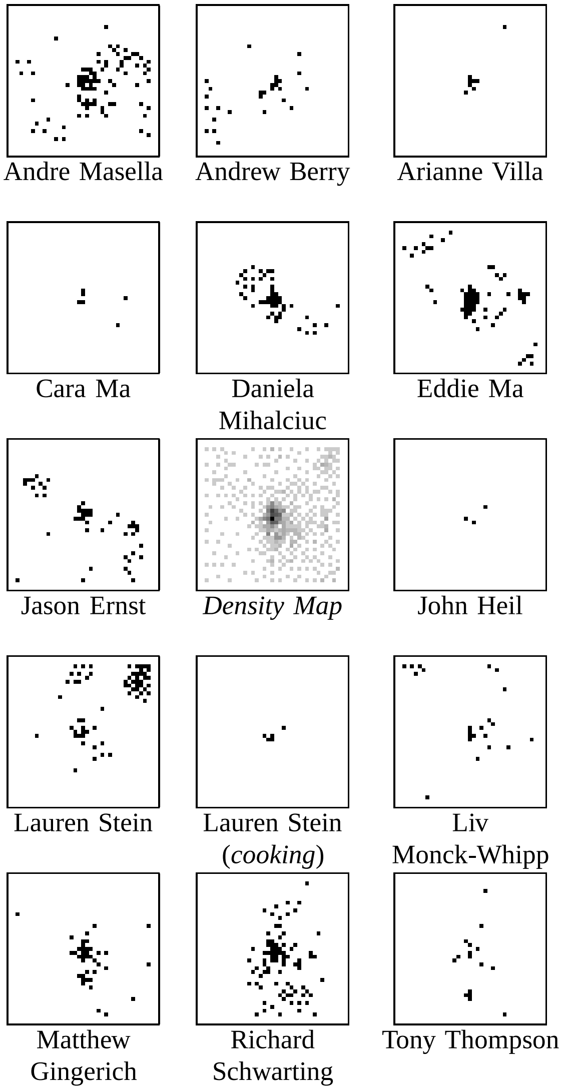

Fig. 3. A representative SOM with random seed 721091. The *Density Map* in the middle of the figure is the view of all posts clustered on this SOM. Each pixel is one cell of the SOM and the darkness is log-proportional to the number of exemplars at each cell ($\log_{10}$(number of exemplars in a cell $+$ 0.3)). The remaining fourteen images are the exemplars of each blog placed into the same SOM without the display of any other exemplars.

IV. CONCLUSIONS

We have used the SOM to cluster the blog posts of several authors. This has shown some success in making meaningful clusters. The graphical results indicate that clusters may have a semantic theme, but will often also incorporate a few posts that deviate from that theme. Clusters that are close to one another would appear as a single large cluster when the solution is not known as depicted in the central SOM image. Due to this behaviour, it appears that the SOM is more successful as a means to sort authors on a gradient of similarity in larger blended regions rather than as a means to create many well-defined discrete clusters. Finally, for an author to produce well-separated unique clusters, a balance must be struck between uniqueness of word choice and also a moderate frequency of use within that choice.

REFERENCES

[1] Mat Buckland. Kohonen's self organizing feature maps, April 2005. http://www.ai-junkie.com/ann/som/som1.html Retrieved 2011 March 8.

[2] Pierre Hansen and Brigitte Jaumard. Cluster analysis and mathematical programming. *Math. Program.*, 79:191–215, October 1997.

[3] Fritz Heckel and Nick Ward. Political blog analysis using bootstrapping techniques. In *Proceedings of the Class of 2005 Senior Conference*, pages 20 – 27, Computer Science Department, Swarthmore College, 2005.

[4] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1997.

[5] Laura Lanzarini, Augusto Villa Monte, and César Estrebou. E-mail processing with fuzzy soms and association rules, 2011.

[6] Daniel Pullwitt. Integrating contextual information to enhance som-based text document clustering, 2002.

[7] Roded Sharan, Adi Maron-Katz, and Ron Shamir. Click and expander: a system for clustering and visualizing gene expression data, 2003.

[8] E. Stamatatos, N. Fakotakis, and Kokkinakis. Computer-based authorship attribution without lexical measures, 2001.