STUDIA INFORMATICA
Nr 1/2(7) Systemy i technologie informacyjne 2006

Map Quality Measurements for GNG and SOM based Document Collection Maps

Mieczysław A. Kłopotek[1], Sławomir T. Wierzchoń [1], Krzysztof Ciesielski[1], Michał Dramiński, [1] Dariusz Czerski[1], Mariusz Kujawiak [2]
[1] Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland
[2] Institute of Computer Science, University of Podlasie,
ul. Sienkiewicza 51, 08-110 Siedlce, Poland

Abstract. The paper presents a proposal of a set of measures for comparison of maps of document collections as well as preliminary results concerning evaluation of their usefulness and expressive power.

Keywords. Search Engines, map of document collection

1 Introduction

Document maps become gradually more and more attractive as a way to visualize the contents of a large document collection.

The process of mapping a document collection to a two-dimensional map is a complex one and involves a number of steps which may be carried out in multiple variants. For example in our search engine BEATCA [1-5], the mapping process consists of the following stages: (1) document crawling (2) indexing (3) topic identification, (4) document grouping, (5) group-to-map transformation, (6) map region identification (7) group and region labeling (8) visualization.

At each of theses stages various decisions can be made implying different views of the document map. For example, the indexing process involves dictionary optimization, which may reduce the documents collection dimensionality and restrict the subspace in which the original documents are placed. Topics identification establishes basic dimensions for the final map and may involve such techniques as SVD analysis, [10], fast Bayesian network learning (ETC [9]) and other. Document grouping may for example involve various variants of growing neural gas (GNG) techniques, [6]. The group-to-map transformation is run in BEATCA based on self-organizing map (SOM) ideas, [7], but with variations concerning dynamic mixing of local and global search, based on diverse measures of local convergence. The visualization involves 2D and 3D variants.

With a strongly parameterized map creation process, the user of BEATCA can accommodate map generation to his particular needs, or even generate multiple maps covering different aspects of document collection.

We are, however, still lacking criteria for comparison of the quality of different maps, that would support the user in decision making. This paper presents our initial effort to establish appropriate measures allowing comparison of the quality of maps generated during map creation process. Section II contains the measure definitions and section III presents evaluation results on a publicly available sample set of documents. Section IV contains some conclusions from our research.

2 Map Measures

Various measures of quality have been developed in the past in the literature, covering diverse aspects of the clustering process.

The clustering process is frequently referred to “learning without a teacher”, or “unsupervised learning”, and is driven by some kind of similarity measure. The term “unsupervised” is not completely reflecting the real nature of learning. In fact, the similarity measure used is not something “natural”, but rather it reflects the intentions of the teacher. So we can say that clustering is a learning process with hidden learning criterion.

The criterion is intended to reflect some esthetic preferences, like: uniform split into groups (topological continuity) or appropriate split of a set of documents with known a priori categorization. As the criterion is somehow hidden, we need tests if the clustering process really fits the expectations. For this reason, the above-mentioned measures of clustering quality have been developed.

2.1 Some popular measures of map quality

WebSOM approach to document clustering is considered as a method of non-linear projection \from a high dimentional space into a low-dimensional one. A projection scheme is expected to preserve spatial relationships between vectors in the input space. A fairly simple measure to assess the quality of projection is the comparison of distances between vectors in the input and an output space. For normalized vectors we can use a simple average square error of distances in both spaces.

$$E_{sq} = \frac{1}{M^2 - M} \sum_{i,j=1,i\neq j}^{M} (d_{ij}^{p} - d_{ij}^{n})^2,$$

where

M – the number of documents

d_{ij}^{p} – distance between documents i and j in a low-dimensional space

d_{ij}^{n} – distance between documents i and j in a highly-dimensional space

Another, highly correlated measure, which does not require normalization of vectors is called Sammon error [12]:

$$E_{Sammon} = \frac{1}{\sum_{i,j=1,j\neq j}^{M} d_{ij}^{n}} \sum_{i,j=1,i\neq j}^{M} \frac{(d_{ij}^{p} - d_{ij}^{n})^2}{d_{ij}^{n}},$$

where
M – the number of documents
d_{ij}^{p} – distance between documents i and j in a low-dimensional space
d_{ij}^{n} – distance between documents i and j in a highly-dimensional space

Both measures are somehow inadequate for document maps, as they assume linear projection, while the WebSOM like projection is obviously non-linear one.
Specially for measuring the SOM map quality, the so called *the average quantization error* and the *topography error*. Measures have been developed.
The *average quantization error* [12] tests how well model vectors approximate documents assigned to them. It is computed as an average distance between document vectors and their closest model vectors.

$$E_q = \frac{1}{M} \sum_{i=1}^{M} \| x_i - m_c \|$$

where
M – the number of documents
x_i – document vector
m_c – winner model vector for document vector x_i

The *topography error* [14], on the other hand, is intended to measures map orderliness: the match between map topology and cluster similarity. The proportion of documents for which the 2 best matching model vectors do not lie in adjacent map units is calculated.

$$E_t = \frac{1}{M} \sum_{i=1}^{M} u(x_i)$$

where
M – document count
$u(x_i) = 1$ when 2 best matching model vectors of the document x_i are not neighbours, =0 otherwise

The values of these measures are known to depend on the data set used. Their mutual proportion tells us a little bit about the compromise compromise between document density approximation and correctness of local proximity of documents representation. The final width of the neighbourhood function, the diversity of documents being mapped and the size of the map influence this proportion. It has

been claimed that the greater the final neighbourhood width the lower value of the *topography error* and greater the *average quantization* error.
An important quality indicator is also *map smoothness* measure (telling, how similar are model vectors of adjacent map cells). The map smoothness may be computed as average distances between adjacent map units over the whole map.
Furthermore, if ione wants to compare results of several map construction alogorithm, it is wise to have a reference map and compare generated maps to it. As maps may be equivalent even under some transformations (shifts, rotations etc.), various sophisticated measures for comparing two SOM maps [15] have been developed. For example, for small maps, the following simplified measure can be applied [16]: The measure below computes the proportion of pairs of documents that are neighbours on one map and are situated farther off on the second.

$$E_{sq} = \frac{1}{M^2 - M} \sum_{i,j=1, i \neq j}^{M} testDist(d_{ij}^{I}, d_{ij}^{II}) \ ,$$

$$testDist(d_{ij}^{I}, d_{ij}^{II}) = \begin{cases} 1 \mapsto (d_{ij}^{I} < 1{,}5 \wedge d_{ij}^{II} > 1{,}5) \vee (d_{ij}^{II} < 1{,}5 \wedge d_{ij}^{I} > 1{,}5) \\ 0 \mapsto otherwise \end{cases}$$

where
M – the number of documents
d_{ij}^{p} – distance between documents i and j on the first map
d_{ij}^{d} – distance between documents i and j on the second map

Kohonen et al. [13] developed an efficient measure of map quality using a priori knowledge about the structure of the data set. It assumes a pre-defined metric of *category distance* Then siome kind of *cluster purity* measure is adopted to establish, if the map construction algorithm reestaboishes intrinsic categorization.

2.2 Map quality measures adopted in our research

We have accommodated for our purposes and investigated the following well known quality measures of clustering (consult e.g. [11] for details)

4001 = *cellErr* - *AverageMapCosine-Quantization*: the average cosine distance between all neighboring cells on the map. Its aim is to measure topological continuity of the map (the lower its value the more "smooth" the model)

$$E_{cellErr} = \frac{1}{|N|} \sum_{i \in N} \frac{1}{|E(i)|} \sum_{j \in E(i)} c(i,j)$$

where N is the set of graph nodes, E(i) is the set of nodes adjacent to the node i and c(i,j) is the cosine distance between nodes i and j.

4002 = *AverageDocumentCosine-Quantization(docErr)* average distance (according to cosine measure) for the learning set between the document and the cell it was classified into. Its aim is to measure the quality of clustering at the level of single cells.

$$E_{docErr} = \frac{1}{|N|} \sum_{i \in N} \frac{1}{|D(i)|} \sum_{d \in D(i)} c(i,d)$$

where D(i) is the set of documents assigned to niode i.

The subsequent measures evaluate a concordance between the clustering and assumed classification.

4003 = *ClusterPurity*: measures "class purity" of a single map cell i, and is equal to $|D_c(i)|/|D(i)|$, where $D_c(i)$ is the number of category c documents assigned to cell i. .

4004 = *ClusterEntropy*: measures entropy of the frequencies in the distribution of individual classes for a cell. I and is the sum over all categories c in the cell i of $-\log(|D_c(i)|/|D(i)|)$

4005 = *AverageWeightedCluster-Purity*: average, weighted by cell density of the map, value of the measure *ClusterPurity*.

4006 = *AverageWeightedCluster-Entropy*: average, weighted by cell density of the map, of the measure *ClusterEntropy*

4007 = *NMI - NormalizedMutual-Information*: meaning approximately the quotient of total class (category) and cluster entropy to the square root of the product of class (category) and cluster entropies for individual clusters .

3 Results

Initial results of experiments were obtained for the Syskill & Webert database, so that their generality still needs to be verified for larger sets. Nonetheless, we believe that these insights are worth deeper analysis.

3.1 A. Comparison of SOM and GNG

The results from Table 1 were based on the following settings:

Experiment #12: GNG with 64 gas cells
Experiment #13: SOM - 8*8 cell map
Experiment #22: GNG with 16 gas cells Experiment #23: SOM - 4*4 cell map

The measure 4001 indicates that in all cases the topology of GNG is more continuous than that of SOM. This may be caused, at least in some cases, by destruction of the SOM net in the initial learning phase; After thematic initialization,

the 4001 measure value is low. So we can say, that this initialization performed as expected. But later on, this value unexpectedly grows even above 0.887 in the first phase of experiment #23. This may indicate that the structure of the map changes.

We can conclude that learning parameters should be carefully chosen; wrong setting may destroy what the thematic initialization had to achieve.

The measure 4002 says that SOM proved to be better for a large map, and GNG for a smaller one. This finding needs a further exploration as the larger maps proved to be better in general (though not always, as subsequent experiments show).

GNG already in the initial stages (with few cells only) induces a relatively good clustering (comparable with SOM at early stages, consisting of much more cells).

Obviously the measure 4002 depends on map size (cell count, cluster count), so it is not very suitable for map and algorithm comparison; only comparison of parameter settings of the same algorithm makes sense.

The measure 4005 allows to draw the next conclusions.

GNG with lower cell number is ranked higher than GNG with higher cell count.

Apparently above-mentioned measures depend on the number of map/gas cells and are the better the closer their number to the natural number of groups in the document collection ((for S&W about 4)

So one could expect they will be better for evaluation of grouping of whole area of the map rather than for the individual cells.

GNG with 16 cells had the highest purity, while SOM with 4*4 cells had the lowest one.

An interesting anomaly is the initial purity in the experiment #13 (SOM 8*8) amounting to 0.74, twice as much as in the remaining experiments.

The measure 4006 shows that: (a) generally entropy seems to be low and without interesting variation, which may be attributed to the low number of documents, and (b) the only distinguishing feature was the high initial entropy for GNG (as there were only two cells).

The measure 4007: (a) yields comparable results for all four experiments, with slightly better values for smaller maps (the reason is similar to the previous one), and (b) the experiment #13 anomaly related to *Avg.ClusterPurity*, is repeated by initial *NormalizedMutualInformation*: at the level of 0.48 (while in the remaining cases it lies at about 0.01 !),

3.2 A Comparison of SOM parameter setting and the initialization procedures

Earlier experiments show that there may exist a natural value for the topology continuity measure (4001) for a given document collection (e.g. for S&W collection it may be about 0.35). Independently of learning parameters this measure converges relatively quickly during learning and stabilizes near to this value. Especially in the first iteration the effect of destruction of cell network is relatively violent – thematic

initialization appears to induce too strong "continuity" for the net (measure at the level of 0.01). There exists also a trade-off between measures 4001 and 4002: lowering the error level 4002 is related to discontinuation of the net in document space (increase of the value of the measure 4001).

Qualitative measures (4002-4007) for the same learning parameters imply comparable results for thematic initialization using ETC and SVD, slightly worse for Naïve Bayes

What is more interesting, 8*8 cells SOM learning with parameter settings (1) *initKernel* = 3 (that is with 49 cells) and (2) *initKernel* = 2 (fewer than 25 cells) yields significantly different results. Learning with a wider neighborhood gives worse results especially in terms of topology (measure 4001) and clustering (4002), smaller differences are observed for clustering-classification measures (4005-4007). Moreover, for larger neighborhood the effect of divergence (in the sense of measure 4002) has been observed in the middle phase of the learning process.

4 Conclusions

Our initial study of quality measures for document maps demonstrates how difficult the problem of finding good measures is. By a good measure we understand one covering many facets of the problem, well reflecting the human perception of the map. Such a measure may on the one hand evaluate the suitability of various map creation processes and their components, on the other hand it may guide selection of appropriate process parameters.

The initial studies show that indeed an apriorical setting of good learning parameters is virtually impossible and a procedure of accommodating them to the current set of documents is necessary. The measures developed so far may be a good starting point.

But also measures oriented directly towards map structure need still to be developed, which would cover beside the aspects referred to in measures 4001-4007 – to mutual position of the cells.

Independence of map scale should also be incorporated into the quality measures.

Apparently measures 4005-4007 seem to be appropriate to evaluate the quality of map areas.

A challenge is to evaluate the quality of GNG to SOM transformation.

Acknowledgement

Research has been partially supported under KBN research grant 4 T11C 026 25 "Maps and intelligent navigation in WWW using Bayesian networks and artificial immune systems"

http://www.ipipan.waw.pl/~klopotek/mak/current_research/KBN2003/KBN2003Translation.htm

References

[l] K. Ciesielski, M. Dramiński, M. Kłopotek, M. Kujawiak, S. Wierzchoń: On some clustering algorithms. To appear in *Proc. Intelligent Information Processing and Web Mining*, Gdansk 2005.

[2] K. Ciesielski, M. Dramiński, M. Kłopotek, M. Kujawiak, S. Wierzchoń: Architecture for graphical maps of Web contents. *Proc. WISIS'2004, Warsaw*

[3] K. Ciesielski, M. Dramiński, M. Kłopotek, M. Kujawiak, S. Wierzchoń: Mapping document collections in non-standard geometries. B. De Beats, R. De Caluwe, G. de Tre, J. Fodor, J. Kacprzyk, S. Zadrożny (eds): *Current Issues in Data and Knowledge Engineering*. Akade-micka Oficyna Wydawnicza EXIT Publ., Warszawa 2004. pp.122-132.

[4] K. Ciesielski, M. Dramiński, M. Kłopotek, M. Kujawiak, S. Wierzchoń: Clustering medical and biomedical texts - document map based approach. Proc. Sztuczna Inteligencja w Inżynierii Biomedycznej SIIB'04, Kraków. ISBN-83--919051-5-2

[5] K. Ciesielski, M. Dramiński, M. Kłopotek, M. Kujawiak, S. Wierzchoń: Crisp versus Fuzzy Concept Boundaries in Document Maps, to appear in Proc DMIN-05

[6] Fritzke, B.A growing neural gas network learns topologies. In G. Tesauro, D.S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge MA, 1995, pp. 625-632.

[7] T. Kohonen, *Self-Organizing Maps*. Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 2001. Third Extended Edition, 501 pages. ISBN 3-540-67921-9, ISSN 0720-678X

[8] M. Kłopotek, M. Dramiński, K. Ciesielski, M. Kujawiak, S.T. Wierzchoń: Mining document maps. Proc. *W1 - Statistical Approaches to Web Mining* (SAWM) of PKDD'04, M. Gori, M. Celi, M. Nanni eds., Pisa, Italy, September 20-24, pp. 87-98.

[9] M.A. Kłopotek: A New Bayesian Tree Learning Method with Reduced Time and Space Complexity. *Fundamenta Informaticae*, 49 (no 4) 2002, IOS Press, pp. 349-367.

[10] Press, W.M., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. *Numerical recipes: The art of scientific computing.* New York, NY: Cambridge University Press, 1986.

[11] Y. Zhao, G. Karypis, Criterion functions for document Clustering: Experiments and analysis, available at URL: http://www-users.cs.umn.edu/~karypis/publications/ir.html.

[12] Sammon Mapping, http://www.eng.man.ac.uk/mech/merg/research/datafusion.org.uk/techniques/sammon.html, November 2003

[13] T. Kohonen, S. Kaski, K. Laguss, J. Sajolarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Transactions on Neural Networks vol. 11 No. 3, May 2000
[14] K. Kiviluoto, Topology Preservation in Self-Organizing Maps, Proceedings of ICANN 96, IEEE International Conference on Neural Networks, 1996.
[15] S. Kasski, K. Lagus, Comparing Self Organizing Maps, Proceedings of ICANN 96, Internetional Conference on Artificial Neural Networks, Lecture Notes in Computer Science vol. 11112, pp.809-814, Springer, Berlin, 1996.
[16] Wilkowski A.: M.Sc. Thesis, Warsaw University of Technology. (supervised by M.A. Klopotek).

Table 1. GNG VERSUS SOM COMPARISON

Abbreviations used (nor explained in the text): docGroup – method of document clustering, ETC – (Edge Tree construction algorithm), init kernel – size of the neighbourhood for SOM learning, IDComponent – learning phase (init – initial, 0 – after first iteration, 63 – after 63rd iteration, final – at the end)

4001 = cellErr 4002 = docErr 4005 = AvgPurity 4006 = AvgEntropy 4007 = NMI	**experiments** 12 / 22 = GNG 13 / 23 = SOM	**settings (12 / 13)** 64 cells init kernel = 2 docGroup = ETC	**settings (22 / 23)** 16 cells init kernel = 1 docGroup = ETC

IDExperiment	IDComponent	IDMeasure	MeasureValue
12	init	4001	2.12554418510535e-011
12	0	4001	0.000433683834039023
12	63	4001	**0.0689259148951177**
13	init	4001	**0.0128107706587762**
13	0	4001	**0.364930347438494**
13	12	4001	0.699183833332539
22	init	4001	2.96089819329382e-011
22	0	4001	0.0065691044856917
22	63	4001	0.0812347284160337
23	init	4001	**5.40997709593446e-011**
23	0	4001	**0.887198888726031**
23	10	4001	0.840901175702208
12	init	4002	0.831972994522146
12	0	4002	0.830728164114876
12	63	4002	**0.592336284145044**
13	init	4002	0.814792895847344
13	0	4002	0.770654366684529
13	12	4002	**0.537887205267935**
22	init	4002	0.842520307370469
22	0	4002	0.822209505999235

22	63	4002	**0.608941513399313**
23	init	4002	0.835195787799028
23	0	4002	0.771169654318995
23	10	4002	**0.672877741405967**
12	init	4005	0.407738095238095
12	final	4005	**0.931547619047619**
13	init	4005	0.744047619047619
13	final	4005	**0.931547619047619**
22	init	4005	0.407738095238095
22	final	4005	**0.964285714285714**
23	init	4005	0.458333333333333
23	final	4005	**0.889880952380952**
12	init	4006	**0.653719933296097**
12	final	4006	0.00228916958286251
13	init	4006	0.00780499500321359
13	final	4006	0.00273635940268996
22	init	4006	**0.657591914930807**
22	final	4006	0.0014569329153533
23	init	4006	0.0749868393800979
23	final	4006	0.0189148951375854
12	init	4007	0.00790486109953081
12	final	4007	**0.503754759938543**
13	init	4007	**0.485035806535169**
13	final	4007	**0.529664324881387**
22	init	4007	0.0159041978936837
22	final	4007	**0.529117996536141**
23	init	4007	0.0550687043240412
23	final	4007	**0.539581385802193**

Table 2. PARAMETER SETTINGS COMPARISON FOR NB, SVD, AND ETC MAP INITIALIZATION METHODS

Abbreviations used (nor explained in the text): NB – naïve Bayes, SVD – Singular Value Decomposition, ETC – Edge Tree construction algorithm

PART I: LARGER NEIGBOURHOODS

measures	**experiments**	**settings**
4001 = cellErr	11 = NB	SOM
4002 = docErr	12 = ETC	64 cells
4005 = AvgPurity	13 = SVD	init kernel = 3 (**49 cells**)
4006 = AvgEntropy		
4007 = NMI		

IDExperiment	**IDComponent**	**IDMeasure**	**MeasureValue**
11	init	4001	0.0113592231770852
11	0	4001	0.138142957234457
11	62	4001	**0.321678224871272**
12	init	4001	0.010615862189708
12	0	4001	0.249322878143704
12	62	4001	**0.368985056529525**
13	init	4001	0.0132719905006483
13	0	4001	0.185522242955072
13	63	4001	**0.364494069144338**
11	init	4002	0.817867157503836
11	0	4002	0.806000148113058
11	62	4002	**0.694115212665557**
12	init	4002	0.831053422158208
12	0	4002	0.816053036016977
12	62	4002	**0.631406207386703**
13	init	4002	0.876018530305631
13	0	4002	0.843069859550306
13	63	4002	**0.628076211312685**
11	init	4005	0.702380952380952
11	final	4005	**0.744047619047619**
12	init	4005	0.711309523809524
12	final	4005	**0.895833333333333**
13	init	4005	0.538690476190476
13	final	4005	**0.895833333333333**
11	init	4006	0.00893566078275286
11	final	4006	0.00814408667952336
12	init	4006	0.0113374065372495
12	final	4006	0.00473600860314676
13	init	4006	0.016012115823195
13	final	4006	0.00443045130059825
11	init	4007	0.436628035060674
11	final	4007	**0.434878194806693**
12	init	4007	0.410706828772287
12	final	4007	**0.511514819865847**

13	init	4007	0.214802039076182
13	final	4007	**0.505484082832231**

PART II: SMALLER NEIGBOURHOODS

measures	**experiments**	**settings**
4001 = cellErr	11 = NB	SOM
4002 = docErr	12 = ETC	64 cells
4005 = AvgPurity	13 = SVD	init kernel = 2 (**25 cells**)
4006 = AvgEntropy		
4007 = NMI		

IDExperiment	**IDComponent**	**IDMeasure**	**MeasureValue**
11	init	4001	0.014647243668883
11	0	4001	0.368257310651787
11	11	4001	**0.621535342633998**
12	init	4001	0.018125388483832
12	0	4001	0.341070701649263
12	11	4001	**0.671669099448811**
13	init	4001	0.0132719875360321
13	0	4001	0.321202771040711
13	12	4001	**0.620604342150223**
11	init	4002	0.805549412366109
11	0	4002	0.760053026336003
11	11	4002	**0.549621061194296**
12	init	4002	0.8002603575029
12	0	4002	0.77659703026613
12	11	4002	**0.552352753756013**
13	init	4002	0.876018542738369
13	0	4002	0.795588877939603
13	12	4002	**0.566813910670896**
11	init	4005	0.735119047619048
11	final	4005	**0.943452380952381**
12	init	4005	0.648809523809524
12	final	4005	**0.875**
13	init	4005	0.538690476190476
13	final	4005	**0.925595238095238**
11	init	4006	0.008556480108359
11	final	4006	0.00247802188983645
12	init	4006	0.0106107581006419
12	final	4006	0.00471652806143332
13	init	4006	0.016012115823195
13	final	4006	0.00271160610523567
11	init	4007	0.484590475200228
11	final	4007	**0.543804837568776**
12	init	4007	0.360629687424118
12	final	4007	**0.459456793512728**
13	init	4007	0.214802039076182
13	final	4007	**0.539688109189256**