Annales UMCS Informatica AI 4 (2006) 45-59

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Incremental document map formation: multi-stage approach

Krzysztof Ciesielski*, Michał Dramiński, Mieczysław A. Kłopotek, Dariusz Czerski, Sławomir T. Wierzchoń

Institute of Computer Science, Polish Academy of Sciences,
Ordona 21, 01-237 Warszawa,Poland

Abstract

The paper presents methodology for the incremental map formation in a multi-stage process of a search engine with the map based user interface[1]. The architecture of the experimental system allows for comparative evaluation of different constituent technologies for various stages of the process. The quality of the map generation process has been investigated based on a number of clustering and classification measures. Some conclusions concerning the impact of various technological solutions on map quality are presented.

1. Introduction

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

The process of mapping a collection to a two-dimensional map is a complex one and involves a number of steps which may be carried out in multiple variants. In our search engine BEATCA [1-6], the mapping process consists of the following stages (see Figure 1): (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, generated by different algorithms.

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 the singular value decomposition analysis (SVD [7]), the fast Bayesian network learning (ETC [8]) and others. Document grouping may involve various variants of

*Corresponding author: *e-mail address*: kciesiel@ipipan.waw.pl

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

growing neural gas (GNG) techniques [9], hierarchical SOM [10] and Artificial Immune Systems [11]. The group-to-map transformation is run in BEATCA based on SOM ideas [12], but with variations concerning dynamic mixing of local and global search, based on diverse measures of local convergence [5]. The visualization involves 2D and 3D variants.

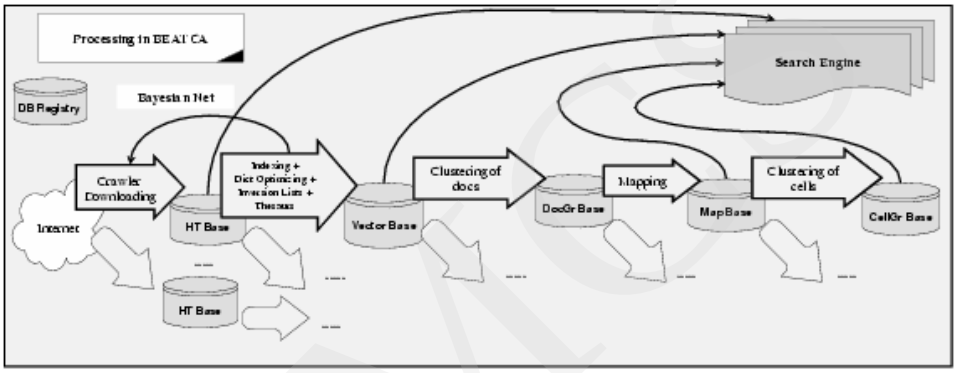

Fig. 1. BEATCA system architecture

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 various aspects of document collection.

The overall complexity of the map creation process, resulting in long run times, as well as the need to avoid "revolutionary" changes of the image of the whole document collection, requires an incremental process of accommodation of new incoming documents into the collection.

Within the BEATCA project we have devoted much effort to enable such a gradual growth. In this study, we investigate vertical (emerging new topics) and horizontal (new documents on current topics) growth of document collection and its effects on the map formation capability of the system.

To ensure intrinsic incremental formation of the map, all the computation-intense stages involved in the process of map formation (crawling, indexing and all the stages of map formation: GNG-based document grouping, model visualization and map region identification) need to be reformulated in terms of incremental growth.

In particular, the Bayesian Network driven crawler is capable of collecting documents around an increasing number of distinct topics. The crawler learning process runs in a kind of horizontal growth loop while it improves its performance with an increasing number of documents collected. It may also grow vertically, as the user can add new topics for search during its run time.

The indexer has been constructed in order to achieve incremental growth and optimization of its dictionary with the growing collection of documents. Query extension capability of the query answering interface, based on the Bayesian

network and the GNG derived dynamic automated thesaurus, accommodates also to the growing document collection. Though the actual clustering algorithms used in our system, like GNG, AIS or fuzzy C-means, are by their nature adaptive, nonetheless their tuning and modification were not a trivial task, especially with respect to our goal to achieve quality of incremental map comparable to the non-incremental one.

Special algorithms for thematic map initialization as well as for identification of document collection topics, based on the GNG, SVD and/or Bayesian networks, lead to stabilization of the overall map. At the same time GNG detects the topic drift and so it may be appropriately visualized, due to plastic clustering approach, as new emerging map regions. It should be stressed at this point, that the map stabilization does not preclude obtaining different views of the same document collection. Our system permits to maintain several maps of the same document collection, obtained via different initializations of the map, and, what is more important, automatically tells the user which of the maps is most appropriate to view the results of his actual query.

In the BEATCA search engine, Bayesian Networks are used in a few critical moments. We found it very useful for initial clustering of documents set. For the map creation phase, a couple of clearly separated clusters is calculated. Such clusters proved to be especially useful for thematic initialization of a clustering model and SOM projection model (the latter is shortly described in section 4).

BN is also used as thesaurus in our system. After the indexing phase in BEATCA, a special dedicated BN is built on all terms in the dictionary[2]. Having collected a relevant set of documents on a given subject, joint information stored in BN and in the main clustering model will constitute context-dependent thesaurus. There is no room for details, so we only note that such thesaurus is used to expand user queries, for the purpose of more precise search in the BEATCA search engine [6].

In the current paper we focus on our incremental version of GNG-based clustering phase of the map creation process. In section 2 we introduce the concept of GNG. In section 3 – our major modification of GNG algorithm: the robust allocation of documents to clusters. In section 4 our original approach to GNG visualization is presented.

To evaluate the effectiveness of the overall incremental map formation process, we compared it to the “from scratch” map formation in our experimental section 5. A brief discussion of related works is presented in section 6. The conclusions from our research work can be found in section 7.

[2]Excluding terms of low clustering quality identified during the dictionary optimization phase [1].

2. Growing Neural Gas approach to clustering of text documents

An efficient solution to the problem of document clustering is offered by the Growing Neural Gas (GNG) network, first presented in [9]. Like Kohonen (SOM) networks [12], GNG can be viewed as a topology learning algorithm. Its aim can be summarized as follows: given some collection of high-dimensional data, find a topological structure that closely reflects the topology of the collection. If we treat a single GNG graph node as a cluster of data then the whole network can be viewed as a meta-clustering structure, where similar groups are linked together by graph edges.

In typical SOM the number of units and topology of the map is predefined. As observed in [9], the choice of SOM structure is difficult, and the need to define a decay schedule for various parameters is problematic.

The GNG network starts learning with a few units[3] and new ones are inserted successively every few iterations. To determine where to insert new units, local error measures are gathered during the adaptation process; a new unit is inserted near the unit which has accumulated maximal error. Interestingly, nodes of the GNG network are joined automatically by links, hence as a result a possibly disconnected graph is obtained, and its connected components can be treated as different data clusters.

The complete GNG algorithm specification and its comparison to numerous other soft competitive methods can be found in [13].

In our approach, objects (text documents as well as graph nodes, described below) are represented in the standard way, i.e. as vectors of the dimension equal to the number of distinct dictionary terms. A single element of so-called *referential vector* represents importance of a corresponding term and is calculated on the basis of the normalized TFxIDF measure. Similarity measure is defined as the cosine of the angle between corresponding vectors.

2.1. Utility factor

Typical problem in web mining applications is that processed data is constantly changing – some documents disappear or become obsolete, while others enter analysis. All this requires models which are able to adapt its structure quickly in response to non-stationary distribution changes. Thus, we decided to adopt and implement GNG with a utility factor model [14].

A crucial concept here is to identify the least useful nodes and remove them from the GNG network, enabling further node insertions in regions where they would be more necessary. The utility factor of each node reflects its contribution to the total classification error reduction. In other words, node utility is

[3]The initial nodes referential vectors are initialized with our broad topic initialization method, briefly described in section 4.

proportional to the expected error growth if the particular node would have been removed. There are many possible choices for the utility factor. In our implementation, the utility update rule of a winning node has been simply defined as $U_s = U_s + error_t - error_s$, where s is the index of the winning node, and t is the index of the second-best node (the one which would become the winner if the actual winning node would be non-existent). Newly inserted node utility is arbitrarily initialized to the mean of two nodes which have accumulated most of the error: $U_r = (U_u + U_v)/2$.

After the utility update phase, a node *k* with the smallest utility is removed if the fraction $error_j/U_k$ is greater than some predefined threshold; where *j* is the node with the greatest accumulated error.

3. Robust winner search in the GNG network

Similarly to the Kohonen algorithm, the most computationally demanding part of the GNG algorithm is the winner search phase. Especially, in application to web documents, where both the text corpus size and the number GNG network nodes is huge, the cost of even a single global winner search phase is prohibitive.

Unfortunately, neither local-winner search method (i.e. searching through the graph edges from some staring node) nor joint-winner search method (our own approach devoted to SOM learning [1]) are directly applicable to the GNG networks. The main reason for this is that a graph of GNG nodes can be unconnected. Thus, the standard local-winner search approach would prevent document from shifting between separated components during the learning process.

A simple modification consists in remembering the winning node for more than one connected component of the GNG graph[4] and conducting in parallel a single local-winner search thread for each component. Obviously, it requires periodical (precisely, once for an iteration) recalculation of connected components, but this is not very expensive[5].

A special case is the possibility of a node removal. When the previous iteration's winning node for a particular document has been removed, search processes (in parallel threads) from each of its direct neighbors in the graph are activated.

We have implemented another method, a little more complex (both in terms of computation time and memory requirements) but, as the experiments show, more accurate. It exploits the data structure known as Clustering Feature Tree [15] to group similar nodes in dense clusters. Node clusters are arranged in the

[4]Two winners are just sufficient to overcome the problem of components separation.

[5]In order of $O(V+E)$, where V is the number of nodes and E is the number of connections (graph edges).

hierarchy and stored in a balanced search tree. Thus, finding the closest (most similar) node for a document requires $O(log_t V)$ comparisons, where V is the number of nodes and t is the tree branching factor (refer to [15]). Amortized tree structure maintenance cost (node insertion and removal) is also proportional to $O(log_t V)$.

4. Adaptive visualization of the model

Despite many advantages over the SOM approach, GNG has one serious drawback: high-dimensional networks cannot be easily visualized. However, we can build Kohonen map on the referential vectors of GNG network, similarly to the case of single documents, i.e. treating each vector as a centroid representing a cluster of documents.

To obtain the visualization that singles out the main topics in the text corpus and reflects the conceptual closeness between topics, the proper initialization of SOM cells is required. We have developed a special initialization method, intended to identify broad topics in the document collection and to improve the stability of the 2D/3D visualization. Briefly, in the first step the centroids of a few main clusters are identified (via fast ETC algorithm [8] and SVD decomposition [7]). Then, we select *fixpoint cells*, spread them uniformly on the map surface and initialize them with the centroid vectors. Finally, we initialize the remaining map cells with *intermediate* topics, calculated as the weighted average of main topics, with the weight proportional to the Euclidean distance from the corresponding fixpoint cells and their density.

After initialization, the map is learned with the standard Kohonen algorithm [12]. Finally, we adopt the so-called *plastic clustering* algorithm [16] for precise adjustation of the position of GNG model nodes on the SOM projection map, so that the distance on the map reflects as close as possible the similarity of the adjacent nodes. The concept is based on the attraction-repulsion technique, similar to the gravity clustering methods, where the nodes are attracted with the force proportional to their similarity and mass (density), while simultaneously they are repelled by graph edges. It should be stressed that the thematic initialization of the map is crucial here to ensure the stability of the final visualization and to emphasize topics which are considered to be user-important [4].

The resulting map is visualization of GNG network with the detail level depending on the SOM size (a single SOM cell can gather more than one GNG node). The user can access the document content via the corresponding GNG node which, in turn, can be accessed via the SOM node – interface here is similar to the hierarchical SOM map case.

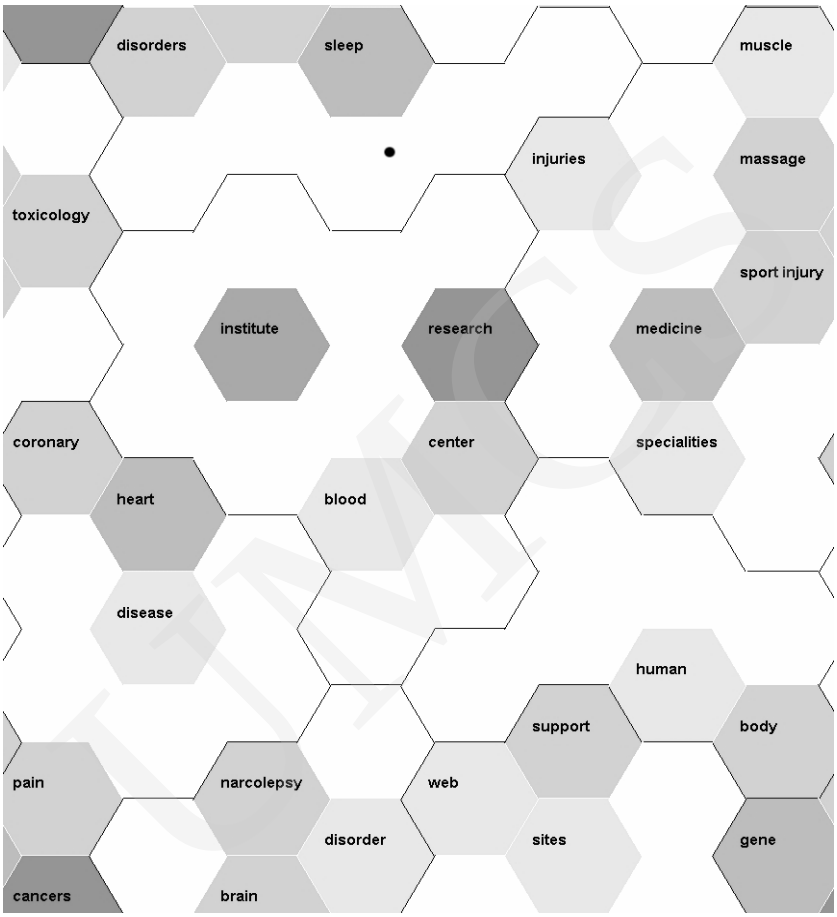

Fig. 2. Example of GNG model visualization for health-related newsgroups

The exemplary map can be seen in Figure 2. The color brightness is related to the number of documents contained in the cell. Each cell containing at least one document is labeled with a few descriptive terms (only one can be seen on the map, the rest is available via the BEATCA search engine). The black lines represent the borders of thematic areas[6]. It is important to stress that this planar representation is in fact a torus surface (which can also be visualized in 3D), so the cells on the map borders are adjacent.

5. Experiments

To evaluate the effectiveness of the overall incremental map formation process, we compared it to the "from scratch" map formation. In this section we

[6]Crisp borders are induced from fuzzy clustering of map nodes, based on the combination of Fuzzy C-Means algorithm and minimal spanning tree; algorithm details can be found in [3].

describe the overall experimental design, quality measures used and the results obtained.

The architecture of our system supports comparative studies of clustering methods at the various stages of the process (i.e. initial document grouping, broad topics identification, incremental clustering, model projection and visualization, identification of thematic areas on the map and its labeling). In particular, we conducted series of experiments to compare the quality and stability of GNG and SOM models for various model initialization methods, winner search methods and learning parameters [5]. In this paper we focus only on evaluation of the GNG winner search method and the quality of the resulting incremental clustering model with respect to the topic-sensitive learning approach.

5.1. Quality measures for the document maps

Various measures of quality have been developed in literature, covering diverse aspects of the clustering process (e.g. [17,18]). The clustering process is frequently referred to as "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 reflects the intentions of the teacher. So we can say that clustering is a learning process with a hidden learning criterion. The criterion is intended to reflect some esthetic preferences, like: uniform split into groups (topological continuity) or appropriate split of documents with known a priori categorization. As the criterion is somehow hidden, we need tests if the clustering process really fits the expectations. In particular, we have accommodated for our purposes and investigated the following well known quality measures of clustering:

- **Average Map Quantization**: the average cosine distance between each pair of adjacent nodes. The goal is to measure topological continuity of the model (the lower this value is, the more "smooth" model is):

$$AvgMapQ = \frac{1}{|N|}\sum_{n\in N}\left(\frac{1}{|E(n)|}\sum_{m\in E(n)} c(n,m)\right),$$

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

- **Average Document Quantization**: the average distance (according to cosine measure) for the learning set between the document and the node it was classified into. The goal is to measure the quality of clustering at the level of a single node:

$$AvgDocQ = \frac{1}{|N|}\sum_{n\in N}\left(\frac{1}{|D(n)|}\sum_{d\in D(n)} c(d,n)\right),$$

where $D(n)$ is the set of documents assigned to the node n.

Both measures have values in the [0,1] interval, the lower values correspond respectively to more "smooth" inter-cluster transitions and more "compact" clusters. To some extent, optimization of one of the measures entails increase of the other one. Still, experiments [5] show that the GNG models are much more smooth than SOM maps while the clusters are of similar quality.

The two subsequent measures evaluate the agreement between the clustering and the a priori categorization of documents (i.e. particular newsgroup in the case of newsgroups messages).

– **Average Weighted Cluster Purity**: the average "category purity" of a node (node weight is equal to its density, i.e. the number of assigned documents):

$$AvgPurity = \frac{1}{|D|}\sum_{n\in N} max_c\left(\left|D_c(n)\right|\right),$$

where D is the set of all documents in the corpus and $D_c(n)$ is the set of documents from category c assigned to the node n.

– **Normalized Mutual Information**: the quotient of the total category and the total cluster entropy to the square root of the product of category and cluster entropies for individual clusters:

$$NMI = \frac{\sum_{n\in N}\sum_{c\in C}\left|D_c(n)\right| \; log\left(\frac{|D_c(n)||D|}{|D(n)||D|}\right)}{\sqrt{\left(\sum_{n\in N}\left|D(n)\right| \; log\left(\frac{|D(n)|}{|D|}\right)\right)\left(\sum_{c\in C}\left|D_c\right| log\left(\frac{|D_c|}{|D|}\right)\right)}},$$

where N is the set of graph nodes, D is the set of all documents in the corpus, $D(n)$ is the set of documents assigned to the node n, D_c is the set of all documents from category c and $D_c(n)$ is the set of documents from category c assigned to the node n.

Again, both measures have values in the [0,1] interval. Roughly speaking, the higher the value is, the better agreement between clusters and a priori categories. At the moment, we are working on the extension of the above-mentioned measures to those covering all aspects of the map-based model quality, i.e. similarities and interconnections between thematic groups both in the original document space and in the toroid map surface space.

5.2. Experimental results

Model evaluations were performed on 2054 documents downloaded from 5 newsgroups with quite well separated main topics (antiques, computers, hockey, medicine and religion). Each GNG network has been trained for 100 iterations,

with the same set of learning parameters, using the previously described winner search methods.

In the main case (depicted with the black line), the network has been trained on the whole set of documents. This case was the reference one for the quality measures of adaptation as well as comparison of the winner search methods.

Figure 3 presents the comparison of a standard global winner search method with our own CF-tree based approach. The local search method is not taken into consideration since, as it has already been mentioned, it is completely inappropriate in the case of unconnected graphs. Obviously, the tree-based local method is invincible in terms of computation time. The main drawback of the global method is that it is not scalable and depends on the total number of nodes in the GNG model.

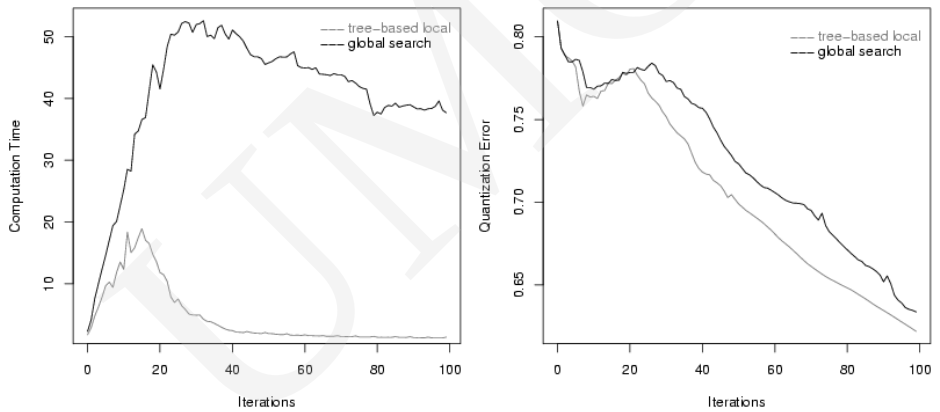

Fig 3. Winner search methods (a) computation time (b) model quality

At first, the result of the quality comparison appeared to be quite surprising. On one hand, the quality was similar, on the other – global search appeared to be worse of the two (!). We have investigated it further and it turned out to be the aftermath of process divergence during the early iterations of the training process. It will be explained using the next example.

In the next experiment, in addition to the main reference case, we had other two cases. During the first 30 iterations the network was trained on 700 documents only. In one of the cases (represented by the red line) the documents were sampled uniformly from all five groups and in the 33^{rd} iteration other 700 uniformly sampled were introduced to training. After the 66^{th} iteration the model was trained on the whole dataset.

In the last case (blue line) initial 700 documents were selected only from two groups. After the 33^{rd} iteration of training, documents from the remaining newsgroups were gradually introduced in the order of their newsgroup membership. It should be noted here that in this case we had a priori information

on the document category (i.e. particular newsgroup). In the general case, we are collecting fuzzy category membership information from the Bayesian Net model.

As expected, in all cases the GNG model adapts quite well to the topic drift. In the non-incremental and the topic-wise incremental cases, the quality of the models were comparable, in terms of Average Document Quantization measure (see figure 5(a)), Average Weighted Cluster Purity, Average Cluster Entropy and Normalized Mutual Information (for the final values see table 1). Also the subjective criteria such as the visualization of both models and the identification of thematic areas on the SOM projection map were similar.

Table 1. Final values of model quality measures

	Cluster Purity	*Cluster Entropy*	*NMI*
non-incremental	0.91387	0.00116	0.60560
topic-wise incremental	0.91825	0.00111	0.61336
massive addition	0.85596	0.00186	0.55306

The results were noticeably worse for the massive addition of documents, even though all covered topics were present in the training from the very beginning and should have occupied specialized thematic areas in the model graph. However, and it can be noticed on the same plot, a complex mixture of topics can pose a serious drawback, especially in the first training iterations. In the non-incremental, reference case, the attempt to cover all topics at once leads the learning process to a local minimum and to subsequent divergence (which, moreover, is quite time-consuming as one can notice in figure 4(a)). As we have previously noticed, the problem of convergence to a local minimum was even more influential in the case of global winner search (figure 3(b)).

However, when we take advantage of the incremental approach, the model ability to separate document categories is comparable for global search and CF-tree based search (Cluster Purity: 0.92232 versus 0.91825, Normalized Mutual Information: 0.61923 versus 0.61336, Average Document Quantization: 0.64012 versus 0.64211).

Figure 4(b) presents the average number of GNG graph edges traversed by a document during a single training iteration. It can be seen that a massive addition causes temporal instability of the model. Also, the above mentioned attempts to cover all topics at once in the case of a global model caused much slower stabilization of the model and extremely high complexity of computations (figure 4(a)). The last reason for such slow computations is the representation of the GNG model nodes. The referential vector in such a node is represented as a balanced red-black tree of term weights. If a single node tries to occupy too big portion of a document-term space, too many terms appear in such

a tree and it becomes less sparse and – simply – bigger. On the other hand, better separation of terms which are likely to appear in various newsgroups and increasing "crispness" of thematic areas during model training leads to highly efficient computations and better models, both in terms of previously mentioned measures and subjective human reception of the results of search queries.

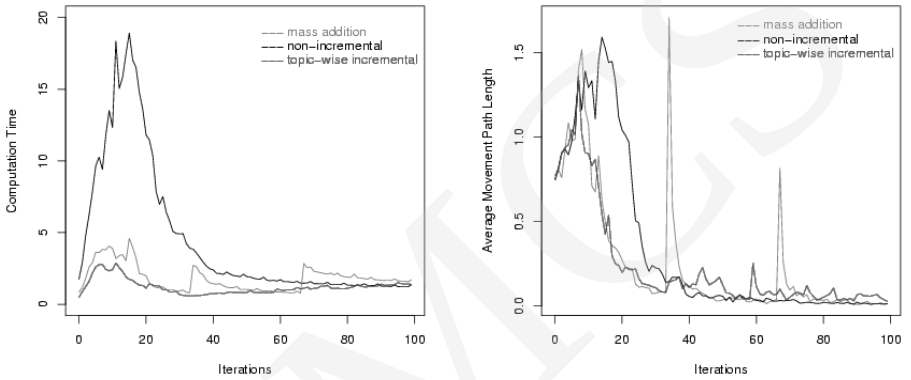

Fig. 4. Computation complexity (a) execution time of a single iteration
(b) average path length of a document

The last figure, 5(b), compares the change in the value of Average Map Quantization measure, reflecting "smoothness" of the model (i.e. continuous shift between related topics). In all three cases the results are almost identical. It should be noted that extremely low initial value of the Average Map Quantization is the result of the model initialization via the broad topics method [1], shortly described in section 4.

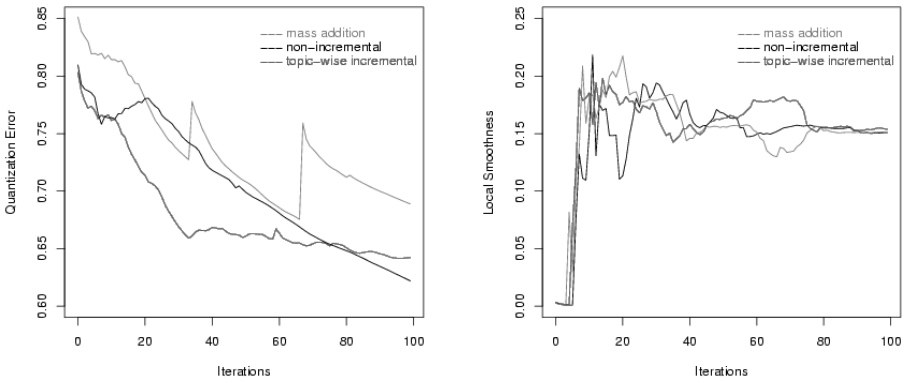

Fig. 5. Model quality (a) Average Document Quantization (b) Average Map Quantization

6. Related works

Modern man faces a rapid growth in the amount of written information. Therefore he needs a means of reducing the flow of information by concentrating on major topics in the document flow. Grouping documents based on similar contents may be helpful in this context as it provides the user with meaningful classes or clusters. Document clustering and classification techniques help significantly in organizing documents in this way. A prominent position among these techniques is taken by the WebSOM (Self Organizing Maps) of Kohonnen and co-workers [12]. However, the overwhelming majority of the existing document clustering and classification approaches rely on the assumption that the particular structure of the currently available static document collection will not change in the future. This seems to be highly unrealistic, because both the interests of the information consumer and of the information producers change over time.

A recent study described in [19] demonstrated deficiencies of various approaches to document organization under non-stationary environment conditions of growing document quantity. The mentioned paper pointed to weaknesses, among others, of the original SOM approach (which itself is adaptive to some extent) and proposed a novel dynamic self-organizing neural model, so-called the Dynamic Adaptive Self-Organising Hybrid (DASH) model. This model is based on an adaptive hierarchical document organization, supported by human-created concept-organization hints available in terms of WordNet.

Other strategies like that of [10,20], attempt to capture the move of topics, enlarge dynamically the document map (by adding new cells, not necessarily on a rectangle map).

We take a different perspective in this paper claiming that the adaptive and incremental nature of a document-map-based search engine cannot be confined to the map creation stage alone and in fact engages all the preceding stages of the whole document analysis process.

7. Conclusions

As indicated e.g. in [19], most document clustering methods, including the original WebSOM, suffer from their inability to accommodate streams of new documents, especially such in which a drift, or even radical change of topic occurs.

Though one could imagine that such an accommodation could be achieved by “brute force” (learning from scratch whenever new documents arrive), but there exists a fundamental technical obstacle for a procedure like that: the processing time. But the problem is more profound and has a “second bottom”: the clustering methods like those of WebSOM contain elements of randomness so

that even re-clustering of the same document collection may lead to a radical change of the view of the documents. The results of this research are concerned with both aspects of adaptive clustering of documents.

The important contribution of this paper is to demonstrate, that the whole incremental machinery not only works, but it works efficiently, both in terms of computation time, model quality abd usability. For the quality measures we investigated, we found that our incremental architecture compares well to non-incremental map learning both under scenario of "massive addition" of new documents (many new documents, not varying in their thematic structure, presented in large portions) and of scenario of "topic-wise-increment" of the collection (small document groups added, but with new emerging topics). The latter seemed to be the most tough learning process for incremental learning, but apparently the GNG application prior to WebSOM allowed for cleaner separation of new topics from those already discovered, so that the quality (e.g. in terms of cluster purity and entropy) was higher under incremental learning than under non-incremental learning.

The experimental results indicate, that the real hard task for an incremental map creation process is a learning scenario where the documents with new thematic elements are presented in large portions. But also in this case the results proved to be satisfactory.

A separate issue is the learning speed in the context of crisp and fuzzy learning models. Apparently, separable and thematically "clean" models allow for faster learning as the referential vectors in the model nodes are smaller (contain fewer non-zero components).

From the point of view of incremental learning under soft-competitive scenario, a crucial factor for the processing time is the winner search method for assignment of documents to neurons. We were capable to elaborate a very effective method of stable, context-dependent winner search which does not deteriorate the overall quality of the final map. At the same time, it comes close to the speed of local search and is not directly dependent on the size of the model.

Our future research will concentrate on exploring further adaptive methods like artificial immune systems [11] for reliable extraction of context-dependent thesauri and adaptive parameter tuning.

References

[1] Kłopotek M., Dramiński M., Ciesielski K., Kujawiak M., Wierzchoń S.T., *Mining document maps*. in Proceedings of Statistical Approaches to Web Mining Workshop (SAWM) at PKDD'04, M. Gori, M. Celi, M. Nanni eds., Pisa, (2004) 87.

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

[3] Ciesielski K., Dramiński M., Kłopotek M., Kujawiak M., Wierzchoń S., *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. Akademicka Oficyna Wydawnicza EXIT Publishing, Warszawa, (2004) 122.

[4] Ciesielski K., Draminski M., Kłopotek M., Kujawiak M., Wierzchoń S.T., *On some clustering algorithms for Document Maps Creation*. to appear in: Proceedings of the Intelligent Information Processing and Web Mining (IIS:IIPWM-2005), Gdańsk, (2005).

[5] Kłopotek M., Wierzchoń S., Ciesielski K., Draminski M., Czerski D., Kujawiak M., *Understanding Nature of Map Representation of Document Collections – Map Quality Measurements*. to appear in Proceeding of International Conference on Artificial Intelligence, Siedlce, (2005).

[6] Ciesielski K., Draminski M., Kłopotek M., Kujawiak M., Wierzchoń S.T., *Crisp versus Fuzzy Concept Boundaries in Document Maps*. to appear in: Proceedings of DMIN-05 Workshop at The 2005 World Congress in Applied Computing, Las Vegas, (2005).

[7] Berry M.W., *Large scale singular value decompositions*. International Journal of Supercomputer Applications, 6(1) (1992) 13.

[8] Kłopotek M, *A New Bayesian Tree Learning Method with Reduced Time and Space Complexity*. Fundamenta Informaticae, IOS Press,49(4) (2002) 349.

[9] Fritzke B., *A growing neural gas network learns topologies*. in: G. Tesauro, D.S. Touretzky, and T.K. Leen (Eds.) Advances in Neural Information Processing Systems 7, MIT Press Cambridge, MA, (1995) 625.

[10] Dittenbach M., Rauber A., Merkl D., *Discovering Hierarchical Structure in Data Using the Growing Hierarchical Self-Organizing Map*. Neurocomputing, Elsevier, ISSN 0925-2312, 48(1-4) (2002) 199.

[11] De Castro L.N., von Zuben F.J., *An evolutionary immune network for data clustering*. SBRN'2000, IEEE Computer Society Press, (2000).

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

[13] Fritzke B., *Some competitive learning methods*. draft available from http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper

[14] Fritzke B., *A self-organizing network that can follow non-stationary distributions*. in: Proceedings of the International Conference on Artificial Neural Networks '97, Springer, (1997) 613.

[15] Zhang T., Ramakrishan R., Livny M., *BIRCH: Efficient Data Clustering Method for Large Databases*. in: Proceedings of ACM SIGMOD International Conference on Data Management, (1997).

[16] Timmis J., *aiVIS: Artificial Immune Network Visualization*. in: Proceedings of EuroGraphics UK 2001 Conference, Univeristy College London, (2001) 61.

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

[18] Halkidi M., Batistakis Y., Vazirgiannis M., *On Clustering Validation Techniques*. Journal of Intelligent Information Systems 17(2-3) (2001) 107.

[19] Hung C., Wermter S., *A Constructive and Hierarchical Self-Organising Model in A Non-Stationary Environment*. International Joint Conference in Neural Networks, (2005).

[20] Rauber A., *Cluster Visualization in Unsupervised Neural Networks*. Diplomarbeit, Technische Universität Wien, Austria, (1996).