Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 77–81

ISBN 978-83-60810-22-4

Growing Hierarchical Self-Organizing Map for searching documents using visual content

Paweł B. Myszkowski
Applied Informatics Institute,
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 51-370 Wrocław, POLAND
email: pawel.myszkowski@pwr.wroc.pl

Bartłomiej Buczek
email: bartlomiej.buczek@op.pl

***Abstract*—This paper presents document search model based on its visual content. There is used hierarchical clustering algorithm - GHSOM. Description of proposed model is given as learning and searching phase. Also some experiments are described on benchmark image sets (e.g. ICPR, MIRFlickr) and created document set. Paper presents some experiments connected with document measures and their influence on searching results. Also in this paper some first results are given and directions of further research are given.**

I. Introduction

THE document search task is connected to large dataset of documents which can be used in many applications e.g. in libraries to find a relevant document similar to one given in a query. Such task can be solved as classification task of data mining domain in knowledge discovery database process. However, classification needs labeling of all documents and this is nearly impossible because of time/cost constraints connected to labeling by human. Thus our work concerns on clustering task in unsupervised learning mode, where structure of data is not given or we barely know anything about it. Also, we decided to use a special type of clustering – hierarchical clustering to build a hierarchy of documents. It gives us a very useful advantage in navigation of document search space to find similar documents navigating between included images that can be connected and further in searching for similar documents basing of its visual aspects. However, our task is specific, where the document is analyzed only in the visual context. So basically our model bases only on visual content of document, not text. The visual content of document, in this stage of research, means only included images, figures, tables and schemas. Our work is alternative to semantic text based document analyses and indeed our approach results would be linked to such method in the complex system.

There are many methods that bases on hierarchical clustering of data for previously artificially created hierarchy by human. There are also some methods, like Growing Hierarchical Self-Organizing Map (*GHSOM*) which creates hierarchy from scratch.

This work is partially financed from the Ministry of Science and Higher Education Republic of Poland resources in 2010-2013 years as a research SYNAT project (System Nauki i Techniki) in INFINITI-PASSIM.

Description of methods and architectures based on similar images search or image category search, clustering or clusterisation can be found in work [13]. Another survey [5] presents various approaches to document layout description, document/images features selection, classification and application. Models are hidden markov model, neural network, k-Nearest Neighbor or rule based approaches.

This work presents our first research results for quality of documents search using *GHSOM*. The 2nd section shows ideas of organizing data as a map and GHSOM as extension of *SOM*. Section 3rd presents proposed approach, architecture and connected processes. Some experiments are included in section 4th, it also shows results of our first research. Last section concludes and describes directions of further research.

II. Growing Hierarchical Self-Organizing Map

Organizing data was firstly proposed in 1982 by Kohonen to explain the spatial organization of the brain's functions. The data is presented as neural network with the aid of adaptation of weights vectors becomes organized [10]. The Self-Organizing Map (*SOM*) is a computational, unsupervised tool to the visualization and analysis of high-dimensional data. There are plenty applications where SOM is used, especially in text mining [10].

A useful way to organize data presents Mükkulainen [3]. His method can be described as Hierarchical Self-Ogranizing Map (*HSOM*) which relays on series of *SOMs* which are placed in layers where the number of *SOMs* in each structure layer depends on number of *SOMs* in previous layers (the upper one). In this model number of structure layers in *HSOM* and dimension of maps are defined *a priori.*

In the learning process (which always starts from top layers to the bottom layer) units weight vectors in *SOMs* of each layer are adapted. The adapt of weight in next layer is only possible when given layer is finished. Final effect of this process is hierarchical clustering of data set. This approach has some advantages:

- smaller number of connections between input and units in *HSOM* layers [3],

- much shorter processing time which comes from the point above and from hierarchical structure of learning process [3].

In *HSOM* there are some necessities:

- definition of maps sizes and number of layers which depends on data set,
- choosing of learning parameters for each layer in HSOM.

There is another *SOM* extension that reduces some above disadvantages. Growing Self-Organizing Map (*GSOM*) is another variant of SOM [1] to solve the map size issue. The model consists of number of units which in learning process grows. The learning process begins with minimal number of units and grows (by adding new units) on the boundary based on a heuristic. To control the growth of the *GSOM* there is special value called Spread Factor (*SF*) [1] - a factor that controls the size of growth. At the beginning of learning process all units are boundary units which means that each unit has the freedom to grow in its own direction. If the unit is chosen to grow it grows in all its free neighboring positions.

However in *HSOM* still exists a problem of *a priori* given architecture. Another approach, Growing Hierarchical Self-Organizing Map (*GHSOM*) solves it.

The architecture of *GHSOM* allows to grow in both a hierarchical and in a horizontal direction [2]. This provides a convenient structure of clustering for large data sets which can be simple navigate through all layers from the top to the bottom.

The learning process begins with a virtual map which consists of one vector initialized as average of all input data [15]. Also for this unit, error (distance) between its weight vector and all data is calculated - this error is global stop parameter in *GHSOM* learning process. Then next layer map (usually initialized by 2x2 [8]) is created below the virtual layer (this newly created *SOM* indeed is a child for unit in virtual layer). From now on, each map grows in size to represent a collection of data at a specified level of detail. The main difference between growing in *GHSOM* from *GSOM* is that the whole row or column of units is added to currently learning *SOM* layer. The algorithm can be shortly described as:

1.For present SOM start learning process and finish it after λ-iterations.

2. For each unit count error (distance) between its weight vector m_i and input patterns $x(t)$ mapped onto this unit in the SOM learning process (this error is called quantization error qe).

3. If the sum of quantization errors is greater or equal than certain fraction of quantization error of parent unit:

$$\sum qe_i \geq \tau_1 * qe_{parent_{unit}}$$

3a. If yes - select unit with biggest error and find neighbor to this unit which is most dissimilar. Go to step 4.

3b. else - stop.

4. Between these two units put row/column.

5. Reset learning parameter and neighbor function for next SOM learning process.
Go to step 1.

When the growth process of layer is finished the units of this map are examined for hierarchical expansion [11]. Basically, units with large quantization error will add a new SOM for the next layer in the *GHSOM* structure according to τ_1 value. More specifically, when quantization error of unit in examined map is greater than fraction τ_2 of global stop parameter, then a new *SOM* (usually initialized by 2x2) is added to structure. The learning process and unit insertion now continue with the newly established *SOMs*. The difference in learning process of new layer is that fraction of data set is used for training corresponding to units mapped into parent unit [2].

The whole learning process of the *GHSOM* is terminated when no more units are required for further expansion. This learning process does not necessarily lead to a balanced hierarchy (the hierarchy with equal depth in each branch) [15]. To summarize, the growth process of the *GHSOM* is guided by two parameters τ_1 and τ_2. The parameter τ_1 controls the growth process in layer (certain fraction in algorithm for expanding in layer for *GHSOM*) and the parameter τ_2 controls hierarchical growth of *GHSOM*.

The clustering of images is needed because, when today's search engine is asked about images usually as an answer comes images which were named by the asked phrase. It would be better if the answer come on the base of what image contains. The content of images can be described by a vector of features - vector of numbers. *GHSOM,* in simply way, bases on vectors of real numbers so images can be successfully used as a data set. In this work the idea was to create a *GHSOM* structure which will hierarchically organize images in groups of similar images. The visual aspects of such clustering (visualization of images in cluster) can be analyzed by human or by some quality measures and ratings.

III. Proposed *GHSOM* based model

Model base on *GHSOM* for searching similar documents in database includes mainly 3 modules: document preprocessing (for visual element extraction and description), *GHSOM* structure and document of indexed documents. The schema of proposed model is given on Fig.1.

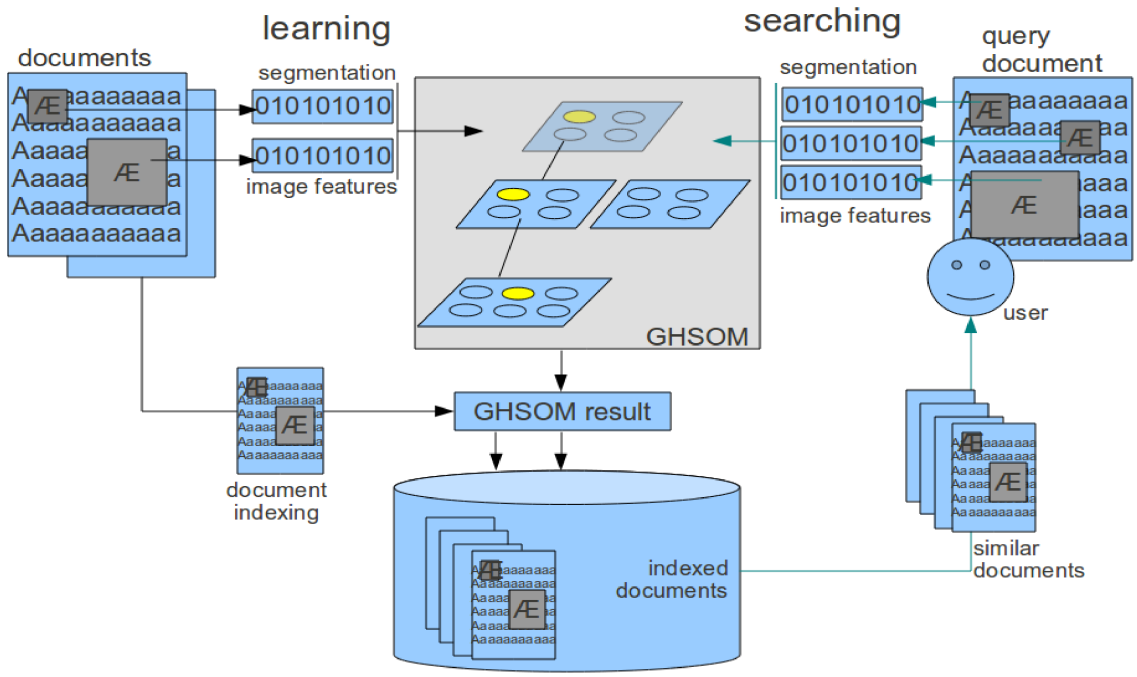

Fig. 1 Proposed document searching model

A. Learning phase

The first stage of such process is preprocessing. Each document is analyzed by getting its visual content: figures, tables, schemas and others. Each visual element is divided in 25 segments (5x5 grid); for each segment image features are calculated (e.g. feature connected to image color, structure or shapes). Each visual element is represented as a sequence of vectors (=segments) and set of visual elements gives the whole representation of document.

Next stage of learning phase uses the hierarchical clustering in *GHSOM* model to organize space of set of documents. Each visual element is presented to *GHSOM* to build hierarchical cluster that would be used in search phase.

In the next step *GHSOM* learning procedure runs and gives hierarchical clustering structure for document indexing. The document is indexed and then registered into base of all indexed documents as vector of output *nodes* from GHSOM structure. The length of vector equals to the total number of *node* units, and each node unit has the own identificator and place in output document vector $\Phi(document_i)$. If its activated value is not zero, e.g. if in *i-th* document the 2nd node has been activated once and 6th has been activated twice output document vector is:

$$\Phi(document_i) := <0,0,1,0,0,0,2,0>.$$

Such representation is intuitive and allows to use document similarity or metrics in simple way (very crucial aspect in searching phase). On the end of learning proces each already indexed document is registered in database, that works the special role in searching phase.

B. Searching phase

The already learned GHSOM can be used for searching similar documents in database of all indexed documents. We assume that the best way of searching is the document query $document_{query}$. The query document has to be converted by preprocessing procedure (finding visual elements, its segmentation and features calculations) and it allows to present to *GHSOM* structure the document representation. Already calculated the output of activated GHSOM units gives $\Phi(document_{query})$ and it becomes the base of similar documents searching in database. According to used document measure (we experimented mainly with euclidean and cosine measures) it returns the sequence of *n* most similar documents in database.

C. Model parameters discussion

The proposed model parametrization is not a complex problem. The more difficult issue is selection of visual images features. In this stage of research we used only color RGB based features, but we plan to extend this representation by adding extra features connected to shape and texture. The GHSOM algorithm uses only two basic values for steering learning process τ_1 and τ_2. Another parameter is connected with database and searching process, where document measure must be given. We experimented with euclidean and canonical cosine measures of documents representation and it gives very promising results.

IV. Experiments and first results

We developed proposed model as programistic platform written in Java for experiments. Our experiments are divided into separate tasks: similar image search and similar document search. The motivation of the first task is connected with series of experiments to basic *GHSOM* model quality verification. Results of second types of experiments give information about effectiveness of the whole model.

A. Similar image searching

The experiments of image hierarchical clustering using GHSOM were based on benchmark images dataset: ICPR[1], IAPR TC12[2] and MIRFlickr[3]. These datasets consist of labeled images that make possible not only image clusterisation but also verification of image distribution in GHSOM hierarchical structure. Our research was concerned on:

- external measure of *GHSOM*: image query to model and verification according to connected with image and *GHSOM* unit keywords,
- internal measure of *GHSOM* as the set of image clusters – we used Davies Bouldin Index (*DBI*) and Dunn Index (*DI*)

Our experiments, detailed in description [4], show that *GHSOM* is a useful tool of image hierarchical clustering. Although we used only RGB color image features, results are very successful. The example of proper *GHSOM* clustering image distribution is presented in Fig. 2.

Fig. 2 Example of GHSOM images clustering – 'buildings' in benchmark ICPR benchmark image dataset

The *GHSOM* application to similar images searching task gives also some hints about further research. There are some badly created cluster and badly classified images (but the percent of them are respectively small). In the future, visual aspects of images in hierarchical clusters created by *GHSOM* can be researched. As well, in this work not only classical Euclidean distance was used (but also L1 and Tchybyszew's measure) but another experiments with different measures for distance should be applied.

[1]ICPR dataset: http://www.cs.washington.edu/research/
[2]IAPR TC12 dataset: http://www.imageclef.org/photodata/
[3]MIRFlikr dataset: http://www.flickr.com/

B. Similar documents searching

The main task of GHSOM in proposed model is similar document searching. We created a dataset 1kPDF that consist of 1000 PDF documents. Each document is in polish language and has at least one visual element. Documents are not labeled but only grouped into 100 or 200 domain groups: sport, motorization, general, architecture and art. Each group is created by other person to make the whole document set more independent. Thus documents in various domains can be alike.

We developed on experiments to examine the search quality of model using given document measure. 25 selected randomly from various groups: 4 documents are connected with architecture, 8 motorization documents, 8 general, 4 sport and 1 concerns art. Search results for cosine document measure for these 25 documents are presented in Table I. The 'general' domain also has high effectiveness (33%). It means that these documents are more relevant in 10 answer documents.

The worst result is observed in art (5%) and sport (2,7%) documents. For art document query situation seems to be problematic in interpretation as there is chosen only one document connected with art domain and result cannot be representative. Also, not clear situation occurs for sport, where only 5% of relevant answers were given. However, selected documents connected to sport are very specific – they are short and often consist of one image. Search result returns on first place similar documents (in sport domain) but some answered documents are connected to others domains.

TABLE I.
RESULTS OF FIRST EXPERIMENTS – COSINE DOCUMENT MEASURE

Q \ A	archi-tect	motor.	general	sport	art
architect.	**25,00%**	33,75%	32,50%	6,25%	2,50%
motor.	31,87%	**26,87%**	35,00%	2,50%	3,75%
general	35,00%	25,64%	**33,12%**	1,25%	5,00%
sport	22,97%	27,02%	36,48%	**2,70%**	10,80%
art	45,00%	30,00%	15,00%	5,00%	**5,00%**

The same set of 25 documents was used to examine Euclidean document measure (see Table II). Such measure seems to be more effective in search similar documents. The

TABLE II.
RESULTS OF FIRST EXPERIMENTS – EUCLIDEAN DOCUMENT MEASURE

Q \ A	archi-tect	motor.	general	sport	art
architect	**21,25%**	40,00%	18,75%	6,25%	13,75%
motor.	5,62%	**63,12%**	11,25%	7,50%	12,50%
general	9,37%	45,62%	**21,25%**	8,12%	15,62%
sport	0,00%	66,25%	7,50%	**11,20%**	15,00%
art	0,00%	35,00%	20,00%	10,00%	**35,00%**

best result was for 'motor' domain – 63%, the worst for sport group (11,2% valid, and 66% answer is motorcycle).

The example query for 'motor' document is presented on Table III, where 9 first most similar documents of 20 are relevant to questioned document, which makes answer very accurate.

TABLE III.
EXAMPLE RESULTS OF 'MOTORYZACJA/620' DOCUMENT QUERY DOCUMENT USING EUCLIDEAN DOCUMENT MEASURE

motoryzacja/656; 17.34935	
motoryzacja/628; 17.52141	
motoryzacja/1293; 17.7200	…
motoryzacja/613; 18.41195	sztuka/1376; 20.09975124
motoryzacja/1269; 18.4661	arch-bud/529; 20.12461179
motoryzacja/1290; 19.39071	**motoryzacja**/627; 20.17424
motoryzacja/1288; 19.51922	**motoryzacja**/1291; 20.1990
motoryzacja/645; 19.748417	arch-bud/572; 20.2731349
motoryzacja/665; 19.899748	arch-bud/569; 20.29778313
arch-bud/552; 20.02498439	ogolne/1085; 20.3224014
ogolne/1115; 20.07485989	ogolne/1060; 20.3469899
…	ogolne/1058; 20.37154878

Our experiments show that model gives very promising results, however it difficult to say which measure is better. Also, results are better than it seems because groups of documents are alike and some documents can be labeled as member of two groups e.g. 'architecture' documents can be included to 'art' group too.

V. SUMMARY AND FURTHER RESEARCH

This paper presents results of our first experiments with *GHSOM* based document searching model. Our approaches use hierarchical clustering of image content of collected documents to search for similar document. Paper presents results of experiments connected to different document measure (Euclidean and classic Cosine one). Thus there is no the best document measure and in conclusion is that there is strong need of extra experiments series. The important factor of future research is the image feature selection. In this stage only *rbg* based features were selected while there are useful features like texture and shape.

Presented series of experiments are initiative and extra experiments are planned for document collections of 10 000 documents (results given in paper bases of 1000 document collection). The larger collection may give the experiments more precision and generalization, which also allows to do various experiments.

REFERENCES

[1] Alahakoon, D., Halgamuge, S. K., Sirinivasan, B.: A Self Growing Cluster Development Approach to Data Mining. Proc. of IEEE Inter.Conf. on Systems, Man and Cybernetics (1998)

[2] Bizzil, S., Harrison, R.F., Lerner, D. N.: The Growing Hierarchical Self-Organizing Map (GHSOM) for analysing multi-dimensional stream habitat datasets. 18th World IMACS / MODSIM Congress (2009)

[3] Blackmore, J., Mükkulainen, R.: Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. Proc. of the IEEE Inter. Conf. on Neural Networks (1993).

[4] Buczek B.M., Myszkowski P.B.: Growing Hierarchical Self-Organizing Map for Images Hierarchical Clustering, ICCCI proc. Conferences (*in printing*), LNCS 2011.

[5] Nawei Chen, Dorothea Blostein.: A survey of document image classification: problem statement, classifier architecture and performance evaluation, IJDAR 10, pp.1–16, (2007)

[6] Chih-Hsiang, C., Chung-Hong, L., Hsin-Chang, Y.:Automatic Image Annotation Using GHSOM. Fourth Inter. Conf. on Innovative Comp.,Infor. and Control (2009)

[7] Fritzke, B.: Some Competitive Learning Methods. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.3885

[8] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity check-ing methods: part II. ACM SIGMOD Record Vol. 31 (3) (2002)

[9] Herbert, J. P., JingTao Yao: Growing Hierarchical Self-Organizing Maps for Web Mining. Proc. of the 2007 IEEE/WIC/ACM Inter. Confer.e on Web Intel. (2007)

[10] Huiskes, M. J., Lew, M. S.: The MIR Flickr Retrieval Evaluation. ACM Inter. Conf. on Multimedia Inf. Retrieval (2008)

[11] Kohonen, T.: Self-organizing maps, Springer-Verlag, Berlin (1995)

[12] Rauber, A., Merkl, D., Dittenbach, M.: The GHSOM: Exploratory Analysis of High-Dimensional Data. IEEE Trans. on Neural Networks(2002)

[13] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang.: Image Retrieval: Ideas, Influences, and Trends of the New Age, ACM Computing Surveys, Vol. 40, No. 2, Article 5 (2008).

[14] Raza A., Usman G., Aasim S.: Data Clustering and Its Applications. http://members.tripod.com/asim_saeed/paper.htm

[15] Vicente, D., Vellido, A.: Review of Hierarchical Models for Data Clustering and Visualization. [In] R.Giraldez et al. (Eds.) Tendencias de la Minera de Datos en Espa. Espa´ola de Minera de Datos (2004)

[16] Hierarchical clustering. http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_hierarchical_clustering.htm

[17] Experiments with GHSOM. http://www.ifs.tuwien.ac.at/~andi/ghsom/experiments.html