Contextual Maps for browsing Huge Document Collections

Krzysztof Ciesielski[1] and Mieczysław A. Kłopotek[2,1]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Ordona 21, 01-237 Warszawa,Poland
[2] Institute of Computer Science, University of Podlasie,
ul. Sienkiewicza 51, 80-110 Siedlce, Poland
kciesiel,klopotek@ipipan.waw.pl

Abstract. The increasing number of documents returned by search engines for typical requests makes it necessary to look for new methods of representation of contents of the results, like document maps. Though visually impressive, doc maps (e.g. WebSOM) are extensively resource consuming and hard to use for huge collections.

In this paper, we present a novel approach, which does not require creation of a complex, global map-based model for the whole document collection. Instead, a hierarchy of topic-sensitive maps is created. We argue that such approach is not only much less complex in terms of processing time and memory requirement, but also leads to a robust map-based browsing of the document collection.

1 Introduction

The rapid growth in the amount of written information prompts for a means of reducing the flow of information by concentrating on major topics in the document flow, including the one on the World Wide Web. Clustering documents based on similar contents may be helpful in this context as it provides the user with meaningful classes or groups. One of the prominent approaches was the WebSOM project, producing a two-dimensional map of document collection, where spacial closeness of documents reflected their conceptual similarity. Hence cluster membership depends not only on other cluster members, but also on the inter-cluster 2D grid structure. This approach results in more intuitive clustering, but also imposes huge computational burden.

A recent study [4] demonstrated also deficiencies of various approaches to document organization (including WebSOM) under non-stationary environment conditions of growing document quantity, in terms of both stability and deficiency. A dynamic self-organizing neural model, so-called Dynamic Adaptive Self-Organising Hybrid (DASH) model, based on an adaptive hierarchical document organization, supported by human-created concept-organization hints available in terms of WordNet, has been proposed out of that study.

In this paper, we propose a novel approach to both the issue of topic drift and scalability. Section 2 explains in detail the concept of so-called contextual maps.

While being based on a hierarchical (three-level) approach, it is characterized by three distinctive features. In our opinion, WebSOM-like clustering is inefficient, because its target 2D grid is too rigid. Hence, we propose first to use growing neural gas (GNG) clustering technique, which is also a structural one, but has a more flexible structure accommodating better to non-uniform similarity constraints. The GNG is then projected onto a 2D map, which is less time consuming than direct WebSOM like map creation. In fact, any other structural clustering like aiNet (artificial immune system approach) could be used instead of GNG. The second innovation is the way we construct the hierarchy: first, we split the documents into (many) independent clusters (which we call "contexts"), then apply structural clustering within them, and in the end cluster structurally the "contexts" themselves. What we gain, is the possibility of drastic dimensionality reduction within the independent clusters (as they are more uniform topically) which accelerates the process and stabilizes it. The third innovation is the way we apply GNG technique. Instead of the classical global search, we invented a mixed global/local search especially suitable for GNG.

Out of these inventions, we gain speed. But not at the expense of final map quality. But what is more, the new techniques deal surprisingly well with non-stationarity of document flow - the issues of topic drift and the rapid growth of document collection. We demonstrate the validity of these claims in section 3. The last section contains some final comments on presented approach and future research directions.

2 Contextual maps

In our work we use well known term vector space approach to document representation. It is a known phenomenon that text documents are not uniformly distributed over that space. Characteristics of frequency distributions of a particular term depend strongly on document location. In our approach we automatically identify groups of similar documents in a preprocessing step. We argue that after splitting documents in such groups, term frequency distributions within each group become much easier to analyze. In particular, it appears to be much easier to select significant and insignificant terms for efficient calculation of similarity measures during map formation step. Such document clusters we call *contextual* groups (or "contexts"). For each contextual group, separate maps are generated. To obtain more informative maps there is a need to balance (during initial contextual clustering) size of each cluster. The number of documents presented on a map cannot be too high due to rapidly growing computational time. On the other hand, ac map should not hold only a few irrelevant documents.

Constraints on cluster size are matched by recurrent divisions and merges of fuzzy document groups, created by a Fuzzy C-Means (ISODATA) algorithm. There is an additional modification in optimized quality criterion that penalizes for inbalanced splits (in terms of cluster size).

In the first step, whole document set is split into a few (2-5) groups. Next, each of these groups is recursively divided until the number of documents inside a group meets required criteria. So we obtain a tree of clusters. In the last phase, groups which are smaller than predefined constraint, are merged to the closest group[3]. Similarity measure is defined as a single-linkage cosine angle between both clusters centroids.

Crucial phase of contextual document processing is the division of terms space (dictionary) into - possibly overlapping - subspaces. In this case it is important to calculate fuzzy membership level, which will represent importance of a particular word or phrase in different contexts (and implicitly, ambiguity of its meaning). Fuzzy within-group membership of the term m_{tG} is estimated as:

$$m_{tG} = \frac{\sum_{d \in G} (f_{td} \cdot m_{dG})}{f_G \cdot \sum_{d \in G} m_{dG}} \tag{1}$$

where f_G is the number of documents in cluster G, m_{dG} is the membership degree of document d in G, f_{td} is the number of occurrences of term t in d.

Next, vector-space representation of a document is modified to take into account document context. This representation increases weights of terms which are significant for a given contextual group and decrease weights of insignificant terms. In the extreme case, insignificant terms are ignored, what leads to the (topic-sensitive) reduction of space dimensionality. To estimate the significance of term in a given context, the following measure is applied:

$$w_{tdG} = f_{td} \cdot m_{tG} \cdot log\left(\frac{f_G}{f_t \cdot m_{tG}}\right) \tag{2}$$

where f_{td} is the number of occurrences of term t in document d, m_{tG} is the degree of membership of term t in group G, f_G is the number of documents in group G, f_t is the number of documents containing term t.

As mentioned, instead of the rigid WebSOM like 2D grid, we use the more flexible GNG [2] model. Main idea behind our approach is to replace a single structural model by a set of independently created contextual models and to merge them together into a hierarchical model. Training data for each model is a single contextual group. Each document is viewed as a standard vector in term-document space, but we use w_{tdG} instead of the $tfidf$ measure.

Notice that in original GNG [2] , like in WebSOM, the most computationally demanding part is the winner search phase. The replacement of global search with local one is not applicable because the GNG graph may not be connected. We propose a simple modification consisting in remembering winning node for more than one connected component of the GNG graph. To increase accuracy, we apply the well-known Clustering Feature Tree [7] to group similar nodes in dense clusters. Node clusters are arranged in the hierarchy and stored in a balanced search tree. Thus, finding 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 [7] for details). Amortized tree structure maintenance cost (node insertion and removal) is also proportional to $O(log_t V)$.

[3] to avoid formation of additional maps which would represent only a few outliers in document collection

To represent visually similarity relation between contexts, additional "global" map is required. Such model becomes a root of contextual maps hierarchy. Main map is created in a manner similar to previously created maps, with one distinction: an example in training data is a weighted centroid of referential vectors of the corresponding "context": $x_i = \sum_{c \in M_i} (d_c \cdot v_c)$, where M_i is the set of cells in i-th contextual model, d_c is the density of the cell and v_c is its referential vector.

Main map cells and regions are labeled with keywords selected by our contextual term quality measure: $Q_{tG} = ln(1 + f_{tG}) \cdot (1 - |EN_{tG} - 0.5|)$, where EN_{tG} denotes normalized entropy[4] of term frequency distibution within the group.

Learning process of the contextual model is to some extent similar to the classic, non-contextual learning. However, it should be noted that each constituent model can be processed independently, even in parallel. Also a partial incremental update of such models is better manageable in terms of model quality, stability and time complexity. The incrementality is in part a consequence of the iterative nature of the learning process. So if new documents come, we can consider the learning process as having been stopped at some stage and it is resumed now with all the documents. We claim that it is not necessary to start the learning process from scratch neither in the case that the new documents "fit" the distribution of the previous ones nor when their term distribution is significantly different. This claim is supported by experimental results presented in the section 3.2. In the section 3.3 we present some thoughts on scalability issues of contextual approach.

3 Experimental results

To evaluate the effectiveness of the presented contextual approach, we compared it to the "from scratch" map formation. The architecture of our visual search engine BEATCA [5] supports comparative studies of clustering methods at the various stages of processing of document collection. In this paper we focus on evaluation of the GNG winner search method and the quality and stability of the resulting incremental clustering model with respect to the topic-sensitive learning approach. Below we describe the overall experimental design, quality measures used and the results obtained. The incrementality study in section 3.2 required manually labeled documents, so the experiments were performed on a subset of widely-used "20 Newgroups" document collection. The scalability study in section 3.3 was based on a collection of more than one million Internet documents, crawled by our topic-sensitive crawler.

3.1 Quality Measures for the Document Maps

A document map may be viewed as a special case of the concept of clustering. One can say that clustering is a learning process with hidden learning criterion. The criterion is intended to reflect some esthetic preferences, like: uniform

[4] entopy divided by the number of the terms in the group

split into groups (topological continuity) or appropriate split of documents with known a priori categorization. As the criterion is hidden, in the literature [8, 1, 3] a number of clustering quality measures have been developed, checking how the clustering fits the expectations. We selected the following ones for our study:

- **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**: 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.
- **Average Weighted Cluster Purity**: average "category purity" of a node (node weight is equal to its density, i.e. the number of assigned documents): $AvgPurity = \frac{1}{|D|}\sum_{n \in N} max_c\left(|D_c(n)|\right)$, 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. Similarly, *Average Weighted Cluster Entropy* measure can be calculated, where $D_c(n)$ term is replaced with the entropy of the categories frequency distribution.
- **Normalized Mutual Information**: the quotient of the entropy with respect to the categories and clusters frequency to the square root of the product of category and cluster entropies for individual clusters [1].

All measures range from 0 to 1. First two describe smoothness of inter-cluster transitions and cluster "compactness" (the lower the better). The other two evaluate the agreement between the clustering and the a priori categorization of documents (i.e. particular newsgroup in case of newsgroups messages). The higher the value is, the better agreement between clusters and a priori categories.

3.2 Incrementality study

Model evaluation were executed on 2054 of 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 our new winner search method.

In the main case (depicted with the black line), 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 1 presents comparison of a standard global winner search method with our own CF-tree based local approach. Standard local search method (used in SOM) is considered since it is completely inappropriate in case of unconnected

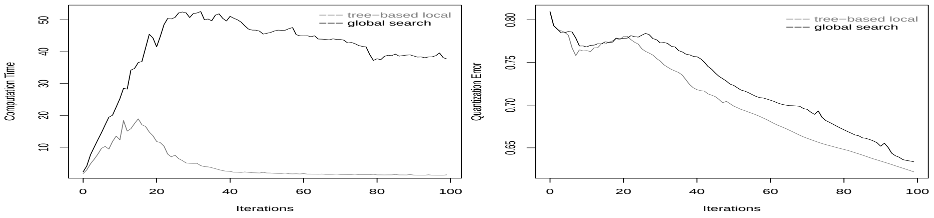

Fig. 1. Winner search methods (a) computation time (b) model quality

graphs. Obviously, 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.

The results seemed to be surprising at first glance. Initially, the quality was similar, later on - 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. We'll explain it later on the example of another experiment.

In the next experiment on topic drift, in addition to the main reference case, we had another two cases. During the first 30 iterations network has been trained on 700 documents only. In one of the cases (light grey line, massive document addition) documents were sampled uniformly from all five groups and in the 33^{rd} iteration another 700 uniformly sampled were introduced to training. After the 66th iteration the model has been trained on the whole dataset.

In the last case (dark grey line, incremental document insertion with topic drift) 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 an a priori information on the categories of documents. In the general case, we are collecting fuzzy category membership information from Bayesian Net model [5].

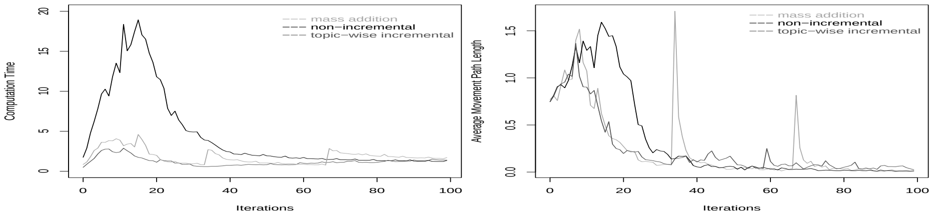

Fig. 2. Computation complexity (a) execution time of a single iteration (b) average path length of a document

As expected, in all cases GNG model adapts quite well to the topic drift. In the global and the topic-wise incremental case, the quality of the models were comparable, in terms of Average Document Quantization measure (see figure 3(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 visualizations of both models and the identification of topical 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 their own, specialized areas in the model. However, it can be noticed on the same plot that a complex mixture of topics can pose a serious drawback, especially in the first training iterations. In the global reference case, the attempt to cover all topics at once leads learning process to a local minimum and to subsequent divergence (what, in fact, is quite time-consuming as one can notice on figure 2(a)).

As we have previously noticed, the above-mentioned difficulties apply also to the case of global winner search (figure 1(b)). The quality of the final models when we take advantage of the incremental approach is almost the same 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).

The figure 2(b) presents 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 case of a global model caused much slower stabilization of the model and extremely high complexity of computations (figure 2(a)). The last reason for such slow computations is the representation of the GNG model nodes. The referential vector in such 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 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 topical 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.

The last figure, 3(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 broad topics method [5].

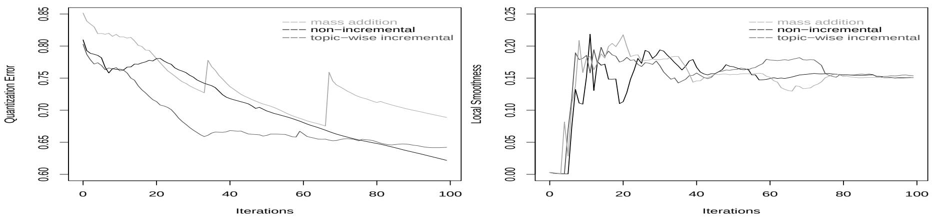

Fig. 3. Model quality (a) Average Document Quantization (b) Average Map Quantization

3.3 Scalability issues

To evaluate scalability of the proposed contextual approach (both in terms of space and time complexity), we built a model for a collection of more than one million documents crawled by our topic-sensitive crawler, starting from several Internet news sites (cnn, reuters, bbc). Resulting model consisted of 412 contextual maps, which means that the average density of a single map was about 2500 documents. Experimental results in this section are presented in series of box-and-whisker plots, which allows to present a distribution of a given evaluation measure (e.g. time, model smoothness or quantization error) over all 412 models, measured after each iteration of the learning process (horizontal axis). Horizontal line represents median value, area inside the box represents 25% - 75% quantiles, whiskers represent extreme values and each dot represents outlier values.

The whole cycle of map creation process took 2 days. It is impressing result, taking into account that Kohonen and his co-workers reported processing times in order of weeks [6]. It should also be noted that the model was built on a single personal computer (Pentium IV HT 3.2 GHz, 1 GB RAM). As it has been stated before, contextual model construction can be easily distributed and parallelized, what would lead to even shorter execution times.

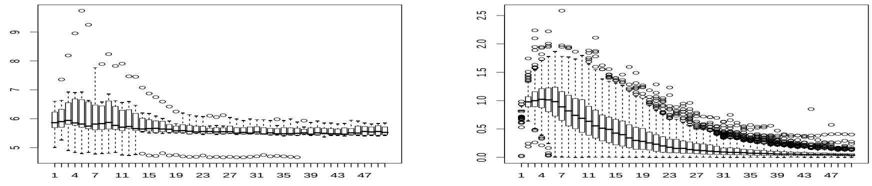

Fig. 4. Contextual model computation complexity (a) execution time of a single iteration (b) average path length of a document

The first observation is the complexity of a single iteration of GNG model learning (Figure 4(a)), which is almost constant, regardless of the increasing size of the model graph. It confirms the observations from section 3, concerning efficiency of the tree-based winner search methods. One can also observe the positive impact of homogeneity of the distribution of term frequencies in documents grouped to a single map cell. Such homogeneity is - to some extent -

acquired by initial split of a document collection into contexts. Another cause of the processing time reduction is the contextual reduction of vector representation dimensionality, described in the section 2.

In the Figure 4(b), the dynamic of the learning process is presented. The average path length of a document is the number of shifts over graph edges when documents is moved to a new, optimal location. It can be seen that model stabilizes quite fast; actually, most models converged to final state in less than 30 iterations. The fast convergence is mainly due to topical initialization. It should also be noted here that the proper topical initialization can be obtained for well-defined topics, which is the case in contextual maps.

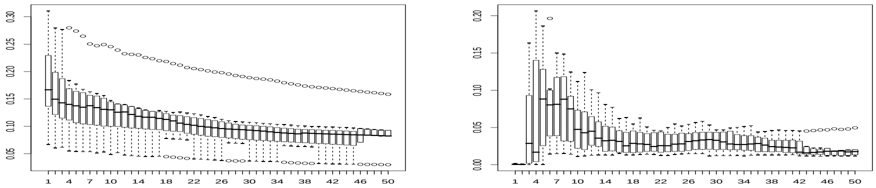

Fig. 5. Contextual model quality (a) Average Document Quantization (b) Average Map Quantization

The Figure 5 presents the quality of the contextual models. The final values of average document quantization (Figure 5(a)) and the map quantization (Figure 5(b)) are low, which means that the resulting maps are both "smooth" in terms of local similarity of adjacent cells and precisely represent documents grouped in a single node. Moreover, such low values have been obtained for moderate size of GNG models (majority of the models consisted of only 20-25 nodes - due to their fast convergence - and represented about 2500 documents each).

4 Concluding remarks

As indicated e.g. in [4], 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. The problem is deeper and has a "second bottom": the clustering methods like those of WebSOM contain elements of randomness so that even re-clustering of the same collection may lead to a radical change of the view of the documents.

From the point of view of incremental learning under SOM, a crucial factor for the processing time is the global winner search for assignment of documents to neurons. We were capable to elaborate a very effective method of mixing global with local winner search which does not deteriorate the overall quality of the final map and at the same time comes close to the speed of local search.

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

We presented the contextual approach, which proved to be an effective solution to the problem of massive data clustering. It is mainly due to: (1) replacement of a flat, global, graph-based meta-clustering structure with a hierarchy of topic-sensitive models and (2) introduction of contextual term weighting instead of standard $tfidf$ weights so that document clusters can be represented in different subspaces of a global vector space. With these improvements, we proposed a scalable approach to mining and retrieval of text data.

Contextual approach leads to many interesting research issues, such as context-dependent dictionary reduction and keywords identification, topic-sensitive document summarization, subjective model visualization based on particular user's information requirements, dynamic adaptation of the document representation and local similarity measure computation. Especially, the user-oriented, contextual data visualization can be a major step on the way to information retrieval personalization in search engines.

Clustering high dimensional data is both of practical importance and at the same time a big challenge, in particular for large collections of text documents. Still, it has to be stressed that not only textual, but also any other high dimensional data (especially characterized by attributes of heterogeneous and correlated distributions) may be clustered using the presented method.

References

1. C. Boulis, M. Ostendorf, Combining multiple clustering systems, Proceedings of 8th European conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2004), LNAI 3202, Springer-Verlag, 2004
2. B. Fritzke, A self-organizing network that can follow non-stationary distributions, in: Proceeding of the International Conference on Artificial Neural Networks '97, Springer, 1997, pp.613-618
3. Halkidi,M., Batistakis,Y., Vazirgiannis,M.: On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), pp.107-145, 2001
4. C. Hung, S. Wermter, A constructive and hierarchical self-organising model in a non-stationary environment, International Joint Conf. on Neural Networks, 2005
5. M. Klopotek, S. Wierzchon, K. Ciesielski, M. Draminski, D. Czerski, Conceptual maps and intelligent navigation in document space (in Polish), to appear in: Akademicka Oficyna Wydawnicza EXIT Publishing, Warszawa, 2006
6. T. Kohonen, S. Kaski, P. Somervuo, K. Lagus, M. Oja, V. Paatero, Self-organization of very large document collections, Helsinki University of Technology technical report, 2003, `http://www.cis.hut.fi/research/reports/biennial02-03`
7. T. Zhang, R. Ramakrishan, M. Livny, BIRCH: Efficient data clustering method for large databases, in: Proceedings of ACM SIGMOD International Conference on Data Management, 1997
8. Y. Zhao, G. Karypis, Criterion functions for document clustering: Experiments and analysis, at `http://www-users.cs.umn.edu/~karypis/publications/ir.html`