Pattern structures for news clustering

Tatyana Makhalova, Dmitry Ilvovsky, Boris Galitsky

School of Applied Mathematics and Information Science, National Research
University Higher School of Economics, Moscow, Russia
Knowledge Trail Incorporated
t.makhalova@gmail.com, dilvovsky@hse.ru, bgalitsky@hotmail.com

Abstract. Usually web search results are represented as long list of document snippets. It is difficult for users to navigate through this collection of text. We propose clustering method that uses pattern structure constructed on augmented syntactic parse trees. In addition, we compare our method to other clustering methods and demonstrate the limitations of the competitive methods.

1 Introduction and related works

Document clustering problem has been widely investigated in many applications of text mining. One of the most important aspects of a text clustering problem is a structured representation of text. The common approach to text representation is the Vector Space Model [1], where the collection or corpus of documents is represented as a term-document matrix. The main drawback of this model is its inability to reflect the importance of words with respect to a document and a corpus. To tackle this issue the weighted scheme based on tf-idf score has been proposed.

However, a term-document matrix built on a large texts collection may be sparse and have high dimensionality. To reduce the feature space one may use PCA, truncated SVD (Latent Semantic Analysis), random projection and other methods. To handle synonyms as similar terms a Generalized Vector Space Model [2, 3], a Topic-based Vector Model [4] and Enhanced Topic-based Vector Space Model [5] were introduced. The most common ways to clustering of a term-document matrix are Hierarchical clustering, k-Means and also Bisecting k-Means.

Graph models are also used for text representation. Document Index Graph (DIG) was proposed by Hammouda [6]. Zamir and Etzioni [7] use suffix tree for representing web snippets, where words are used instead of characters. The more sophisticated model based on n-grams was introduced in [8].

In this paper, we consider a particular application of document clustering: representation of web search results that could make it easier for users to find the information they are looking for [9]. Clustering snippets on salient phrases (i.e. key phrases that characterize a cluster) are described in [10, 11]. But the most promising approach for document clustering is conceptual clustering, because it allows to obtain overlapping clusters and to organize them into a hierarchical structure as well [12–17]. We present an approach to select the most significant clusters based on pattern structures [18]. This approach was introduced in [19]. The main idea is to construct a hierarchical structure of clusters using a reduced representation of syntactic trees with discourse relations between them. Leveraging discourse information allows to combine news articles not only by keyword similarity but by broader topicality and writing styles as well.

2 Clustering based on pattern structure

Parse Thickets Parse thicket [19] is defined as a set of parse trees for each sentence augmented with a number of arcs, reflecting inter-sentence relations. In this work we use parse thickets based on a limited set of relations: coreferences [20], Rhetoric structure relations [21] and Communicative Actions [22]. More information could be found in [19].

FCA A formal context is a triple (G, M, I), where G and M be sets, called the set of objects and attributes, respectively. Let I be a relation $I \subseteq G \times M$ between objects and attributes, i.e. $(g, m) \in I$ if the object g has the attribute m. The derivation operator $(\cdot)^{'}$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A^{'} = \{m \in M | \forall g \in A : gIm\}$$

$$B^{'} = \{g \in G | \forall m \in B : gIm\}$$

$A^{'}$ is the set of attributes common to all objects of A and $B^{'}$ is the set of objects sharing all attributes of B. The double application of $(\cdot)^{'}$ is a closure operator, i.e., $(\cdot)^{''}$ is extensive, idempotent and monotone. Sets $(A)^{''}$ and $(B)^{''}$ are said to be closed. A formal concept is a pair (A, B), where $A \subseteq G$, $B \subseteq M$ and $A^{'} = B$, $B^{'} = A$. A and B are called the formal extent and the formal intent, respectively.

Pattern Structure and Projections Pattern Structures are generalization of formal contexts, where objects are described by more complex structures, rather than a binary data. A pattern structure [18] is defined as a triple $(G, (D, \sqcap), \delta)$, where G is a set of objects, $(D, \sqcap)$ is a complete meet-semilattice of descriptions and $\delta : G \to D$ is a mapping an object to a description. The Galois connections between set of objects and their descriptions are defined as follows:

$$A^{\Box} := \sqcap_{g \in A} \delta(g) \text{ for } A \subseteq G$$

$$d^{\Box} := \{g \in G | d \sqsubseteq \delta(g)\} \text{ for } d \in D$$

A pair (A, d) for which $A^{\Box} = d$ and $d^{\Box} = A$ is called a pattern concept.

A projection ψ is a kernel operator, i.e. it is monotone ($x \sqsubseteq y \Rightarrow \psi(x) \sqsubseteq \psi(y)$), contractive ($\psi(x) \sqsubseteq x$), and idempotent ($\psi(\psi(x)) = \psi(x)$). The mapping $\psi : D \to D$ is used to replace $(G, (D, \sqcap), \delta)$ by $(G, (D_\psi, \sqcap_\psi), \psi \circ \delta)$, where $D_\psi = \{d \in D | \exists d' \in D : \psi(d') = d\}$.

In our case, *an original paragraph of text* and *parse thickets constructed from this paragraph* correspond to *an object* and *a description of pattern concepts* respectively. To improve efficiency and decrease time complexity we use projection instead of a parse thicket itself. Projection on a parse thicket is defined as a set of its maximal sub-trees and the intersection operator takes the form of pairwise intersection of elements within noun and verb phrase groups.

3 Reduced pattern structures

A pattern structure constructed from the collection of short texts usually has a huge number of concepts. To reduce the computational costs and improve the interpretability of pattern concepts we introduce several metrics that are described below.

Average and Maximal Pattern Score The average and maximal pattern score indices are meant to assess how meaningful is the common description of texts in the concept. The higher the difference of text fragments from each other, the lower their shared content is. Thus, meaningfulness criterion of a pattern concept $\langle A, d\rangle$ is

$$Score^{max}\left\langle A, d\right\rangle := \max_{chunk \in d} Score\left(chunk\right)$$

$$Score^{avg}\left\langle A, d\right\rangle := \frac{1}{|d|}\sum_{chunk \in d} Score\left(chunk\right)$$

The score function $Score\left(chunk\right)$ estimates description d using its weights for different parts of speech.

Average and Minimal Pattern Loss Score This scores estimate how much information contained in the description of a text is lost with respect to the original text. The average pattern loss score calculates the average loss of a cluster content with respect to texts in this cluster, while minimal pattern score loss represents a minimal loss of content among all texts included in a concept.

$$ScoreLoss^{min}\left\langle A, d\right\rangle := 1 - \frac{Score^{max}\left\langle A, d\right\rangle}{\min_{g \in A} Score^{max}\left\langle g, d_g\right\rangle}$$

$$ScoreLoss^{avg}\left\langle A, d\right\rangle := 1 - \frac{Score^{avg}\left\langle A, d\right\rangle}{\frac{1}{|d|}\sum_{g \in A} Score^{max}\left\langle g, d_g\right\rangle}$$

We use a reduced pattern structure. We propose to create exactly meaningful pattern concepts. For arbitrary sets of texts A_1 and A_2, corresponding descriptions d_1, d_2 and candidate for a pattern concept $\langle A_1 \cup A_2\ , d_1 \cap d_2\rangle$ need to satisfy the following constrains

$$ScoreLoss^{*}\left\langle A_1 \cup A_2\ , d_1 \cap d_2\right\rangle \leq \theta$$

$$Score^{*}\left\langle A_1 \cup A_2\ , d_1 \cap d_2\right\rangle \geq \mu_1 \min\left\{Score^{*}\left\langle A_1\ , d_1\right\rangle, Score^{*}\left\langle A_2\ , d_2\right\rangle\right\}$$

$$Score^* \langle A_1 \cup A_2\ , d_1 \cap d_2 \rangle \leq \mu_2 \max \{Score^* \langle A_1\ , d_1 \rangle, Score^* \langle A_2\ , d_2 \rangle\}$$

The first constraint provides condition for the construction of concepts with meaningful content, while two other constrains ensure that we do not use concepts with similar content.

4 Experiments

In this section we consider two examples for the proposed clustering method. The first one corresponds to the case when clusters are overlapping and distinguishable, the second one is the case of non-overlapping clusters.

4.1 User Study

In the most cases it is quite difficult to identify disjoint classes for a text collection [23]. To confirm this, we conducted experiments similar to the experiment scheme described in [11]. We took web snippets obtained by querying the Bing search engine API and asked a group of four experts to label ground truth for them. We performed news queries related to world's most pressing news (for example, "fighting Ebola with nanoparticles", "turning brown eyes blue", "F1 winners", "read facial expressions through webcam", "2015 ACM awards winners") to make labeling of data easier for the experts.

According to the experts, it was difficult to determine partitions, while overlapping clusters naturally stood out. As a result, in the case of non-overlapping clusters we usually got a small number of large classes or a sufficiently large number of classes consisting of 1-2 snippets. More than that, for the same set of snippets we obtained quite different partitions.

We used the Adjusted Mutual Information score to estimate pairwise agreement of non-overlapping clusters, which were identified by the experts. This metric allows one to estimate agreement of two clustering results with correction for randomness partition.

$$MI_{adj} = \frac{MI\left(U,V\right) - E\left[MI\left(U,V\right)\right]}{max\left(H(U),H(V)\right) - E\left[MI\left(U,V\right)\right]}$$

where U and V are partitions of the news set, $MI(U,V)$ - the mutual information between them and $E\left[MI\left(U,V\right)\right]$ is the expected mutual information between two random clusterings.

To study the behavior of the conventional clustering approach we consider 12 short texts on news query "The Ebola epidemic". Tests are available by link [1].

Experts identify quite different non-overlapping clusters. The pairwise Adjusted Mutual Information score was in the range of 0,03 to 0,51. Next, we

[1] `https://drive.google.com/file/d/0B7I9HM34b_62TEFtUTRqdzdqWjA/view?usp=sharing`

compared partitions to clustering results of the following clustering methods: k-means clustering based on vectors obtained by truncated SVD (retaining at least 80% of the information), hierarchical agglomerative clustering (HAC), complete and average linkage of the term-document matrix with Manhattan distance and cosine similarity, hierarchical agglomerative clustering (both linkage) of tf-idf matrix with Euclidean metric. In other words, we turned an unsupervised learning problem into the supervised one. The accuracy score for different clustering methods is represented in Figure 1. Curves correspond to the different partitions that have been identified by people.

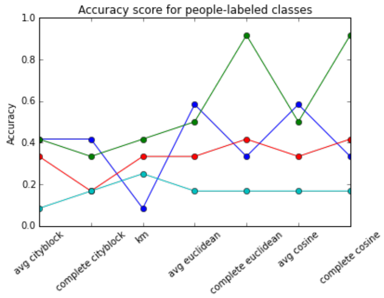


Fig. 1: Classification accuracy of clustering results and "true" clustering (example 1). Four lines are different news labeling made by people. The y-axis values for fixed x-value correspond to classification accuracy of a clustering method for each of the four labeling

As it was mentioned earlier, we obtain inconsistent "true" labeling. Thereby the accuracy of clustering differs from labeling made by evaluators. This approach doesn't allow to determine the best partition, because a partition itself is not natural for the given news set. For example, consider clusters obtained by HAC based on cosine similarity (trade-off between high accuracy and its low variation): 1-st cluster: 1,2,7,9; 2-nd cluster: 3,11,12; 3-rd cluster: 4,8; 4-th cluster: 5,6; 5-th cluster: 10.

Almost the same news 4, 8, 12 and 9, 10 are in the different clusters. News 10, 11 should be simultaneously in several clusters (1-st, 5-th and 2-nd,3-rd respectively).

4.2 Examples of pattern structures clustering

To construct hierarchy of overlapping clusters by the proposed methods, we use the following constraints: $\theta = 0,75$, $\mu_1 = 0,1$ and $\mu_2 = 0,9$. The value of θ limits the depth of the pattern structure (the maximal number of texts in a cluster), put differently, the higher θ, the closer should be the general intent of clusters. μ_1 and μ_2 determine the degree of dissimilarity of the clusters on different levels of the lattice (the clusters are prepared by adding a new document to the current one).

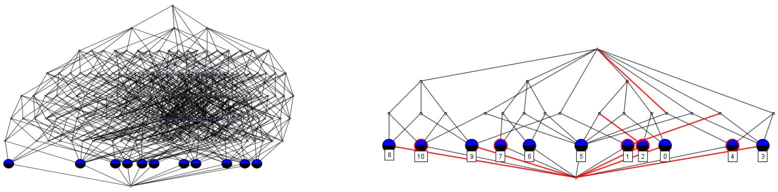
We consider the proposed clustering method on 2 examples. The first one was described above, it corresponds to the case of overlapping clusters, the second one is the case when clusters are non-overlapping and distinguishable. Texts of the second example are available by link [2]. Three clusters are naturally identified in this texts.

The cluster distribution depending on volume are shown in Table 1. We got 107 and 29 clusters for the first and the second example respectively.

Text number	1	2	3	4	5	6
Example 1	12	34	33	20	7	1
Example 2	11	15	3	0	0	0

Table 1: The clusters volume distribution for non-overlapping clusters (example 1) and overlapping clusters (example 2)

In fact, this method is an agglomerative hierarchical clustering with overlapping clusters. Hierarchical structure of clusters provides browsing of texts with similar content by layers. The cluster structure is represented on Figure 2. The top of the structure corresponds to meaningless clusters that consist of all texts. Upper layer consists of clusters with large volume.



(a) pattern structure without reduction

(b) reduced pattern structure

Fig. 2: The cluster structure (example 2). The node on the top corresponds to the “dummy” cluster, high level nodes correspond to the big clusters with quite general content, while the clusters at lower levels correspond to more specific news.

Clustering based on pattern structures provides well interpretable groups. The upper level of hierarchy (the most representative clusters for example 1) consists of the clusters presented in Table 2.

MaxScore	Cluster (extent)	MaxScore	Cluster (extent)	MaxScore	Cluster (extent)
7,8	{3, 11, 12}	3,8	{1, 2, 3, 7, 9}	3,2	{3, 9, 11}
4,1	{4, 8, 11}	3,3	{2, 4, 11}	2,8	{3, 10}
3,8	{1, 5, 11}	3,3	{2, 11}	2,4	{1, 2, 6, 9, 10}
3,8	{1, 11}	3,3	{5, 6}	2,3	{1, 5, 6}

Table 2: Scores of representative clusters

We also consider smaller clusters and select those for which adding of any object (text) dramatically reduces the $MaxScore$ $\{1, 2, 3, 7, 9\}$ and $\{5, 6\}$. For

[2] https://drive.google.com/file/d/0B7I9HM34b_62czFlZ29zZl9kblk/view?usp=sharing

other nested clusters significant decrease of $MaxScore$ occurred exactly with the an expansion of single clusters.

For the second example we obtained 3 clusters that corresponds to "true" labeling.

Our experiments show that pattern structure clustering allows to identify easily interpretable groups of texts and significantly improves text browsing.

5 Conclusion

In this paper, we presented an approach that addressed the problem of short text clustering. Our study shows a failure of the traditional clustering methods, such as k-means and HAC. We propose to use parse thickets that retain the structure of sentences instead of the term-document matrix and to build the reduced pattern structures to obtain overlapping groups of texts. Experimental results demonstrate considerable improvement of browsing and navigation through a texts set for users. Introduced indices *Score* and *ScoreLoss* both improve computing efficiency and tackle the problem of redundant clusters.

An important direction for future work is to take into account synonymy and to compare the proposed method to similar approach that use key words instead of parse thickets.

Acknowledgments

The project is being developed by the "Methods of web corpus collection, analysis and visualisation" research and study group under guidance of prof. B.Mirkin (grant 15 - 05 - 0041 of Academic Fund Program).

References

1. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
2. Wong, S.M., Ziarko, W., Wong, P.C.: Generalized vector spaces model in information retrieval. In: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1985) 18–25
3. Tsatsaronis, G., Panagiotopoulou, V.: A generalized vector space model for text retrieval based on semantic relatedness. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics (2009) 70–78
4. Becker, J., Kuropka, D.: Topic-based vector space model. In: Proceedings of the 6th International Conference on Business Information Systems. (2003) 7–12
5. Polyvyanyy, A., Kuropka, D.: A quantitative evaluation of the enhanced topic-based vector space model. (2007)
6. Hammouda, K.M., Kamel, M.S.: Document similarity using a phrase indexing graph model. Knowledge and Information Systems **6** (2004) 710–727

7. Zamir, O., Etzioni, O.: Web document clustering: A feasibility demonstration. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 46–54
8. Schenker, A., Bunke, H., Last, M., Kandel, A.: Clustering of web documents using graph representations. In: Applied Graph Theory in Computer Vision and Pattern Recognition. Springer (2007) 247–265
9. Galitsky, B.: Natural language question answering system: Technique of semantic headers. Advanced Knowledge International (2003)
10. Zamir, O., Etzioni, O.: Grouper: a dynamic clustering interface to web search results. Computer Networks **31** (1999) 1361–1374
11. Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: Learning to cluster web search results. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2004) 210–217
12. Galitsky, B., Ilvovsky, D., Kuznetsov, S., Strok, F.: Finding maximal common subparse thickets for multi-sentence search. In Croitoru, M., Rudolph, S., Woltran, S., Gonzales, C., eds.: Graph Structures for Knowledge Representation and Reasoning. Volume 8323 of Lecture Notes in Computer Science. Springer International Publishing (2014) 39–57
13. Cole, R., Eklund, P., Stumme, G.: Document retrieval for e-mail search and discovery using formal concept analysis. Applied artificial intelligence **17** (2003) 257–280
14. Koester, B.: Conceptual knowledge retrieval with fooca: Improving web search engine results with contexts and concept hierarchies. In: Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. Springer (2006) 176–190
15. Messai, N., Devignes, M.D., Napoli, A., Smail-Tabbone, M.: Many-valued concept lattices for conceptual clustering and information retrieval. In: ECAI. Volume 178. (2008) 127–131
16. Carpineto, C., Romano, G.: A lattice conceptual clustering system and its application to browsing retrieval. Machine Learning **24** (1996) 95–122
17. Strok, F., Galitsky, B., Ilvovsky, D., Kuznetsov, S.: Pattern structure projections for learning discourse structures. In Agre, G., Hitzler, P., Krisnadhi, A., Kuznetsov, S., eds.: Artificial Intelligence: Methodology, Systems, and Applications. Volume 8722 of Lecture Notes in Computer Science. Springer International Publishing (2014) 254–260
18. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: Conceptual Structures: Broadening the Base. Springer (2001) 129–142
19. Galitsky, B., Ilvovsky, D., Kuznetsov, S., Strok, F.: Matching sets of parse trees for answering multi-sentence questions. Proc. Recent Advances in Natural Language Processing (RANLP 2013), Bulgaria (2013)
20. Lee, H., Recasens, M., Chang, A., Surdeanu, M., Jurafsky, D.: Joint entity and event coreference resolution across documents. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics (2012) 489–500
21. Mann, W.C., Thompson, S.A.: Discourse description: Diverse linguistic analyses of a fund-raising text. Volume 16. John Benjamins Publishing (1992)
22. Searle, J.R.: Speech acts : an essay in the philosophy of language. Cambridge University Press (1969)
23. Galitsky, B., de la Rosa, J.L.: Concept-based learning of human behavior for customer relationship management. Information Sciences **181** (2011) 2016–2035