News clustering approach based on discourse text structure

Tatyana Makhalova
National Research University
Higher School of Economics
Moscow, Russia
t.makhalova@gmail.com

Dmitry Ilvovsky
National Research University
Higher School of Economics
Moscow, Russia
dilvovsky@hse.ru

Boris Galitsky
Knowledge Trail Incorporated
San Jose, USA
bgalitsky@hotmail.com

Abstract

A web search engine usually returns a long list of documents and it may be difficult for users to navigate through this collection and find the most relevant ones. We present an approach to post-retrieval snippet clustering based on pattern structures construction on augmented syntactic parse trees. Since an algorithm may be too slow for a typical collection of snippets, we propose a reduction method that allows us to construct a reduced pattern structure and make it scalable. Our algorithm takes into account discourse information to make clustering results independent of how information is distributed between sentences.

1 Introduction and related works

The document clustering problem was widely investigated in many applications of text mining. One of the most important aspects of the text clustering problem is a structural representation of texts. A common approach to the text representation is a vector space model (Salton et al., 1975), 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 a word with respect to a document and a corpus. To tackle this issue the weighted scheme based on tf-idf score has been proposed. Also, a term-document matrix built on a large texts collection may be sparse and have a high dimensionality. To reduce feature space, PCA, truncated SVD (Latent Semantic Analysis), random projection and other methods have been proposed. To handle synonyms as similar terms the general Vector Space Model (Wong et al., 1985; Tsatsaronis and Panagiotopoulou, 2009), topic-based vector model (Becker and Kuropka, 2003) and enhanced topic-based vector space model (Polyvyanyy and Kuropka, 2007) were introduced. The most common ways to clustering 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 (2004). Zamir and Etzioni (1998) use suffix tree for representing web snippets, where words are used instead of characters. A more sophisticated model based on n-grams was introduced in Schenker et al. (2007).

In this paper, we consider a particular application of document clustering, it is a representation of web search results that could improve navigation through relevant documents. Clustering snippets on salient phrases is described in (Zamir and Etzioni, 1999; Zeng et al., 2004). But the most promising approach for document clustering is a conceptual clustering, because it allows to obtain overlapping clusters and to organize them into a hierarchical structure as well (Cole et al., 2003; Koester, 2006; Messai et al., 2008; Carpineto and Romano, 1996). We present an approach to selecting most significant clusters based on a pattern structure (Ganter and Kuznetsov, 2001). An approach of extended representation of syntactic trees with discourse relations between them was introduced in (Galitsky et al., 2013). Leveraging discourse information allows to combine news articles not only by keyword similarity but by broader topicality and writing styles as well.

The paper is organized as follows. Section 2 introduces a parse thicket and its simplified representation. In section 3 we consider approach to clustering web snippets and discuss efficiency issues. The illustrative example is presented in section 4. Finally, we conclude the paper and discuss some research perspectives.

Proceedings of the First Workshop on Computing News Storylines, pages 16–20,
Beijing, China, July 31, 2015.

2 Clustering based on pattern structure

Parse Thickets Parse thicket (Galitsky et al., 2013) is defined as a set of parse trees for each sentence augmented with a number of arcs, reflecting inter-sentence relations. In present work we use parse thickets based on limited set of relations described in (Galitsky et al., 2013): coreferences (Lee et al., 2012), Rhetoric structure relations (Mann and Thompson, 1992) and Communicative Actions (Searle, 1969).

Pattern Structure with Parse Thickets simplification To apply parse thickets to text clustering tasks we use pattern structures (Ganter and Kuznetsov, 2001) that 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 $\delta : G \rightarrow D$ is a mapping an object to a description. The Galois connection between set of objects and their descriptions is also defined as follows:

$$A^{\diamond} := g \in A \prod \delta(g)$$

$$d^{\diamond} := \{g \in G | d \sqsubseteq \delta(g)\}$$

for $A \subseteq G$, for $d \in D$

A pair $\langle A, d \rangle$ for which $A^{\diamond} = d$ and $d^{\diamond} = A$ is called a pattern concept. In our case, A is the set of news, d is their shared content.

We use AddIntent algorithm (van der Merwe et al., 2004) to construct pattern structure. On each step, it takes the parse thicket (or chunks) of a web snippet of the input and plugs it into the pattern structure.

A pattern structure has several drawbacks. Firstly, the size of the structure could grow exponentially on the input data. More than that, construction of a pattern structure could be computationally intensive. To address the performance issues, we reduce the set of all intersections between the members of our training set (maximal common sub-parse thickets).

3 Reduced pattern structure

Pattern structure constructed from a 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 the common description of texts in the concept is. The higher the difference of text fragments from each other, the lower their shared content is. Thus, meaningfulness criterion of the group of texts is

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

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

The score function $Score(chunk)$ estimates chunks on the basis of parts of speech composition.

Average and Minimal Pattern Score loss Average and minimal pattern score loss describe how much information contained in text is lost in the description with respect to the source texts. Average pattern score loss expresses the average loss of shared content for all texts in a concept, while minimal pattern score loss represents a minimal loss of content among all texts included in a concept.

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

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

We propose to use a reduced pattern structure. There are two options in our approach. The first one - construction of lower semilattice. This is similar to iceberg concept lattice approach (Stumme et al., 2002). The second option - construction of concepts which are different from each other. Thus, for arbitrary sets of texts A_1 and A_2, corresponding descriptions d_1 and d_2 and candidate for a pattern concept $\langle A_1 \cup A_2, d_1 \cap d_2 \rangle$ criterion has the following form

$$Score^{max}\langle A_1 \cup A_2, d_1 \cap d_2 \rangle \geq \theta$$
$$Score^{*}\langle A_1 \cup A_2, d_1 \cap d_2 \rangle \geq$$
$$\mu_1 min\{Score^{*}\langle A_1, d_1 \rangle, Score^{*}\langle A_2, d_2 \rangle\}$$
$$Score^{*}\langle A_1 \cup A_2, d_1 \cap d_2 \rangle \leq$$
$$\mu_2 max\{Score^{*}\langle A_1, d_1 \rangle, Score^{*}\langle A_2, d_2 \rangle\}$$

The first constraint provides the 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 the proposed clustering method on 2 examples. 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 some cases it is quite difficult to identify disjoint classes for a text collection. To confirm this, we conducted experiments similar to the experiment scheme described in (Zeng et al., 2004). We took web snippets obtained by querying the Bing search engine API and asked a group of four assessors 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 assessors.

In most cases, according to the assessors, 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 people.

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

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

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).

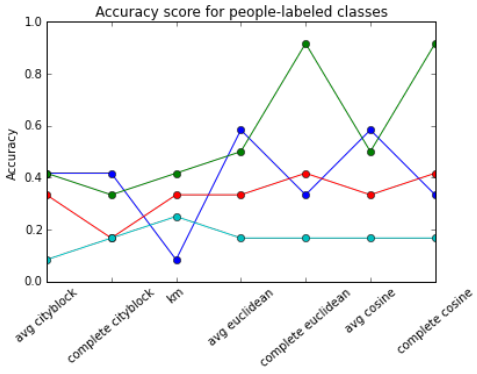


Figure 1: Classification accuracy of clustering results and "true" clustering (example 1). Four lines are different news labeling made by people. The y-axis values for fixed x-value correspond to classification accuracy of a clustering method for each of the four labeling

4.2 Examples of pattern structures clustering

To construct hierarchy of overlapping clusters by the proposed methods, we use the following constraints: $\theta = 0,25$, $\mu_1 = 0,1$ 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 sec-

[1] https://github.com/anonymously1/CNS2015/blob/master/NewsSet1

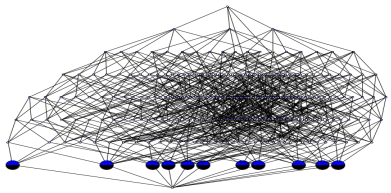
ond example are available by link [2]. Three clusters are naturally identified in this texts.

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.

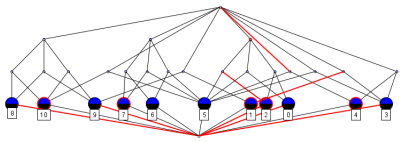
Text number	Clusters number	
	Example 1	Example 2
1	12	11
2	34	15
3	33	3
4	20	0
5	7	0
6	1	0

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

Figure 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.

[2] https://github.com/anonymously1/CNS2015/blob/master/NewsSet

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

References

Jörg Becker and Dominik Kuropka. 2003. Topic-based vector space model. In *Proceedings of the 6th International Conference on Business Information Systems*, pages 7–12.

Claudio Carpineto and Giovanni Romano. 1996. A lattice conceptual clustering system and its application to browsing retrieval. *Machine learning*, 24(2):95–122.

Richard Cole, Peter Eklund, and Gerd Stumme. 2003. Document retrieval for e-mail search and discovery using formal concept analysis. *Applied artificial intelligence*, 17(3):257–280.

Boris Galitsky, Dmitry Ilvovsky, Sergey Kuznetsov, and Fedor Strok. 2013. Matching sets of parse trees for answering multi-sentence questions. *Proc. Recent Advances in Natural Language Processing (RANLP 2013), Bulgaria*.

Bernhard Ganter and Sergei O Kuznetsov. 2001. Pattern structures and their projections. In *Conceptual Structures: Broadening the Base*, pages 129–142. Springer.

Khaled M Hammouda and Mohamed S Kamel. 2004. Document similarity using a phrase indexing graph model. *Knowledge and Information Systems*, 6(6):710–727.

Bjoern Koester. 2006. 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*, pages 176–190. Springer.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. 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*, pages 489–500. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1992. *Discourse description: Diverse linguistic analyses of a fund-raising text*, volume 16. John Benjamins Publishing.

Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smail-Tabbone. 2008. Many-valued concept lattices for conceptual clustering and information retrieval. In *ECAI*, volume 178, pages 127–131.

Artem Polyvyanyy and Dominik Kuropka. 2007. A quantitative evaluation of the enhanced topic-based vector space model.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Adam Schenker, Horst Bunke, Mark Last, and Abraham Kandel. 2007. Clustering of web documents using graph representations. In *Applied Graph Theory in Computer Vision and Pattern Recognition*, pages 247–265. Springer.

John Rogers Searle. 1969. *Speech acts : an essay in the philosophy of language*. Cambridge University Press.

Gerd Stumme, Rafik Taouil, Yves Bastide, Nicolas Pasquier, and Lotfi Lakhal. 2002. Computing iceberg concept lattices with titanic. *Data & knowledge engineering*, 42(2):189–222.

George Tsatsaronis and Vicky Panagiotopoulou. 2009. 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*, pages 70–78. Association for Computational Linguistics.

Dean van der Merwe, Sergei Obiedkov, and Derrick Kourie. 2004. Addintent: A new incremental algorithm for constructing concept lattices. In Peter Eklund, editor, *Concept Lattices*, volume 2961 of *Lecture Notes in Computer Science*, pages 372–385. Springer Berlin Heidelberg.

SK Michael Wong, Wojciech Ziarko, and Patrick CN Wong. 1985. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25. ACM.

Oren Zamir and Oren Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54. ACM.

Oren Zamir and Oren Etzioni. 1999. Grouper: a dynamic clustering interface to web search results. *Computer Networks*, 31(11):1361–1374.

Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. 2004. Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM.