# Full-Text Clustering Methods for Current Research Directions Detection*

© Dmitry Devyatkin © Ilya Tikhomirov © Alexander Shvets ©Oleg Grigoriev
Institute for systems analysis of RAS,
Moscow
devyatkin@isa.ru, shvets@isa.ru, tih@isa.ru, oleggpolikvart@yandex.ru

© Konstantin Popov
Engelhardt Institute of Molecular Biology of RAS,
Moscow
konstantin.v.popov@gmail.com

## Abstract

The paper contains a brief overview of full-text clustering methods for current research directions detection. A novel full-text clustering method is proposed. A dataset is created and experimental results are verified by problem domain "Regenerative medicine" experts with PhD degrees. The proposed method is well applicable for research directions detection according to experimental results. Finally, prospects and drawbacks of the proposed method are discussed.

*The research is supported by Russian Foundation for Basic Research, project №14-29-05008-ofi_m

## 1 Introduction

Commonly methods for research directions detection are based on different clustering methods [4]. The problem is that all these methods require tuning to a research area and dataset. Common clustering evaluation metrics are not applicable for directions detection. They are often based on degree of insularity of clusters [1, 12]. But criteria for cluster building are weakly formalized in directions detection task. Several approaches use empirical estimates, determined by experts, for solving this problem [6]. However, obtaining of these estimates is complicated and nondeterministic process, which requires interaction between data-mining experts and experts of analyzed research area. Moreover, criteria for inclusion of some scientific paper to a research direction depends on research area of this paper, thus an opinion of experts in this research area should be taken into consideration. Therefore, according to specialty of the problem a semi-automatically approach, that uses a small labeled part of analyzed dataset for training clustering methods, is proposed in this paper. Let $D=\{d_1,d_2,..,d_n\}$ is a corpus of scientific papers; where $n$ is a number of documents in this collection and $C=\{c_1,c_2,..,c_n\}$ is a set of previously known research directions in $D$. Suppose $D_c \subset D$ is a small incomplete set of papers that are in $C$. The goal is to get full set of research directions $C_f$ and allocation of papers to this set. The approach consists in applying a quality function that can estimate difference between clustering method results and given distribution $C$. Then methods of constrained optimization are used for tuning parameters of the clustering method. Unlike well-known classifier training task, the complete directions set $C_f$ is a priori unknown. So it is necessary to find a method, which can be tuned in appliance with this approach properly.

The primary goal of this paper is to present an improved method for research directions detection, which can process datasets semi-automatically. Besides, we present results of comparison between proposed method and the state-of-the-art clustering methods such as Birch [19], affinity propagation [5] and their combinations with topical latent Dirichlet allocation model [3].

## 2 Related work

Consider some methods for science directions detection. There are three groups of methods, used for research direction detection. All of them use clustering methods, but differ in the applied measure of similarity detection: the co-citation measure [4], the text measure or the hybrid measure. The last one usually is based on an assessment of co-citations of papers, on the intelligent analysis of texts and on allocation of significant papers [6].

In [9] definition of the prospective research directions is carried out using co-word analysis, i.e. papers similarity. This approach is close to generative topic distribution clustering methods. Application of this method for detection of regularities and trends in the area of information security is shown. The SCI base [13] is used as data source; keywords are collected from the list of the terms allocated by authors of papers. These keywords were normalized. To check dynamic

**Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015**

changes in the area of information security, authors offer the whole set of methods for marking of papers. They conclude that in the area of information security there is constant directions set, and at the same time new directions emerge regularly. The main disadvantage of this method is that it uses only structured authors' keywords. This approach leads to decreasing of methods objectivity; because authors' keywords often distinguish from the real terms of a paper.

In this work we will use full-text clustering approaches because of their universality: they do not need citation databases for proper results. Let us discuss several clustering methods.

For example, affinity propagation is one of common clustering methods. In this method a set of papers is considered as a network of connected nodes, and the similarity of papers corresponds to the weight of edges in this network. The method is based on mechanism of "passing messages" between nodes of the network [5]. Affinity propagation detects "exemplars" - papers in the input dataset that are representative for clusters. The network nodes send to each other messages until a set of exemplars and clusters gradually emerges. Potential drawback of this method is initially exemplars selection. One cannot provide initially exemplars distribution for all directions, so the exemplars can be selected inaccurately. In addition, the approach can lead to inaccuracies in cases when it is impossible to identify an exemplar which describes all the papers in the cluster properly [10]. Unlike affinity propagation, proposed method uses terms descriptors to describe clusters. These terms correspond to all papers of a cluster (due to the using of the topic importance measure for the terms weighting) [14] that leads to better quality of clustering.
Another common clustering method is Birch [19]. This algorithm builds a weighted balanced tree (CF-tree) in a single pass through the data. The tree stores information about sub-clusters in leaves of the tree. During the clustering each paper is added to the existing leaf, or a new leaf is created. Birch also can be used in stream clustering when the number of documents to be processed can be theoretically infinite.

Specialized methods for text clustering are worth mentioning. In [14] well-scalable full-text clustering approach for detecting research directions was proposed. Hash functions are widely used in this method in order to get good performance. This method has insufficient coverage of a dataset: there are large amount of papers, which do not belong to any of the clusters. In this study we improve classification decision rule for better coverage. We also refine extracting of cores of clusters for better clustering quality.

The methods of topic modelling based on the generating models are often applied for full-text clustering. For example, Latent Dirichlet Allocation (LDA) [3] is applied for clustering of large amounts of papers. The method is an improvement of Latent Semantic Indexing. It assumes that each object is presented in several classes which distribution is included into parametrical family of Dirichlet distribution. A drawback of the method is instability to input data, that makes the results uninterpretable [8, 18]. We use LDA as a preliminary step for clustering by well-known methods. This applies to significantly reduce a feature space length and so to increase the common clustering methods performance and quality.

# 3 Method description

## 3.1 Fast algorithm for similar document search

The cornerstone of proposed full-text clustering method is similar document search, previously discussed in [15]. This algorithm use inverted index structures for input data. Most of the full-text search algorithms apply similar structures for improving search performance [17]. Let $w$ is a term (a word or a phrase) from a paper, $t$ – weight of the term. We use well-known $tf \cdot idf$ measure for weighting terms in papers. It is a product of term frequency function $tf$ and inverse document frequency function $idf$ .

$$tf(w,d) = \frac{n_i}{\sum_k n_k}$$

Where $n_i$ is number of times term $w$ occurs in the paper $d$ , $\sum_k n_k$ is a number of all terms occurrences in $d$ .

$$idf(w,D) = \log \frac{|D|}{|d_i \supseteq w|}$$

Where $|D|$ is the number of papers in corpus $D$, $|d_i \supseteq w|$ is the number of papers from $D$ that have term w and the logarithm base does not matter.

Define direct index of papers as a function that returns a set of different terms and their weights from the paper $d$ .

$$D_{idx}(d) = \{\langle w_1, t_1\rangle, \langle w_2, t_2\rangle, .., \langle w_n, t_n\rangle\}$$

Where $n$ is the number of different terms in the document.

Define inverted index of papers as a function that returns a set of different papers, in which given term $w$ is occurred and weights of this term in these papers: $I_{idx}(w) = \{\langle d_1, t_1\rangle, \langle d_2, t_2\rangle, .., \langle d_n, t_n\rangle\}$ . The algorithm of the fast similar document search is the following.

1. Get terms of the paper $d$ from its direct index $D_{idx}(d)$ .
2. Retrieve list of papers containing the terms got in step 1 from inverted index $I_{idx}(w)$ .
3. Filter papers that are weakly intersected by the terms of the input paper.
4. Get lists of terms for retrieved papers from $D_{idx}(d)$.
5. Calculate Manhattan [2] distance between the lists of terms got in step 4. We chose this distance because it is fast to calculate and

provides quality comparable to more complicated metrics like cosine distance [16].

6. Filter papers with high distance to input paper according to predefined threshold *H*.

In step 3 of this algorithm we cut weakly similar papers without calculating a distance measure that significantly improves the algorithm. Steps 1-4 can be performed in parallel, which improves the performance.

### 3.2 Clustering algorithm

Suppose a descriptor of a cluster is a set of terms, relevant to this cluster. Let an initial core of a cluster is a subset of highly similar papers, which is used for generation of a descriptor. A descriptor is built based on full-text of these papers. Let $Core(d)$ is a function that returns a tuple $\langle c, m \rangle$ for given paper $d$, where $c$ is an identifier for initial core of cluster and *m* is a numeric characteristic of paper *d* in initial core *c*. Characteristic *m* is used for decreasing of distance threshold *H* during building of initial core, that prevents merging of initial cores. If paper *d* is not presented in any initial core, the function $Core(d)$ returns empty tuple. Direct index of descriptors is a function that returns a set of different terms and their weights from the descriptor of cluster $CD_{idx}(c) = \{\langle w_1, t_1 \rangle, \langle w_2, t_2 \rangle, .., \langle w_n, t_n \rangle\}$ where *n* is count of different terms in the descriptor. Inverted index of descriptors is a function that returns a set of descriptors in which given term *w* is occurred and weights of this term in these descriptors: $CI_{idx}(w) = \{\langle c_1, t_1 \rangle, \langle c_2, t_2 \rangle, .., \langle c_n, t_n \rangle\}$. We use topic importance measure $\Delta I$ for weighting terms of descriptors. This measure for term *w* from corpus *D* and cluster *c* is calculated as follows:

$$\Delta(w,c,D) = idf(w,D) - idf(w,D_c)$$
$$\Delta I(w,c,D) = \Delta(w,c,D) \cdot X(\Delta(w,c,D)),$$

where $D_c$ is a set of papers from cluster c and $X()$ is a Heaviside step function. Due to the topic importance measure, descriptor of cluster consists mostly of terms specific for papers of the cluster.

Let $H_a$ and $\alpha$ are clustering thresholds, affecting on insularity of clusters.

The proposed full-text clustering method contains two parts. The first part could be performed independently on different nodes of computer network. This part conventionally can be called "initial cores of clusters detection". The steps are the following.

1. Index the input corpus $D$.
2. While $D$ is not empty, exclude a paper $d$ from $D$. If $D$ is empty, turn to step 6. If $Core(d)$ returns empty tuple, then turn to step 3, else turn to step 4.
3. Generate new cluster core index *c*. Add new tuple $\langle c,1 \rangle$ with $m$=1 to $Core$. Then turn to step 5.
4. Get identifier *c* for current initial core and *m* for current paper. Reduce $m = \alpha \cdot m$.
5. Search similar papers by the fast algorithm with threshold $H = H_a \cdot m$. Add it to $Core$ with current value of *m*. Then turn to step 2.
6. Build descriptors $CD_{idx}(c)$ and $CI_{idx}(w)$ for each initial core $c$.

The second part is performed in a selected "supervisor" node and consists of the following steps.

1. Search similar descriptors by the fast algorithm using $CD_{idx}$ and $CI_{idx}$. Then merge these descriptors. In practice this step can be useful, if the first section executes on several computer nodes.
2. Classify all papers. For each retrieved descriptor use steps 2-6 from the fast similar search algorithm and resulting $CD_{idx}$ as a list of target terms. To prevent fuzzy clustering each paper is labelled by identifier of its cluster and by value of its similarity to descriptor of this cluster. Classifier uses these labels to include each paper in the most similar cluster.

The advantage of the proposed method is its distributed nature thereby full-text clustering of large amounts of papers can be implemented. That is necessary for detecting current research directions represented in wide research area. Another benefit consists in using indexes which have incremental structure, so adding a number of papers to the corpus $D$ does not involve full re-clustering.

## 4 Experiment

### 4.1 Dataset description

For experiment a dataset of papers of the research area "Regenerative medicine" is created and verified by experts with PhD degree. The dataset contains 112 well-cited publicly accessible papers from 2000 to 2014 distributed into 4 research directions. We apply widely used linguistic analysis library Freeling [11] to retrieve normalized terms (words and noun phrases without stop words) from text of papers in our study. These terms are included in the dataset. The dataset is available on demand, and can be extended and used according to the BSD license.

| Descriptor | Size | Average publication year |
|---|---|---|
| Brain stroke | 23 | 2004 |
| Chimeric antigen receptor cell therapy | 25 | 2009 |
| Induced pluripotent stem cells | 41 | 2011 |
| Wound healing burns | 21 | 2007 |

### 4.2 Experiment setup

| Descriptor | Terms |
|---|---|
| Brain stroke | stroke, brdu test, behavioral, endothelial cell, cell brdu, psa-ncam, cortical cell, reactive, regeneration neuronal, endogenous, neurogenesis, endostatin, cortical, expansion nonhematopoietic, ischemic neuron |
| Chimeric antigen receptor cell therapy | pbls, receptor chimeric, cd19-speci, immunity, antitumor, adoptive ifn, malignancy b-cell, aapcs, protein fusion, infusion cell |
| Induced pluripotent stem cells | gene signature, photoreceptor, epithelium retinal, retinal, suppressor, tumor, cell ips, technology ips, retinal cell, ips-derived, ipscs pigmented |
| Wound healing burns | wound chronic, wound heal, keratinocyte, epidermal cell, fusenig keratin, epidermis region, |

We use the semi-automatic approach for tuning clustering methods that can avoid empirical parameters setup. Clustering parameters are calculated by maximization of the quality function with boundaries $0 < H_a < 1$ and $0.7 < \alpha < 0.99$ (for the proposed method). For other methods we set boundaries according to constraints of these methods. We apply commonly used F-measure function for classification methods assessment. This function based on precision and recall.

Recall $R$ is defined as the ratio of the number of correctly classified papers to size of class in the train set:

$$R = \frac{t}{t+c},$$

where $t$ is a number of correctly classified papers; $c$ is a number of papers which the method carried to other classes.

Precision $P$ is defined as the ratio of number of correctly classified papers to the total of classified papers:

$$P = \frac{t}{t+i},$$

where $t$ is a number of correctly classified papers; $i$ is a number of incorrectly classified papers.
Then the f-measure calculates as follows:

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R},$$

where the $\beta$ parameter setting precision priority over recall. In our experiment, $\beta = 1$. We use a controlled random search method [7] for optimization of quality function.

| Measure | Affinity propagation | Birch | LDA + Affinity propagation | LDA + Birch | Proposed method |
|---|---|---|---|---|---|
| Precision | 0.85 | 0.75 | 0.7 | 0.85 | 0.83 |
| **Recall** | **0.28** | **0.27** | **0.49** | **0.61** | **0.71** |
| F-measure | 0.42 | 0.4 | 0.57 | 0.71 | 0.76 |

### 4.3 Experiment result

Thus, we show that proposed method can be tuned properly using a part of analyzed dataset. Results show that proposed method has sufficient quality of clustering (Table 2). Generated terms represent research directions from the dataset. Table 3 shows experiment results using cross-validation technique. Both qualitative and quantitative assessment shows that the proposed method produces relatively good results. In opposite, the affinity propagation method returns unsatisfactory results. Obviously this method is not suitable for the solution of research direction detection problem. Possibly it relates to incorrect exemplars selection due to the missing of the initial distribution for exemplars.

We suggest that our method overcomes other examined methods due to applying Manhattan distance for paper similarity estimation, to using of topic importance measure for determination of terms of clusters and due to implementation of linguistic analysis for terms extraction as well. That leads to more precise estimation of similarity measure between papers and, consequently, to better quality of research directions detection.

## 5 Conclusion and future work

In this study we investigate various methods for detection of current research directions, which are based on clustering. We found that it is necessary to create a method that can be tuned successfully in a small labelled part of analyzed dataset.

In this paper we present the improved full-text clustering method for detection of research directions. Experiment results show that proposed method is more applicable for research directions detection, than the other widely used methods. The semi-automatic approach for tuning of clustering method demonstrates its performance also. Besides, it was shown that full-text clustering methods work imprecisely for thematically heterogeneous research directions. This paper is considered as an initial study that will be extended in a further research. Moreover, the further development of proposed method consists in implementation of hybrid metrics for clustering. These metrics can be used not only texts similarity, but also co-citations and co-authorships of papers in dataset. It is necessary to continue work with experts for extension of our experimental dataset, because hybrid clustering methods cannot process small datasets properly.

## References


[1] Aletras N. and Stevenson M. Evaluating topic coherence using distributional semantics // Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) –Long Papers. – 2013. – pp. 13-22.

[2] Black P. E. Manhattan distance // Dictionary of Algorithms and Data Structures. – 2006. – Vol. 18.

[3] Blei D. M., Ng A. Y. and Jordan M. I. Latent dirichlet allocation // The Journal of machine Learning research. – 2003. – Vol. 3. – pp. 993–1022.

[4] Cobo M. J. et al. Science mapping software tools: Review, analysis, and cooperative study among tools. // Journal of the American Society for Information Science and Technology. – 2011. – Vol. 62(7). – pp. 1382–1402.

[5] Frey B. J. and Dueck D. Clustering by passing messages between data points. // Science. – 2007. – Vol. 315(5814). – pp. 972-976.

[6] Glanzel W. Bibliometric methods for detecting and analysing emerging research topics. // El profesional de la informacion. – 2012. – Vol. 21(2). – pp. 194-201.

[7] Kaelo P. and Ali M. Some variants of the controlled random search algorithm for global optimization // J. Optim. Theory Appl. – 2006. – Vol. 130(2). – pp. 253-264.

[8] Koltcov S., Koltsova O., Nikolenko S. Latent dirichlet allocation: stability and applications to studies of user-generated content //Proceedings of the 2014 ACM conference on Web science. – ACM, 2014. – pp. 161-165.

[9] Lee W. H. How to identify emerging research fields using scientometrics: An example in the field of information security // Scientometrics. – 2008. – Vol. 76(3). – pp. 503–525.

[10] Leone M. et al. Clustering by soft-constraint affinity propagation: applications to gene-expression data //Bioinformatics. – 2007. – Vol. 23 (20). – pp. 2708-2715.

[11] Padró Lluís and Stanilovsky Evgeny. FreeLing 3.0: Towards Wider Multilinguality // Proceedings of the Language Resources and Evaluation Conference (LREC 2012). – Istanbul, 2012.

[12] Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of computational and applied mathematics. – 1987. – Vol. 20. – pp. 53-65.

[13] Science Citation Index by Thomson Reuters http://thomsonreuters.com/science-citation-index-expand, 2015.

[14] Shvets A. et al. Proceedings of the Science and Information Conference // Detection of Current Research Directions Based on Full-Text Clustering. – London, 2015.

[15] Sochenkov I. Relational-situational data structures, algorithms and methods for search and analytical tasks solving [in Russian] . PhD thesis. Institute for systems analysis of RAS, Moscow, 2014.

[16] Suvorov R. E. and Sochenkov I. V. Method for detecting relationships between sci-tech documents based on topic importance characteristic. [In Russian] ISA RAS, Moscow, 2013. – Vol. 1. – pp. 33-40.

[17] Stein B. Principles of hash-based text retrieval //Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2007. – C. 527-534

[18] Vorontsov K. and Potapenko A. Additive regularization of topic models // Machine Learning. – 2014. – pp. 1-21.

[19] Zhang T., Ramakrishnan R., and Livny M. BIRCH: an efficient data clustering method for very large databases. // ACM SIGMOD Record. - ACM, 1996. – Vol. 25(2). – pp. 103-114.